A NEURAL OSCILLATOR MODEL OF PRIMITIVE AUDITORY GROUPING

Guy J. Brown and Martin Cooke

Department of Computer Science, University of Sheffield
Regent Court, 211 Portobello Street, Sheffield S1 4DP,
United Kingdom

Email: g.brown@dcs.shef.ac.uk, m.cooke@dcs.shef.ac.uk

Published in the Proceedings of the IEEE Signal Processing Society Workshop on Applications of Signal Processing to Audio and Acoustics, New York, October 15-18, 1995, pp. 53-56.

ABSTRACT

A computational model of primitive auditory scene analysis is described, in which the grouping of peripheral frequency channels is signalled by the pattern of temporal synchronisation in a network of neural oscillators. It is demonstrated that the model is able to group acoustic components according to their fundamental frequencies and onset times. Implications for models of pitch perception are discussed.

1. INTRODUCTION

Bregman's [2] recent account of audition holds that an *auditory scene analysis* is performed on the complex mixture of sounds reaching the ears. This process consists of two conceptually distinct stages. In the first stage, sound is decomposed into a collection of sensory elements (features). Subsequently, features that are likely to have arisen from the same acoustic source are grouped to form a perceptual whole, or 'stream'. Although Bregman's account finds much support in the psychophysical literature, the physiological mechanisms which underlie auditory grouping are far from clear. In particular, there have been few attempts to address the question of how groups of features are represented and communicated within the auditory system. This issue is the subject of the computational modelling study described in this article.

Physiological studies suggest that features are encoded in the higher auditory nuclei by maps of cells, in which frequency and some other parameter are represented on orthogonal axes (see [4] for a review). In normal listening environments, a number of sound sources will be simultaneously active and therefore the activity in these maps will represent the superimposed responses to several acoustic sources. Hence, auditory grouping requires that the responses of neurons which code for features of the same acoustic source should be linked. This issue has been called the *binding problem* [9].

One solution to the binding problem is the grouping of feature detectors according to the coherence (synchrony) of their temporal responses. Essentially, this scheme involves the multiplexing of neural activity; firing *rate* indicates the probability that a feature is present, whereas firing *synchrony* represents grouping [9]. This hypothesis is supported by recent physiological studies, which suggest that visual stimuli initiate synchronised neural oscillations in disparate regions of the visual cortex [8].

With the exception of the recent work of Wang [14], there have been few attempts to exploit temporal synchronisation in models of auditory grouping. In this paper, we describe an auditory model in which the grouping of peripheral channels is signalled by the pattern of temporal synchronisation within a network of neural oscillators. It is demonstrated that the oscillator network is able to group acoustic components which have the same fundamental frequency or have a common onset time.

2. THE MODEL

The proposed model consists of four stages; peripheral auditory processing, periodicity analysis (correlogram), oscillator network and attentional searchlight. For further details see [3].

2.1. Auditory Periphery

The frequency selective properties of the basilar membrane are modelled by a bank of gammatone filters [13], where each filter simulates the frequency response of a particular point along the cochlear partition. The studies described here employ a bank of 32 filters, with centre frequencies distributed between 100 Hz and 2 kHz on an ERB-rate scale [7]. The output of each filter is converted to a probabilistic representation of auditory nerve firing activity by the Meddis [10] model of inner hair cell transduction.

2.2. Correlogram

Recently, theories of pitch perception have been proposed which integrate periodicity information across resolved and unresolved harmonic regions. Models of this type are able to account for many psychophysical pitch phenomena, and are also able to explain the finding that a difference in fundamental frequency between two complex sounds can assist their perceptual segregation [11]. Here, we adopt one of this class of pitch models known as the *correlogram*, in which periodicities in the temporal fine structure of auditory nerve firing patterns are identified by an autocorrelation analysis at each characteristic frequency [4]. For an auditory filter channel f, the running autocorrelation a_f at time t and lag τ is given by

$$a_f(t,\tau) = \sum_{i=0}^{\infty} r_f(t-T)\, r_f(t-T-\tau)\, h(T) \tag{1}$$

where

$$T = i\,dt \tag{2}$$

Here, dt is the sample period (0.0625ms), $h(T)$ is a rectangular window of width 20 ms and r_f is the probability of firing activity in the auditory nerve, derived from the Meddis hair cell model.

Equation 1 was computed for values of τ between 0 ms and 20 ms in steps of *dt*.

