A COMPUTATIONAL MODEL OF AMPLITUDE MODULATION PROCESSING IN THE HIGHER AUDITORY SYSTEM

G.J. Brown and M.P. Cooke

Department of Computer Science
University of Sheffield

ABSTRACT - A novel speech processing technique is presented, which is based on the concept of a modulation map. This term describes the way in which the higher auditory system codes amplitude modulation rate as spatially distributed peaks of activity within a neural array. The map has applications in grouping and pitch analysis, and will form part of an integrated model of auditory processing. A simple pitch detector based on the map shows a superior performance in noise compared to a conventional autocorrelation analysis.

INTRODUCTION

Many biologically significant sounds, including human speech, are characterized by temporal variations in amplitude and frequency. It seems likely, therefore, that the ability of the auditory system to code and analyze these amplitude modulations (AM) and frequency modulations (FM) is of basic importance.

Support for this hypothesis has come from physiological studies, which suggest that the neural responses to AM and FM stimuli become progressively specialized at successively higher levels of the auditory pathway (Kay, 1982). In the auditory nerve, complex sounds are coded as temporal and spatial variations in the pattern of neural firings. Consequently, there does not appear to be any specificity for modulated sounds at the level of the auditory periphery. Beyond the auditory nerve, however, in the higher auditory system, many neurons show selectivity for particular rates and directions of amplitude and frequency change. Indeed, many cortical neurons do not respond do static tones at all.

Recent physiological studies suggest that some neurons which respond to particular rates of AM are organized topographically, forming structures that we shall refer to as *modulation maps* (Schreiner and Langner, 1988). It is proposed that modulation maps offer a novel computational means of representing amplitude fluctuations in the speech signal, and we present a computer simulation of a map which will form part of an integrated model of auditory processing. We also report our results from using the map as a basis for extracting the fundamental frequency of digitally recorded speech.

MODELLING THE HIGHER AUDITORY SYSTEM

Although the auditory periphery is understood in some detail, our knowledge of the physiological mechanisms of higher auditory processing is relatively incomplete and fragmented. Clearly, this lack of detailed physiological information has implications for the way in which the higher auditory system can be modelled. In this section, we describe some of the approaches that have been taken and assess their validity.

Feature Extraction Models

Many physiological investigations of higher auditory function have attempted to attribute feature extracting properties to single neurons, based on a knowledge of the pattern of their frequency or temporal response. This work has inspired several computational studies, such as the ON cell model employed by Wu et al. (1989) to detect articulatory-acoustic events. The problem with this approach is that the choice of 'features' which are 'extracted' is largely arbitrary, because we have a poor understanding of the transforms that the auditory system performs on the acoustic stimulus (Whitfield, 1979). Until we have a sound knowledge of the way in which the higher auditory system organizes the inputs from its constituent cells, feature extraction models remain very speculative.

Detailed Physiological Models

Although there is a poor understanding of their function, the neural structure of the first few higher auditory nuclei is now quite well documented. Thus, it is possible to model these areas as networks of simulated neurons that are connected according to the known neural circuitry. This approach has been adopted by Pont (1990), who describes a computational model of the dorsal cochlear nucleus (DCN). The model avoids the difficulties of attributing specific functions to auditory neurons, because it considers the response of the network as a whole. However, even at the level of the DCN, the physiological data is still quite incomplete and the model represents a simplified and probably inaccurate view of higher auditory processing. Furthermore, it is difficult to extend this approach to modelling areas of the auditory system beyond the DCN. Although there is data describing the response properties of neurons at higher levels, there is very little information concerning their connectivity.

Representational Models

Given the limitations of the approaches described above, it seems that the best way to proceed is one which does not assume functions for single cells, or which attempts to replicate the physiology at a detailed level. Rather, *we should be concerned with modelling transformations that the higher auditory system might apply to the acoustic stimulus.* The nature of the higher auditory representation should be guided by our knowledge of the cell types that provide the basis for the sensory analysis, and of the important features in the speech signal that the auditory system is likely to preserve. Essentially, our approach is to ask 'What is it sensible for the auditory system to compute?'. In this respect, our rationale is similar to the computational approach to vision pioneered by Marr (1982).

