Separation of Speech from Interfering Sounds Based on Oscillatory Correlation

DeLiang L. Wang, *Associate Member, IEEE,* and Guy J. Brown

***Abstract*— A multistage neural model is proposed for an auditory scene analysis task—segregating speech from interfering sound sources. The core of the model is a two-layer oscillator network that performs stream segregation on the basis of oscillatory correlation. In the oscillatory correlation framework, a stream is represented by a population of synchronized relaxation oscillators, each of which corresponds to an auditory feature, and different streams are represented by desynchronized oscillator populations. Lateral connections between oscillators encode harmonicity, and proximity in frequency and time. Prior to the oscillator network are a model of the auditory periphery and a stage in which mid-level auditory representations are formed. The model has been systematically evaluated using a corpus of voiced speech mixed with interfering sounds, and produces improvements in terms of signal-to-noise ratio for every mixture. The performance of our model is compared with other studies on computational auditory scene analysis. A number of issues including biological plausibility and real-time implementation are also discussed.**

***Index Terms*—Auditory scene analysis, harmonicity, oscillatory correlation, speech segregation, stream segregation.**

I. Introduction

IN practically all listening situations, the acoustic waveform reaching our ears is composed of sound energy from multiple environmental sources. Consequently, a fundamental task of auditory perception is to disentangle this acoustic mixture, in order to retrieve a mental description of each sound source. In an influential account, Bregman [6] describes this aspect of auditory function as an *auditory scene analysis* (ASA). Conceptually, ASA may be regarded as a two-stage process. The first stage (which we term "segmentation") decomposes the acoustic mixture reaching the ears into a collection of sensory elements. In the second stage ("grouping"), elements that are likely to have arisen from the same environmental event are combined into a perceptual structure termed a *stream* (an auditory stream roughly corresponds to an object in vision). Streams may be further interpreted by higher-level processes for recognition and scene understanding.

Over the past decade, there has been a growing interest in the development of computational systems which mimic ASA (see [13] for a review). Most of these studies have been motivated by the need for a front-end processor for robust automatic speech recognition in noisy environments. Early work includes the system of Weintraub [57], which attempted to separate the voices of two speakers by tracking their fundamental frequencies (see also the nonauditory work of Parsons [40]). More recently, a number of multistage computational models have been proposed by Cooke [12], Mellinger [35], Brown and Cooke [7], and Ellis [16]. Generally, these systems process the acoustic input with a model of peripheral auditory function, and then extract features such as onsets, offsets, harmonicity, amplitude modulation and frequency modulation. Scene analysis is accomplished by symbolic search algorithms or high-level inference engines that integrate a number of features. Recent developments of such systems have focussed on increasingly sophisticated computational architectures, based on the multiagent paradigm [37] or evidence combination using Bayesian networks [26]. Hence, although reasonable performances are reported for these systems using real acoustic signals, the grouping algorithms employed tend to be complicated and computationally intensive.

Currently, computational ASA remains an unsolved problem for real-time engineering applications such as automatic speech recognition. Given the impressive advance in speech recognition technology in recent years, the lack of progress in computational ASA now represents a major hurdle to the application of speech recognition in unconstrained acoustic environments.

The current state of affairs in computational ASA stands in sharp contrast to the fact that humans and higher animals can perceptually segregate sound sources with apparent ease. It seems likely, therefore, that computational systems which are more closely modeled on the neurobiological mechanisms of hearing may offer performance advantages over current approaches. This observation—together with the motivation for understanding the neurobiological basis of ASA—has prompted a number of investigators to propose neural-network models of ASA. Perhaps the first of these was the neural-network model described by von der Malsburg and Schneider [52]. In an extension of the *temporal correlation* theory proposed earlier by von der Malsburg [51], they suggested that neural oscillations could be used to represent auditory grouping. In their scheme, a set of auditory elements forms a perceptual stream if the corresponding oscillators are synchronized (phase locked with no phase lag), and are desyn-

Manuscript received June 16, 1998; revised January 11, 1999. This work was mainly undertaken while G. J. Brown was a visiting scientist at the Center for Cognitive Science, The Ohio State University. The work of D. L. Wang was supported in part by an NSF Grant (IRI-9 423 312) and an ONR Young Investigator Award (N00014-96-1-0767). The work of G. J. Brown was supported by EPSRC under Grant GR/K18962.

D. L. Wang is with the Department of Computer and Information Science and Center for Cognitive Science, The Ohio State University, Columbus, OH 43210-1277 USA.

G. J. Brown is with the Department of Computer Science, University of Sheffield, Sheffield S8 0ET, U.K.

Publisher Item Identifier S 1045-9227(99)03831-X.

chronized from oscillators that represent different streams. On the basis of this representation, Wang [53], [55] later proposed a neural architecture for auditory organization (see also Brown and Cooke [9] for a different account also based on oscillations). Wang's architecture is based on new insights into locally excitatory globally inhibitory networks of relaxation oscillators [49], which take into consideration the topological relations between auditory elements. This *oscillatory correlation* framework [55] may be regarded as a special form of temporal correlation. Recently, Brown and Wang [10] gave an account of concurrent vowel separation based on oscillatory correlation.

The oscillatory correlation theory is supported by neurobiological findings. Galambos *et al.* [20] first reported that auditory evoked potentials in human subjects show 40 Hz oscillations. Subsequently, Ribary *et al.* [42] and Llinás and Ribary [29] recorded 40 Hz activity in localized brain regions, both at the cortical level and at the thalamic level in the auditory system, and demonstrated that these oscillations are synchronized over widely separated cortical areas. Furthermore, Joliot *et al.* [25] reported evidence directly linking coherent 40-Hz oscillations with the perceptual grouping of clicks. These findings are consistent with reports of coherent 40-Hz oscillations in the visual system (see [46] for a review) and the olfactory system (see [18] for a review). Recently, Maldonado and Gerstein [30] observed that neurons in the auditory cortex exhibit synchronous oscillatory firing patterns. Similarly, deCharms and Merzenich [15] reported that neurons in separate regions of the primary auditory cortex synchronize the timing of their action potentials when stimulated by a pure tone. Also, Barth and MacDonald [2] have reported evidence suggesting that oscillations originating in the auditory cortex can be modulated by the thalamus, and that these synchronous oscillations are underlain by intracortical interactions.

Currently, however, the performance of neural-network models of ASA is quite limited. Generally, these models have attempted to reproduce simple examples of auditory stream segregation using stimuli such as alternating pure-tone sequences [9], [55]. Even in [10], which models the segregation of concurrent vowel sounds, the neural network operates on a single time frame and is therefore unable to segregate time-varying sounds.

Here, we study ASA from a neurocomputational perspective, and propose a neural network model that is able to segregate speech from a variety of interfering sounds, including music, "cocktail party" noise, and other speech. Our model uses oscillatory correlation as the underlying neural mechanism for ASA. As such, it addresses auditory organization at two levels; at the functional level, it explains how an acoustic mixture is parsed to retrieve a description of each source (the ASA problem), and at the neurobiological level, it explains how features that are represented in distributed neural structures can be combined to form meaningful wholes (the *binding problem*). We note that the binding problem is inherent in Bregman's notion of a two-stage ASA process, although it is only briefly discussed in his account [6].

In our model, a stream is formed by synchronizing oscillators in a two-dimensional time-frequency network. Lateral connections between oscillators encode proximity in frequency and time, and link oscillators that are stimulated by harmonically related components. Time plays two different roles in our model. One is "external" time in which auditory stimuli are embedded; it is explicitly represented as a separate dimension. Another is "internal" time, which embodies oscillatory correlation as a binding mechanism. The model has been systematically evaluated using a corpus of voiced speech mixed with interfering sounds. For every mixture, an increase in signal-to-noise ratio (SNR) is obtained after segregation by the model.

The remainder of this article is organized as follows. In the next section, the overall structure of the model is briefly reviewed. Detailed explanations of the auditory periphery model, mid-level auditory representations, neural oscillator network and resynthesis are then presented. A systematic evaluation of the sound-separation performance of the model is given in Section VII. Finally, we discuss the relationship between our neural oscillator model and previous approaches to computational ASA, and conclude with a general discussion.

II. Model Overview

In this section we give an overview of the model and briefly explain each stage of processing. Broadly speaking, the model comprises four stages, as shown in Fig. 1. The input to the model consists of a mixture of speech and an interfering sound source, sampled at a rate of 16 kHz with 16 bit resolution. In the first stage of the model, peripheral auditory processing is simulated by passing the input signal through a bank of cochlear filters. The gains of the filters are chosen to reflect the transfer function of the outer and middle ears. In turn, the output of each filter channel is processed by a model of hair cell transduction, giving a probabilistic representation of auditory nerve firing activity which provides the input to subsequent stages of the model.

