Proceedings of the Institute of Acoustics

THE AUDITORY SPEECH SKETCH

M. P. Cooke & P. D. Green

University of Sheffield, Department of Computer Science, Sheffield, England.

1. INTRODUCTION

The auditory system appears to group together sound components which are likely to have arisen from the same acoustic source. The aim of the Auditory Speech Sketch Project is to model some of the processes which are thought to underlie this complex task. The Auditory Speech Sketch (ASS) is a collection of representations which result from the application of grouping processes to data derived from a model of the auditory periphery. The lower levels of the ASS consist of what are believed to be perceptually important representations of synchrony, onsets and modulation. Most of the work to date has been to elaborate such early descriptions. More recently, higher levels of representation have been developed; specifically, time-frequency representations of synchronous activity are grouped together according to harmonic constraints. Thus, the ASS can be seen as a hierarchical structure in which higher levels represent components from a lower level which have been grouped according to some criterion.

This paper is organised as follows: first, the various theoretical threads which underlie the work are mentioned, leading to a rather different view of speech analysis than that embodied in most current Automatic Speech Recognition (ASR) systems. Next, an example of processing in the ASS from signal to symbolic description is given. In the third section, an algorithm for harmonic interpretation is applied to the ASS as a demonstration of the way in which a symbolic description can facilitate models of auditory grouping processes. Finally, other work at Sheffield relating to the ASS is summarised.

2. MOTIVATION

The Auditory Speech Sketch Project, which began in 1988, was motivated by work in experimental studies of auditory grouping (for a review, see Bregman, [1]). Bregman and others have suggested, with a good deal of experimental support, that the auditory system carries out a 'scene analysis' in order to determine which parts of the complex mixture reaching the ears belong together. Components are more likely to be grouped if they share some property such as occupation of similar time intervals or if they are modulated in a common manner. The key notion is that the perception of acoustic sources appears to be mediated by the formation of auditory streams, and that ongoing interpretation of the signal is made with respect to these streams. Streams may have a similar role to that played by objects in vision.

In order to model some of these grouping processes directly, it is clear that we have to construct representations of signals couched in the same descriptive vocabulary as that used by the experimentalist. For example, *explicit* characterisations of such things as harmonics, onsets, offsets and local modulation in amplitude and frequency may be required. The notion of computing explicit representations to describe aspects of auditory data stems from Marr's ideas in computational vision (Marr, [17]), which in turn have been adopted for speech in the work of Green and his colleagues [15].

These factors suggest a computational architecture for ASR substantially different from those predominant in the current crop of recognition structures. Some ways in which a system based on adoption of the streaming theory will differ from conventional approaches include the following:

the role of relational structure: Relations between elements across frequency and time are emphasised, in constrast with frame-based descriptions. For instance, the so-called 'spectral integration force' responds to onset synchrony of components (Dannenbring & Bregman, [12]), whilst a sequential grouping principle attempts to group successive elements in time on the basis of frequency proximity (Tougas & Bregman, [21]). Further examples can be attributed to the general notion of 'common fate', where components which behave in the same way in time tend to be grouped into a single stream. In all these respects, the relations between components are at least as important as the components themselves.

identification is made after stream formation: Measurements (eg. of timbre) appear to be made after stream formation. For instance, a spectral shape which ordinarily gives rise to one vowel percept can be perceived as a different vowel if there is evidence that part of it belongs to a separate stream (eg. Darwin, [13]). The implication is that an ASR system might postpone its spectral description until grouping into streams has been achieved. This notion is taken up in the work of Crawford & Cooke [9].

signal decoding is an active process: Streaming indicates that signal decoding may be an active process, in which streams are formed in parallel and compete for components. Possible interpretations of the data appear to interact (Bregman & Tougas, [2]). Notions of disjoint assignment (Bregman & Rudnicky, [3]) and the role of duplex perception (Ciocca & Bregman, [5]) come into play here.

auditory illusions: The overriding concern of the auditory system appears to be to find consistent explanations of the incoming evidence. This might mean that the auditory system makes assumptions about the acoustic data in order to maintain a coherent percept (eg. Dannenbring, [11]).

