A Neural Oscillator Model of Auditory Attention

Stuart N. Wrigley and Guy J. Brown

Speech and Hearing Research Group, Department of Computer Science,
University of Sheffield, Regent Court, 211 Portobello Street, Sheffield S1 4DP, UK
{s.wrigley,g.brown}@dcs.shef.ac.uk
http://www.dcs.shef.ac.uk/~stu
Tel: +44 (0) 114 222 1879 Fax: +44 (0) 114 222 1810

Abstract. A model of auditory scene analysis is proposed, which incorporates an attentional mechanism and is implemented using a network of neural oscillators. The core of the model is a two-layer neural oscillator network which performs stream segregation and selection on the basis of oscillatory correlation. A stream is represented by a sychronised oscillator population, whereas different streams are represented by desynchronised oscillator populations. The output of the model is an 'attentional stream' describing which parts of the auditory scene are in the attentional foreground. The model simulates a number of perceptual phenomena including two tone streaming and the capture of a tone from a harmonic complex.

1 Introduction

In typical listening situations, a mixture of sounds reaches our ears: for example, a party with multiple concurrent conversations, a musical recording or a busy urban environment. Despite this, human listeners can selectively attend to a particular acoustic source, suggesting that they can separate the complex mixture into its constituent components. It has been convincingly argued that the acoustic signal is subjected to a similar form of scene analysis as occurs in vision [2]. Such *auditory scene analysis* (ASA) takes place in two conceptual stages. Firstly, the signal is decomposed into discrete sensory elements. These are then recombined into perceptual *streams* on the basis of the likelihood of them having arisen from the same physical source.

In common usage, the term *attention* usually refers to both selectivity and capacity limitation. It is widely accepted that conscious perception is selective; in other words, we perceive only a small fraction of the information impinging upon the senses. The second phenomenon - that of capacity limitation - can be illustrated by the fact that two tasks when performed individually pose no problem; however, when they are attempted simultaneously, they become very difficult. It would appear, therefore, that attention is a finite resource [5].

In this study we propose a conceptual model [14] of how auditory attention influences auditory scene analysis. Following [7], we propose that attention can be split

G. Dorffner, H. Bischof, and K. Hornik (Eds.): ICANN 2001, LNCS 2130, pp. 1163–1170, 2001.

into two mechanisms: unconscious and conscious allocation (*exogenous* and *endogenous*, respectively). One or more exogenous processes are responsible for grouping of stimuli reaching the ears. The perceptual organisations of these processes are directed to the endogenous selection mechanism. It is at this stage that both conscious decisions and salient information about the incoming organisations are combined, and a selection made which allows for the occasional over-ruling of endogenous mechanisms by exogenous attention (e.g. a sudden loud bang or other 'survival situations'). Only a single perceptual stream can be selected by the mechanism, which simulates the capacity limitation of attention.

The model is based upon the oscillatory correlation theory ([12]; see also [10]), in which neural oscillators representing a single stream are synchronised, and are desynchronised from oscillators representing other streams. The model consists of two stages: peripheral processing and a network of neural oscillators.

2 Peripheral Auditory Processing

Peripheral auditory processing is simulated by a bank of bandpass filters, followed by a model of inner hair cell transduction. Specifically, the frequency selectivity of the basilar membrane is modelled by a gammatone filterbank, in which the output of each filter represents the frequency response of the membrane at a specific position [6]. The gammatone filter is based on an analytical approximation to physiological measurements of auditory nerve impulse responses. The gammatone filter of order n and centre frequency f_0 Hz is given by

$$gt(t) = t^{n-1} e^{-2\pi bt} \cos(2\pi f_0 t + \phi) H(t) \ . \tag{1}$$

where ϕ represents the phase and $H(t)$ is the unit step (Heaviside) function for which $H(t) = 1$ if $t \geq 0$ and $H(t) = 0$ otherwise. The bandwidth b is set to the equivalent rectangular bandwidth (ERB), a psychophysical measurement of critical bandwidth in human subjects [4]. A bank of 32 filters were used with centre frequencies equally distributed on the ERB scale between 50 Hz and 5 kHz. The gain of each filter is adjusted to simulate the pressure gains of the outer and middle ears.

Our oscillator network requires an estimate of firing rate in each auditory filter channel as its input. Accordingly, the envelope of each gammatone filter response is computed, and this is half-wave rectified and compressed to give an approximation of auditory nerve firing rate.

3 Oscillator Network

The three conceptual stages of attentionally modulated computational auditory scene analysis (segmentation, grouping, attentional stream selection) occur within an oscillatory correlation framework (Fig. 1A).