The second stage of the model produces so-called "mid-level" auditory representations (see also Ellis and Rosenthal [17]). The first of these, the *correlogram*, is formed by computing a running autocorrelation of the auditory nerve activity in each filter channel. Correlograms are computed at 10-ms intervals, forming a three-dimensional volume in which time, channel center frequency and autocorrelation lag are represented on orthogonal axes (see the lower left panel in Fig. 1). Additionally, a "pooled" correlogram is formed at each time frame by summing the periodicity information in the correlogram over frequency. The largest peak in the pooled function occurs at the period of the dominant fundamental frequency (F0) in that time frame; the third stage of the model uses this information to group acoustic components according to their F0's. Further features are extracted from the correlogram by a cross-correlation analysis. This is motivated by the observation that filter channels with center frequencies that are close to the same harmonic or formant exhibit similar patterns of periodicity. Accordingly, we compute a running cross-correlation between adjacent correlogram channels, and this provides the basis for segment formation in the third stage of the model.

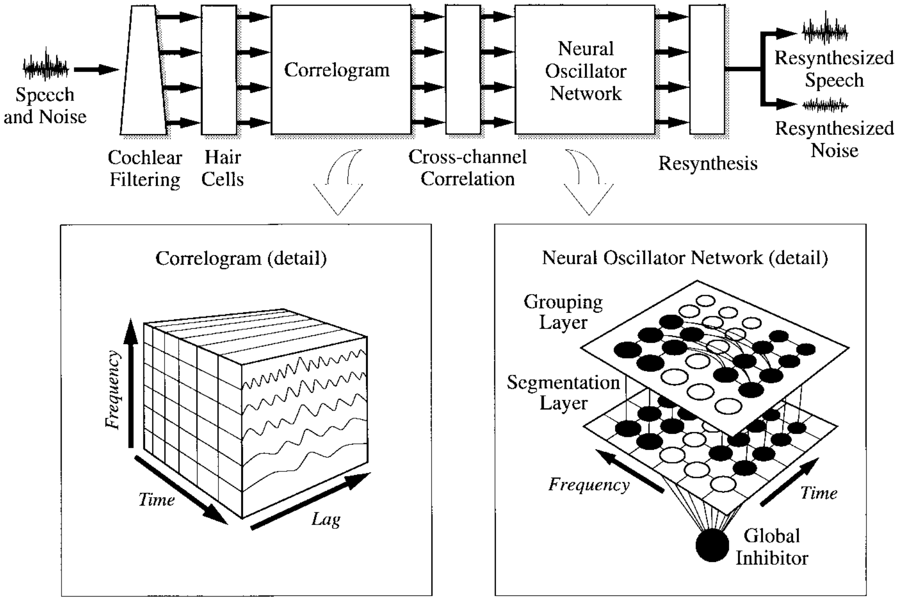


Fig. 1. Schematic diagram of the model. A mixture of speech and noise is processed in four main stages. In the first stage, simulated auditory nerve activity is obtained by passing the input through amodel of the auditory periphery (cochlear filtering and hair cells). Mid-level auditory representations are then formed (correlogram and cross-channel correlation map). Subsequently, a two-layer oscillator network performs grouping of acoustic components. Finally, are synthesis path allows the separation performance to be evaluated by listening tests or computation of signal-to-noise ratio.

The third stage comprises the core of our model, in which auditory organization takes place within a two-layer oscillator network (see the lower right panel of Fig. 1). The first layer produces a collection of segments that correspond to elementary structures of an auditory scene, and the second layer groups segments into streams.

The first layer is a locally excitatory globally inhibitory oscillator network (LEGION) composed of relaxation oscillators. This layer is a two-dimensional network with respect to time and frequency, in which the connection weights along the frequency axis are derived from the cross-correlation values computed in the second stage. Synchronized blocks of oscillators (*segments*) form in this layer, each block corresponding to a connected region of acoustic energy in the time-frequency plane. Different segments are desynchronized. Conceptually, segments are the atomic elements of a represented auditory scene; they capture the evolution of perceptually-relevant acoustic components in time and frequency. As such, a segment cannot be decomposed by further processing stages of the model, but it may group with other segments in order to form a stream.

The oscillators in the second layer are linked by two kinds of lateral connections. The first kind consist of mutual excitatory connections between oscillators within the same segment. The formation of these connections is based on the input from the first layer. The second kind consist of lateral connections between oscillators of different segments, but within the same time frame. In light of the time-frequency layout of the oscillator network, these connections along the frequency axis are termed *vertical connections* (see Fig. 1). Vertical connections may be excitatory or inhibitory; the connections between two oscillators are excitatory if their corresponding frequency channels either both agree or both disagree with the F0 extracted from the pooled correlogram for that time frame; otherwise, the connections are inhibitory. Accordingly, the second layer groups a collection of segments to form a "foreground" stream that corresponds to a synchronized population of oscillators, and puts the remaining segments into a "background" stream that also corresponds to a synchronized population. The background population is desynchronized from the foreground population. Hence, the second layer embodies the result of ASA in our model, in which one sound source (foreground) and the rest (background) are separated according to a F0 estimate.

The last stage of the model is a resynthesis path, which allows an acoustic waveform to be derived from the time-frequency regions corresponding to a group of oscillators. Resynthesized waveforms can be used to assess the performance of the model in listening tests, or to quantify the SNR after segregation.

III. Auditory Periphery Model

It is widely recognized that peripheral auditory frequency selectivity can be modeled by a bank of bandpass filters with overlapping passbands (for example, see Moore [36]). In this study, we use a bank of "gammatone" filters [41] which have an impulse response of the following form:

$$g_i(t) = t^{n-1}\exp(-2\pi b_i t)\cos(2\pi f_i t + \phi_i)H(t) \quad (1 \le i \le N). \quad (1)$$

Here, N is the number of filter channels, n is the filter order and H is the unit step function (i.e., $H(x) = 1$ for $x \ge 0$, and zero otherwise). Hence, the gammatone is a causal filter with an infinite response time. For the ith filter channel, f_i is the center frequency of the filter (in Hz), ϕ_i is the phase (in radians) and b_i determines the rate of decay of the impulse response, which is related to bandwidth. We use an implementation of the fourth-order gammatone filter proposed by Cooke [12], in which an impulse invariant transform is used to map the continuous impulse response given in (1) to the digital domain. Since the segmentation and grouping stages of our model do not require the correction of phase delays introduced by the filterbank, we set $\phi_i = 0$.

Physiological studies of auditory nerve tuning curves [39] and psychophysical studies of critical bandwidth [21] indicate that auditory filters are distributed in frequency according to their bandwidths, which increase quasilogarithmically with increasing center frequency. Here, we set the bandwidth of each filter according to its equivalent rectangular bandwidth (ERB), a psychophysical measurement of critical bandwidth in human subjects (see Glasberg and Moore [21])

$$ERB(f) = 24.7(4.37f/1000 + 1). \tag{2}$$

More specifically, we define

$$b_i = 1.019\,ERB(f_i) \tag{3}$$

and use a bank of 128 gammatone filters (i.e., $N = 128$) with center frequencies equally distributed on the ERB scale between 80 Hz and 5 kHz. Additionally, the gains of the filters are adjusted according to the ISO standard for equal loudness contours [24] in order to simulate the pressure gains of the outer and middle ears.

Our use of the gammatone filter is consistent with a neurobiological modeling perspective. Equation (1) provides a close approximation to experimentally derived auditory nerve fiber impulse responses, as measured by de Boer and de Jongh [14] using a reverse-correlation technique. Additionally, the fourth-order gammatone filter provides a good match to psychophysically derived "rounded-exponential" models of human auditory filter shape [41]. Hence, the gammatone filter is in good agreement with both neurophysiological and psychophysical estimates of auditory frequency selectivity.

In the final stage of the peripheral model, the output of each gammatone filter is processed by the Meddis [32] model of inner hair cell function. The output of the hair cell model is a probabilistic representation of firing activity in the auditory nerve, which incorporates well-known phenomena such as saturation, two-component short-term adaptation and frequency-limited phase locking.

IV. Mid-Level Auditory Representations

There is good evidence that mechanisms similar to those underlying pitch perception can contribute to the perceptual segregation of sounds which have different F0's. For example, Scheffers [43] has shown that the ability of listeners to identify two concurrent vowels is improved when they have different F0's, relative to the case in which they have the same F0. Similar findings have been obtained by Brokx and Nooteboom [5] using continuous speech.

