Proceedings of the Institute of Acoustics

EXTRACTING DESCRIPTIONS OF AMPLITUDE MODULATION FROM AN AUDITORY MODEL: A COMPARATIVE STUDY

G J Brown & M P Cooke

University of Sheffield, Department of Computer Science, Sheffield, UK

1. INTRODUCTION

The vocalisations produced by most animals, including humans, exhibit temporal variations in amplitude and frequency. It seems likely, therefore, that a fundamental task of the auditory system must be to code and analyse these amplitude modulations (AM) and frequency modulations (FM) in some way. This hypothesis has been supported by physiological studies, which suggest that neural responses to complex sounds become progressively specialised at successively higher levels of the auditory pathway. In the auditory nerve, AM and FM are coded as temporal and spatial variations in the pattern of neural firings. Consequently, there appears to be no specificity for modulated sounds at the level of the auditory periphery. However, beyond the auditory nerve, in the higher auditory system, many neurons show selectivity for particular rates and directions of amplitude and frequency change (Moller [12]; Rees & Moller [16]). Indeed at the highest level of the auditory system, the auditory cortex, many neurons do not respond to static tones at all (Whitfield & Evans [24]).

Higher auditory neurons that are tuned to a particular frequency of AM, with regard to their degree of synchronisation or average firing rate, can be described as having a best modulation frequency or BMF (Schreiner & Urbas [18]). In the inferior colliculus of the cat, there is evidence for a topographic organisation of neurons according to their BMF, such that AM is represented by the spatial distribution of activity within a neural array (Schreiner & Langner [19]). Here, we refer to such an arrangement as a *modulation map*.

It is proposed that a modulation map offers a novel computational means of representing amplitude fluctuations in the speech signal, and we present a computer simulation of the 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.

2. APPROACHES TO MODELLING THE HIGHER AUDITORY SYSTEM

In comparison with the auditory periphery, 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. Below, we discuss some of the approaches that have been taken and assess their validity.

2.1 Feature Extraction Models

Much of the physiological literature on the higher auditory system attempts to classify single neurons according to the pattern of their frequency or temporal response. Based on this information, some workers have hypothesised that particular cell types are extracting specific features from the acoustic stimulus. A typical example is the ON cell model employed by Wu *et al.* [25] to detect articulatory-acoustic events.

The problem with this approach is that the choice of 'features' which are 'extracted' is largely arbitrary, because the transforms that the higher auditory system performs on the acoustic stimulus are poorly understood (Whitfield [23]). This difficulty is reflected in the diverse range of functions that have been suggested for cell types in areas such as the cochlear nucleus (Godfrey *et al.* [8]). Until we have a sound knowledge

of the way in which the higher auditory system organises the input from such cells, feature extraction models remain purely speculative. Consequently, the concept of the auditory system as a hierarchical collection of feature detectors is losing favour. Intuitively, it seems unlikely that in an auditory system consisting of many millions of neurons, specific feature extracting tasks will be assigned to small numbers of a particular cell type. Processing tasks are more likely to be distributed across a large number of cells.

2.2 Detailed Physiological Models

The structure of the cochlear nucleus, the first of the higher auditory nuclei, is now quite well documented. Consequently, some workers have attempted to model this area as a network of neuron models that are connected according to the known neural circuitry. This approach has been adopted by Pont [13], who describes a computational model of the dorsal cochlear nucleus (DCN). This 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. For example, the model described by Pont ignores the input from adjacent cochlear nuclei into the DCN.

Also, there are problems with extending this method beyond the cochlear nucleus. Although there is data describing the response properties of neurons beyond the DCN, there is very little information concerning their connectivity. Clearly, there is a limit to the applicability of this approach until the detailed physiology of higher auditory nuclei is known.

2.3 Representational Models

Given the problems of the two approaches described above, it seems that the best way to proceed is not to attach a function to single cells, or to attempt 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 the important features of speech that the auditory system is likely to preserve.

Several models of auditory processing have been described which use concepts from neurophysiology. These include the cross-correlation model described by Deng *et al.* [7], and the lateral inhibition network of Shamma [21]. Other models propose higher auditory representations of the synchrony between auditory nerve firings, such as the synchrony/mean rate model of Seneff [20] and the synchrony strand analysis described by Cooke [5].

In this work, we adopt a representational approach that is based on the concept of a modulation map.

3. MODULATION MAPS

A concept that has recently emerged in neuroscience is the *computational map*, a term which describes the transformation of information into the topography of a neural array (Knudsen *et al.* [10]). 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 organisation enables rapid processing of information, 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 represent 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. Whilst many have been identified in birds and bats, which have highly specialised hearing, maps have also been identified in other animals such as the cat. It seems, therefore, that the computational map is a fairly general concept of neuronal organisation.