Several representational models of auditory processing have been described in the literature, such as the models of synchrony between auditory nerve firings proposed by Seneff (1988) and Cooke (1990). In this work, we propose a representational approach that is based on the concept of a modulation map.

COMPUTATIONAL MAPS

The computational map is a principle of neuronal organization which describes the transformation of information into the topography of a neural array (Knudsen *et al.*, 1982). Anatomically, a map consists of a parallel array of neural processors that are tuned to slightly different values of the same parameter, so that there is a systematic, place-coded representation of the parameter across the map. This organization enables information to be processed very rapidly, and codes it into a form that can be processed by simple schemes of connectivity, such as lateral inhibition. It is likely that computational maps are the most efficient way that the brain can represent and process information.

Most of the maps identified so far have been in sensory areas, including several in the auditory system. Auditory maps seem to conform to a general framework in which best frequency is represented on one plane, and a mapped parameter is represented in an orthogonal plane. While many auditory maps have been identified in birds and bats, which have highly specialized hearing, they have also been found in other animals such as the cat. It seems, therefore, that computational mapping is a fairly general principle of neuronal organization.

Below, we describe some of the maps found in the higher auditory system and discuss their potential for deriving useful computational representations of the speech signal.

Maps of Auditory Space.

Neurons in the midbrain of the barn owl only respond to sounds when they originate from a small area of auditory space, called a receptive field (Knudsen *et al.*, 1982). Furthermore, the units are topographically organized according to the elevation and azimuth of their receptive fields, so that they form a physiological map of auditory space. It appears that azimuth is localized by the difference in the time of arrival of a stimulus at the two ears, and elevation is localized by the difference in the intensity of a stimulus at the two ears. A low resolution map of auditory space has also been identified in the guinea-pig (King & Palmer, 1983).

A representation of auditory space would form a good basis for the separation of simultaneous speakers from a stereo recording. We feel that the possibility of incorporating a map of auditory space into a speech recognizer merits further research, but do not pursue this idea here.

Maps of Best Intensity.

Many neurons in the higher auditory system receive inhibitory inputs from other cells, the strength of which depends on the stimulus intensity. Consequently, a best intensity can be defined for many higher auditory neurons, which is the stimulus intensity which elicits the maximal firing rate. In the auditory cortex of the echo-locating bat, there is a topographic representation of best intensities (Suga & Manabe, 1982). Best intensity maps have limited interest for the auditory modeler, because they have not been identified in animals other than the bat. It is likely, therefore, that intensity maps have a special significance related to echo-location.

Modulation Maps.

The response of neurons to AM and FM as a function of modulation frequency changes at different levels of the higher auditory system. At the level of the cochlear nucleus, the response of most cells can be described by a lowpass function. However, neurons at higher levels typically have a bandpass response to modulation, so that it is possible to determine a best modulation frequency (BMF) to which the cell is tuned. The maximum BMF observed decreases at progressively higher levels of the auditory pathway, being 500-600Hz at the cochlear nucleus and less than 20 Hz at the cortex (Rees & Moller, 1983). This suggests that information about low modulation frequencies is extracted at some stage in the higher auditory system, and is probably coded into another form.

Schreiner and Langner (1988) suggest that information about AM rate is recoded into the form of a neural array. They describe a map of AM rate in the inferior colliculus of the cat, which we refer to as a modulation map. The inferior colliculus consists of sheets of cells, called laminae, in which all the neurons are tuned to a similar best frequency. Within each lamina, neurons with a similar BMF are systematically arranged into contours, with high BMFs represented in the middle of the lamina and low BMFs represented on the circumference. Thus, there is a two-dimensional arrangement of neurons in which frequency is represented on one axis and BMF on the other. The distribution of modulation frequencies present at a particular best frequency is coded as peaks of activity in the neural map.