The building block of the network is a single relaxation oscillator [11], which is defined as a reciprocally connected excitatory variable x and inhibitory variable y:

$$\dot{x} = 3x - x^3 + 2 - y + I + S + \eta \ . \tag{2}$$

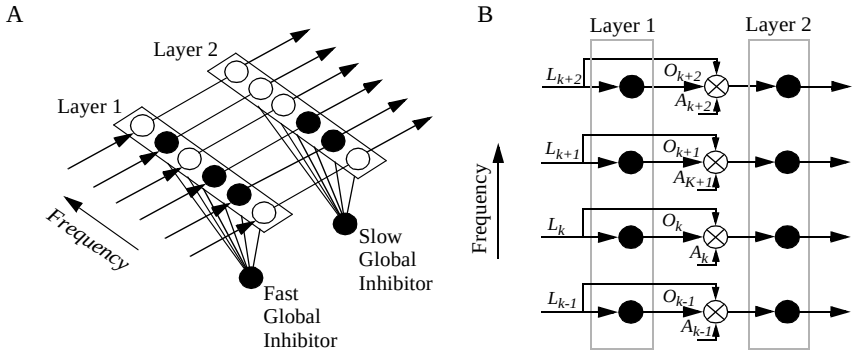

Fig. 1. A. A one-dimensional, two-layer oscillator network performs segmentation and grouping (layer one) followed by stream selection (layer two) to produce an 'attentional stream'. **B.** Section of the network showing how the output of the layer one activity, stimulus intensity and attention highlight are combined. For clarity, excitatory and inhibitory connections within layers are not shown.

$$\dot{y} = \varepsilon\left[\gamma\left(1 + \tanh\frac{x}{\beta}\right) - y\right]. \tag{3}$$

Here, ε, γ and β are parameters, I represents the oscillator's external input, S represents the coupling from other oscillators in the network, and η is a noise term. Ignoring S and η, constant I defines a relaxation oscillator with two time scales. The x-nullcline is a cubic function and the y-nullcline is a sigmoid function. When $I>0$, the two nullclines intersect only at the middle branch of the cubic in which case the oscillator exhibits periodic activity (a stable limit cycle). When $I<0$, the nullclines intersect at a stable fixed point and no activity occurs. The oscillator activity is therefore stimulus dependent.

3.1 Layer One: Segmentation and Grouping

The first layer of the network is based upon a one-dimensional locally excitatory globally inhibitory oscillator network (LEGION [8]) composed of relaxation oscillators. Synchronised blocks of oscillators (*segments*) form in this layer in which each segment corresponds to a contiguous region of acoustic energy. The input I to layer one is the estimate of auditory nerve firing rate in each channel. To restrict the spread of activation across frequency, the firing rate is adjusted such that rates below a threshold value are set to zero.

Layer one is a one-dimensional network of neural oscillators with a global inhibitor. The coupling term S in (2) is given by:

$$S = \sum_{k \neq i} W_{ik} H(x_k - \theta_x) - W_z H(z - \theta_z). \tag{4}$$

Here, W_{ik} is the connection weight between oscillators i and k, and H is the Heaviside function. The parameter θ_x is a threshold above which an oscillator can affect others in the network and W_z is the weight of inhibition from the global oscillator z whose activity is defined as:

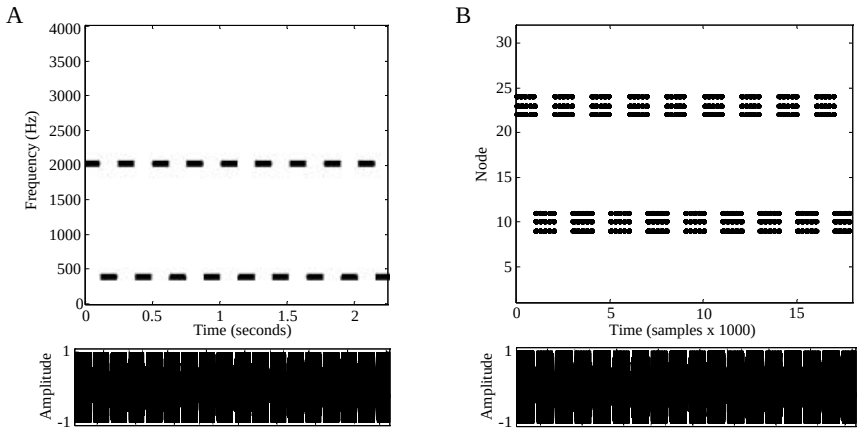

Fig. 2. A. Stimulus spectrogram showing tones at 400Hz and 2kHz. **B.** Layer 1 output showing segments corresponding to tones centred on channels 10 and 23. The time-domain representation of the stimulus is shown below each plot for reference. The parameter values are: ε=0.1, γ=6.0, β=0.1.