Schreiner and Langner [19] 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.

We believe that computational modelling of modulation maps is a novel means of representing amplitude and frequency modulations in the speech signal. It is proposed that the maps should be used in two ways.

3.1 Fundamental Frequency Extraction

Given the mapped representation of AM rates across the auditory nerve fibre population, it is possible to estimate the fundamental frequency of a speech stimulus. 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. Note, however, that this is a 'signal processing' approach which does not reflect the psychophysical properties of periodicity (residue) pitch.

3.2 Formation of Auditory Objects

It is usually the case that speech is heard against a background of other, conflicting, 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 amplitude and frequency modulations (Bregman *et al.* [2]; McAdams [11]).

The grouping of spectral components into contributions from the same source at any particular time instant is defined as *simultaneous grouping* by Bregman et al. [3]. Spectral components are more likely to be grouped together if they share a common AM rate, independent of whether they are harmonically related. Clearly, a map of AM rate would provide the information about the modulation rates present at different best frequencies which is necessary for this process. Additionally, a map of AM rate could form the basis for an investigation of the mechanisms of comodulation masking release. This term describes the ability of the auditory system to separate out a masked tone by correlating amplitude fluctuations in the background spectrum.

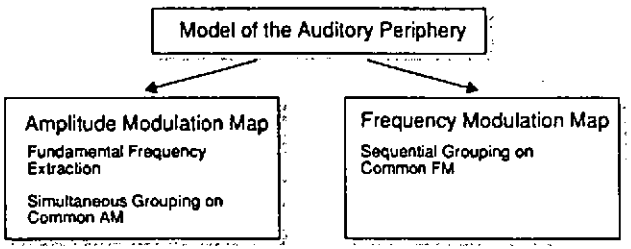

Figure 1. Summary of proposals for using modulation maps in a model of auditory processing.

Bregman *et al.* also define a second type of grouping, *sequential grouping*, which determines which spectral components have arisen over time from the same source (this is rather similar to the correspondence prob-

lem in vision). A map of FM could provide a description of the way in which spectral components are moving in time, which would form a basis for this type of grouping.

A summary of these proposals is given in Figure 1. We intend to incorporate AM and FM maps into an integrated model of auditory processing (Cooke *et al.* [6]), which will embody a wider range of grouping principles together with a post-streaming analysis.

4. THE MODEL

The model presented here is based firmly on the concept of a map for AM rate, but ignores some of the physiological details for reasons of efficiency. Neural maps are highly redundant, and contain many cells that are broadly tuned to similar values of the mapped parameter (Knudsen *et al.* [10]). This redundancy confers resistance to the topological variations that occur in biological neural networks, but is not a necessary feature of a computer model. Hence, we use a much smaller number of modulation detectors.

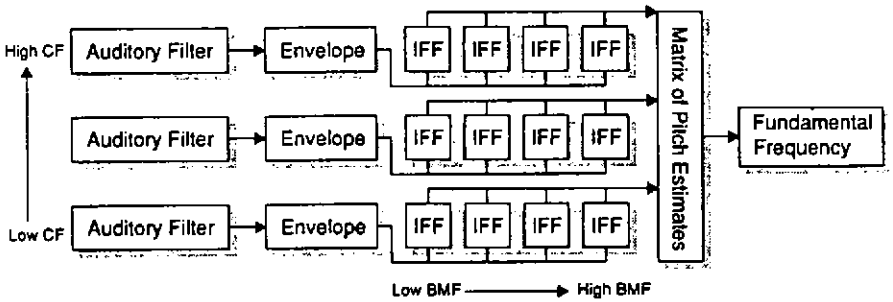

Figure 2. Schematic diagram of the model.

The structure of the model is shown in Figure 2. 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 [4]). We use an IIR filter 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 [1]), 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).

Our model of the periphery does not include a hair cell transduction stage, because adaptation of the auditory nerve has a negligible effect on the subsequent modulation extraction. 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, to reflect the physiological range of BMFs in the inferior colliculus (Schreiner and Langner [19]). Rather than using, for example, 450 filters tuned in 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. This is achieved by employing a form of the gammatone filter (which is used simply for convenience), from which the instantaneous frequency is calculated as

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

Here, *R(t)* and *I(t)* are the outputs of the real and imaginary parts of the gammatone filter (a derivation of this equation is given in Cooke [5]). By extracting the frequency components of each envelope in this way, we effectively determine the modulation frequencies present in each channel.

5. FUNDAMENTAL FREQUENCY EXTRACTION USING THE MODEL

Fundamental frequency is an important parameter for many speech processing applications, such as speaker verification and identification systems. This is reflected in the large number of algorithms that have been proposed for fundamental frequency estimation (Hess [9]).