Accordingly, the second stage of our model identifies periodicities in the simulated auditory nerve firing patterns. This is achieved by computing a *correlogram*, which is one member of a class of pitch models in which periodicity information is combined from resolved (low-frequency) and unresolved (high-frequency) harmonic regions. The correlogram is able to account for many classical pitch phenomena [33], [47]; additionally, it may be regarded as a functional description of auditory mechanisms for amplitude-modulation detection, which have been shown to exist in the auditory mid-brain [19]. Other workers have employed the correlogram as a mechanism for segregating concurrent periodic sounds with some success (for example, see Assmann and Summerfield [1]; Meddis and Hewitt [34]; Brown and Cooke [7]; Brown and Wang [10]).

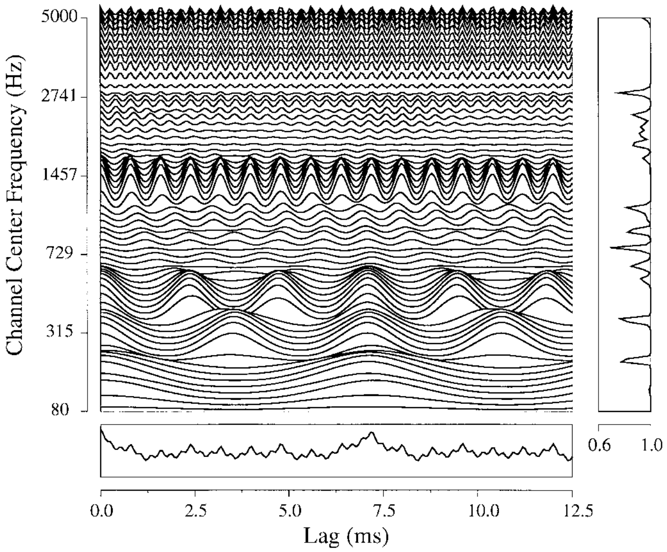


Fig. 2. A correlogram of a mixture of speech and trill telephone, taken at time frame 45 (i.e., 450 ms after the start of the stimulus). The large panel in the center of the figure shows the correlogram; for clarity, only the autocorrelation function of every second channel is shown, resulting in 64 filter channels. The pooled correlogram is shown in the bottom panel, and the cross-correlation function is shown on the right.

A correlogram is formed by computing a running autocorrelation of the simulated auditory nerve activity in each frequency channel. At a given time step j, the autocorrelation $A(i, j, \tau)$ for channel i with a time lag τ is given by

$$A(i,j,\tau) = \sum_{k=0}^{K-1} r(i,j-k)r(i,j-k-\tau)w(k). \tag{4}$$

Here, r is the output of the hair cell model (i.e., the probability of a spike occurring in the auditory nerve) and w is a rectangular window of width K time steps. We use $K = 320$, corresponding to a window width of 20 ms. The autocorrelation lag τ is computed in L steps of the sampling period Δt, between 0 and $L - 1$. Here we use $L = 201$, corresponding to a maximum delay of 12.5 ms; this is appropriate for the current study, since the F0 of voiced speech in our test set does not fall below 80 Hz. Equation (4) is computed for M time frames, each taken at intervals of 10 ms (i.e., at intervals of 160 steps of the time index j). Hence, the correlogram is a three-dimensional volume of size $N \times M \times L$ in which each element $A(i, j, \tau)$ represents the auditory nerve firing rate for a frequency channel i at time step j and autocorrelation lag τ (see the lower left panel of Fig. 1).

For periodic sounds, a characteristic "spine" appears in the correlogram which is centred on the lag corresponding to the stimulus period (see Fig. 2). This pitch-related structure can be emphasized by summing the channels of the correlogram across frequency, yielding a "pooled" correlogram. Formally, we define the pooled correlogram $s(j,\tau)$ at time frame j and lag τ as follows:

$$s(j,\tau)=\sum_{i=1}^{N}A(i,j,\tau). \tag{5}$$

Several studies [47], [33] have demonstrated that there is a close correspondence between the position of the peak in the pooled correlogram and perceived pitch. Additionally, the height of the peak in the pooled correlogram may be interpreted as a measure of pitch strength. A pooled correlogram is shown in the lower panel of Fig. 2 for one time frame of a mixture of speech and trill telephone. In this frame, the F0 of the speech is close to 139 Hz, giving rise to a peak in the pooled correlogram at 7.2 ms. Note that periodicities due to the telephone ring (which dominate the high-frequency region of the correlogram and a band at 1.4 kHz) also appear as regularly spaced peaks in the pooled function.

It is also apparent from Fig. 2 that correlogram channels which lie close to the same harmonic or formant share a very similar pattern of periodicity (see also Shamma [45]). This redundancy can be exploited in order to group channels of the correlogram that are excited by the same acoustic component (see also Brown and Cooke [7]). Here, we quantify the similarity of adjacent channels in the correlogram by computing a cross-channel correlation metric. Specifically, each channel i at time frame j is correlated with the adjacent channel $i+1$ as follows:

$$C(i,j)=\frac{1}{L}\sum_{\tau=0}^{L-1}\hat{A}(i,j,\tau)\hat{A}(i+1,j,\tau)\quad(1\le i\le N-1). \tag{6}$$

Here, $\hat{A}(i,j,\tau)$ is the autocorrelation function of (4) which has been normalized to have zero mean and unity variance (this ensures that $C(i,j)$ is sensitive only to the pattern of periodicity in the correlogram, and not to the mean firing rate in each channel). The right panel of Fig. 2 shows $C(i,j)$ for the speech and telephone example. It is clear that the correlation metric provides a good basis for identifying harmonics and formants, which are apparent as bands of high cross-channel correlation. Similarly, adjacent acoustic components are clearly separated by regions of low correlation.

Our mid-level auditory representations are well supported by the physiological literature. Neurons that are tuned to preferred rates of periodicity are found throughout the auditory system (for example, see [19]). Furthermore, Schreiner and Langner [44] have presented evidence that frequency and periodicity are systematically mapped in the inferior colliculus, a region of the auditory mid-brain. Inferior colliculus neurons with the same characteristic frequency are organized into layers, and neurons within each layer are tuned to a range of periodicities between 10 Hz and 1 kHz. Additionally, separate iso-frequency layers are connected by interneurons [38]. Hence, it appears that the neural architecture of the inferior colliculus is analogous to the correlogram described here, and that physiological mechanisms exist for combining periodicity information across frequency regions (as in the computation of our pooled correlogram function). Similarly, Carney [11] has identified neurons which receive convergent inputs from auditory nerve fibers with different characteristic frequencies. These neurons appear to behave as cross-correlators, and hence they might be functionally equivalent to the cross-channel correlation mechanism described here.

V. Grouping and Segregation by a Two-Layer Oscillator Network

In our model, the two conceptual stages of ASA (segmentation and grouping) take place within an oscillatory correlation framework. This approach has a number of advantages. Oscillatory correlation is consistent with neurophysiological findings, giving our model a neurobiological foundation. In terms of functional considerations, a neural-network model has the characteristics of parallel and distributed processing. Also, the results of ASA arise from emergent behavior of the oscillator network, in which each oscillator and each connection is easily interpreted. The use of neural oscillators gives rise to a dynamical systems approach, where ASA proceeds as an autonomous and dynamical process. As a result, the model can be implemented as a real-time system, a point of discussion in Section IX.

The basic unit of our network is a single oscillator, which is defined as a reciprocally connected excitatory variable x_{ij} and inhibitory variable y_{ij}. Since each layer of the network takes the form of a two-dimensional time-frequency grid (see Fig. 1), we index each oscillator according to its frequency channel (i) and time frame (j)

$$\dot{x}_{ij}=3x_{ij}-x_{ij}^{3}+2-y_{ij}+I_{ij}+S_{ij}+\rho \tag{7a}$$

$$\dot{y}_{ij}=\varepsilon(\gamma(1+\tanh(x_{ij}/\beta))-y_{ij}). \tag{7b}$$

Here, I_{ij} represents external stimulation to the oscillator, S_{ij} denotes the overall coupling from other oscillators in the network, and ρ is the amplitude of a Gaussian noise term. In addition to testing the robustness of the system, the purpose of including noise is to assist desynchronization among different oscillator blocks.