$$\dot{z} = \sigma_\infty - z\,. \tag{5}$$

Here, the input to the global inhibitor $\sigma_\infty = 1$ if $x_k \geq \theta_z$ for at least one oscillator k, otherwise $\sigma_\infty = 0$. Similar to θ_x in (4), θ_z acts a threshold above which an oscillator can affect the global inhibitor. If the global inhibitor receives supra-threshold stimulation from at least one oscillator, $z \rightarrow 1$.

Initially, all oscillators have the same phase (all frequencies belong to a single stream) - an assumption supported by psychophysical evidence which suggests that perceptual fusion is the default perceptual organisation [2].

An excitatory connection weight is placed between adjacent oscillators whose corresponding auditory nerve firing rates exceed a threshold value to generate segments. This allows contiguous regions of acoustic energy to be represented by sychronised groups of oscillators.

Figure 2B shows the response over time of the first layer to a repeating sequence of alternating high and low frequency tones (Fig. 2A). Each dot indicates that the oscillator for that channel is active at that time step. From this, it can be seen that each tone generates a synchronised block of oscillators.

3.2 Layer Two: Attentional Stream Selection

The oscillators in the first layer are connected to oscillators in the second layer by excitatory links. The strength of these connections is modulated by endogenous attention, thus allowing multiple frequency regions to dominate the network. The second layer is also a LEGION, but in this case the global inhibitor operates on a much

slower timescale. This causes layer two to act as a winner take all (WTA) network (see [12]). The activity of the layer two global inhibitor z_s is defined as:

$$\dot{z}_s = [\alpha - z_s]^+ - cz_s \ . \tag{6}$$

Here, $[n]^+ = n$ if $n \geq 0$ and $[n]^+ = 0$ otherwise; α is the total activity of the layer two network. This combined with the small value of c result in a quick rise and slow decay for the inhibitor. It is the slow decay that allows this network to behave as a winner take all network. When a group of oscillators are in the active phase, they feed input to the slow inhibitor. The slow inhibitor through its quick rise and slow decay maintains a level of inhibition that must be overcome by a group if the group is to oscillate. The central idea for object selection is that one group sets the level of slow inhibition, which can then be overcome by that group only. The key is to choose an appropriate value for c in (6).

Input to oscillator k in the layer is a weighted version of the corresponding layer one output:

$$I_k = [L_k - T_k]^+ O_k \ . \tag{7}$$

Here, T_k is the attentional threshold which is related to the endogenous interest at frequency k; O_k is the activity of oscillator k in layer one and L_k is the thresholded firing rate (see section 3.1) of the stimulus at frequency k (Fig. 1B).

An important property highlighted by [1] is the build up over time of the two tone streaming effect. It is proposed that the level of endogenous interest builds over time in relation to the input to the network. Attention can be reoriented consciously to other segregated streams (without further buildup); when input ceases, endogenous interest decays such that a second presentation of a stimulus will require a second buildup of streaming. Such behaviour is achieved by use of a leaky integrator with slow rise and decay time constants to weight the endogenous interest:

$$T_k = (1 - A_k)S \ . \tag{8}$$

Here, A_k is the endogenous interest at frequency k and S is the leaky integrator defined as:

$$\dot{S} = d([\sigma_\infty - S]^+ - [1 - H(\sigma_\infty - S)]aS) \ . \tag{9}$$

Here, small values of a and d result in a slow rise and slow decay for the integrator. The time-varying activity of the layer two network can be thought of as an *attentional stream*: a trace through time highlighting which stream is in the attentional foreground at any one time.

L_k is included in (7) to model the over-riding of endogenous attention by loud acoustic events. In normal circumstances, the endogenously chosen group (using the A_k) is the only group that can overcome the level of inhibition present in the network. However, should a sufficiently loud stimulus appear outside the area of interest defined by A_k, this will overcome the inhibition and so re-orientate attention exogenously.