2.3. Oscillator Network

In our scheme, it is assumed that source segregation is achieved by selective attention to the neural activity in groups of auditory filter channels. Grouping is encoded by the pattern of temporal synchronisation in a fully connected network of sine circle maps [1]. Each circle map models the phase dynamics of a neural oscillator, which signals the grouping between its corresponding auditory filter channel and other channels. For clarity, we refer to units in the network as 'neurons', although each is intended to reflect the activity of a collection of neurons rather than a single cell. The sine circle map is given by

$$\varphi(x) = x + \Omega + \tfrac{k}{2\pi}\sin(2\pi x) + \eta \pmod 1 \tag{3}$$

where the noise term η represents equally distributed random numbers in the interval $[0,10^{-9}]$. The new phase $\theta_i(t+1)$ of the neuron for channel *i* is computed by applying the circle map on the old phase $\theta_i(t)$ and on an input value υ_i, weighted by a coupling strength κ:

$$\theta_i(t+1) = \frac{1}{1+\kappa}\left[\varphi(\theta_i(t)) + \kappa\varphi(\upsilon_i(t))\right] \tag{4}$$

The input υ_i is related to the phases of the other neurons in the network, weighted by a coupling matrix *W*:

$$\upsilon_i(t) = \frac{\sum_j W_{ij}\theta_j}{\sum_j W_{ij}} \tag{5}$$

With parameter values κ=1.5, Ω=0.618 and *k*=5.0, neurons in the network exhibit chaotic oscillations. When the coupling strength W_{ij} between two neurons *i* and *j* is high, the cells show an identical phase response. Similarly, neurons that are weakly coupled give rise to uncorrelated time series.

Learning in the network is determined by periodicity detection in the correlogram. Initially, all neurons are strongly coupled (W_{ij}=1, for all i≠j). At intervals of 1 ms, a correlogram is computed and the autocorrelation functions for all channels are summed. The lag at which the largest peak occurs in this summary function, τ_p, is taken to be the pitch period of the sound source (see also [4]):

$$\tau_p = \max_{\tau} \sum_f a_f(t,\tau) \tag{6}$$

The channels of the correlogram are sampled at this lag, giving an estimate $a_f(t,\tau_p)$ of the strength to which each channel *f* is responding to the pitch period τ_p at time *t*. These pitch strengths are used to modify the coupling between neurons in the oscillator network, according to the following learning rule:

$$W_{ij}(t+1) = \Phi\left(1 - \lambda\left[1 - W_{ij}(t)\right] - \gamma\left|a_i(t,\tau_p) - a_j(t,\tau_p)\right|\right) \tag{7}$$

Here, γ determines the learning rate and λ determines the rate at which W_{ij} returns to its resting level. The function $\Phi(x)$, given by

$$\Phi(x) = \begin{cases} x & 0 \le x \le 1 \\ 1 & x > 1 \\ 0 & x < 0 \end{cases} \tag{8}$$

confines the coupling strength to the range [0,1]. Parameter values were set by inspection (γ=5×10^{-6}, λ=0.95).

The learning rule given in equation 7 embodies a principle that is closely related to the Gestalt heuristic of 'common fate'. This term describes the tendency to group sensory elements which change in the same way at the same time [2]. Similarly, our learning rule ensures that the coupling is reduced between channels that are not responding to the same fundamental frequency at the same time.

2.4. Attentional Searchlight

The last stage of the model is an attentional mechanism, inspired by Crick's [6] proposal for an attentional 'searchlight'. The searchlight takes the form of rapid bursts of firing in a subset of thalamic cells. When the thalamic neurons fire in synchrony with the oscillations of a neuronal group, that group becomes the attentional 'foreground' and other groups are relegated to the 'background'.

Currently, the attentional searchlight is not explicitly implemented in our computer model. Instead, we consider the temporal correlation between the activity of pairs of neurons in the oscillator network. The correlation between two time series $X(t)$ and $Y(t)$ is given by

$$C(X,Y) = \frac{\sum x(t)\,y(t)}{\sqrt{\sum x^2(t)\,\sum y^2}} \tag{9}$$

where $x(t)=X(t)-\langle X\rangle$ and $y(t)=Y(t)-\langle Y\rangle$. If a pair of neurons in the oscillator network are strongly coupled, the temporal correlation of their responses *C* will be high. Hence, an attentional searchlight that is synchronised to one of the cells will inevitably be synchronised to the other. Similarly, *C* will be low between neurons that are weakly coupled; in this case, the searchlight may synchronise to one cell or the other, but not both simultaneously. In Bregman's [2] terms, high temporal correlation between neurons indicates 'temporal coherence' and low correlation indicates 'streaming'.