Currently, there are no physiological reports of a map for FM rate. However, since FM appears to be a parameter of some relevance to the auditory system, it is quite possible that such maps exist.

APPLICATION OF MODULATION MAPS IN A MODEL OF AUDITORY PROCESSING

We believe that computational modelling of modulation maps is a novel means of representing amplitude and frequency modulations in the speech signal, which is consistent with our representational approach to auditory modelling. It is proposed that the maps should form part of an integrated model of auditory processing (Cooke *et al.*, 1990), and will be used in two ways.

Fundamental Frequency Extraction

From the mapped representation of AM rates across the auditory nerve fibre population, it is possible to estimate the fundamental frequency of speech. This can be achieved by building a distribution of AM rates over all best frequencies, within a time frame, and selecting the modulation rate which occurs most often.

Formation of Auditory Objects

Speech is usually heard against a background of other sounds. Consequently, the auditory system must be able to separate out those spectral components which belong to the same voice. One way in which it appears to do this is by grouping components which have common rates of AM and FM (Bregman *et al.*, 1985; McAdams, 1984).

Bregman *et al.* (1990) define two types of grouping. *Simultaneous grouping* describes the grouping of spectral components into contributions from the same source at any particular time instant. Spectral components are more likely to be grouped simultaneously if they share a common AM rate, independent of whether they are harmonically related. Clearly, a map of AM rate would provide information about the modulation rates present at different best frequencies which is necessary for this process. A second type of grouping, *sequential grouping*, determines which spectral components have arisen over time from the same source. A map of FM would provide a description of the way in which spectral components are moving in time, which would form a basis for this type of grouping. Our previous work has implemented a neural network which demonstrates sensitivity to FM sounds, and it is anticipated that a map can be modelled from a collection of these units (Brown and Cooke, 1990a).

THE MODEL

Here, we present a model which is based firmly on the concept of a map for AM rate, but ignores some of the physiological details for reasons of efficiency. In particular, we use a much smaller number of modulation detectors than the physiological map. Neural maps are highly redundant, and contain many cells that are tuned to similar values of the mapped parameter (Knudsen *et al.*, 1982).This redundancy confers resistance to the topological variations that occur in biological neural networks, but is not a necessary feature of a computer model.

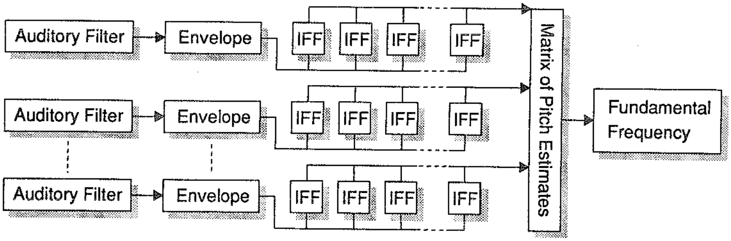

Figure 1. Schematic diagram of the model, showing three auditory filter channels and four instantaneous frequency filter (IFF) channels.

The structure of the model is shown in Figure 1. The auditory periphery is modelled by a bank of bandpass filters tuned to frequencies between 100Hz and 5000Hz, which simulate the effects of the mechanical filtering action of the basilar membrane at a number of points along its length (Cooke, 1989). We use bandpass filters based on the gammatone function. This is an analytic expression for the impulse response of an auditory nerve fibre derived from reverse correlation (deBoer and deJongh, 1978), and has the form

$$g(t) \propto t^{n-1} exp(-2\pi bt) \cos(2\pi w_{cf} t + \phi) \qquad \text{for } (t \geq 0)$$

Here, n is the order of the filter, b is the bandwidth, w_{cf} is the centre frequency and ϕ is the phase (in radians). Because adaptation of the auditory nerve has a negligible effect on the subsequent modulation extraction, we do not include a hair cell transduction stage in our model of the periphery. Instead, we extract the instantaneous envelope of each gammatone filter output.