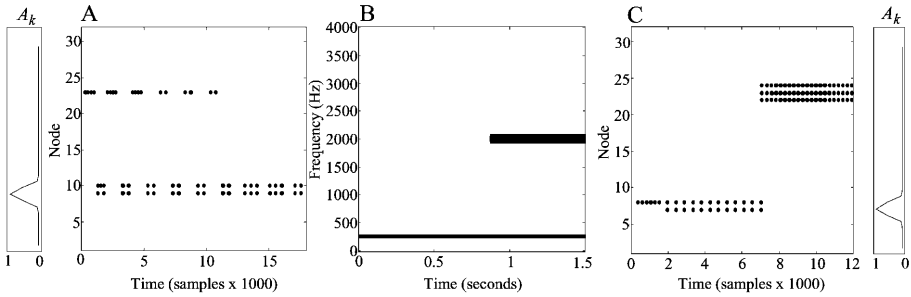

Fig. 3. A. The attentional stream of layer two shows the build-up of streaming [1] with the low frequency stream dominating by time step 11000. **B.** Stimulus spectrogram showing a tone at 300Hz and a much louder tone at 2kHz. **C.** The attentional stream of layer two shows the exogenously redirection of attention by the onset of the louder tone. Attention is directed toward the low frequency tone. Attentional interest A_k is shown beside each layer two diagram.

Figure 3A shows the response over time of the second layer to the repeating alternating tone sequence used earlier (Fig. 2). Again, each dot indicates that the oscillator for that channel is active at that time step. The relationship between frequency proximity and temporal proximity has been studied extensively using the two tone streaming phenomenon [9]. The more distant in frequency two tones are, the more likely it is that they will segregate into two streams with one dominating the other. Similarly, as presentation rate increases, tones of similar frequency group together. From the diagram, it can be observed that activity for the low frequency tones is sustained whereas activity for the high frequency tones gradually decreases until all high frequency oscillator activity ceases. A build-up of streaming has occurred culminating in the low frequency tones being in the attentional foreground. In this case, attention was oriented via the A_k values in (8) to the low frequency tones.

Figure 3C shows how the attentional stream can be subconsciously re-directed by the onset of a new, loud stimulus (Fig. 3B). Attention is directed toward the low frequency tone but when the loud tone begins, this overrides the conscious decision and the new tone becomes the attention foreground. This is achieved by having a very low level of attentional interest across all frequencies which can only be overcome by loud stimuli.

Figure 4 shows the layer two responses to the stimulus used in [3] in which captor tones are used to affect the percept of a two tone complex. Listeners reported that the higher frequency tone of the complex was captured when the frequency separation between the captor tones and the complex was small. This result can been seen in the responses of Figure 4 in which the captor tones capture the high frequency complex tone into a separate stream only for small frequency separations.

4 Discussion

A model of attentionally modulated auditory scene analysis has been described, in which stream segregation and selection occur within an oscillatory framework. The

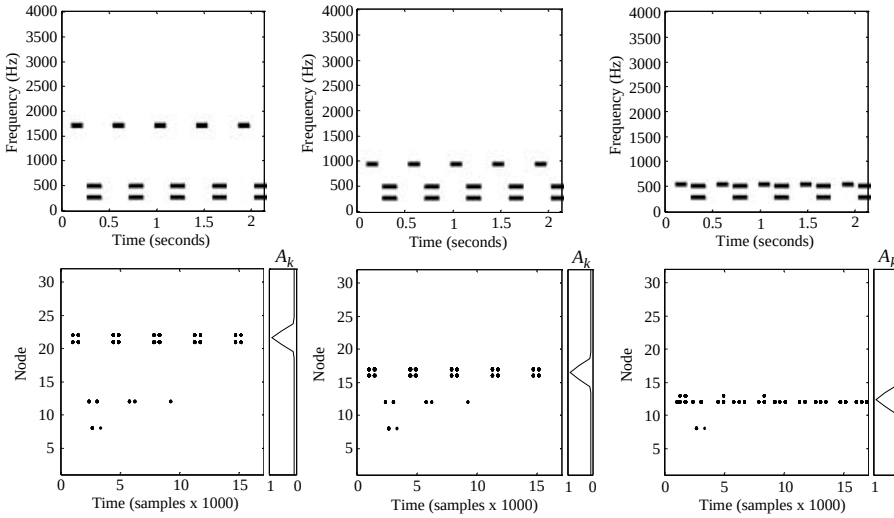

Fig. 4. The attentional stream of layer two (bottom) in response to the stimuli used in [3] (top). Attentional interest A_k is shown to the right of each layer two diagram.