streams as objects: Whilst the bulk of work in ASR is frame-based (what has been termed the 'bacon-slicer approach'), systems based on streaming theory are more naturally though of as object-based. For example, experiments described in Dannenbring & Bregman [12] indicate that auditory streaming reduces a listener's ability to judge relationships between components of different streams. An appreciation of object-based processing allows novel computational approaches, one of which is outlined in section 4.

explanations are possible: In contrast with much of the current work in low-level speech processing, we expect to be able to construct adequate *explanations* of the incoming signal. Indeed, the goal of the work is precisely that; given an arbitrary complex mixture of speech and other sources, the system should develop consistent explanations, each of which accounts for some proportion of the analysed data.

3. FROM SIGNAL TO SYMBOLS

The processes and representations currently employed in the Auditory Speech Sketch are outlined in Figure 1 below. In this section, brief descriptions of the auditory model and the processes which enable construction of early symbolic descriptions are given.

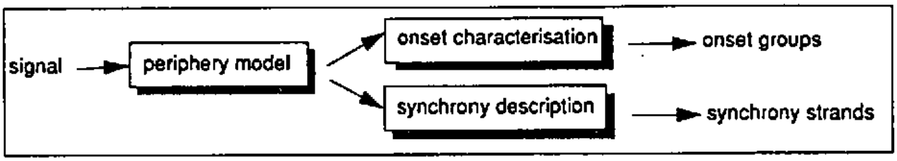

Figure 1: signal processing to symbolic primitives in the ASS

3.1 Peripheral processing

The signal is first procesed by a model of the auditory periphery (Cooke, [6]). Briefly, the peripheral filtering action is accomplished by a bank of gammatone filters (Patterson *et al.*, [19]). The output of each filter is characterised by its instantaneous frequency and amplitude. The instantaneous envelope undergoes a non-linear static compression based on an expression relating stimulus level to inner hair cell receptor potential (Crawford & Fettiplace, [10]). The compressed signal forms the input to a model of hair cell/nerve-fibre transduction. The model is analytic for constant input levels and provably additive for ideal signals.

3.2 Synchrony strands

The instantaneous frequency measures are combined to form temporally-extensive descriptions of synchronous activity in the filter outputs. The process operates in three stages (Figure 2, left panel). First, the dominant frequency in each auditory channel is calculated by median-smoothing the instantaneous frequency estimates. This takes place each millisecond. Each frame of smoothed dominant frequency estimates will contain, in general, a high degree of redundancy since large numbers of filters will be responding to the same stimulus component. The next stage attempts to provide a summary of this synchronous activity within a single millisecond, using a simple but powerful ordering constraint (Cooke, [7]). The resulting tokens represent groups of channels with similar characteristics - *place-groups*. The final stage is to aggregate place-groups over time in a manner akin to formant-tracking (except here we are operating over a finer time-scale and can recover from tracking errors). The resulting symbolic representation forms one aspect of the ASS, namely, *synchrony strands* (Figure 3, top). Strands produced in this way tend to represent harmonics, formants, or, in the region of 1000-1500 Hz, some mixture of the two (the representation of F2 in the auditory system is something most workers appear to finesse, to the extent of choosing inappropriate frequency scales in which F2 is not resolved into harmonics).