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. This is achieved by summing the instantaneous amplitude 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 $\rho(f)$ with the largest value.

This approach bears some similarity to the correlogram proposed by Slaney and Lyon [22], which accounts for many of the attributes of psychophysical pitch. However, our map is purely a 'signal processing' approach that extracts fundamental frequency *per se*, and does not attempt to model the mechanisms of periodicity pitch. Consequently, although the map correctly predicts the pitch of harmonic complexes, it does not predict the psychophysical pitch of inharmonic stimuli.

Here, we test how the accuracy of the fundamental frequency estimates derived from the modulation map degrades in noisy conditions. For the purposes of a preliminary study, comparisons are made with an autocorrelation pitch detector. This is based on the autocorrelation method described by Rabiner and Shafer [14], and includes center clipping and simple logic to detect continuity errors such as pitch doubling. A voiced/unvoiced decision is made by thresholding the peak in the autocorrelation function. The modulation map does not currently employ a voicing detector, so only those frames that are declared voiced by the autocorrelation pitch detector are used in the comparison.

The comparison of the two techniques was made over a selection of 32 utterances from the TIMIT database, consisting of two male and two female speakers from each of the 8 dialect regions. Pitch contours were derived for the clean utterance, and for five noisy conditions. Random noise was added to each utterance to give signal-to-noise ratios (SNR) ranging from 10 dB to -10 dB, in 5 dB steps. The pitch tracks from both the autocorrelation and modulation map were median smoothed, using a window size of 3. In all tests, the autocorrelation pitch contour for the clean utterance was used as the reference for the comparison.

6. RESULTS

In order to quantify the error between the two pitch extraction techniques, we consider gross errors rather than small variations in the fundamental frequency contours. For a particular frame, a gross error is considered to have occurred when the quantity

EXTRACTING AMPLITUDE MODULATION FROM AN AUDITORY MODEL

$$E = 100 \times \frac{|f_{test} - f_{ref}|}{f_{ref}}$$

exceeds 10%. Here, f_{test} is the map or autocorrelation estimate for the pitch of the noisy utterance, and f_{ref} is a reference autocorrelation pitch estimate for the clean utterance, in which obvious continuity errors have been corrected manually. The percentage gross error for the whole pitch contour is then given by the sum of all the individual errors divided by the number of voiced frames. A frame length of 30 ms is used, and pitch estimates are calculated every 10 ms (hence there is a 20 ms overlap).

Currently, 16 auditory filter channels are used in the modulation map. Using more channels improves the smoothness of the pitch contours very slightly, but this is not significant enough to justify the extra computational expense.

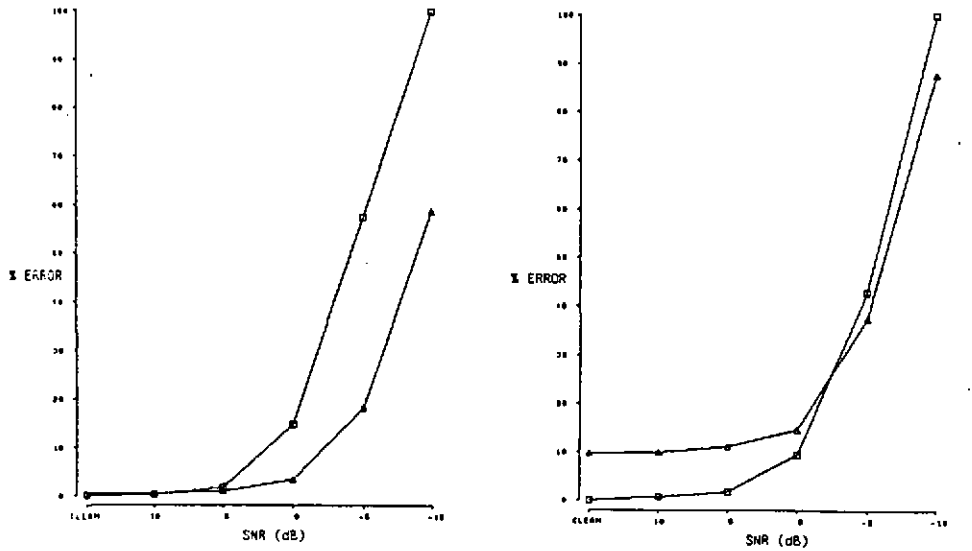

Figure 3. Deterioration of the map (triangles) and autocorrelation (squares) pitch detectors for the male (left) and female (right) speaker sets.

