A Computational Model of Auditory Selective Attention

Stuart N. Wrigley, *Member, IEEE,* and Guy J. Brown

***Abstract*—The human auditory system is able to separate acoustic mixtures in order to create a perceptual description of each sound source. It has been proposed that this is achieved by an auditory scene analysis (ASA) in which a mixture of sounds is parsed to give a number of perceptual streams, each of which describes a single sound source. It is widely assumed that ASA is a precursor of attentional mechanisms, which select a stream for attentional focus. However, recent studies suggest that attention plays a key role in the formation of auditory streams. Motivated by these findings, this paper presents a conceptual framework for auditory selective attention in which the formation of groups and streams is heavily influenced by conscious and subconscious attention. This framework is implemented as a computational model comprising a network of neural oscillators, which perform stream segregation on the basis of oscillatory correlation. Within the network, attentional interest is modeled as a Gaussian distribution in frequency. This determines the connection weights between oscillators and the attentional process, which is modeled as an attentional leaky integrator (ALI). Acoustic features are held to be the subject of attention if their oscillatory activity coincides temporally with a peak in the ALI activity. The output of the model is an "attentional stream," which encodes the frequency bands in the attentional focus at each epoch. The model successfully simulates a range of psychophysical phenomena.**

***Index Terms*—Attention, auditory scene analysis, binding problem, neural oscillator, temporal correlation.**

I. Introduction

IN ORDER to make sense of a complex environment, we are able to selectively attend to the features corresponding to a single object (be it visual or auditory). For example, although a mixture of sounds usually reaches the ears, a human listener can "pick out" a particular acoustic source from the mixture, such as a voice or a musical instrument. Such an ability is demonstrated by Cherry's well known study on the separability of concurrent conversations [1, p.976]. He found that when two messages were recorded by the same speaker and replayed simultaneously to a listener, *"the result is a babel, but nevertheless the messages may be separated."* This ability is colloquially termed the *cocktail party effect*.

To explain this, Bregman [2] proposes that the acoustic environment is subjected to an *auditory scene analysis* (ASA), which takes place in two stages. First, the signal is decomposed into a number of discrete sensory elements. Those elements likely to have arisen from the same acoustic source are then recombined into a perceptual *stream*, in a process termed *auditory grouping*. Therefore, a stream can be considered to be a cognitive representation of a sound source.

Bregman makes a distinction between two different, but complementary, mechanisms involved in auditory grouping. The first is *primitive grouping*, in which decisions on how to group sensory elements are made in a purely data-driven manner. Primitive grouping principles are believed to be innate and are well described by the Gestalt principles of perceptual organization such as common fate and good continuity [3]. In contrast, a second grouping mechanism, termed *schema-driven grouping*, employs prior knowledge of commonly experienced acoustic stimuli such as speech and music. For example, a schema can process ambiguous speech before conscious perception occurs, even when the disambiguating word occurs much later in the sentence [4].

II. Auditory Selective Attention

In common usage, the term "attention" usually refers to both selectivity and capacity limitation. It is widely accepted that conscious perception is selective, and that it encompasses 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 difficult. This occurs even when the two tasks are not physically incompatible, such as reading a book and listening to the radio. In turn, this leads to the common conclusion that attention is a finite resource.

Attention can be directed to a site of interest identified by some form of cueing. Mondor and Bregman [5] investigated this ability using a paradigm in which listeners were asked to indicate whether a target tone was longer or shorter in duration than a cue tone. It was found that performance declined as frequency separation between the cue and target tones increased. This implies that judgements about specific features of an auditory stimulus may be facilitated by orienting attention to the frequency at which the stimulus occurred. It was also found that increasing the duration of the cue-target interval improved performance, suggesting that a finite amount of time is required before attention is fully allocated to a particular frequency region. Similar findings were obtained by [6] when using tones of differing spatial location rather than differing frequency.

The "shape" of attentional deployment has also been the subject of research. Two general classes of model have been proposed to describe the focus of attention. Spotlight models propose that attention is allocated to a discrete range of frequencies with an even distribution within this range [7]. The edges of this spotlight are characterized by a sharp demarcation between

Manuscript received June 2, 2003; revised December 15, 2003.

The authors are with the Department of Computer Science, University of Sheffield, Sheffield S1 4DP, U.K. (e-mail: s.wrigley@dcs.shef.ac.uk; g.brown@dcs.shef.ac.uk).

Digital Object Identifier 10.1109/TNN.2004.832710

attended and unattended frequencies. Alternatively, the attentional focus may be defined as a gradient [8] in which the density of the attentional resources is greatest at the cued frequency and declines gradually with frequency separation from the focal point of attention. The latter is supported by psychophysical evidence [5], [6] in which listener performance was related to frequency separation: as frequency separation increased, so too did the response time. A model incorporating a spotlight of attention with abrupt changes between attended and unattended frequencies could not account for this result.

Finally, it has been suggested that auditory (and visual) attention may be oriented by two different mechanisms, which rely on differing amounts of conscious intervention by the listener [9]. The *exogenous* system is considered to take place automatically under pure stimulus control; attention is drawn to the site of the stimulus. *Endogenous* attention is considered to be under control of the listener, such that attention can be consciously oriented to a particular site or percept (such as a voice). In other words, the studies investigating frequency sensitivity [5] and spatial sensitivity effects [6] described above are, in fact, examining the allocation of endogenous attention.

A. Attention and Auditory Grouping

Consider the cocktail party effect [1] in which a listener has the task of following a conversation in a noisy environment. It is likely that the process of selective attention is assisted by the speaker's voice having some acoustic properties which separate it from the other voices. Because these factors are similar to ones involved in primitive stream segregation, for example, differences in fundamental frequency (F0), it can be argued that stream segregation is a form of selective attention. Bregman [2] rejects this view; rather, he regards stream segregation as being largely the result of grouping by a preattentive mechanism. In support of this, Bregman cites an experiment by Bregman and Rudnicky [10] in which the central part of the stimulus was a four tone pattern FABF (Fig. 1). Listeners were given the task of judging whether A and B formed an ascending or descending pair. In the absence of flanking tones F, listeners found the task easy. However, in the presence of tones F, the pattern formed a single stream and the AB subpattern was found to be very difficult to extract. When a sequence of capturing tones C were included preceding and following and F tones, they captured the latter into a new stream. Thus, tones A and B were separated into a different stream and their relative ordering was again found easy to judge. Bregman and Rudnicky argued that even though the stream of captor tones C was not attended to (listeners were concentrating on the occurrence of tones A and B) it was still able to capture tones F. The implication is that stream segregation can occur without attention.

However, Jones *et al.* [11] suggest an alternative explanation; such stimuli could be grouped according to temporal pattern. Specifically, the effect [10] could be explained in terms of rhythm-based grouping since the captor tones are not only related in frequency to the flanking tones but are also related rhythmically. Indeed, further trials in which the rhythm of the sequence was adapted were found to alter the ease with which and AB tones could be segregated from the rest of the sequence. Jones *et al.* [11, p.1071] concluded that *"temporal predictability*

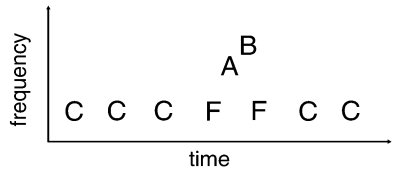

Fig. 1. Tone sequence used by Bregman and Rudnicky [10].

may be a prerequisite to the establishment of stream segregation based on frequency relationships." However, Bregman [2] argues that such a conclusion may be too strong; he notes that the tone sequence is influenced by both primitive grouping principles and exogenous attention. Bregman argues that rhythmic information "assists the selection process *directly* rather than indirectly through its effects on primitive grouping" [2, p.445]. In other words, rhythm acts as a schema at the endogenous processing level rather than at the exogenous processing level.

Recent work by Carlyon *et al.* [12] brings Bregman and Rudnicky's preattentive theory into question. Their study aimed to manipulate attention more rigorously by presenting a tone sequence monaurally. When attention was to be oriented away from the tone sequence, subjects were required to perform a competing task in the contralateral ear. Specifically, a 21 s alternating tone sequence [13] was presented to the left ear in which the frequency separation was sufficient for stream segregation to occur after a certain time period [14].

In the "baseline" condition, no stimulus was presented to the right ear. Subjects were instructed to indicate whether they heard a galloping rhythm or two separate streams. In the "two-task" condition, a series of bandpass filtered noise bursts were presented to the right ear for the first 10 s of the stimulus. During this period, subjects were instructed to ignore the tones in the left ear and simply concentrate on labeling the noise bursts as either approaching (linear increase in amplitude) or departing (the approaching burst reversed in time). Subsequently, subjects switched their attention to the tone sequence. Consistent with [14], subjects heard a single stream at the beginning of each sequence with an increased tendency to hear two streams as the sequence progressed in time. However, for the two-task condition the amount of streaming after 10 s (during which period listeners had been concentrating on labeling the noise bursts) was similar to that at the beginning of the baseline sequence—in the absence of attention, streaming had not built up. It is this important new finding that motivates our development of an attentionally driven ASA system.