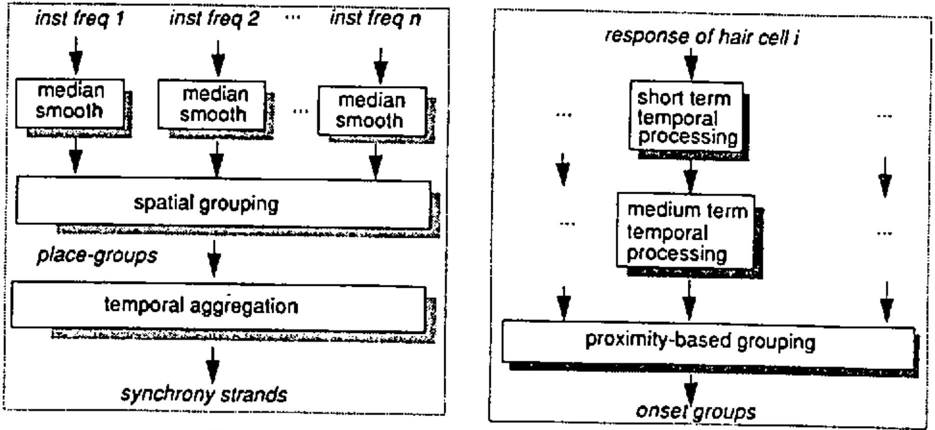

Figure 2: synchrony strand (left) and onset group (right) formation

An attempt to assess the completeness of synchrony strands as a general speech representation has been made via resynthesis. If each strand is treated as a time-varying specification of frequency and amplitude, then a trivial additive synthesis procedure can be adopted (Cooke, [7]). The results of informal (but fairly wide-ranging) listening tests indicate that speech synthesised from strands is not only highly intelligible but, in many cases, virtually indistinguishable from the original. This is the case for male/female speech from 4 databases, whispered speech (perhaps surprisingly) and speech with added white noise. In fact, the resynthesised speech with this kind of noise sounds rather more intelligible than the original.

3.3 Onset groups

The synchrony strand representation is complemented by an initial characterisation of onset responses from the hair cell outputs in the periphery model. The starting point is a determination of peaks in the hair cell responses. We recognise that such peaks sometimes correspond to onsets of signal components, whilst at other times reflect local variations in signal level. Such local variations may themselves be due to amplitude modulation caused by the interaction of harmonics within a filter. Both signal onsets and regularities related to envelope modulation are likely to be useful aspects of any auditory description. We attempt to extract both types of peak using 2 stages of temporal processing via a neuronal model based loosely on that of Segundo *et al.*[20].The first stage provides short-term temporal processing employing a time-constant of about 2 ms. This fits with various estimates of temporal interval detection ability (reviewed in Moore, [18]). Intervals between peaks which result from this stage can provide a crude estimate of amplitude modulation rate; abetter approach is described in Brown & Cooke [4].The second stage is identical to the first except that a time constant of 40 ms is used. Finally, a relatively simple grouping of such onsets across channels is employed based on proximity of onsets across time. The characterisation in terms of such *onset-groups* can be seen in the lower panel of Figure 3.

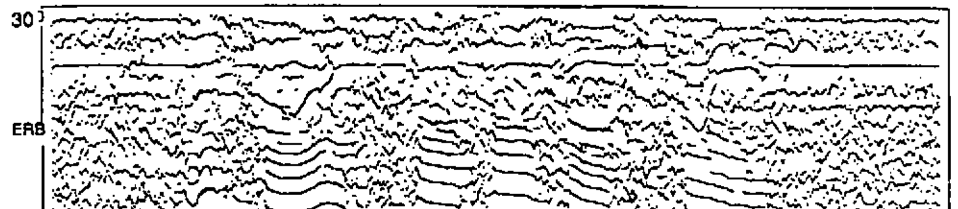

Figure 3: Onset groups (bottom) and synchrony strands (top) produced in response to the utterance "*This time its tar tar tart to read*" (female speaker).

4. HARMONIC LABELLING in the ASS

As an illustration of a process which may be used to group ASS primitives, we have considered the task of labelling strands which represent harmonics. The process described below might be considered as employing a 2-dimensional time-varying harmonic sieve. However, the approach outlined here is data-driven, starting with the most dominant strand (during strand formation, the effective number of channels which appear to be responding to the same component are summed, thus giving a dominance measure for each strand which is a function of its length and its prominence).