We choose ε to be a small positive number. Thus, if coupling and noise are ignored and I_{ij} is a constant, (7) defines a typical relaxation oscillator with two time scales, similar to the van der Pol oscillator [50]. The x-nullcline, i.e., $\dot{x}_{ij}=0$, is a cubic function and the y-nullcline is a sigmoid function. If $I_{ij}>0$, the two nullclines intersect only at a point along the middle branch of the cubic with β chosen small. In this case, the oscillator gives rise to a stable limit cycle for all sufficiently small values of ε, and is referred to as *enabled* [see Fig. 3(A)]. The limit cycle alternates between *silent* and *active* phases of near steady-state behavior, and these two phases correspond to the left branch (LB) and the right branch (RB) of the cubic, respectively. The oscillator is called active if it is in the active phase. Compared to motion within each phase, the alternation between the two phases takes place rapidly, and it is referred to as *jumping*. The parameter γ determines the relative times that the limit cycle spends in the two phases—a larger γ produces a relatively shorter active phase. If $I_{ij}<0$, the two nullclines

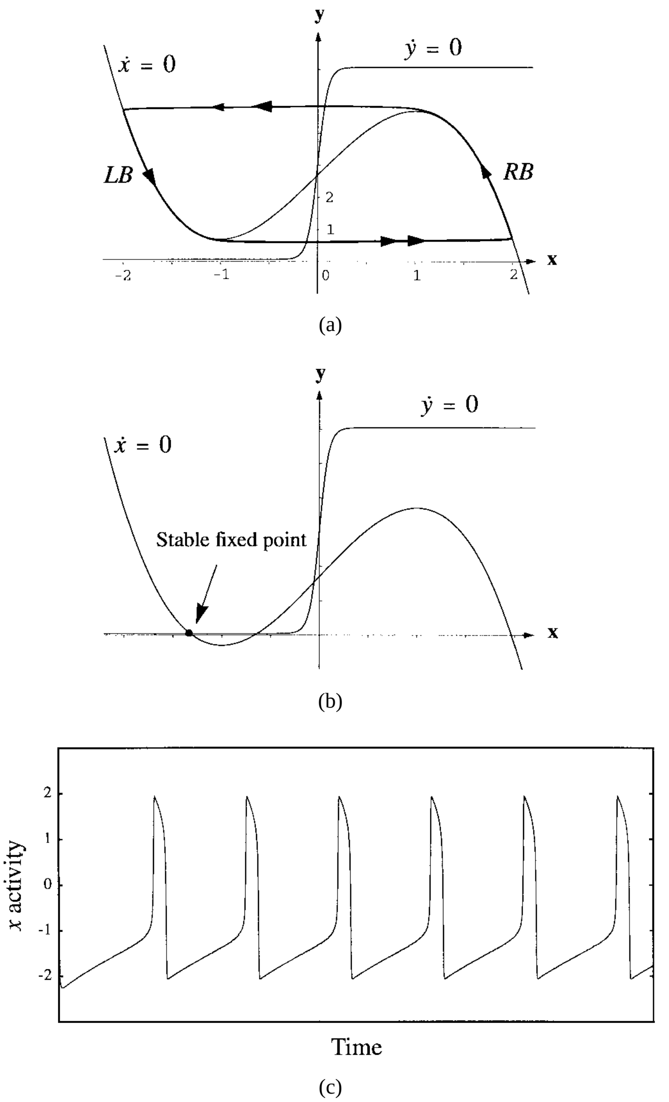


Fig. 3. Nullclines and trajectories of a single relaxation oscillator. (a) Behavior of an enabled oscillator. The bold curve shows the limit cycle of the oscillator, whose direction of motion is indicated by arrowheads. LB and RB indicate the left branch and the right branch of the cubic. (b) Behavior of an excitable oscillator. The oscillator approaches the stable fixed point. (c) Temporal activity of the oscillator. The x value of the oscillator is plotted. The parameter values are: $I = 0.8$, $\rho = 0.02$, $\varepsilon = 0.04$, $\gamma = 9.0$, and $\beta = 0.1$.

of (7) intersect at a stable fixed point on LB of the cubic [see Fig. 3(b)]. In this case no oscillation occurs, and the oscillator is called *excitable*, meaning that it can be induced to oscillate. We call an oscillator *stimulated* if $I_{ij} > 0$, and unstimulated if $I_{ij} \leq 0$. It should be clear, therefore, that oscillations in (7) are stimulus-dependent.

The above definition and description of a relaxation oscillator follows Terman and Wang [49]. The oscillator may be interpreted as a model of action potential generation or oscillatory burst envelope, where x represents the membrane potential of a neuron and y represents the level of activation of a number of ion channels. Fig. 3(c) shows a typical trace of x activity.

A. First Layer: Segment Formation

In the first layer of the network, *segments* are formed—groups of synchronised oscillators that trace the evolution of an acoustic component through time and frequency. Segments may be regarded as atomic elements of the auditory scene, in the sense that they cannot be decomposed by later stages of processing.

The first layer is a two-dimensional time-frequency grid of oscillators with a global inhibitor (see Fig. 1). Accordingly, S_{ij} in (7) is defined as

$$S_{ij} = \sum_{kl \in N(i,j)} W_{ij,kl} H(x_{kl} - \theta_x) - W_z H(z - \theta_z) \qquad (8)$$

where $W_{ij,kl}$ is the connection weight from an oscillator (i,j) to an oscillator (k,l) and $N(i,j)$ is the set of nearest neighbors of the grid location (i,j). Here, $N(i,j)$ is chosen to be the four nearest neighbors, and θ_x is a threshold, which is chosen between LB and RB. Thus an oscillator has no influence on its neighbors unless it is in the active phase. The weight of the neighboring connections along the time axis is uniformly set to one. The weight of vertical connections between an oscillator (i,j) and its neighbor $(i+1,j)$ is set to one if the cross-correlation $C(i,j)$ exceeds a threshold θ_c; otherwise it is set to zero. Here, we set $\theta_c = 0.985$ for all the following simulations.

W_z in (8) is the weight of inhibition from the global inhibitor z, defined as

$$\dot{z} = \sigma_\infty - z \qquad (9)$$

where $\sigma_\infty = 1$ if $x_{ij} \geq \theta_z$ for at least one oscillator (i,j), and $\sigma_\infty = 0$ otherwise. Hence θ_z is another threshold. If $\sigma_\infty = 1$, $z \rightarrow 1$.

Small segments may form which do not correspond to perceptually significant acoustic components. In order to remove these noisy fragments from the auditory scene, we follow [56] by introducing a lateral potential, p_{ij}, for oscillator (i,j), defined as

$$\dot{p}_{ij} = (1 - p_{ij}) H \left[\sum_{kl \in N_p(i,j)} H(x_{kl} - \theta_x) - \theta_p \right] - \varepsilon p_{ij} \quad (10)$$

where $N_p(i,j)$ is called the potential neighborhood of (i,j), which is chosen to be the left neighbor $(i, j-1)$ and the right neighbor $(i, j+1)$. θ_p is a threshold, chosen to be 1.5. Thus if both the left and right neighbor of (i,j) are active, p_{ij} approaches one on a fast time scale; otherwise, p_{ij} relaxes to zero on a slow time scale determined by ε.

The lateral potential, p_{ij}, plays its role through a gating term on I_{ij} of (7a). In other words, (7a) is now replaced by

$$\dot{x}_{ij} = 3x_{ij} - x_{ij}^3 + 2 - y_{ij} + I_{ij} H(p_{ij} - \theta) + S_{ij} + \rho. \qquad (7a1)$$

With p_{ij} initialized to one, it follows that p_{ij} will drop below the threshold θ in (7a1) unless (i,j) receives excitation from its entire potential neighborhood.

Through lateral interactions in (10), the oscillators that maintain high potentials are those that have both their left and right neighbors stimulated. Such oscillators are called *leaders*. Besides leaders, we distinguish *followers* and *loners*. Followers are those oscillators that can be recruited to jump by leaders, and loners are those stimulated oscillators which belong to noisy fragments. Loners will not be able to jump up beyond a short initial time, because they can neither become leaders and thus jump by themselves, nor be recruited because they are not near leaders. We call the collection of all noisy regions corresponding to loners the *background*, which is generally discontiguous.

An oscillator at grid location (i, j) is stimulated if its corresponding input $I_{ij} > 0$. Some channels of the correlogram may have a low energy at particular time frames, indicating that they are not being excited by an acoustic component. The oscillators corresponding to such time-frequency locations do not receive an input; this is ensured by setting an energy threshold θ_a. It is evident from (4) that the energy in a correlogram channel i at time j corresponds to $A(i, j, 0)$, i.e., the autocorrelation at zero lag. Thus, we define the input I_{ij} as follows:

$$I_{ij} = \begin{cases} 0.2, & \text{if } A(i,j,0) > \theta_a \\ -5, & \text{otherwise.} \end{cases} \tag{11}$$

Here, we set $\theta_a = 50$, which is close to the spontaneous rate of the hair cell model.

Wang and Terman [56] have proven a number of mathematical results about the LEGION system defined in (7)–(10). These analytical results ensure that loners will stop oscillating after an initial brief time period; after a number of oscillation cycles a block of oscillators corresponding to a significant region will synchronize, while oscillator blocks corresponding to distinct regions will desynchronize from each other. A significant region corresponds to an oscillator block that can produce at least one leader. The choice of $N_p(i, j)$ in (10) implies that a segment, or a significant region, extends at least for three consecutive time frames. Regarding the speed of computation, the number of cycles required for full segregation is no greater than the number of segments plus one.