Furthermore, it has been argued that Bregman and Rudnicky's experiment [10] was flawed because the listener did not have a competing attentional task to perform; despite the listener having been instructed to only concentrate on the A and B tones, there was no other task to distract the listener's attention from the C tones. Indeed, [12, p.115] notes that, *"it seems likely that listeners were in fact attending to the C tones, as they were the only sounds present at the time, and there was no other task competing for attention."*

In an attempt to clarify the role of attention in auditory scene analysis, we have developed a conceptual framework for auditory attention (Section III), and partially implemented it as a computational auditory model (Section IV). We demonstrate

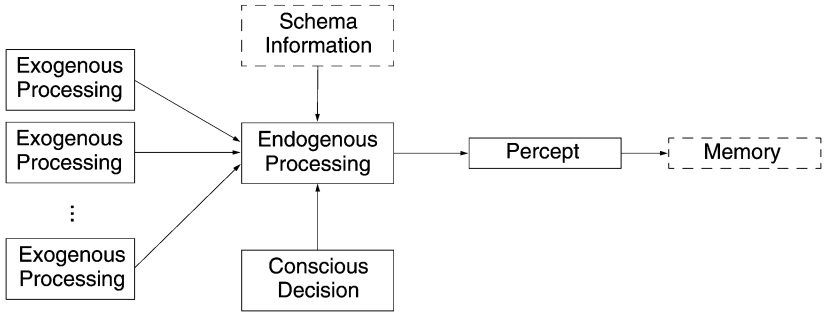

Fig. 2. Structure of the proposed conceptual framework for attentional processing. Note that endogenous processing is required before exogenous perceptual organizations can be perceived and encoded into memory. Only processing stages shown in solid boxes are implemented in the current computer model.

that the model is able to account for a number of psychophysical phenomena, including Carlyon *et al.*'s binaural stream segregation experiments and other findings relating to the perceptual segregation of a mistuned harmonic from a complex tone.

III. Conceptual Framework

The conceptual framework employed here [15] assumes that a number of exogenous processes can operate simultaneously, and that endogenous attention is required to allow the outcome of these processes to be perceived and encoded into memory (Fig. 2).

In this framework, exogenous processes are responsible for performing primitive grouping of individual features within the stimulus. These groups are then passed to the endogenous processing stage, which takes into account the conscious decision of the listener, changes in the stimulus and schema information to form an "attentional stream," which is a time-varying conscious percept.

Each of these endogenous factors compete to produce a single stream. Indeed, it is possible for a subject's attentional focus to be reoriented to an alternative grouping, contrary to their conscious preference. For example, an important schema such as one's name can overrule a conscious decision. A subject's attention can be drawn from one conversation to another in which their name was spoken [16]. However, experimental findings suggest that such schema only encode small amounts of strongly salient information. When subjects are instructed to learn arbitrary pieces of information, awareness of their presence in an unattended stream is significantly reduced [17].

Another form of unconscious redirection of attention is the startle reflex [18] which occurs in response to a loud, unexpected sound. In this situation, exogenous information about the gross change in the stimulus forces endogenous attention to be directed to the new sound without regard for the listener's conscious preference.

Schema information can also be used to aid the grouping of the exogenous processing outputs and form the attentional stream. In particular, schemas can encapsulate semantic information about grammar and contextual meaning. For example, despite a subject's conscious intention to shadow one of two concurrent utterances in a particular ear, Treisman [19] found that when the two sentences switched ears, the subject shadowed the original sentence. This implies that at the stage of endogenous processing, schema information related to the sentence semantics is being employed to overrule the listener's conscious decision to shadow a particular ear.

Although schema information is included in our conceptual framework (Fig. 2), computational modeling of schema-based processing is a challenging problem which lies beyond the scope of the current study. It should be emphasized, therefore, that schema information is not included in the computational model presented in Section IV.

A. Binding Problem

One difficulty involved in producing a computational solution to the ASA problem is the lack of a strong link between Gestalt theories of perception [2] and the underlying physiological processes. The neurophysiological mechanisms underlying auditory stream formation are poorly understood and it is not fully known how groups of features are coded and communicated within the auditory system.

In order to perceive a unified representation of an object, the brain must be able to correctly associate all the different types of features (e.g., location, color, texture, pitch, etc.) derived from that object. Such associations, or *bindings*, are even more important when the stimulus consists of more than one object, in which case the brain must avoid incorrectly associating features from different objects (illusory conjunctions, e.g., [20]). Furthermore, it is also known that features within the same modality (such as auditory, olfactory, visual, etc.) can be encoded in widely distributed, spatially discontiguous, regions of the brain [21].

It is this representational complexity which lies at the heart of the *binding problem*: how does the brain, confronted with many features, encoded in many different regions, draw them all together to form a perceptual whole? Here, we adopt an approach to this problem that is based on the concept of an assembly—a large number of spatially distributed neurons [22].

With a distributed representation it is necessary to be able to distinguish a neuron as belonging to one assembly or another. Therefore, the responses of related neurons must be labeled as such. One possible solution to this problem is the proposal of von der Malsburg [23], [24] that assemblies are labeled by temporal synchronization of neural responses. In this scheme, each assembly is identified as a group of neurons with synchronised

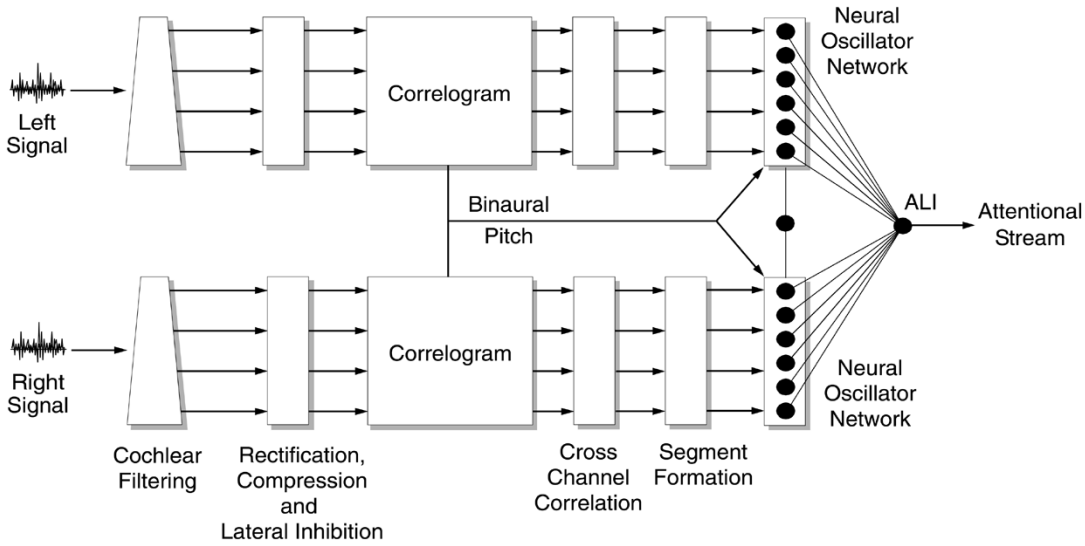

Fig. 3. Schematic diagram of the binaural model (the attentional leaky integrator is labeled ALI).

firing patterns, and the responses of neurons in different assemblies are desynchronised.

However, the computational expense of evaluating synchrony between multiple spike trains is high; to alleviate this problem, [25] proposed a mechanism in which the mean discharge response of a pool of cells is represented by an *oscillator*. In this manner, groups of features form streams if their oscillators are synchronised and the oscillations of unrelated streams are desynchronised.

Physiological support for the temporal correlation framework can be found in studies which show that neurons can sychronise their discharges [26]–[28]. Evidence that synchronised activity encodes salient information is supported by studies in which human subjects displayed a high correlation between perception and neuronal response synchronization [29], [30]. Additionally, engagement in cognitive tasks has been found to increase oscillations in task-dependent cortical areas [31], [32]. Similar increases in oscillatory activity have been observed in states of focused attention [33]. Hence, oscillations appear to be implicated in both feature binding and attentional processing.

IV. Computational Model

The computational model is comprised of three main stages (Fig. 3). The input to the model is a two-channel audio signal sampled at a rate of 8 kHz. The first stage of the model simulates peripheral auditory processing for each ear.

The second stage of the model extracts periodicity information by means of a correlogram [34], which allows primitive grouping to be performed on the basis of harmonicity. The periodicity information from each ear is then combined to produce a "binaural" F0 estimate [35, p.208] to enable binaural harmonicity grouping.

The third stage of our model is a one-dimensional (1-D) neural oscillator network in which auditory grouping and segregation take place. The network is based upon the locally excitatory globally inhibitory oscillator network (LEGION) described by Wang and Terman [36]. A cross-channel correlation mechanism identifies contiguous regions of acoustic activity in the correlogram, which correspond to Bregman's [2] notion of "sensory elements"; we use the term *segments*. These are encoded in the network by synchronised blocks of oscillators, which are established by local excitatory connections. Information from the correlogram is then used to group these segments on the basis of their conformity with the F0 estimate. Long range excitatory connections promote these oscillator blocks to synchronise to form an oscillatory "group." Blocks of oscillators which are not harmonically related desynchronise from each other.