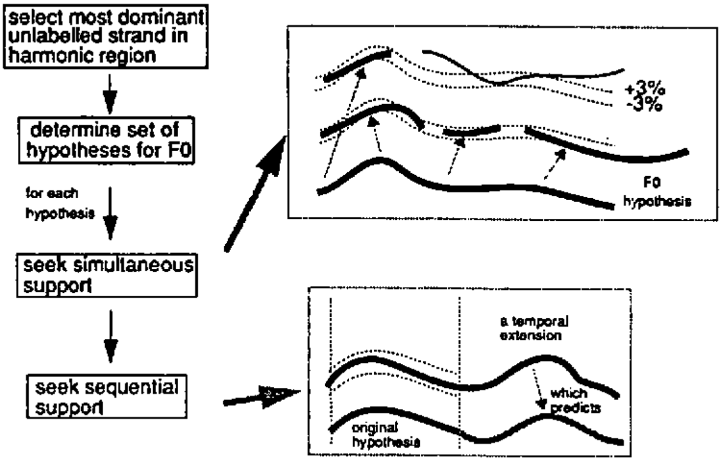
select most dominant unlabelled strand in harmonic region
determine set of hypotheses for F0
for each hypothesis
seek simultaneous support
seek sequential support
+3%
-3%
F0 hypothesis
a temporal extension
which predicts
original hypothesis

Figure 4: One cycle of harmonic labelling, showing simultaneous (top right) and sequential (bottom right) hypothesis propagation.

The algorithm begins by finding the most dominant strand below 1500 Hz, which is usually a fairly long, low-harmonic. A set of hypotheses about F0 is then generated, based on the location in frequency of the strand. For example, if the most dominant strand covers the frequency region 250-350 Hz, it is likely to be either the fundamental, or the first, second, third or fourth harmonic. It is unlikely, based on the existence region for pitch, to be a higher harmonic. A set of hypotheses is set up to represent each one of these cases. Hypotheses are placed on an agenda.

4.1 Simultaneous grouping

The algorithm proceeds to develop each item on the agenda by seeking first simultaneous, then sequential, support, as illustrated in Figure 4. For example, if the hypothesis is that the dominant strand is the second harmonic, then a time-varying sieve would be set up, with its 2nd harmonic aligned with the dominant strand. Since strands live in frequency and time, we can be quite strict about those strands which fit into the sieve. Currently, a figure of 80% overlap between strand and sieve is required for a strand to support the hypothesis. Coincidental matches in individual frames are therefore ruled out.

A scoring mechanism is used in which the total possible support for a hypothesis (assuming all strands in the region where harmonics are resolved are harmonics of the estimated F0) is calculated, and used to normalise the actual support found during this stage. Thus, each hypothesis obtains a score which conveniently represents its coverage of the data. Clearly, if we are dealing with fully voiced, single speaker speech, we would expect to get hypotheses which account for a significant amount of the data (this is what we actually find).

4.2 Sequential propagation

Hypotheses which have undergone a stage of simultaneous grouping are placed back on the agenda (in score order, for future pruning, although in the current system we fully explore all hypotheses). The second form of support gathering is via sideways (sequential) propagation, as depicted in the lower right panel of Figure 4. Since strands occur asynchronously, it is likely that, following the simultaneous grouping phase, strands with portions outside the temporal extent of the original hypothesis will be found. These (both on the left and right of the original) form new hypotheses which are then be subject to simultaneous grouping. In this way, hypotheses are propagated throughout the utterance.

It is worth noting that, at this stage, it is possible to find conflicting predictions of F0 outside the original region of the hypothesis. This means that we can recover from tracking errors made during strand formation (although the conflict-resolution step has yet to be implemented in the model).

4.3 An illustration: labelling harmonics from concurrent synthetic vowels

The process described above works well on all the single speaker material we have examined. It is of more interest to see an example of its performance on more complex stimuli such as simultaneous vowels. Briefly, two vowels (/iy/ with a fundamental of 100 Hz and /aw/ on 150 Hz) were generated via the Klatt synthesiser (Klatt, [16]), and their waveforms were summed. The combined waveform formed the input to the model, resulting in the strands shown in the central part of Figure 5.