3. SIMULATION RESULTS

In the remainder of this article, we present results from two simulations using the neural oscillator model (further results are also reported in [3]). The first demonstrates that the oscillator network is able to signal grouping by common fundamental frequency, and the second demonstrates grouping by common onset. The parameter values given in Section 2 were used for both simulations.

3.1. Grouping by Common Fundamental

The segregation of concurrent harmonic sounds can be likened to the sifting of partials through a 'harmonic sieve', which has 'holes' at integer multiples of its fundamental frequency. Moore *et al*. [12] have quantified the width of the holes in the harmonic sieve using a mistuning paradigm. They presented listeners with a harmonic complex in which one component was mistuned, so that its frequency was not an integer multiple of the fundamental. For mistunings of up to 3% of the harmonic frequency, the partial made a normal contribution to the pitch of the complex. Components mistuned by more than 3% began to be rejected by the harmonic sieve, and made a smaller contribution to the perceived pitch. Also, listeners heard partials that were mistuned by more than 2-3% as a separate sound source.

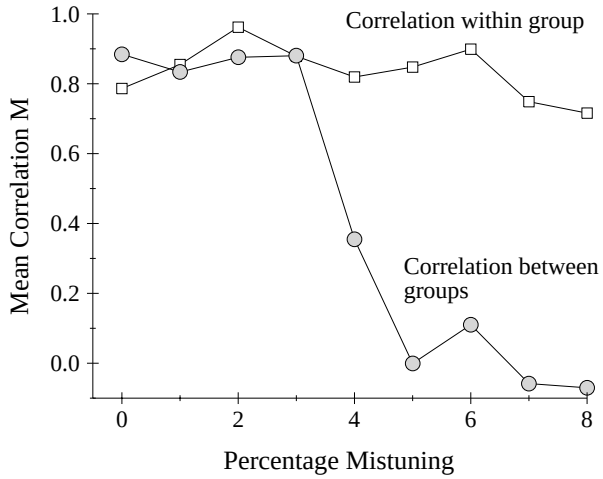

Figure 1: Model simulation showing the effect of mistuning on the mean correlation within and between groups in the neural oscillator network. See the text for details.

Figure 1 illustrates the response of the model to a stimulus of 90 ms duration, which consisted of the first 12 harmonics of a 155 Hz fundamental. The 4th partial of the stimulus was mistuned by a percentage of its harmonic frequency. The correlation between groups (grey circles) was computed by averaging the temporal correlation C between the oscillator channel centred on the 4th partial and the oscillator channels centred on the other harmonics of the complex. Formally,

$$M_{between} = \frac{1}{11} \sum_{i=1..3,5..12} C(h_i, h_4) \tag{10}$$

where h_i is the last 30 ms of the oscillator response for the peripheral channel centred on harmonic i of the stimulus. Similarly, the within-group correlation (open squares) is given by

$$M_{within} = \frac{1}{10} \sum_{i=2,3,5..12} C(h_i, h_1) \tag{11}$$

It is clear from Figure 1 that the model is in qualitative agreement with the psychophysical data; mistunings of up to 3% maintain a high correlation between neurons in the oscillator network, indicating that all harmonics of the complex have been allocated to the same perceptual group. At mistunings of greater than 3% however, there is a significant reduction in correlation between the oscillator centred on the 4th partial and the remaining oscillators; this indicates that the mistuned 4th harmonic has been perceptually segregated from the other components of the complex.

The response of the oscillator network to a mistuned component can be explained by consideration of the learning rule in equation 6. Recall that the amount of weight change between two neurons is related to the difference in peak height in their corresponding correlogram channels, measured at the pitch period of the source. For a small mistuning, the channels of the correlogram near to the 4th harmonic still exhibit a peak at the pitch period of the complex; hence, the weights in the oscillator network remain unchanged. For larger mistunings, however, this peak is reduced in height and therefore the weights to the oscillators in the region of the 4th harmonic are decreased. Consequently, the oscillators coding the 4th partial desynchronise from the rest of the network, indicating perceptual segregation. This desynchronisation is clearly illustrated in Figure 2, which shows the response of the oscillator network for the 8% mistuning condition.