Each oscillator feeds activity to the *attentional leaky integrator* (ALI), which is the core of our attentionally motivated stream segregation mechanism. The output of the ALI is the attentional stream as defined in Section III. The connection weights between the network and the ALI are modulated by endogenous processes including "conscious" preference. Such conscious preference is modeled as a Gaussian distribution across frequency consistent with [5]. Initially, connection weights are maximal for all channels so that the default state of organization is grouping [2]. In this initial condition, all segments and groups contribute to the attentional stream. Over a period of time, connection weights adapt to the shape of the endogenous attentional focus. In this situation, only oscillators under the attentional focus can influence the ALI. In terms of the computational model, the attentional stream is defined as containing all frequencies whose oscillators are synchronously active with the ALI. The use of synchrony to encode stream selection allows entire harmonic groups to contribute to the attentional stream, even though some harmonics of the group lie outside the attentional focus.

A. Peripheral Auditory Processing

The frequency selectivity of the basilar membrane is modeled by a bank of 128 gammatone filters [37] distributed in frequency between 50 Hz and 3.5 kHz on the ERB scale [38]. Each filter simulates the response of the basilar membrane at a specific position along its length.

The gammatone filter of order n and centre frequency f Hz is given by

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

where ϕ is phase, b is related to bandwidth, n is the order of the filter, and $H(t)$ is the unit step (Heaviside) function defined as $H(t) = 1$ if $t \geq 0$, $H(t) = 0$ otherwise. We use $n = 4$, for which the magnitude characteristic of the gammatone filter exhibits a good fit to human auditory filter shapes [39]. The gain of each filter is adjusted according to the ISO standard for equal loudness contours [40] in order to simulate the pressure gains of the outer and middle ears.

The auditory nerve response is approximated by half-wave rectifying and square root compressing the output of each filter [41]. Finally, to improve the frequency resolution and reduce cross channel activity spread, the simulated auditory nerve response at each time instant is convolved with a difference of Gaussians "mexican hat" kernel.

B. Pitch and Harmonicity Analysis

Pitch information is extracted from the auditory nerve responses by computing the autocorrelation of each channel to form a correlogram [34]. The correlogram may be regarded as a computational implementation of Licklider's coincidence model [42], [43] [44]. The resulting two-dimensional (2-D) representation has channel centre frequency and autocorrelation lag on orthogonal axes. At time t, the autocorrelation of channel i with lag τ for ear e is given by

$$A_e(i,t,\tau) = \sum_{k=0}^{P-1} r_e(i,t-k)r_e(i,t-k-\tau)w(k). \quad (2)$$

Here, $r_e(i,t)$ is the auditory nerve activity in channel i at time t for ear e. The autocorrelation for channel i is computed using a 25 ms ($P = 200$) rectangular window w [45, p.1151], with lag steps equal to the sampling period (0.125 ms), up to a maximum lag of 20 ms.

Research into the effect of harmonic mistuning on the pitch of a complex tone [46] suggests that the perceived pitch changes even when the mistuned harmonic is presented to the contralateral ear. This suggests that the auditory system can assess whether a tone in one ear should be grouped with a contralateral sound on the basis of harmonicity. Hence, it is assumed that binaural harmonicity grouping proceeds in the following manner: an overall pitch estimate from both ears is calculated and harmonic grouping is performed using this pitch estimate. Specifically, a summary function s_b is formed by summing all the channels of the left- and right-ear correlograms

$$s_b(t,\tau) = \sum_{e\in\{\text{left,right}\}} \sum_{i=0}^{N-1} A_e(i,t,\tau) \quad (3)$$

where $A_e(i,t,\tau)$ is the autocorrelation of channel i at time t with lag τ for ear e and N is the number of channels in each correlogram. Typically, a large peak occurs in the summary function at a lag corresponding to the fundamental period of the stimulus. Here, we select this peak as the one with the shortest lag whose height is larger than 80% of the energy in the correlogram frame (which corresponds to the autocorrelation at zero lag).

C. Segment Identification

The correlogram for each ear can also be used to identify formant and harmonic regions from their patterns of periodicity, because contiguous regions of the filterbank respond to the same spectral component. Such contiguous areas of acoustic energy are used to form *segments* [47], which are identified by computing the cross correlation between adjacent channels of the correlogram

$$C_e(i) = \frac{1}{\tau_{\max}} \sum_{\tau=0}^{\tau_{\max}-1} \hat{A}_e(i,t,\tau)\hat{A}_e(i+1,t,\tau). \quad (4)$$

Here, $\hat{A}_e(i,t,\tau)$ is the autocorrelation function of (2) which has been normalized to have zero mean and unity variance. This ensures that $C_e(i)$ is only sensitive to periodicity in the correlogram, and not to the mean firing rate of each channel. $\tau_{\max}$ is the maximum autocorrelation lag in samples ($\tau_{\max} = 160$; equivalent to 20 ms).

Once the cross correlation $C_e(i)$ has been computed, it is necessary to decide a "similarity score" by which adjacent channels are deemed to be sufficiently similar to be grouped together to form a segment. This is achieved by applying a threshold to the energy-weighted cross correlation. Adjacent channels whose cross correlations are above a certain threshold form a segment; specifically, channels i and $i+1$ are said to contribute to a segment when

$$C_e(i)A_e(i,t,0) > \theta_s \quad (5)$$

where θ_s is the segment membership threshold. The cross correlation is energy-weighted in order to increase the contrast between spectral peaks and spectral dips. A high threshold would result in a small number of segments as few adjacent channels would be nearly identical; as the threshold is lowered, so too is the similarity requirement and so similar adjacent channels form segments. These segments are encoded by a binary mask, which is unity when a channel contributes to a segment and zero otherwise.

In order to deal with noise stimuli, an alternative segment formation strategy is used since, by definition, periodicity information cannot be obtained from the correlogram for channels containing noise. Instead, the instantaneous frequency of each gammatone filter is used [48]. In response to a pure tone, a channel's instantaneous frequency over time will be stable. However, in response to noise, it exhibits significant fluctuations. This property can be exploited by calculating the inverse variance of the instantaneous frequency in each channel; responses to periodic signals produce low-signal variance and, hence, high-inverse variance. When weighted by channel energy (obtained from the correlogram), a large peak indicates periodic activity in that channel; a smaller peak indicates noise activity.

The segment estimation process, therefore, occurs in two stages. First, periodic segments are identified, i.e., channels for which a peak exists in the energy weighted inverse variance function that exceeds a given "tonal" threshold θ_t. All channels

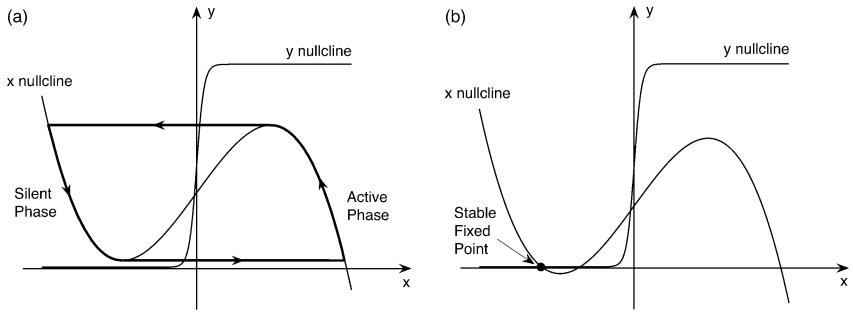

Fig. 4. Nullclines of a single oscillator. (a) The bold line shows the limit cycle of an enabled oscillator whose direction of motion is indicated by the arrow heads. (b) Disabled oscillator approaches a stable fixed point.

under such peaks are said to constitute a tonal segment; this information is used to confirm the periodic segments identified by the cross-correlation technique described above. These peaks are then removed from the energy weighted inverse variance function. The final stage identifies noise segments: channels for which any remaining peaks in the function exceed a "noise" threshold θ_n.

To further improve the clarity of the segment representation, each segment is replaced by one of fixed frequency channel spread. This sharpens the across-frequency responses, thus reducing the tendency for segments which are close in frequency to merge together. This is especially important for spectral components at higher frequencies where the bandwidth of the peripheral filters are relatively large in comparison to the frequency separation between harmonics. Since this model investigates how individual acoustic elements are influenced by grouping rules and attentional mechanisms, it is important that each is represented by an individual segment where possible.

D. Neural Oscillator Network

The neural oscillator network consists of an array of 128 oscillators and is based upon LEGION [36]. Within LEGION, oscillators are synchronised by placing local excitatory links between them. Additionally, a global inhibitor receives excitation from each oscillator, and inhibits every oscillator in the network. This ensures that only one block of synchronised oscillators can be active at any one time. Hence, separate blocks of synchronised oscillators (segments) arise through the action of local excitation and global inhibition.