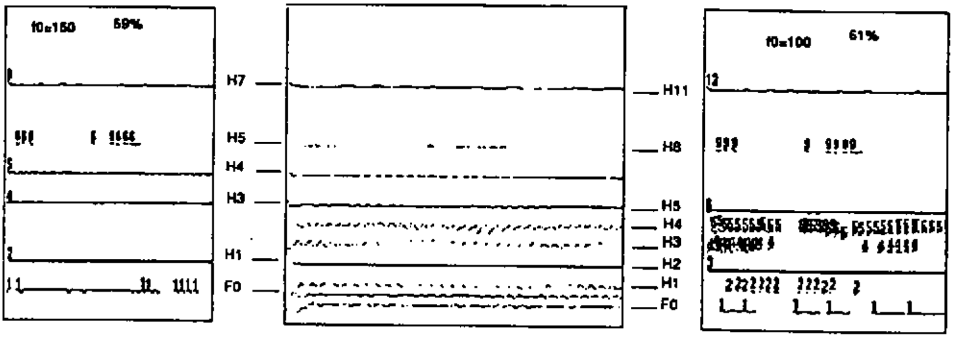

Figure 5: Labelling harmonics from simultaneous synthetic vowels. Strands (centre) together with locations of F0 and harmonics attributable to 150 Hz vowel (left of centre) and 100 Hz vowel (right of centre). Right panel shows top scoring interpretation of data, whilst left panel shows the next best interpretation. Numbers on strands show which *multiple* of the fundamental they represent (subtract one from this to get the harmonic number !).

The top two hypotheses, which account for 61% and 59% of the data respectively, are shown on the right and left of Figure 5. The leftmost panel correctly identifies those strands which support the 150 Hz fundamental, whilst the right panel identifies those for the 100 Hz fundamental. Thus, the top two hypotheses are the correct ones. A total of 10 other hypotheses were generated, all of which scored rather less than 30%. In most cases, these weaker hypotheses were supported by strands which formed some subset of those utilised by one or other of the two correct solutions. This suggests a possible mechanismfor dealing with low-scoring hypotheses. Ideally, the set of hypotheses can be considered as forming a partial ordering defined by the subset relation on sets of supporting strands. If a hypothesis accounts for some subset of the data accounted for by some other hypothesis, then it will fall lower in the ordering than the correct hypothesis (Occam's razor). Hence, the system, because of its explanatory descriptive base, might handle octave errors and the like.

5. FUTURE AND RELATED WORK

This paper has provided a basic description of the ASS and illustrated a single grouping approach which it has been possible to implement. Much more work is required to develop a generalised framework in which a collection of different grouping processes can reside. For that work, consideration must be given to competition between hypotheses and an interpretation of the grouped structures where components are shared. In this regard, the role of the principle of disjoint assignment, and its antithesis, duplex perception, must be assessed. It is possible that the notion of a partial ordering of hypotheses is sufficiently powerful to serve as a computational structure for more complex interactions of grouping principles. Further, the scoring scheme, based on the amount of data explained by the hypothesis appears to provide a powerful mechanism for scheduling hypotheses.

In parallel with these activities, further elaborations of the early levels of the ASS will be required. We do not, as yet, know how robust the onset groups are since they are not currently part of any grouping process. Similarly, the rather crude measure of amplitude modulation rate derived from the short-term temporal processing of spikes needs to be rigorously assessed. A good test would be to use it as the basis for grouping strands which represent formants, using the /ru/-/li/ data of Darwin& Gardner [14].

Other work at Sheffield will feed into the Auditory Speech Sketch Project, and make use of the streaming algorithms embodied in the work. Brown & Cooke [4] report on the notion of using physiologically-based maps as a computational representation of certain signal parameters such as amplitude and frequency modulations. In particular, we hope to develop a better temporal aggregation stage through the use of a map of frequency modulation. Crawford & Cooke [9] are tackling the further issue of how large-scale spectral integration might lead to important normalisations in phonetic systems. In that approach, integration is seen as a post-streaming process, so it is natural to see any modules developed in that work as following on from the streaming described in the current paper. We are attempting to develop an integrated approach to modelling auditory processes; our current thinking on this is contained in Cooke, Crawford & Brown [8]. There is still a great deal of work which needs to be done, both in the experimental sciences to support models, and from the computational viewpoint of how to extract appropriate descriptions and coordinate exploration of the auditory scene.