We use the LEGION algorithm described in [55] and [56] for all of our simulations, because integrating a large system of differential equations is very time-consuming. The algorithm follows the major steps in dynamic evolution of the differential equations, and maintains the essential characteristics of the LEGION network, such as two time scales and properties of synchrony and desynchrony. The derivation of the algorithm is straightforward and will not be discussed here. A major difference between the algorithm and the dynamics is that the algorithmic version does not exhibit a *segmentation capacity*, which refers to the maximum number of segments that can be separated by a LEGION network. It is known that a LEGION network, with a fixed set of parameters, has a limited capacity [56]. Given that many segments may be formed at this oscillator layer, we choose the algorithmic version for convenience in addition to saving computing time. The following parameters are either incorporated into algorithmic steps or eliminated: θ, ρ, ε, γ, β, θ_x and θ_z.

As an example, Fig. 4 shows the results of segmentation by the first layer of the network for a mixture of speech and trill telephone (one frame of this mixture was shown in Fig. 2). The size of the network is 128 × 150, representing 128 (N) frequency channels and 150 (M) time frames. The parameter W_z is set to 0.5. Each segment in Fig. 4 is represented by a distinct gray-level; the system produces 94 segments plus the background, which consists of small components lasting just one or two time frames. Not every segment is discernible in Fig. 4 due to the large number of segments. Also, it should be noted that although all segments are shown together in Fig. 4, each arises during a unique time interval in accordance with the principle of oscillatory correlation (see Figs. 6 and 7 for an illustration).

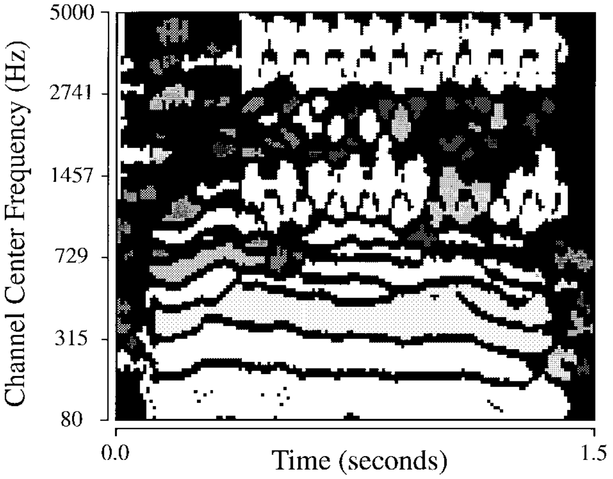


Fig. 4. The result of segment formation for the speech and telephone mixture, generated by the first layer of the network. Each segment is indicated by a distinct gray-level in a grid of size 128 (frequency channels) by 150 (time frames). Unstimulated oscillators and the background are indicated by black areas. In this case, 94 segments are produced.

B. Second Layer: Grouping

The second layer is a two-dimensional network of laterally connected oscillators without global inhibition, which embodies the grouping stage of ASA. An oscillator in this layer is stimulated if its corresponding oscillator in the first layer is either a leader or a follower. Also, the oscillators initially have the same phase, implying that all segments from the first layer are assumed to be in the same stream. More specifically, all stimulated oscillators start at the same randomly placed position on LB [see Fig. 3(a)]. This initialization is consistent with psychophysical evidence suggesting that perceptual fusion is the default state of auditory organization [6]. The model of a single oscillator is the same as in (7), except that x_{ij} is changed slightly to

$$\dot{x}_{ij} = 3x_{ij} - x_{ij}^3 + 2 - y_{ij} + I_{ij}[1 + \mu H(p_{ij} - \theta)] + S_{ij} + \rho. \tag{7a2}$$

Here μ is a small positive parameter. The above equation implies that a leader with a high lateral potential gets a slightly higher external input. We choose $N_p(i, j)$ and θ_p [see (10)] so that leaders are only those oscillators that correspond to part of the longest segment from the first layer. How to select a particular segment, such as the largest one, in an oscillator network was recently addressed in [54]. With this selection mechanism it is straightforward to extract the longest segment from the first layer. Because oscillators have the same initial phase on LB, leaders with a higher external input have a higher cubic (see Fig. 3), and thus will jump to RB first.

The coupling term S_{ij} in (7a2) consists of two types of lateral coupling, but does not include a global inhibition term

$$S_{ij} = S_{ij}^e + S_{ij}^v. \tag{12}$$

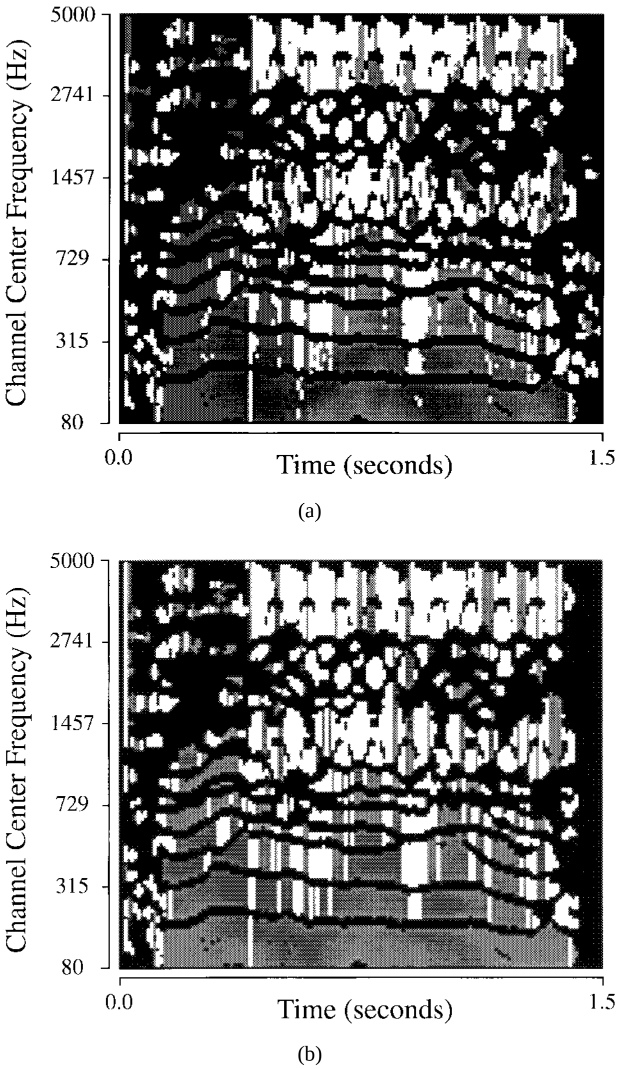
Here S_{ij}^e represents mutual excitation between the oscillators within each segment. Specifically, $S_{ij}^e = W_{e1}$ if the active oscillators from the same segment occupy more than half of the length of the segment; otherwise $S_{ij}^e = W_{e2}$ if there is at least one active oscillator from the same segment.

The coupling term S_{ij}^v denotes vertical connections between oscillators corresponding to different frequency channels and different segments, but within the same time frame. At each time frame, a F0 estimate from the pooled correlogram (5) is used to classify frequency channels into two categories: a set of channels, P, that are consistent with the F0, and a set of channels that are not. More specifically, given a delay τ_m at which the largest peak occurs in the pooled correlogram, for each channel i at time frame j, $i \in P$ if

$$A(i,j,\tau_m)/A(i,j,0) > \theta_d. \tag{13}$$

Note that (13) amounts to classification on the basis of an energy threshold, since $A(i,j,0)$ corresponds to the energy in channel i at time j. Our observations suggest that this method is more reliable than conventional peak detection, since low-frequency channels of the correlogram tend to exhibit very broad peaks (see Fig. 2). The delay τ_m can be found by using a winner-take-all network, although for simplicity we apply a maximum selector in the current implementation. The threshold θ_d is chosen to be 0.95. Note that (13) is applied only to a channel whose corresponding oscillator belongs to a segment from the first layer, and not to a channel whose corresponding oscillator is either a loner or unstimulated. As an example, Fig. 5(a) displays the result of channel classification for the speech and telephone mixture. In the figure, gray pixels correspond to the set P, white pixels correspond to the set of channels that do not agree with the F0, and black pixels represent loners or unstimulated oscillators.