The building block of the network is a single oscillator, which consists of a reciprocally connected excitatory unit and inhibitory unit whose activities are represented by x and y, respectively

$$\dot{x} = 3x - x^3 + 2 - y + I_0 \tag{6}$$

$$\dot{y} = \varepsilon\left[\gamma\left(1 + \tanh\frac{x}{\beta}\right) - y\right]. \tag{7}$$

Here ε, γ, and β are parameters, and I_o represents the input to the oscillator. The x-nullcline ($\dot{x} = 0$) is a cubic function and the y-nullcline ($\dot{y} = 0$) is a sigmoid function (see Fig. 4). When $I_o > 0$, the two nullclines intersect at a point along the middle branch of the cubic [Fig. 4(a)] and give rise to a stable periodic orbit provided that ε is sufficiently small. In this situation, the oscillator is said to be *enabled*. The solution of an enabled oscillator alternates between a phase of high-x values (*active phase*) and a phase of low-x values (*silent phase*); transitions between these two phases occur on a much faster time scale compared to time spent in active and silent phases. When $I_o < 0$, the nullclines intersect on the left branch of the cubic [Fig. 4(b)] and produce a stable fixed point at a low value of x. When the oscillator is in this state of equilibrium, it is said to be *disabled*. The parameter γ can be used to adjust the amount of time an oscillator spends in the two phases: a smaller value of γ results in a shorter active phase duration. It is clear, therefore, that oscillations are stimulus dependent: they are only observed when the external input to the oscillator is greater than zero. Because it has two timescales, the oscillator in (6) and (7) belongs to the family of relaxation oscillators. It is related to both the van der Pol oscillator [49] and to simplifications of the Hodgkin–Huxley equations for action potential generation in nerve membrane [50]–[52]. The system may be regarded as a model for the behavior of a single neuron in which x represents the membrane potential of the cell and y represents the inhibitory ion channel activation, or as a mean field approximation to a group of reciprocally connected excitatory and inhibitory neurons.

The input I_o to oscillator i is a combination of three factors: external input I_r, network activity, and global inhibition

$$I_o = I_r - W_z S(z, \theta_z) + \sum_{k \neq i} W_{ik} S(x_k, \theta_x). \tag{8}$$

Here, W_{ik} is the connection strength between oscillators i and k, and x_k is the activity of oscillator k. 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 inhibitor z. Similar to θ_x, θ_z acts as a threshold above which the global inhibitor can affect an oscillator. S is a squashing function which compresses oscillator activity to be within a suitable range

$$S(m, \theta) = \frac{1}{1 + e^{-K(m-\theta)}} \tag{9}$$

where K determines the steepness of the sigmoidal function. The activity of the global inhibitor is defined as

$$\dot{z} = H\left(\sum_k S(x_k, \theta_x) - 0.1\right) - z \tag{10}$$

where H is the Heaviside function.

E. Segment Formation and Primitive Grouping

Oscillators within a segment are synchronized by excitatory connections. The external input (I_r) of an oscillator which is a member of a segment is set to I_{high} (0.2) otherwise it is set to I_{low} (-5).

A further set of connections are made between segments if a majority of channels in each segment are consistent with the pitch estimate as derived above. The autocorrelation for each channel in the segment is inspected. If the ratio of channel energy to autocorrelation value at the pitch lag is above a certain threshold θ_c (0.8) the channel is classified as being consistent with the pitch estimate [47]. It is this tolerance in the measure of harmonicity that allows the model to account for the perceptual grouping of harmonics which have been mistuned by limited amounts [46]. In other words, channel i is consistent with a fundamental period of τ_0 when

$$\frac{A_e(i,t,\tau_0)}{A_e(i,t,0)} > \theta_c. \tag{11}$$

If the majority of segment channels are consistent, the entire segment is said to be consistent with the pitch estimate.

It is at this stage that the *old-plus-new* heuristic is incorporated into the model. The old-plus-new heuristic refers to the auditory system's tendency to *"interpret any part of a current group of acoustic components as a continuation of a sound that just occurred"* [2]. In our model, *age trackers* are attached to each channel of the oscillator array. The age trackers are leaky integrators defined as

$$\dot{B}_k = d_B\left(g_B[M_k - B_k]^+ - [1 - H(M_k - B_k)]c_B B_k\right). \tag{12}$$

Here, B_k is the age of the channel, M_k is the (binary) value of the segment mask at channel k; small values of c_B and d_B result in a slow rise (d_B) and slow decay (c_B) for the integrator. g_B is a gain factor. $[n]^+ = n$ if $n \geq 0$ and $[n]^+ = 0$ otherwise. These parameters have the values $d_B = 0.001$, $c_B = 5$, $g_B = 3$. Excitatory links are placed between harmonically related segments only if the two segments are of similar age. The age of a segment is defined as

$$AS = \frac{1}{Q}\sum_{k\in\text{segment}} B_k \tag{13}$$

where Q is the number of channels in the segment. Two segments are considered to be of similar age if

$$|AS_1 - AS_2| < \theta_a \tag{14}$$

where AS_1 and AS_2 are the ages of the two segments and the threshold θ_a (0.1) defines the degree of similarity in age between the two segments.

Consider two segments that start at the same time. The age trackers for their constituent channels all begin receiving input at the same time and continue to receive the same input: the values of the leaky integrators will be the same. However, if the two segments start at different times, the age trackers for the earlier segment will have already built up to a positive value when the second segment starts (whose age trackers will be initially at zero): the two ages will be different.

F. Attentional Process

Each output channel of the oscillator array is connected to the attentional leaky integrator (ALI) by excitatory links (Fig. 3), and the strength of these connections is modulated by endogenous attention. Input to the ALI is a weighted version of the oscillator array output

$$\dot{a}li = I_{\text{ALI}} - ali \tag{15}$$

where I_{ALI} is defined as

$$I_{\text{ALI}} = H\left(\sum_k H(x_k)\left[\left(\frac{\alpha_k}{\theta_\alpha}\right) - T_k\right]^+ - \theta_{\text{ALI}}\right). \tag{16}$$

Here, H is the Heaviside function and x_k is the activity of oscillator k in the array. The parameter θ_{ALI} is a threshold above which oscillator array activity can influence the ALI. α_k is the envelope of the gammatone filter response to the stimulus at channel k. θ_α is a normalizing factor which determines how intense a stimulus needs to be to overcome the conscious attentional interest.

T_k is the attentional threshold which is related to the endogenous interest at channel k. In order to model the findings of Carlyon *et al.* [12] in which listeners are instructed to attend to sounds in a particular ear, we incorporate a mechanism by which attention can be allocated to specific ears

$$T_k = (1 - A_{\text{ear}}A_k)\,L. \tag{17}$$

Here, A_k is the endogenous attentional interest at channel k, A_{ear} represents the ear preference, ranging from 0 to 1, and L is the leaky integrator defined as

$$\dot{L} = d_L\left(g_L[R - L]^+ - [1 - H(R - L)c_L L]\right). \tag{18}$$

Here, small values of c_L and d_L result in a slow rise (d_L) and slow decay (c_L) for the integrator. g_L is a gain factor. These parameters have the values $d_L = 0.0005$, $c_L = 5$, $g_L = 3$. R is given by

$$R = H(x_{\max}) \tag{19}$$

where $x_{\max}$ is the largest output activity of the oscillator array.

In accordance with the experimental findings of [5], the attentional interest is modeled as a Gaussian.

$$A_k = \max_{A_k} e^{\frac{-((k-p)^2}{2\sigma_{\text{ALI}}{}^2)}}. \tag{20}$$

A_k is the attentional interest at frequency channel k, $\max_{Ak}$ is the maximum value that A_k can attain, p is the channel at which the peak of attentional interest occurs, and σ_{ALI} determines the width of the peak. In order to allow segments which are outside of the attentional interest peak, but are sufficiently intense, to overrule the "conscious" attentional selection, the A_k vector must be nonzero on both sides of the peak. Hence, a minimum A_k value of $\min_{Ak}$ is enforced

$$A_k = \begin{cases} \min_{A_k}, & A_k < \min_{A_k} \\ A_k, & \text{otherwise.} \end{cases} \tag{21}$$

In the model, a segment or group of segments are considered to be attended to if their oscillatory activity coincides tempo-

rally with a peak in the ALI activity. In other words, their connection strengths to the ALI must be sufficiently large to promote activity within the ALI. Initially, the connection weights between all oscillators in the array and the ALI are strong, and hence, all segments feed large amounts of excitation to the ALI. This means that initially all segments contribute to the attentional stream representing the default grouping state of fusion [2].

During sustained activity in the oscillator array, these weights relax toward the A_k interest vector such that strong weights exist for channels of high attentional interest and low weights exist for channels of low-attentional interest. This relaxation toward the A_k interest vector is achieved by the use of the leaky integrator L. Thus, after a finite period of time, oscillators which are desynchronised from those within the attentional interest (e.g., because they are harmonically unrelated) will have low-connection weights to the ALI and will be unlikely to overcome the θ_{ALI} threshold required to influence the ALI. Such "relaxation" of the connection weights toward the attentional interest vector models the period of build-up observed in auditory streaming [14]. ALI activity will only coincide with oscillator activity within the attentional interest peak and any perceptually related (i.e., synchronised) activity outside the A_k peak. All other activity will occur within a trough of ALI activity. This behavior allows both individual tones and harmonic complexes to be attended using only a single A_k peak.