Figure 3 shows a plot of the gross percentage error between the two techniques, for the male and female speaker sets. For male speakers, the modulation map clearly shows a better performance in noisy conditions. For the female speaker set, the pitch estimates from the map have a higher initial error than the autocorrelation, but still degrade less in noise. Currently, we are unable to explain the poor initial performance of the map for female speakers, but it may be due to the use of autocorrelation as a reference.

7. FURTHER WORK

Further work will concentrate on implementing a map for FM sounds, and on making the AM map more physiologically accurate. For example, Rees & Palmer [15] report that the BMF of modulation sensitive neurons changes depending on the intensity of the stimulus and the presence of noise. Eventually, the models will be incorporated into an integrated simulation of auditory processing.

8. ACKNOWLEDGEMENTS

GJB is supported by SERC CASE award 88501484 and British Telecom Research Laboratories. MPC is supported by SERC award GR/E 42754. The assistance of Dr. William Millar is gratefully acknowledged.

9. REFERENCES

[1] E deBOER & H R deJONGH,'On cochlear encoding: Potentialities and limitations of the reverse-correlation technique', *JASA*, 63 pp115-135 (1978)

[2] A S BREGMAN, J ABRAMSON, P DOEHRING & C J DARWIN,'Spectral integration based on common amplitude modulation', *Perception & Psychophysics*, 37 pp483-493 (1985)

[3] A S BREGMAN, R LEVITAN, & C LIAO,'Fusion of auditory components: Effects of the frequency of amplitude modulation', *Perception & Psychophysics*, 47(1) pp68-73 (1990)

[4] M P COOKE,'The auditory periphery: Physiology, function and a computer model', Department of Computer Science Research Report, University of Sheffield (1989)

[5] M P COOKE,'Synchrony strands: An early auditory time-frequency representation', Department of Computer Science Research Report, University of Sheffield (1990)

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

[7] L DENG, C D GEISLER, & S GREENBERG,'A composite model of the auditory periphery for the processing of speech', *J Phonetics*, 16 pp93-108 (1988)

[8] D A GODFREY, N Y S KIANG, & B E NORRIS,'Single unit activity in the posteroventral nucleus of the cat', *J Comp Neur*, 162 pp247-268 (1975)

[9] W J HESS,'Pitch determination of speech signals: Algorithms and devices', Springer-Verlag, New York (1983)

[10] E I KNUDSEN, S duLAC, & S D ESTERLY,'Computational maps in the brain', *Ann Rev Neurosci*, 10 pp41-65 (1975)

[11] S McADAMS,'Spectral fusion, spectral parsing and the formation of auditory images', Ph.D. thesis, Stanford University (1984)

[12] A R MOLLER,'Dynamic properties of the responses of single neurons in the cochlear nucleus of the rat', *J Physiology*, 259 pp63-82 (1976)

[13] M J PONT,'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 (1990)

[14] L R RABINER & R W SHAFER,'Digital processing of speech signals', Prentice-Hall, New Jersey (1978)

[15] A REES & A R PALMER,'Neuronal responses to amplitude modulated and pure tone stimuli in the guinea pig inferior colliculus, and their modification by broadband noise', *JASA*, 85(5) pp1978-1994 (1989)

[16] A REES & A R MOLLER,'Responses of neurons in the inferior colliculus of the rat to AM and FM tones', *Hearing Research*, 10 pp301-330 (1983)

[18] C E SCHREINER & J V URBAS,'Representation of amplitude modulation in the auditory cortex of the cat. I. The anterior auditory field (AAF)', *Hearing Research*, 21 pp227-241 (1986)

[19] C E SCHREINER & G LANGNER,'Periodicity coding in the inferior colliculus of the cat. II. Topographical organization', *J Neurophysiology*, 60(6) pp1823-1840 (1988)

[20] S SENEFF,'A joint synchrony/mean-rate model of auditory speech processing', *J Phonetics*, 16 pp55-76 (1988)

[21] S A SHAMMA,'Speech processing in the auditory system II. Lateral inhibition and the central processing of speech evoked activity in the auditory nerve', *JASA*, 75(5) pp1622-1632 (1985)

[22] M SLANEY & R F LYON,'A perceptual pitch detector', *Proc ICASSP90*, pp357-360 (1990)

[23] I C WHITFIELD,'The object of the sensory cortex', *Brain Behav Evol*, 16 pp129-154 (1979)

[24] I C WHITFIELD & E F EVANS,'Responses of auditory cortical neurons to stimuli of changing frequency', *J Neurophysiology*, 28 pp655-672 (1965)

[25] Z L WU, P ESCUDIER & J L SCHWARTZ,'Specialized physiology-based channels for the detection of articulatory-acoustic events: A preliminary scheme and its performance', *Proc ICASSP89*, pp2013-2016 (1989)