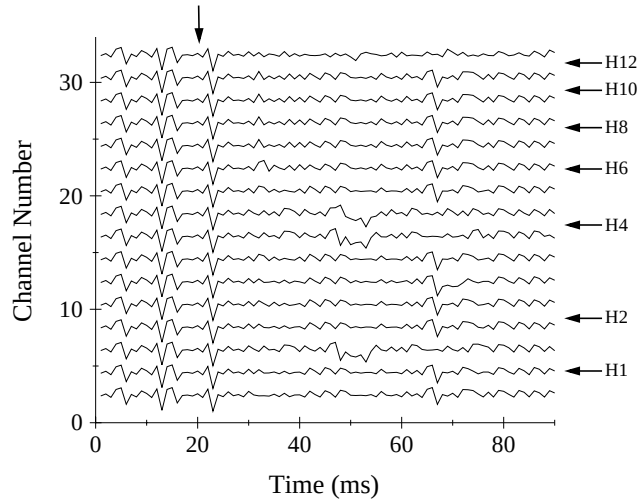

Figure 2: Neural oscillator response to a stimulus consisting of the first 12 partials of a 155 Hz fundamental, in which the 4th partial (H4) has been mistuned by 8% of its harmonic frequency. The stimulus was preceded by 20 ms of silence, during which all oscillators remain synchronised. Following the onset of the harmonic complex (arrow), oscillators near to the mistuned component rapidly desynchronise from the other neurons in the network.

3.2. Grouping by Common Onset

The role of onset asynchrony in perceptual grouping has been investigated by Ciocca and Darwin [5]. They asked listeners to judge the pitch of a harmonic complex in which one of the resolved harmonics was mistuned. When the mistuned partial started 160 ms before the other components of the complex, it made a reduced contribution to the perceived pitch. Its contribution was abolished if it started more than 300 ms before. Further, it was confirmed that this effect was due to perceptual grouping rather than peripheral adaptation.

Figure 3 illustrates the response of the oscillator model to stimuli similar to those used by Ciocca and Darwin. The stimuli consisted of the first 12 harmonics of a 155 Hz fundamental, and had a duration of 90 ms. The 4th partial started between 50ms and 300 ms before the other components. The figure shows the mean correlation M between oscillators in the same group and in different groups, using the metrics given in equations 10 and 11. It is clear that the oscillators coding the 4th harmonic desynchronise from the other neurons in the network as the onset asynchrony is increased, indicating that the leading partial is perceptually segregated. This result follows directly from the learning rule given in equation 6, which reduces the weight between neural oscillators that do not exhibit a peak in their corresponding correlogram channels at the same time.

4. DISCUSSION

A model of primitive auditory grouping has been described, in which groups are signalled by the pattern of temporal synchronisation in a network of neural oscillators. The model is in qualitative

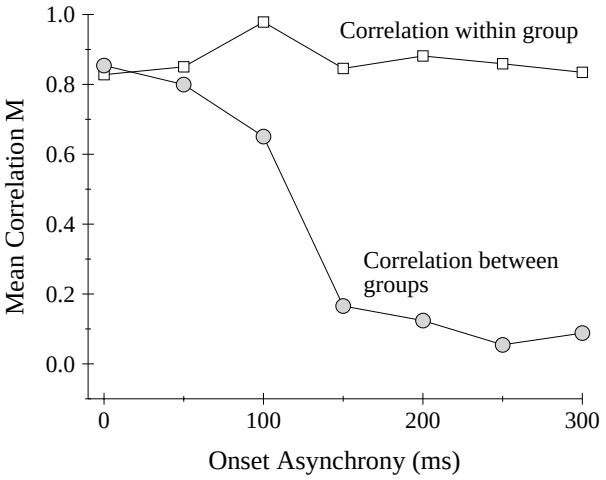

Figure 3: Model simulation showing the effect of onset asynchrony on the mean correlation within and between groups in the neural oscillator network. See the text for details.

agreement with psychophysical findings on perceptual grouping by onset asynchrony and common fundamental frequency. Related studies have also shown that a neural oscillator model is able to explain auditory grouping by frequency proximity and temporal proximity [3].