G. Attentional Reset

Listeners in Carlyon *et al.*'s study [12] were instructed to concentrate on a stimulus in a particular ear. Initially, this was a sequence of noise bursts with increasing or decreasing amplitude; the task was to classify the amplitude ramp. After a period of time subjects were told to concentrate on a sequence of alternating tones in the other ear. In this situation, listeners showed a greatly reduced amount of streaming relative to the situation in which they were allowed to listen to the alternating tones from the beginning of the task.

In our model, one ear is given a higher attentional weighting than the other by making A_{ear} high for the attended ear and low for the other. When an ear dominance (A_{ear}) change is detected, we propose that the model's attentional mechanism "resets," the leaky gain factor L on the A_k vector described in (17), used to model the build up of the attentional effect, is reset to zero. Hence, after the change, the attentional interest requires time to build up before streaming can be observed.

H. Summary

To summarize, the model consists of three core stages: auditory peripheral processing, periodicity analysis, and a neural oscillator network. Once the audio signal has been processed to simulate cochlear filtering and auditory nerve encoding, periodicity information is extracted using the correlogram. Such information allows both the F0 of the signal to be extracted, and identifies areas of contiguous periodic activity which correspond to Bregman's [2] concept of acoustic elements. Further analysis of the cochlear filtering outputs allow noise segments (i.e., nonperiodic regions of energy) to be identified.

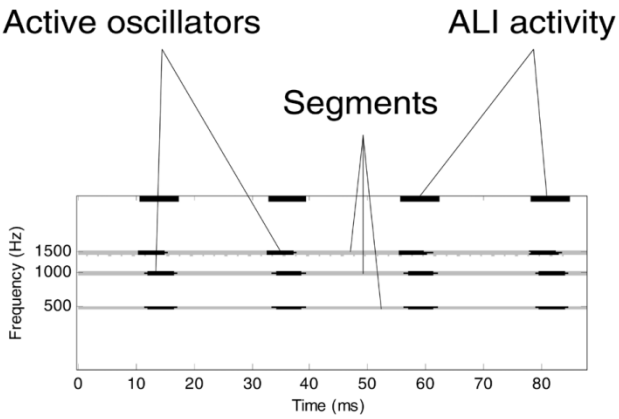

Fig. 5. Sample output of the oscillator array in response to a three-harmonic complex tone with a F0 of 500 Hz and a duration of 90 ms.

Within the network, each oscillator in the array corresponds to a particular frequency channel; segments are created by placing excitatory connections between the relevant oscillators, which cause them to synchronise. These segments are then grouped on the basis of common harmonicity by using further excitatory connections between constituent oscillators. Each oscillator is connected to the ALI by means of connections whose strengths are modulated by "conscious" attentional interest: maximum strength occurs at the frequency of highest interest and spreads over frequency in a Gaussian manner. Selective attentional focus is only observed over a period of time in order to model the time course of auditory stream buildup [14]. Only the activity of oscillators representing frequency channels of high-attentional interest can influence the ALI, and hence, any activity synchronised with the ALI is said to contribute to the attentional stream. Furthermore, when attention is moved in space, a resetting of the attentional buildup occurs.

V. Evaluation

The output of the model is evaluated by inspecting the time course of the neural oscillator array activity and the ALI. This information allows the behavior of the grouping and attentional processes to be compared with the findings of psychophysical studies. Before presenting the results, we describe the format in which the oscillator array and ALI outputs will be presented.

The activity of the oscillator array over the time course of a stimulus is represented in a pseudospectrogram format, as shown in Fig. 5. Channel centre frequency is represented on the ordinate and time is represented on the abscissa. Pixels at each time-frequency location in the diagram may take one of three values. Gray pixels denote stimulated oscillators which receive an input I_{high}, black pixels denote oscillators in their active phase, and white pixels correspond to unstimulated oscillators (i.e., those receiving an input I_{low}). The activity of the ALI is shown along the top of the diagram: a black block indicates that the ALI is active. Any oscillators which are temporally synchronised (vertically aligned) with the ALI are considered to be in the attentional foreground. Such diagrams are constructed on a sample-by-sample basis: after each stimulus sample has been processed, the state of the oscillator array is recorded and forms a vertical slice in the diagram.

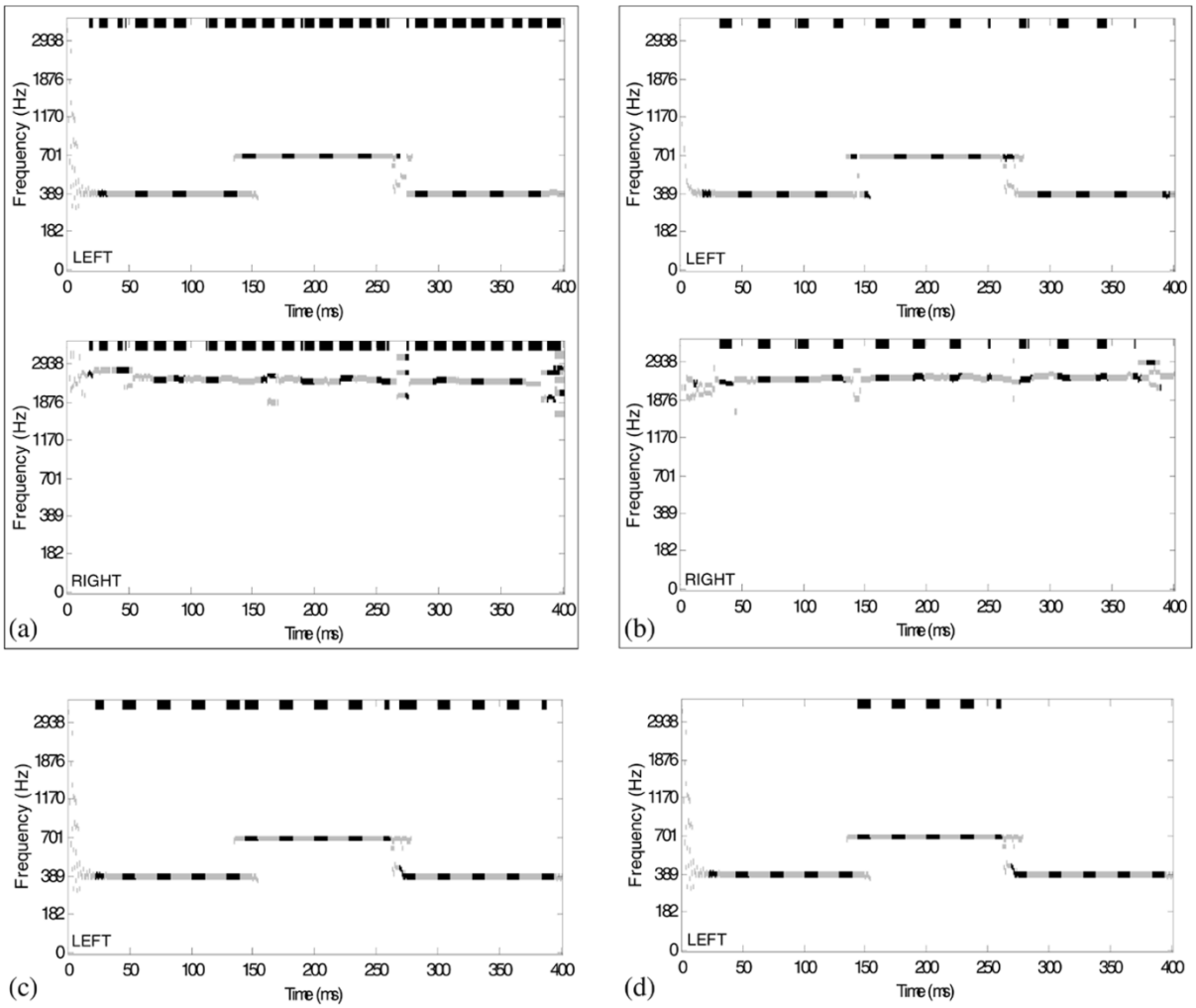

Fig. 6. Model response to the Carlyon *et al.* [12] stimulus after varying time periods: (a) 1 s, (b) 8 s, (c) 11 s, and (d) 20 s. In each condition, the response of the left ear network is shown in the top panel and the right ear network is shown in the lower panel. Note that the noise bursts in the right ear cease after 10 s, and hence, the empty right ear network activity plots after 10 s (c and d) are not shown. Segments whose oscillator activity is temporally aligned with ALI activity are considered to be in the attentional foreground.

A. *Two-Tone Streaming With Attentional Distractor*

The stimulus used by Carlyon *et al.* [12] and described in Section II was used as input to the model, and the corresponding output from the oscillator network is shown in Fig. 6. The conscious movement of attention from the right ear (noise bursts) to the left ear (alternating tone sequence) was simulated by altering A_{ear} after 10 s. For the first half of the stimulus $A_{\text{left}} = 0$ and $A_{\text{right}} = 1$, causing attention to be directed toward the right ear; for the remainder of the stimulus $A_{\text{left}} = 1$ and $A_{\text{right}} = 0$, so that attention is directed toward the left ear.