The classification process described above operates on channels, rather than segments. As a result, channels within the same segment at a particular time frame may be allocated to different pitch categories [see, for example, the bottom segment in Fig. 5(a)]. Once segments are formed, our model does not allow them to be decomposed; hence, we enforce a rule that all channels of the same frame within each segment must belong to the same pitch category as that of the majority of channels. After this conformational step, vertical connections are formed such that, at each time frame, two oscillators of different segments have mutual excitatory links if the two corresponding channels belong to the same pitch category; otherwise they have mutual inhibitory links. Furthermore, $S_{ij}^v = W_i$ if (i,j) receives an input from its inhibitory links—this occurs when some active oscillators have inhibitory connections with (i,j). Otherwise, $S_{ij}^v = W_e$ if (i,j) receives any excitation from its vertical excitatory links. After the lateral connections are formed, the oscillator network is numerically solved using a recently proposed method, called the singular limit method [28], for integrating relaxation oscillator networks.



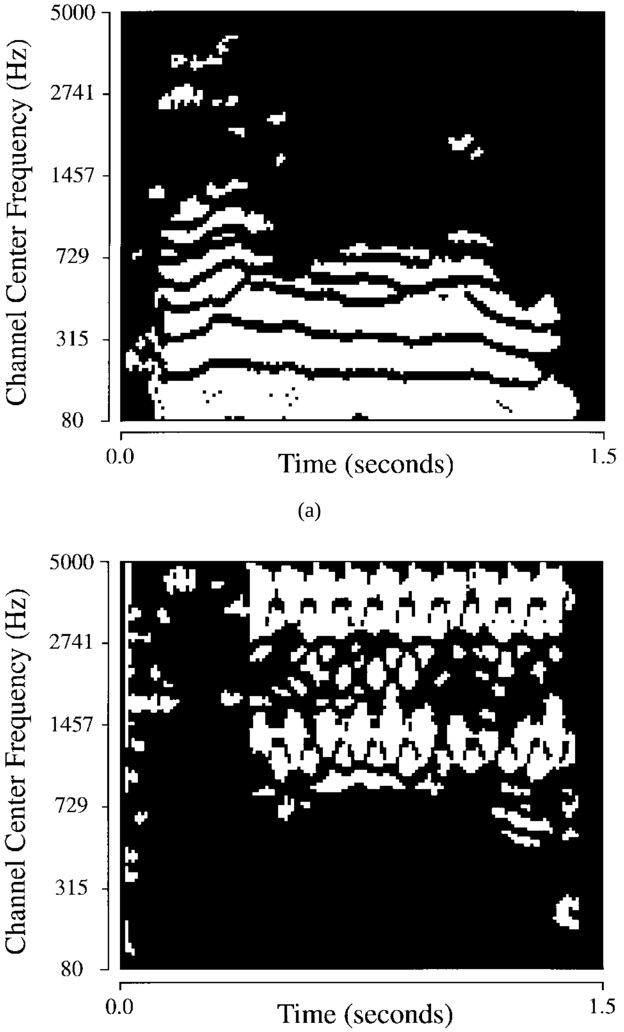
(a)

(b)

Fig. 5. (a) Channel categorization of all segments in the first layer of the network, for the speech and telephone mixture. Gray pixels represent the set P, and white pixels represent channels that do not agree with the F0. (b) Result of channel categorization after conformation and trimming by the longest segment.

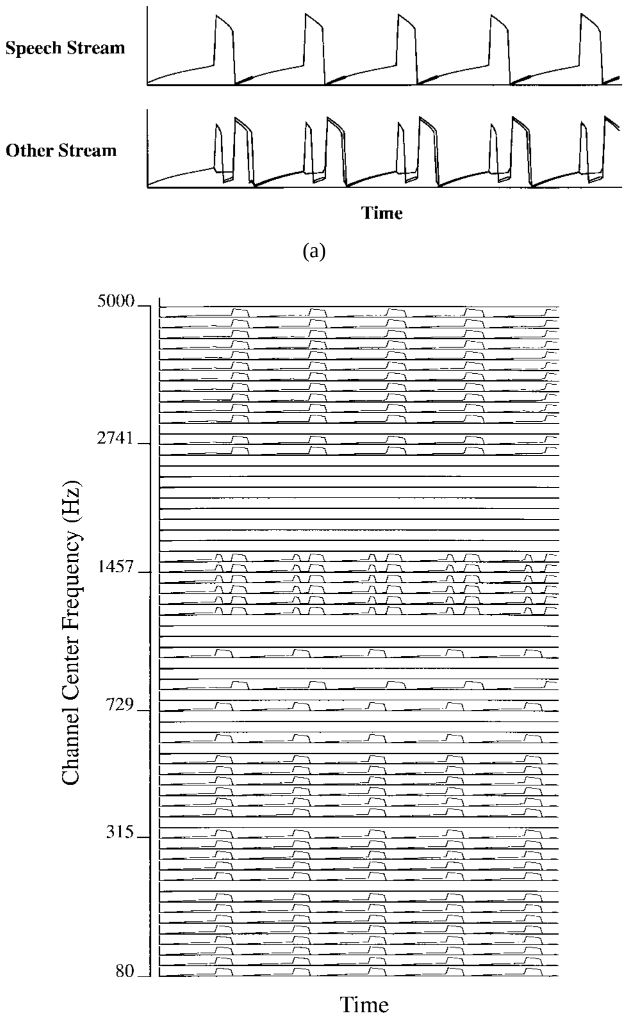
At present, our model does not address sequential grouping; in other words, there is no mechanism to group segments that do not overlap in time. Lacking this mechanism, we limit operation of the second layer to the time window of the longest segment. In our particular test domain, as indicated in Fig. 4, the longest segment extends through much of the entire window due to our choice of speech examples that are continuously voiced sentences. Clearly, sequential grouping mechanisms would be required in order to group a sequence of voiced and unvoiced speech sounds. Fig. 5(b) shows the results of channel classification for the speech and telephone mixture after conformation and trimming by the longest segment.

We now consider the response of the second layer to the speech and telephone mixture. The second layer has the same size as the first layer, and in this case it is a network of 128 × 150 oscillators. The following parameter values are



(b)

Fig. 6. The result of separation for the speech and telephone mixture. (a) A snapshot showing the activity of the second layer shortly after the start of simulation. Active oscillators are indicated by white pixels. (b) Another snapshot, taken shortly after (a).



(b)

Fig. 7. (a) Temporal traces of every enabled oscillator in the second layer for the speech and telephone mixture. The two traces show the combined activities of two oscillator blocks corresponding to two streams. (b) Temporal traces of every other oscillator at timeframe 45 (cf. Fig. 2). The normalized x activities of the oscillators are displayed. The simulation was conducted from $t = 0$ to $t = 24$.

used: $\mu = 0.01$; $\theta = 0.95$; $W_{e1} = 4.0$; $W_{e2} = 0.1$; $W_e = 0.5$; and $W_i = -0.5$. With the initialization and lateral connections described earlier, the network quickly (in the first cycle) forms two synchronous blocks, which desynchronize from each other. Each block represents a stream extracted by our model. Fig. 6 shows two snapshots of the second layer. Each snapshot corresponds to the activity of the network at a particular time, where a white pixel indicates an active oscillator and a black pixel indicates either a silent or excitable oscillator. Fig. 6(a) is a snapshot taken when the oscillator block (stream) corresponding primarily to segregated speech is in the active phase. Fig. 6(b) shows a subsequent snapshot when the oscillator block (stream) corresponding primarily to the telephone is in the active phase. This successive "pop-out" of streams continues in a periodic fashion.

Recall that, while the speech stream is grouped together due to its intrinsic coherence (i.e., all acoustic components belonging to the speech are modulated by the same F0), the telephone stream is formed because no further analysis is performed and all oscillators start in unison. In this particular example, a further analysis using the same strategy would successfully group segments that correspond to the telephone source because the telephone contains a long segment throughout its duration [see Fig. 5(b)]. However, unlike Brown and Cooke [7] we choose not to do further grouping since intruding signals often do not possess such coherence (for example, consider the noise burst intrusion described in Section VII). Since our model lacks an effective sequential grouping mechanism, further analysis would produce many streams of no perceptual significance. Our strategy of handling the second stream is in line with the psychological process of figure-ground separation, where a stream is perceived as the foreground (figure) and the remaining stimuli are perceived as the background [36].

To illustrate the entire segregation process, Fig. 7 shows the temporal evolution of the stimulated oscillators. In Fig. 7(a), the activities of all the oscillators corresponding to one stream are combined into one trace. Since unstimulated oscillators remain excitable throughout the simulation process, they are excluded from the display. The synchrony within each stream and desynchrony between the two streams are clearly shown. Notice that the narrow active phases in the lower trace of Fig. 7(a) are induced by vertical excitation, which is not strong enough to recruit an entire segment to jump up. This narrow (also relatively lower) activity is irrelevant when interpreting segregation results, and can be easily filtered out. Notice also that perfect alignment between different oscillators of the same stream is due to the use of the singular limit method. To illustrate the oscillator activities in greater detail, Fig. 7(b) displays the activity of every other oscillator at time frame 45; this should be compared with the correlogram in Fig. 2 and the snapshot results in Fig. 6.