ACKNOWLEDGEMENTS

Aspects of this work have benefitted from discussions with Christophe D'Allesandro, Guy Brown, Steve Beet and Malcolm Crawford. The authors are supported by SERC Image Interpretation Initiative grant GR/E42754.

REFERENCES

[1] A S BREGMAN, *Auditory Scene Analysis*. Cambridge, MA: MIT Press (1990).

[2] A S BREGMAN & Y TOUGAS, 'Propagation of constraints in auditory organisation', *Perception & Psychophysics*, 46(4), pp. 395-396 (1989).

[3] A S BREGMAN & A RUDNICKY,'Auditory segregation: stream or streams ?',*J. Exp. Psych: Human Percep-*

tion & Performance, **1**, pp. 263-267 (1975).

[4] G J BROWN & M P COOKE, 'Extracting descriptions of amplitude modulation from an auditory model: a comparative study', these proceedings (1990).

[5] V CIOCCA & A S BREGMAN,'The effects of auditory streaming on duplex perception', *Perception & Psychophysics*, **46**(1), pp. 39-48 (1989).

[6] M P COOKE, 'The auditory periphery: physiology, function and a computer model', University of Sheffield Department of Computer Science Research Report, CS-89-32 (1989).

[7] M P COOKE, 'Synchrony Strands: An early auditory time-frequency representation', University of Sheffield Department of Computer Science Research Report, CS-90-05 (1990).

[8] M P COOKE, M D CRAWFORD & G J BROWN,'An integrated treatment of auditory knowledge in a model of speech analysis', SST-90, Melbourne (1990).

[9] M D CRAWFORD & M P COOKE,'A computational study of large-scale integration', these proceedings (1990).

[10] A D CRAWFORD & R FETTIPLACE, 'Nonlinearities in the response of turtle hair cells', *J. Physiol.*, **315**, pp. 317-338, (1981).

[11] G L DANNENBRING,'Perceived auditory continuity with alternately rising and falling frequency transitions', *Can. J. Psychol.*, **30**, pp. 99-114 (1976).

[12] G L DANNENBRING & A S BREGMAN,'Stream segregation and the illusion of overlap', *J. Exp. Psych: Human Perception & Performance*, **2**, (1976).

[13] C J DARWIN, 'Perceiving vowels in the presence of another sound: Constraints on formant perception', *J. Acoustic. Soc. Am.*, **76**(6) , pp. 1636-1647 (1984).

[14] C J DARWIN & R B GARDNER,'Perceptual separation of speech from concurrent sounds', in: *The psychophysics of speech perception*, M E H Schouten (ed), Martinus Nijhoff (1987).

[15] P D GREEN, G J BROWN, M P COOKE, M D CRAWFORD & A J SIMONS, 'Bridging the gap between signals and symbols in speech recognition', in: *Speech, hearing and language processing*, W A Ainsworth (ed), JAI Press (1990).

[16] D H KLATT, 'Software for a cascade/parallel formant synthesiser', *J. Acoust. Soc. Am.*, **67**(3), pp. 971-995 (1980).

[17] D MARR, *Vision*, Freeman (1982).

[18] B C J MOORE, *An introduction to the psychology of hearing*, 3rd edition, Academic Press (1989).

[19] R D PATTERSON, I NIMMO-SMITH, J HOLDSWORTH & P RICE,'An efficient auditory filterbank based on the GammaTone function', Meeting of the Speech Group of the Institute of Acoustics, RSRE, Dec. (1987).

[20] J P SEGUNDO, D H PERKEL, H WYMAN, H HEGSTAD & G P MOORE, 'Input-output relations in computer-simulated nerve cells', *Kybernetic*, **12**, pp. 157-171 (1968).

[21] Y TOUGAS & A S BREGMAN, 'Auditory streaming and the continuity illusion', *Perception & Psychophysics*, **47**, pp. 121-126 (1990).