Information about AM rate is extracted from the periphery by processing the envelope of each auditory nerve channel with a parallel array of bandpass filters. These are tuned to frequencies between 50Hz and 500Hz, in order to reflect the physiological range of BMFs in the inferior colliculus (Schreiner and Langner, 1988). Rather than using, for example, 450 filters tuned to 1Hz increments within this range, the computational load can be reduced by employing 10 filters with overlapping bandwidths, and calculating the instantaneous frequency to which each is responding. Given that *R(t)* and *I(t)* are the outputs of the real and imaginary parts of the gammatone filter (which is used simply for convenience), instantaneous frequency is given by

$$\upsilon(t) = \frac{1}{2\pi}\left(w_{cf} + \frac{I(t)\frac{d}{dt}R(t) - R(t)\frac{d}{dt}I(t)\,I(t)}{I^2(t) + R^2(t)}\right)$$

A derivation of this equation is given in Cooke (1990). By extracting the frequency components of each envelope in this way, we effectively determine the modulation frequencies present in each channel.

FUNDAMENTAL FREQUENCY EXTRACTION USING THE MODEL

One possible application of the model is the estimation of fundamental frequency, a parameter which is important for many speech processing applications such as speaker verification and identification. An estimate of fundamental frequency can be derived from the modulation map by forming a distribution of the modulation frequencies that occur within a time frame, and selecting the rate which occurs most often. We do this by summing the instantaneous amplitudes of each instantaneous frequency filter (IFF) into a bin that corresponds to the frequency, *f*, at which it is responding.

$$\rho(f) = \sum_{0}^{t} IFF_f(t) \qquad \text{for } (50 \le f \le 500)$$

Over the time frame, the fundamental frequency will correspond to the bin ρ*(f)* with the largest value. Note that this is purely a 'signal processing' approach to fundamental frequency extraction, which does not attempt to model the psychophysical properties of periodicity (residue) pitch. Consequently, although the map correctly predicts the pitch of harmonic complexes, it does not predict the pitch of inharmonic stimuli.

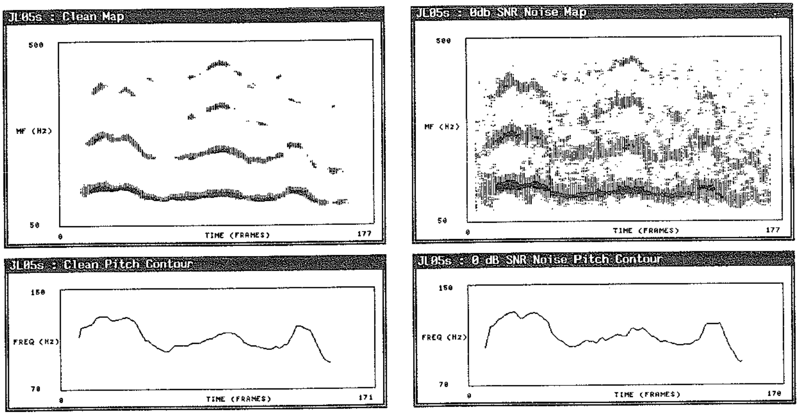

Figure 2. Modulation map (top) and pitch contour derived from the map (bottom) for an utterance in clean and noisy conditions.

Figure 2 shows the output of the map for the (voiced) utterance 'our rule will allow you a lawyer', in clean conditions and with random noise added to give a SNR of 0 dB. In both cases, the fundamental and its first harmonic are clearly delineated. The lower diagrams show pitch contours derived from the map in each case, which suggest that the representation is quite resistant to noise. The map is robust even to a SNR of -10 dB, but we cannot illustrate these results here due to limitations imposed by the quality of the greyscale display.