To improve clarity, Fig. 6 shows four representative excerpts from the network output: the state of the oscillator array after 1 s, 8 s, 11 s, and, finally, after 20 s of the stimulus. Fig. 6(a) shows the network after 1 s of the stimulus. At this stage, attention is directed toward the noise bursts. Since only 1 s has elapsed, the attention buildup has not reached a sufficiently high level to exclude activity from the "ignored" tones. However, after a period of time, sufficient buildup occurs and the ALI is only influenced by the noise segment, indicating that only the noise burst is contained in the attentional stream [Fig. 6(b)]. After 10 s, the simulated switch of attention from the right ear to the left ear is made and the attentional "reset" occurs. Following this switch, the ALI is once again influenced by all the segments present [Fig. 6(c)] until the buildup period has elapsed. Once the attentional buildup has occurred, streaming is observed since attention is now directed toward the high-frequency tone within the alternating tone sequence. This is apparent from the ALI activity, which is synchronous only with the oscillators corresponding to the high-frequency tone. It should be noted that an attentional reset is not caused by abrupt movements in frequency since neither A_{left} nor A_{right} is altered.

B. *Segregation of an Harmonic From a Complex Tone*

Darwin *et al.* [46] investigated the effect of a mistuned harmonic upon the pitch of a 12 component complex tone. As the degree of mistuning of the fourth harmonic increased toward 4%, the shift in the perceived pitch of the complex also increased. This effect was less pronounced for mistunings of more than 4%; beyond 8% mistuning, little pitch shift was observed. This suggests that the pitch of a complex tone is calculated using only those channels which belong to the corresponding stream. When the harmonic is subject to mistunings below 8%, it is

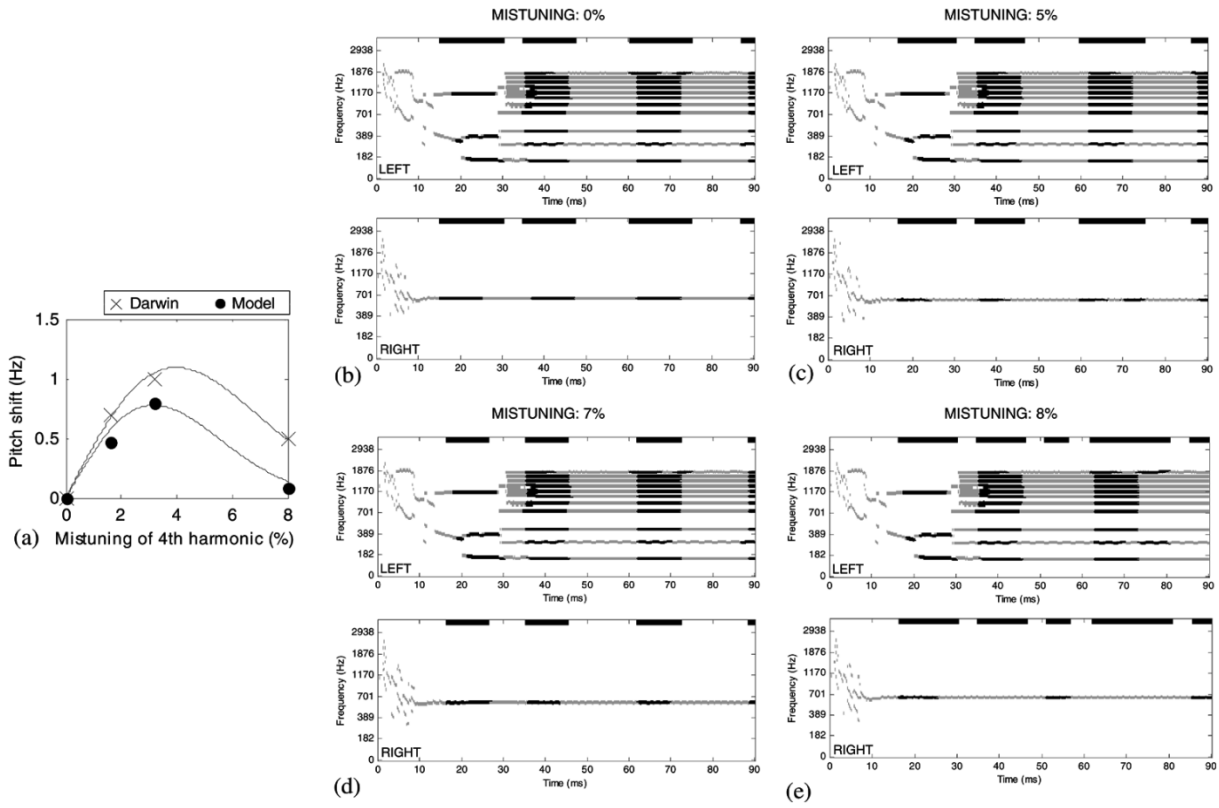

Fig. 7. (a) Monaural pitch shift versus degree of mistuning. A Gaussian derivative is fitted to each data set. Experimental data from [46]. (b)–(e) Model response to a 12-harmonic complex tone (155 Hz fundamental frequency) whose contralaterally presented fourth harmonic is mistuned by varying degrees: (b) 0%, (c) 5%, (d) 7% and (e) 8%. For mistunings of less than 8% (b)–(d), oscillator activity corresponding to the mistuned harmonic is temporally synchronised to that of the complex: the mistuned harmonic is perceptually grouped with the complex. However, once mistuning reaches 8% (e), oscillator activity corresponding to the mistuned harmonic is temporally desynchronised from that of the complex: two groups have emerged, indicating that the mistuned harmonic has been segregated from the rest of the complex. Response of the left ear network is shown in the top panel and the right ear network in the lower panel in each case.

grouped with the rest of the complex and so can affect the pitch percept. For mistuning greater than 8%, the lack of influence on the pitch percept exerted by the mistuned harmonic implies that it has been perceptually segregated from the complex. In other words, for mistunings below 8%, a single group exists; mistunings beyond 8% result in two groups.

The pitch shifts reported in [46] and the shifts made by the model are in good agreement, as shown in Fig. 7(a). The pitch of the complex was calculated by creating a summary correlogram (as described in Section IV-C) using frequency channels contained within the complex tone group.

Darwin *et al.* [46] also investigated the effect of a contralaterally presented mistuned harmonic upon the pitch of a 12-component complex tone. As in the monaural situation, the degree of mistuning of the fourth harmonic influenced the shift in the perceived pitch of the complex for mistunings of up to 8%; mistunings of more than 8% had little effect on the perceived pitch. These results imply that harmonics presented contralaterally are assessed to see if they match an overall binaural pitch estimate and are grouped accordingly. Fig. 7 shows the model response to a 12-harmonic complex with a F0 of 155 Hz in which the mistuned fourth harmonic is presented contralaterally to the remainder of the complex. In Fig. 7, the 12 harmonics, represented as segments, can clearly be seen. The activities of their corresponding neural oscillators exhibit temporal synchrony, indicating that all the harmonics have been grouped. Similar behavior is observed when the harmonic is mistuned by 5% and 7%: synchronization occurs and the harmonic is considered to be perceptually grouped with the complex even at this relatively high degree of mistuning. When the degree of mistuning reaches 8%, the fourth harmonic is segregated from the rest of the complex; the oscillators corresponding to the fourth harmonic are temporally desynchronised from the remaining oscillators [Fig. 7(e)]. Two distinct perceptual groups are now apparent, one containing the fourth harmonic and the other containing the remainder of the complex tone. ALI activity displayed at the top of the diagrams in Fig. 7(e) indicates that both the complex and the segregated harmonic are attended to; subsequently, attentional buildup will occur so that one or the other could become the subject of attentional focus.

C. Old-Plus-New Heuristic

The effect of mistuning is almost eliminated when the mistuned harmonic is "captured" from the complex by preceding tones at the same frequency [46]. In this situation, no matter how small the mistuning, the harmonic is segregated from the complex and does not influence the pitch percept. Fig. 8(a) shows the capture of the fourth harmonic even when there is no mistuning. During the 550 ms before onset of the complex tone, the age tracker activities B_k for the captor tone channels build up. When the complex tone begins, there is a significant age difference between the frequency channels stimulated by the fourth

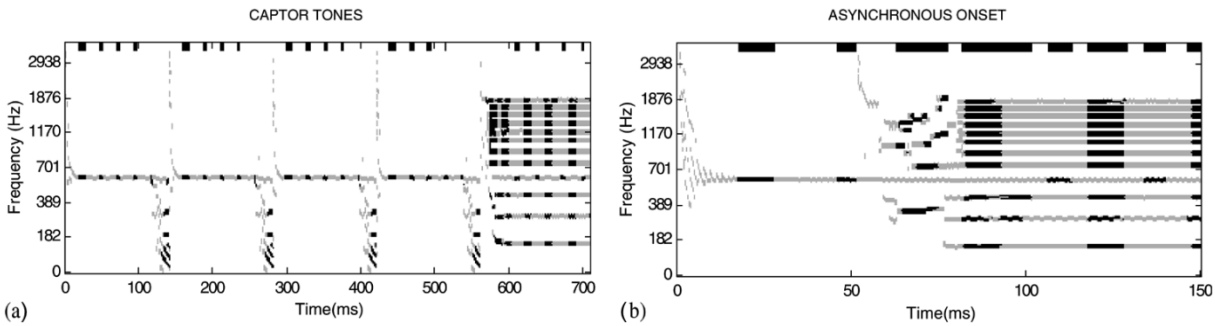

Fig. 8. (a) Model response to a 12-harmonic complex tone whose 4th harmonic is preceded by four "captor" tones. Note that despite the fourth harmonic being harmonically related to the complex, it is segregated (indicated by temporal desynchronization) by virtue of the captor tones. (b) Response of the model to a 12-harmonic complex tone with a F0 of 155 Hz, whose fourth harmonic begins 50 ms before the remainder of the complex.

harmonic and those stimulated by the remainder of the complex. Such a difference prevents excitatory harmonicity connections from being made between the fourth harmonic and the remaining harmonics. It can be seen in the diagram that the oscillator activities corresponding to the fourth harmonic temporally desynchronise from those of the remainder of the complex tone.