model is based on the previous neural oscillator work of Wang and colleagues [11][13] but incorporates a number of novel factors. First, we use a one-dimensional layer of oscillators, in contrast to the two-dimensional time-frequency grid used in Wang's model. The latter suffers from conceptual difficulties. The input to Wang's two-dimensional network is sampled at regular intervals, such that the network produces a finite duration 'snapshot' of the auditory scene at each epoch. It is unclear how these snapshots should be integrated to produce a time-varying estimate of stream content. The model proposed here produces an attentional stream in continuous time; thus, there is no need to deal with the difficult problem of combining overlapping temporal snapshots. Furthermore, sequential grouping in Wang's model is only possible via lateral connections on the time axis, which do not have a known physiological correlate. Sequential grouping is an emergent property of the model described here; segments occurring over time which appear in the attentional stream are implicitly grouped.

The main contribution of this study is the incorporation of attentional mechanisms within the oscillatory framework. Little work has been conducted on auditory models with an attentional component. Wang's 'shifting synchronisation' theory [11] assumes that '*attention is paid to a stream when its constituent oscillators reach their active phases*'. Such stream multiplexing contradicts experimental findings [2], which show that listeners perceive one dominant stream. Furthermore, his theory cannot explain how attention can be redirected by a sudden stimulus. Such an event would be encoded by Wang's network as an individual stream which would be multiplexed as normal - with no attentional emphasis. As the shifting synchronisation theory allows attention to alternate quickly between each stream in the entire temporal snapshot, it is possible for attention to switch from a stream at the beginning of the snapshot to one at the end (later in time) and back again. This suggests a confusion in Wang's model between 'oscillator

time' and 'real time'. Attention exists in real time - but in Wang's system it shifts in oscillator time.

The model proposed here accounts for a number of psychophysical phenomena including the streaming of alternating tone sequences [9] and its associated build-up over time [1]. An attentional stream can also be re-directed by the onset of a new, loud stimulus, and the model demonstrates other perceptual phenomena such as the capture of a tone from a complex [3]. Current work is concentrating on using a pitch analysis technique to allow segments to be created by cross-channel correlation and groups to be formed by common periodicity. The model will then be expanded to account for binaural streaming phenomena.

We would like to thank DeLiang Wang and the three anonymous reviewers for comments on a previous version of this paper. The work of S. N. Wrigley is supported by the University of Sheffield Hossein Farmy Scholarship.

References

[1] Anstis, S., Saida, S.: Adaptation to auditory streaming of frequency-modulated tones. Journal of Experimental Psychology: Human Perception Performance **11** (1985) 257-271

[2] Bregman, A.S.: Auditory Scene Analysis. The Perceptual Organization of Sound. MIT Press (1990)

[3] Bregman, A.S., Pinker, S.: Auditory streaming and the building of timbre. Canadian Journal of Psychology **32**(1) (1978) 19-31

[4] Glasberg, B.R., Moore, B.C.J.: Derivation of auditory filter shapes from notched-noise data. Hearing Research **47** (1990) 103-138

[5] Pashler, H.E.: The Psychology of Attention. MIT Press (1998)

[6] Patterson, R. D., Nimmo-Smith, I., Holdsworth, J., Rice, P.: APU Report 2341: An Efficient Auditory Filterbank Based on the Gammatone Function, Cambridge: Applied Psychology Unit (1988)

[7] Spence, C.J., Driver, J.: Covert spatial orienting in audition: exogenous and endogenous mechanisms. Journal of Experimental Psychology: Human Perception and Performance **20**(3) (1994) 555-574

[8] Terman, D., Wang, D.L.: Global competition and local cooperation in a network of neural oscillators. Physica D **81** (1995) 148-176

[9] van Noorden, L.P.A.S.: Temporal coherence in the perception of tone sequences. Doctoral thesis, Institute for Perceptual Research, Eindhoven, NL (1975)

[10] von der Malsburg, C., Schneider, W.: A neural cocktail-party processor. Biological Cybernetics **54** (1986) 29-40

[11] Wang, D.L.: Primitive auditory segregation based on oscillatory correlation. Cognitive Science **20** (1996) 409-456

[12] Wang, D.L.: Object selection based on oscillatory correlation. Neural Networks **12** (1999) 579–592

[13] Wang, D.L., Brown, G.J.: Separation of speech from interfering sounds based on oscillatory correlation. IEEE Transactions on Neural Networks **10** (1999) 684-697

[14] Wrigley, S.N.: A Model of Auditory Attention. Technical Report CS-00-07, Department of Computer Science, University of Sheffield, UK (2000). Available from http://www.dcs.shef.ac.uk/~stu/pubs.html