A notable feature of the model is that the neural oscillators are strongly coupled in their resting state (see Figure 2). Hence, the default condition of organisation in the model is fusion; all components of the stimulus are assumed to have originated from the same acoustic source unless there is a reason to segregate them. There is some empirical evidence to support this stance; for example, a burst of white noise contains random cues for fusion and segregation, but is perceived as a single coherent source [2].

The simulations reported in Section 2 illustrate an intriguing paradox. In the first example, a partial was excluded from an auditory group when it was sufficiently mistuned; that is, pitch analysis determined perceptual grouping. However, in the second case perceptual grouping determined pitch; a partial of a harmonic complex that started earlier than the other components was excluded from the pitch percept. It appears, then, that pitch analysis can both follow auditory grouping and contribute to it.

Figure 4 shows a modification of the neural oscillator model which might explain this phenomenon. This scheme employs a processing loop, in which coupling adjustments based on the current pitch period are fed forward to the oscillator network, and channel weights are fed back to the correlogram to influence the computation of the next pitch period. The channel weights fed back from the network would reflect the contribution of each channel to the group in the attentional foreground, which in turn would be determined by the thalamic searchlight. Work on this model is currently in progress.

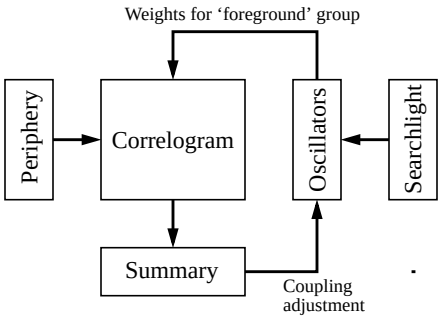

Figure 4: A modified neural oscillator model.

REFERENCES

1. Bauer, M. and Martienssen, W., "Coupled circle maps as a tool to model synchronisation in neural networks", *Network*, Vol. 2, 1991, pp. 345-351.

2. Bregman, A.S., *Auditory Scene Analysis*, MIT Press, Cambridge, 1990.

3. Brown, G.J. and Cooke, M.P., "Temporal synchronisation in a neural oscillator model of primitive auditory stream segregation", *IJCAI Workshop on Computational Auditory Scene Analysis*, Montreal, 19-20th August, 1995.

4. Brown, G.J. and Cooke, M.P., "Computational auditory scene analysis", *Computer Speech and Language*, Vol. 8, 1994, pp. 297-336.

5. Ciocca, V., and Darwin, C.J., "Effects of onset asynchrony on pitch perception: Adaptation or grouping?", *Journal of the Acoustical Society of America*, Vol. 93, 1993, pp. 2870-2878.

6. Crick, F., "Function of the thalamic reticular complex: The searchlight hypothesis", *Proceedings of the National Academy of Sciences USA*, Vol. 81, 1984, pp. 4586-4590.

7. Glasberg, B. and Moore, B.C.J., "Derivation of auditory filter shapes from notched-noise data", *Hearing Research*, Vol. 47, 1990, pp. 103-138.

8. Gray, C.M., König, P., Engel, A.K. and Singer, W., "Oscillatory responses in cat visual cortex exhibit inter-columnar synchronisation which reflects global stimulus properties", *Nature*, Vol. 338, 1989, pp. 334-337.

9. Malsburg, von der C., and Schneider, W., "A neural cocktail party processor", *Biological Cybernetics*, Vol. 54, 1986, pp. 29-40.

10. Meddis, R., "Simulation of auditory-neural transduction: Further studies", *Journal of the Acoustical Society of America*, Vol. 83, 1988, pp. 1056-1063.

11. Meddis, R. and Hewitt, M.J., "Modelling the identification of concurrent vowels with different fundamental frequencies", *Journal of the Acoustical Society of America*, Vol. 91, 1992, pp. 233-245.

12. Moore, B.C.J., Glasberg, B. and Peters, R.W., "Relative dominance of individual partials in determining the pitch of complex tones", *Journal of the Acoustical Society of America*, Vol. 77, 1985, pp. 1853-1860.

13. Patterson, R.D., Holdsworth, J., Nimmo-Smith, I. and Rice, P., *Implementing a Gammatone Filterbank*, APU Report 2341, Cambridge, 1988.

14. Wang, D., "Modelling global synchrony in the visual cortex by locally coupled neural oscillators", *Proceedings of the 15th Annual Conference of the Cognitive Science Society*, Boulder, CO, 1993, pp. 1058-1063.