The old-plus-new heuristic can be further demonstrated by starting the fourth harmonic before the rest of the complex. Fig. 8(b) shows the output of the model when the fourth harmonic is subject to a 50 ms onset asynchrony. During this time, the age trackers of channels excited by the fourth harmonic increase to a significantly higher value than those of the remaining harmonics. This is the same mechanism by which captor tones, in the previous example, caused the harmonic to segregate. Once again, this difference in segment activity age prevents excitatory connections being made between the fourth harmonic and the other harmonically related segments. Thus, the leading harmonic is desynchronised from the rest of the complex and two groups are formed. However, after a period of time, the importance of the onset asynchrony decreases as the channel ages approach their maximal values. Once this occurs, there is no longer any evidence to prevent excitatory links from being made between the fourth harmonic and the rest of the complex. Grouping by harmonicity then occurs for all segments: the complex and the early harmonic synchronise to form a single stream. This is consistent with experimental data [53] in which the effect of segregation on the pitch shift of a complex due to asynchronous onset is reduced for stimuli of longer duration.

VI. Summary and Discussion

A model of auditory streaming has been presented in which the allocation of attention lies at the heart of the stream formation process. We propose a conceptual framework in which a number of different processes are responsible for forming auditory streams. Firstly, we make the distinction between exogenous and endogenous attention. Exogenous, unconscious, attention is responsible for performing primitive grouping of individual features within the stimulus. These groups are then passed to the endogenous processing stage, which takes into account the conscious decision of the listener, changes in the stimulus input and schema information to form an "attentional stream." It is at this stage of processing that attentional allocation influences stream formation. It is proposed that schema information is used to both aid the grouping of the exogenous processing outputs [19], and to act as a form of detector which causes salient information in any group to reorient conscious attention [16]. Furthermore, the conscious decision of the listener can be overruled by, and attention forcibly reoriented to, intense sounds that occur unexpectedly in the environment. This mimics the startle reflex [18].

We have implemented this conceptual framework as a computational auditory model. The core of the model is a 1-D neural oscillator network based on LEGION [36]. However, it differs from LEGION in that it is one-, rather than 2–D and incorporates long range excitatory connections between oscillators. Channels responding to the same spectral event are encoded in the network by locally excitatory connections to form "segments." Further sets of excitatory connections are placed between individual segments whose periodicity conforms with the F0 estimate. The activities of oscillators with excitatory connections between them synchronise temporally to form an oscillatory "group." Blocks of oscillators which are not linked by excitatory connections desynchronise from each other.

In contrast to previous computational models of auditory streaming, attention plays a crucial role in the stream formation and segregation process of our model. We argue that, on the basis of psychophysical research [12], distinct streams are not formed unless attention is directed toward particular (groups of) features. To this end, our model employs an ALI and a representation of attentional allocation across frequency, A_k. A_k corresponds to the conscious preference of the listener and is modeled by a Gaussian distribution in accordance with psychophysical findings [5]. Furthermore, we argue that this conscious preference cannot be deployed instantaneously: it is subject to a buildup over time. Hence, the degree to which grouped frequency channels can influence the ALI is determined by the timecourse of the A_k buildup. The output of the ALI describes the frequency content of the "attentional stream" at each epoch. The use of synchrony allows harmonic groups, most of whose harmonics may be outside of the attentional focus, to contribute to the attentional stream simply by attending to one harmonic. In Section V we demonstrated that this mechanism can explain both auditory streaming [13] and the associated build-up of streaming over time [14].

In addition to this, it has been shown that when the ear of presentation of an alternating tone sequence is switched, a second buildup period is required immediately after the switch [12],

[14], [15]. We contend that the buildup of attentional efficacy is subject to a "reset" when abrupt changes in the stimulus location are consciously detected and tracked. In other words, the attentional buildup is reset following an abrupt change of spatial attentional preference. However, such a reset is not precipitated by abrupt movements in frequency.

The old-plus-new heuristic has been incorporated into the model and influences the ability of the network to form harmonically related groups. If a segment is deemed to be "older" than other harmonically related segments, it is prevented from becoming a member of that group. In Bregman's account [2], the old-plus-new heuristic is treated as a "high-level" grouping principle similar to common fate or harmonicity. However, we adopt an approach in which the old-plus-new heuristic arises from low-level mechanisms.

Previous neural oscillator models of auditory grouping [47], [54] have represented auditory activity within a time-frequency grid. Each point in the grid is associated with a neuron that has an oscillating response pattern and the time dimension is created by a system of delay lines. In contrast to these studies, the network presented here is 1-D; it has a frequency axis, but no explicit time axis. This is preferable, since there is little physiological evidence for the arrays of delay lines necessary to produce such a long neural time axis.

Furthermore, attempts to incorporate attention into a network with an explicit time axis might lead to unrealistic properties. For example, Wang's *shifting synchronization theory* [54] states that *"attention is paid to a stream when its constituent oscillators reach their active phases."* This implies that the process of sound segmentation and grouping and the process of attentional selection are intimately linked. Once a stream has been formed, it will be attended to when its associated set of oscillators reach the active phase. Since each synchronised block of oscillators become active in a repeating sequence, attention quickly alternates between each different stream at different positions on the time-frequency grid. Hence, stream multiplexing occurs and all streams are perceived as equally salient at all times. This contradicts experimental findings [2] which show that listeners tend to perceive one stream as dominant. Furthermore, this theory cannot explain how attention may 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. Our 1-D network processes the stimulus on a sample by sample basis and, in turn, produces an estimate of the segments present and which of these are contributing to the attentional stream at any epoch. Hence, the timecourse of auditory organization and attentional influence are implicit in the network output.

An assumption of the model is that attentional buildup resets following a spatial movement of attention. Recent experimental findings support such an assumption, and show that the degree of reset observed in listeners is related to the degree of spatial movement of attention [15]. The model also assumes that an exogenous overruling of attention (such as a startle reflex) would not affect the attentional buildup, provided that the listener's endogenous attention was not subsequently moved. This arises from that fact that there would be no change in the ear dominance (A_{ear}) and, hence, no reset of the attentional buildup would occur. This prediction remains to be tested.

Finally, we note that the proposed model relies exclusively on temporal codes and processing. Specifically, pitch information is extracted from the temporal fine structure of auditory nerve firing patterns, and both grouping and attention are encoded temporally within a neural oscillator network.

To conclude, we have described a model which successfully simulates a number of auditory grouping phenomena within a framework in which attention plays a pivotal role. Future research must continue to investigate the role of attention in ASA if we hope to produce computational models which will mimic the human perception of sound with any accuracy.

REFERENCES

[1] E. C. Cherry, "Some experiments on the recognition of speech, with one and two ears," *J. Acoust. Soc. Amer.*, vol. 25, pp. 975–979, 1953.

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

[3] K. Koffka, *Principles of Gestalt Psychology*. Orlando, FL: Harcourt Brace Jovanovich, 1936.

[4] R. M. Warren and R. P. Warren, "Auditory illusions and confusions," *Sci. Amer.*, vol. 223, no. 12, pp. 30–36, 1970.

[5] T. A. Mondor and A. S. Bregman, "Allocating attention to frequency regions," *Percep. Psychophys.*, vol. 56, no. 3, pp. 268–276, 1994.

[6] T. A. Mondor and R. J. Zatorre, "Shifting and focusing auditory spatial attention," *J. Exp. Psychol. Human*, vol. 21, no. 2, pp. 387–409, 1995.

[7] Y. Tsal, "Movements of attention across the visual field," *J. Exp. Psychol. Human*, vol. 9, no. 4, pp. 523–530, 1983.

[8] G. J. Andersen and A. F. Kramer, "Limits of focused attention in three-dimensional space," *Percep. Psychophys.*, vol. 53, no. 6, pp. 658–667, 1993.

[9] C. J. Spence and J. Driver, "Covert spatial orienting in audition: Exogenous and endogenous mechanisms," *J. Exp. Psychol. Human*, vol. 20, no. 3, pp. 555–574, 1994.

[10] A. S. Bregman and A. Rudnicky, "Auditory segregation: Stream or streams," *J. Exp. Psychol. Human*, vol. 1, pp. 263–267, 1975.

[11] M. R. Jones, G. Kidd, and R. Wetzel, "Evidence for rhythmic attention," *J. Exp. Psychol. Human*, vol. 7, pp. 1059–1073, 1981.