As illustrated in Figs. 6 and 7, stream formation arises from the emergent behavior of our two-layer oscillator network, which has so far been explained in terms of local interactions. What does the oscillator network compute at the system level? The following description attempts to provide a brief outline. Recall that all stimulated oscillators in the second layer start synchronized, and through lateral potentials some leaders emerge from the longest segment. The leaders with a small additional input [see (7a2)] are the first to jump up within a cycle of oscillations. When the leaders jump to the active phase, they recruit the rest of the segment to jump up. With the leading segment on RB, vertical connections from the leading segment exert both excitation and inhibition on other segments. If a majority of the oscillators (in terms of time frames) in a segment receive excitation from the leading segment, not only will the oscillators that receive excitation jump to the active phase, but so will the rest of the segment that receives inhibition from the leading segment. This is because of strong mutual excitation (W_{e1}) within the segment induced by the majority of the active oscillators. On the other hand, if a minority of the oscillators receive excitation from the leading segment, only the oscillators that receive direct excitation tend to jump to the active phase. This is because mutual excitation within the segment is weak (W_{e2}) and it cannot excite the rest of the oscillators. If these oscillators jump to RB, they will stay on RB for only a short period of time because, lacking strong mutual excitation within the segment, their overall excitation is weak. In Fig. 7(a), these are the oscillators with a narrow active phase. Additionally, the inhibition that a majority of the oscillators receive serves to desynchronize the segment from the leading one. When the leading segment and the others it recruits—which form the first stream—jump back, the release of inhibition allows those previously inhibited oscillators to jump up, and they in turn will recruit a whole segment if they constitute a majority within a segment. These segments form the second stream, which is the complement of the first stream. These two streams will continue to be alternately activated, a characteristic of oscillatory correlation. The oscillatory dynamics reflect the principle of "exclusive allocation" in ASA, meaning that each segment belongs to only one stream [6].

VI. Resynthesis

The last stage is a resynthesis path, which allows an acoustic waveform to be reconstructed from the time-frequency regions corresponding to a stream. Resynthesis provides a convenient mechanism for assessing the performance of a sound separation system, and has previously been used in a number of computational ASA studies (for example, see [57]; [12]; [7]; [16]). We emphasize that, although we treat resynthesis as a separate processing stage, it is not part of our ASA model and is used for the sole purpose of performance evaluation.

Here, we use a resynthesis scheme that is similar in principle to that described by Weintraub [57]. Recall that the second layer of our oscillator network embodies the result of auditory grouping; blocks of oscillators representing auditory streams "pop-out" in a periodic fashion. For each block, resynthesis proceeds by reconstructing a waveform from only those time-frequency regions in which the corresponding oscillators are in their active phase. Hence, the plots of second-layer oscillator activity in Fig. 6 may be regarded as time-frequency "masks," in which white pixels contribute to the resynthesis and black pixels do not (see also Brown and Cooke [7]).

Given a block of active oscillators, the resynthesized waveform is constructed from the output of the gammatone filterbank as follows. In order to remove any across-channel phase differences, the output of each filter is time-reversed, passed through the filter a second time, and time-reversed again. Subsequently, the phase-corrected filter output from each channel is divided into 20-ms sections, which overlap by 10 ms and are windowed with a raised cosine. Hence, each section of filter output is associated with a time-frequency location in the oscillator network. A binary weighting is then applied to each section, which is unity if the corresponding oscillator is in its active phase, and zero if the oscillator is silent or excitable. Finally, the weighted filter outputs are summed across all channels of the filterbank to yield a resynthesized waveform.

For each of the 100 mixtures of speech and noise described in Section VII, the speech stream has been resynthesized after segregation by the system. Generally, the resynthesized speech is highly intelligible and is reasonably natural. The highest quality resynthesis is obtained when the intrusion is narrowband (1-kHz tone, siren) or intermittent (noise bursts). The resynthesis is of lower quality when the intrusion is continuous and wideband (random noise, "cocktail party" noise).

VII. Evaluation

A resynthesis pathway allows sound separation performance to be assessed by formal or informal intelligibility testing (for example, see [48] and [12]). Alternatively, the segregated output can be assessed by an automatic speech recognizer [57]. However, these approaches to evaluation suffer some disadvantages; intelligibility tests are time-consuming, and the interpretation of results from an automatic recognizer is complicated by the fact that auditory models generally do not provide a suitable input representation for conventional speech recognition systems [4].

Here, we use resynthesis to quantify segregation performance using a well-established and easily interpreted metric; SNR. Given a signal waveform s and noise waveform n, the

SNR in dBs is given by

$$\mathrm{SNR} = 10\log_{10}\left(\sum_j s^2(j) \Big/ \sum_j n^2(j)\right). \tag{14}$$

The model has been evaluated using a corpus of 100 mixtures of speech and noise previously employed by Cooke [12] and Brown and Cooke [7]. The mixtures are obtained by adding the waveforms of each of ten intrusions to each of ten voiced utterances (five sentences spoken by two male speakers). The intrusions consist of synthetic sounds (1 kHz tone, noise bursts, random noise, siren), environmental sounds (trill telephone, "cocktail party" noise, rock music) and speech (one male utterance and two female utterances).

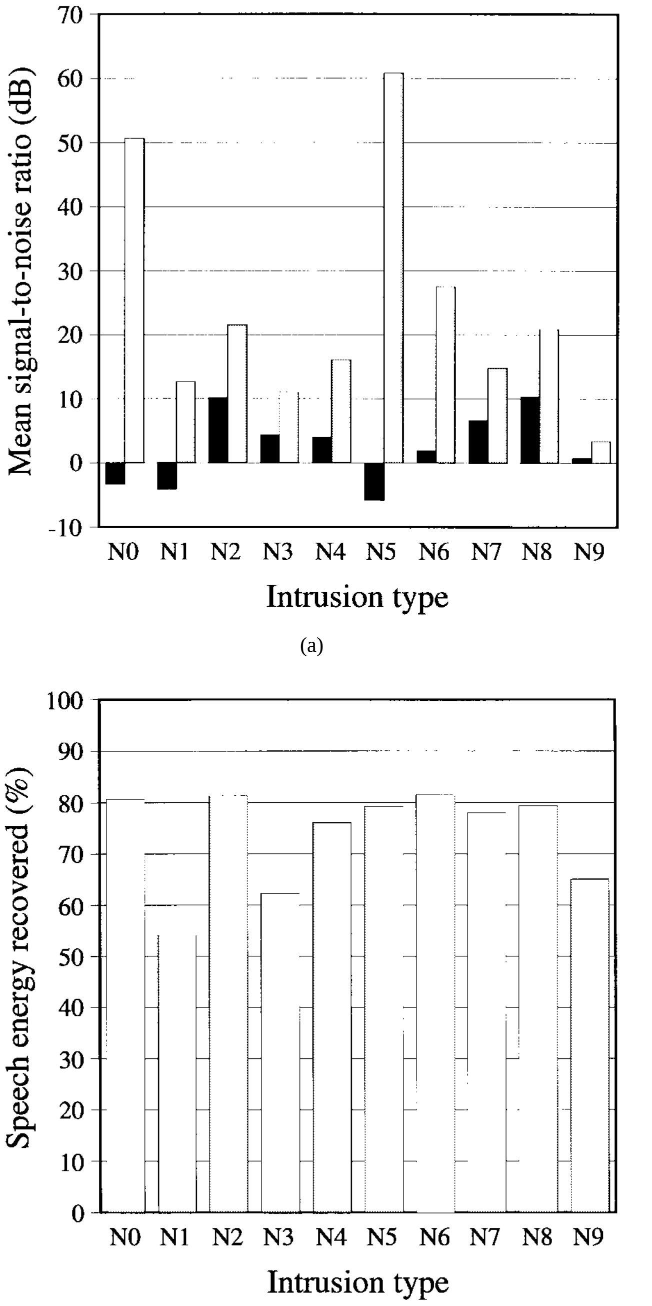
Since separate speech and noise waveforms are available, (14) can be computed before segregation by the model. Additionally, our resynthesis process allows the SNR to be computed *after* segregation by the model, so that performance can be quantified as a change in SNR. This is possible because the resynthesis pathway is linear [i.e., it consists of two passes of gammatone filtering, and the gammatone filter (1) is linear]. Hence, the resynthesis process R satisfies the property of superposition, and we can write

$$R(s+n) = R(s) + R(n). \tag{15}$$

Given a block B of active oscillators which correspond to a stream, (15) implies that the proportion of signal in the stream can be obtained by resynthesizing the signal waveform from B, and the proportion of noise in the stream can be obtained by resynthesizing the noise waveform from B. Hence, separate signal and noise waveforms can be obtained after segregation by the model, and the postsegregation SNR can be computed using (14).