Currently, we use 16 auditory filter channels in our implementation of the model. Using more filter channels improves the smoothness of the pitch contours, but this effect is not significant enough to justify the extra computational expense. A more detailed study of the performance of the map in noise is given in Brown and Cooke (1990b). We find that the quality of the pitch contours derived from the map degrades less in noise than those calculated by a conventional autocorrelation technique.

ACKNOWLEDGEMENTS

GJB is supported by SERC CASE award 88501484, a SERC travel grant and British Telecom Research Laboratories. MPC is supported by SERC award GR/E 42754. Thanks are due to Malcolm Crawford, Phil Green and William Millar for their support.

REFERENCES

deBoer, E. & deJongh, H.R. (1978) *On cochlear encoding: Potentialities and limitations of the reverse-correlation technique*, J. Acoust. Soc. Am., 63, 115-135.

Bregman, A.S., Abramson, J., Doehring, P. & Darwin, C.J. (1985) *Spectral integration based on common amplitude modulation*, Perception & Psychophysics, 37, 483-493.

Bregman, A.S., Levitan, R. & Liao, C. (1990) *Fusion of auditory components: Effects of the frequency of amplitude modulation*, Perception & Psychophysics, 47(1), 68-73.

Brown, G.J. & Cooke, M.P. (1990a) *Modelling modulation maps in the central auditory system*, Brit. J. Audiol., 24(3), 196.

Brown, G.J. & Cooke, M.P. (1990b) *Extraction of amplitude modulation from an auditory model: A comparative study*, Proc IOA, Windermere.

Cooke, M.P. (1989) *The auditory periphery: Physiology, function and a computer model*, Department of Computer Science Research Report, University of Sheffield.

Cooke, M.P. (1990) *Synchrony strands: An early auditory time-frequency representation*, Department of Computer Science Research Report, University of Sheffield.

Cooke, M.P., Crawford, M.D. & Brown, G.J. (1990) *An integrated treatment of auditory knowledge in a model of speech processing*, these proceedings.

Kay, R.H. (1982) *Hearing of modulation in sounds*, Physiological Reviews, 62(3), 894-975.

King, A.J. & Palmer, A.R. (1983) *Cells responsive to free field auditory stimuli in guinea-pig superior colliculus: Distribution and response properties*. J. Physiology, 342, 361-381.

Knudsen, E.I., duLac, S. & Esterly, S.D. (1982) *Computational maps in the brain*, Ann. Rev. Neurosci., 10, 41-65.

Marr, D. (1982) *Vision*, (W.H. Freeman and Company: New York).

McAdams, S. (1984) *Spectral fusion, spectral parsing and the formation of auditory images*, Ph.D. thesis, Stanford University.

Pont, M.J. (1990) *The role of the dorsal cochlear nucleus in the perception of voicing contrasts in english stop consonants: A computational modelling study*, Ph.D. thesis, University of Southampton.

Rees, A. & Moller, A.R. (1983) *Responses of neurons in the inferior colliculus of the rat to AM and FM tones*, Hear. Res., 10, 301-330.

Schreiner, C.E. & Langner, G. (1988) *Periodicity coding in the inferior colliculus of the cat. II. Topographical organization*, J. Neurophysiology, 60(6), 1823-1840.

Seneff, S. (1988) *A joint synchrony/mean-rate model of auditory speech processing*, J. Phonetics, 16, 55-76.

Suga, N. & Manabe, T. (1982) *Neural basis of amplitude spectrum representation in auditory cortex of the mustached bat*, J. Neurophysiology, 47(2), 225-255.

Whitfield, I.C. (1979) *The object of the sensory cortex*, Brain Behav. Evol., 16, 129-154.

Wu, Z.L., Escudier, P. & Schwartz, J.L. (1989) *Specialized physiology-based channels for the detection of articulatory-acoustic events: A preliminary scheme and its performance*, ICASSP-89, 2013-2016.