[12] R. P. Carlyon, R. Cusack, J. M. Foxton, and I. H. Robertson, "Effects of attention and unilateral neglect on auditory stream segregation," *J. Exp. Psychol. Human*, vol. 27, no. 1, pp. 115–127, 2001.

[13] L. P. A. S. van Noorden, "Temporal coherence in the perception of tone sequences," Ph.D. dissertation, Inst. Perceptual Res., Eindhoven, The Netherlands, 1975.

[14] S. Anstis and S. Saida, "Adaptation to auditory streaming of frequency-modulated tones," *J. Exp. Psychol. Human*, vol. 11, pp. 257–271, 1985.

[15] S. N. Wrigley, "A Theory and computational model of auditory selective attention," Ph.D. dissertation, Dept. Comput. Sci., Univ. Sheffield, Sheffield, U.K., 2002.

[16] N. Moray, "Attention in dichotic listening: Affective cues and the influence of instructions," *Q. J. Exp. Psychol.*, vol. 11, pp. 56–60, 1959.

[17] M. E. Dawson and A. M. Schell, "Electrodermal responses to attended and nonattended significant stimuli during dichotic listening," *J. Exp. Psychol. Human*, vol. 8, pp. 315–324, 1982.

[18] J. T. Winslow, L. A. Parr, and M. Davis, "Acoustic startle, prepulse inhibition, and fear-potentiated startle measured in rhesus monkeys," *Biol. Psychiatry*, vol. 51, no. 11, pp. 859–866, 2002.

[19] A. Treisman, "Contextual cues in selective listening," *Q. J. Exp. Psychol.*, vol. 77, pp. 533–546, 1960.

[20] A. Treisman and H. Schmidt, "Illusory conjunctions in the perception of objects," *Cogn. Psychol.*, vol. 14, pp. 107–141, 1982.

[21] M. S. Livingstone and D. H. Hubel, "Segregation of form, color, movement, and depth: Anatomy, physiology, and perception," *Science*, vol. 240, pp. 740–749, 1988.

[22] C. von der Malsburg, "Am I thinking assemblies?," in *Proc. Trieste Meeting Brain Theory*, G. Palm and A. Aertsen, Eds., Berlin, Germany, 1986.

[23] ——, "The correlation theory of brain function.," Max Planck Inst. Biophys. Chem., Göttingen, Germany, Rep. 81-2, 1981.

[24] P. M. Milner, "A model for visual shape recognition," *Psychol. Rev.*, vol. 81, pp. 521–535, 1974.

[25] C. von der Malsburg and W. Schneider, "A neural cocktail-party processor," *Biol. Cybern.*, vol. 54, pp. 29–40, 1986.
[26] C. M. Gray and W. Singer, "Stimulus-specific neuronal oscillations in orientation columns of cat visual cortex," *Proc. Nat. Acad. Sci. USA*, vol. 86, pp. 1698–1702, 1989.
[27] C. M. Gray, P. Koenig, A. K. Engel, and W. Singer, "Oscillatory responses in cat visual cortex exhibit inter-columnar synchronization which reflects global stimulus properties," *Nature*, vol. 338, pp. 334–337, 1989.
[28] M. Joliot, U. Ribary, and R. Llinás, "Human oscillatory brain activity near 40 Hz coexists with cognitive temporal binding," *Proc. Nat. Acad. Sci. USA*, vol. 91, pp. 11748–11751, 1994.
[29] G. Tononi and G. M. Edelman, "Consciousness and complexity," *Science*, vol. 282, pp. 1846–1851, 1998.
[30] G. Tononi, R. Srinivasan, D. P. Russell, and G. M. Edelman, "Investigating neural correlates of conscious perception by frequency-tagged neuromagnetic responses," *Proc. Nat. Acad. Sci. USA*, vol. 95, pp. 3198–3203, 1998.
[31] C. Tallon-Baudry, O. Bertrand, F. Peronnet, and J. Pernier, "Induced gamma band activity during the delay of a visual short-term memory task in humans," *J. Neurosci.*, vol. 18, pp. 4244–4254, 1998.
[32] K. Keil, M. M. Müller, W. J. Ray, T. Gruber, and T. Elbert, "Human gamma band activity and perception of a Gestalt," *J. Neurosci.*, vol. 19, pp. 7152–7161, 1999.
[33] F. Pulvermüller, N. Birbaumer, W. Lutzenberger, and B. Mohr, "High-frequency brain activity: Its possible role in attention, perception and language processing," *Prog. Neurobiol.*, vol. 52, pp. 427–445, 1997.
[34] R. Meddis and M. Hewitt, "Virtual pitch and phase sensitivity of a computer model of the auditory periphery: I. Pitch identification," *J. Acoust. Soc. Amer.*, vol. 89, no. 6, pp. 2866–2882, 1991.
[35] B. C. J. Moore, *An Introduction to the Psychology of Hearing—4th Edition*. New York: Academic, 1997.
[36] D. L. Wang and D. Terman, "Locally excitatory globally inhibitory oscillator networks," *IEEE Trans. Neural Networks*, vol. 6, pp. 283–286, Jan. 1995.
[37] R. D. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice, "An efficient auditory filterbank based on the gammatone function," Applied Psychol. Unit, Univ. Cambridge, Cambridge, U.K., APU Rep. 2341, 1988.
[38] B. R. Glasberg and B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing Res.*, vol. 47, pp. 103–138, 1990.
[39] R. D. Patterson and B. C. J. Moore, "Auditory filters and excitation patterns as representations of frequency resolution," in *Frequency Selectivity in Hearing*, B. C. J. Moore, Ed. New York: Academic, 1986, pp. 123–177.
[40] *Normal Equal-Loudness Level Contours*, ISO Standard 226.
[41] J. O. Pickles, *An Introduction to the Physiology of Hearing*, 2nd ed. New York: Academic, 1988.
[42] J. C. R. Licklider, "A duplex theory of pitch perception," *Experientia*, vol. 7, no. 4, pp. 128–134, 1951.
[43] ——, "Three auditory theories," in *Psychology: A Study of a Science*, S. Koch, Ed. New York: McGraw-Hill, 1959, pp. 41–144.
[44] E. C. Cherry, "Two ears—But one world," in *Sensory Communication*, W. A. Rosenblith, Ed. Cambridge, MA: MIT Press, 1961, pp. 99–117.
[45] P. F. Assmann, "Modeling the perception of concurrent vowels: Role of formant transitions," *J. Acoust. Soc. Amer.*, vol. 100, no. 2, pp. 1141–1152, 1996.
[46] C. J. Darwin, R. W. Hukin, and B. Y. Al-Khatib, "Grouping in pitch perception: Evidence for sequential constraints," *J. Acoust. Soc. Amer.*, vol. 98, no. 2, pp. 880–885, 1995.
[47] D. L. Wang and G. J. Brown, "Separation of speech from interfering sounds based on oscillatory correlation," *IEEE Trans. Neural Networks*, vol. 10, pp. 684–697, May 1999.
[48] M. P. Cooke, *Modeling Auditory Processing and Organization*. Cambridge, U.K.: Cambridge Univ. Press, 1993.
[49] B. van der Pol, "On relaxation oscillations," *Philos. Mag.*, vol. 2, no. 11, pp. 978–992, 1926.
[50] A. L. Hodgkin and A. F. Huxley, "A quantitative description of membrane current and its application to conduction and excitation in the nerve," *J. Physiol.*, vol. 117, pp. 500–544, 1952.
[51] R. FitzHugh, "Impulses and physiological states in models of nerve membrane," *Biophys. J.*, vol. 1, pp. 445–466, 1961.
[52] J. Nagumo, S. Arimoto, and S. Yoshizawa, "An active pulse transmission line simulating nerve axon," *Proc. IEEE*, vol. 50, pp. 2061–2070, 1962.
[53] C. J. Darwin and V. Ciocca, "Grouping in pitch perception: Effects of onset asynchrony and ear of presentation of a mistuned component," *J. Acoust. Soc. Amer.*, vol. 91, no. 6, pp. 3381–3390, 1992.
[54] D. L. Wang, "Primitive auditory segregation based on oscillatory correlation," *Cogn. Sci.*, vol. 20, pp. 409–456, 1996.

Stuart N. Wrigley (M'03) received the B.Sc. and Ph.D. degrees in computer science from the University of Sheffield, Sheffield, U.K., in 1998 and 2002, respectively.

He is currently a Research Associate on the EU Multimodal Meeting Manager (M4) Project within the Department of Computer Science, University of Sheffield. His research interests include auditory selective attention, short-term memory, feature binding, binaural hearing, and multisource signal analysis.

Guy J. Brown received the B.Sc. degree in applied science from Sheffield Hallam University, Sheffield, U.K., in 1988 and the Ph.D. and M.Ed. degrees in computer science from the University of Sheffield, Sheffield, U.K., in 1992 and 1997, respectively.

He is currently a Senior Lecturer in Computer Science at the University of Sheffield. He has a long-standing interest in computational models of auditory perception, and also has research interests in robust automatic speech recognition, spatial hearing, and music technology.