Fig. 8(a) shows the SNR before and after segregation by the model. The SNR was similar for each utterance in the same noise condition, and hence the results are expressed as a mean SNR (i.e., an average over the ten utterances in each noise condition). Relative to the SNR of the original mixture, an improvement in SNR is obtained after segregation by the model for each type of noise intrusion. Dramatic improvements in SNR are obtained when the interfering noise is narrowband (1 kHz tone and siren); these intrusions tend to be represented as a single segment because of their compact spectral structure, and hence they can be segregated very effectively from the speech source. We emphasize that the same set of parameter values is used for the entire corpus of 100 mixtures. Our results are robust to considerable parameter variations.

Of course, SNR does not indicate the *intelligibility* of the resynthesized speech signal. For example, the model could retrieve a small proportion of the speech energy and totally reject the noise; this would give a very high SNR, but the resynthesized speech would be unintelligible. Accordingly, we complement the SNR metric with a measure of the percentage of speech energy recovered from each acoustic mixture [Fig. 8(b)]. The recovered speech signal is produced by masking the original speech signal, i.e., before it is mixed, with the segregated speech stream [see Fig. 6(a) for example]. Again, results are presented as an average over the ten utterances in each noise condition. Taken together, Fig. 8(a) and 8(b) provide a good indication of the intelligibility of the resynthesized speech. Intelligibility is high when the intrusion is narrowband (e.g., 1-kHz tone), as indicated by the high SNR after segregation and the high percentage of speech energy recovered. Similarly, the intelligibility of the resynthesized speech is relatively poor when the intrusion is wideband (e.g., random noise); in such cases, the SNR and percentage of speech energy recovered are both low.



(a)

(b)

Fig. 8. (a) Signal to noise ratio before (black bar) and after (gray bar) segregation by the model, for voiced speech mixed with ten different intrusions (N0 = 1 kHz tone; N1 = random noise; N2 = noise bursts; N3 = "cocktail party" noise; N4 = rock music; N5 = siren; N6 = trill telephone; N7 = female speech; N8 = male speech; N9 = female speech). (b) Percentage of speech energy recovered from each mixture after segregation by the model.

VIII. Comparison with Other Models

A multistage sound separation system has previously been described by Brown and Cooke [7]. Their system consists of four stages, the first of which models the auditory periphery. In the second stage, a collection of "auditory maps" extract information about periodicity, frequency transitions, onsets, and offsets (these correspond to the mid-level auditory representations described here). Information from the auditory maps is used to construct a symbolic representation of the auditory scene in the third stage of their model. More specifically, the auditory scene is represented as a collection of elements, each of which traces the movement of a spectral peak through time and frequency. In the final stage of the Brown and Cooke model, a search strategy is employed which groups elements according to their fundamental frequency, onset time, and offset time.

Clearly, the initial stages of our model bear a close resemblance to the Brown and Cooke scheme. However, there are substantial differences in our two approaches. The way in which segment formation and grouping of segments are performed is significantly different at the algorithmic level. For example, their method relies on comparison of "local" pitch contours of individual elements to compute pitch-based grouping, whereas ours is based on "global" pitch estimates. Conceptually, the model described here is more strongly motivated by neurobiological findings. It embodies a more principled computational framework—oscillatory correlation—in which segmentation and grouping arise from oscillatory dynamics.

We note that our simulation results (Fig. 8) are comparable with those of Brown and Cooke, who evaluated their system using the same set of 100 acoustic mixtures described in Section VII. Both of our systems show the same pattern of SNR improvement across noise conditions. Hence, our neural oscillator model is able to match the performance of their symbol-based system, but is computationally simpler and better suited to real-time implementation.

Our system has similar advantages over other symbol-based approaches, such as the blackboard-based systems of Klassner *et al.* [27], Ellis [16], and Godsmark and Brown [22], and the multiagent architecture of Nakatani *et al.* [37]. All of these systems require complex control strategies to coordinate the grouping of acoustic components. However, our model could benefit from the wider representational vocabulary used in these models. Currently, our mid-level auditory representations do not provide good descriptions of noise clouds and transient clicks; Ellis [16], [17] describes representations of such acoustic components, together with a method of resynthesizing from them. Consequently, his resynthesis pathway is of a higher quality than that described here. In our model, segments are formed only from periodic components in the acoustic input; noisy and impulsive regions are allocated to a "background" stream (see Section V-A), and hence do not contribute to resynthesis.

Our approach also differs substantially from other neural-network models of auditory segregation. Beauvois and Meddis [3] and McCabe and Denham [31] have both described neural architectures which model the perception of alternating pure-tone sequences. However, neither offer a general account of ASA; although they are able to explain the grouping of a sequence of tones, they lack a mechanism for grouping simultaneous components (for example, harmonics of the same F0). Furthermore, these models represent auditory grouping through the *spatial* separation of neural activity; for example, in the model of McCabe and Denham, each stream is represented by a separate neural array (see also Grossberg [23] for a similar approach). In contrast, our model represents auditory grouping by a *temporal* coding, in the form of oscillatory correlation.

IX. Discussion and Conclusion

A significant feature of the multistage model proposed here is that every stage has a neurobiological foundation. The peripheral auditory model is based upon the gammatone filter, which is derived from physiological measurements of auditory nerve impulse responses. Similarly, our mid-level auditory representations are consistent with the neurophysiology of the higher auditory system. Overall, the model is based on a framework—oscillatory correlation—which is supported by recent neurobiological findings.

As illustrated in Fig. 7, segregation in the oscillatory correlation representation is performed in time; after segregation, each stream pops out at a distinct time from the network and different streams alternate in time. While auditory segregation in a spatial representation (e.g., one layer for each stream as in [31] and [23]) requires an explicit assumption of how many streams are in an auditory scene, our representation needs no prior assumption about the number of streams because oscillatory correlation is capable of temporal multiplexing in terms of stream segregation. As a result, our representation is more flexible and parsimonious.

Currently, our model lacks a mechanism for sequential grouping (i.e., it is unable to group acoustic events that are separated in time, such as a sequence of voiced and unvoiced speech sounds). There are a number of ways in which sequential grouping could be implemented within the neural oscillator framework. For instance, a sequence of acoustic events could be allocated to the same stream if they had an F0 in the same average pitch range. This extension to our model could be readily implemented, since F0 information is available in the pooled correlogram. Additionally, sounds could be grouped sequentially by virtue of their spatial location or timbre (computational techniques for extracting timbral information have been described by Brown and Cooke [8] and Godsmark and Brown [22]).

Our multistage model is entirely bottom-up (see Fig. 1), and does not include any top-down processing. Such bottom-up processing corresponds to *primitive* segregation [6]. It is clear that ASA is also influenced by attention and prior knowledge, so-called *schema-based* organization [6]. Little computational study has been directed to schema-based grouping and segregation. We expect that overall computational performance of speech-related segregation tasks, such as the one addressed here, will improve with an effective mechanism for schema-based organization.

Our model can potentially be implemented as a real-time system. The first two stages—peripheral processing and mid-level processing (see Fig. 1)—can be readily turned to real-time implementation because the processing involves only local time windows, and the computations for each frequency channel can be performed in parallel. Given its rate of synchronization and desynchronization, our two-layer oscillator network may be extended to a real-time system. Three issues need to be addressed in real-time implementation. The first one is how external stimuli map to the network in real time. One possible realization is to use systematic time delays to maintain a recent history of the auditory input [55], say 150 time frames as used in the speech and telephone mixture. Consistent with the shifting synchronization theory that Wang proposed to explain primitive stream segregation [55], such an architecture implies that oscillator populations corresponding to different streams shift on the oscillator network as the stimuli unfold in time. The second issue is how the past ASA result influences current processing. Again, consider the speech and telephone example. During the duration of the entire utterance, there are time intervals within which the two sound sources are well separated, and those within which the two cannot be separated properly (see Fig. 5). In a real-time system, the segregation decision at a particular time instant should be based not only on the auditory information at that time, but also the segregation decisions in the recent past. How this is done in a way that enhances overall segregation performance is at the heart of the issue. The third issue is how connections in the oscillator network are set up in real time. Setting up local connections in the first layer is straightforward. For the second layer, both mutual excitatory connections within each segment and vertical connections between different segments must be set up quickly, based on the input both from the first oscillator layer and from the pooled correlogram. This calls for a mechanism of fast changing synapses [51].

Our oscillator network computes ASA in a parallel and distributed fashion, where each oscillator behaves autonomously and in parallel with all the other oscillators in the network. With the above issues for real-time implementation resolved, there is a real possibility that the oscillator network, with its continuous-time dynamics, can be implemented on an analog VLSI chip. This feature is particularly attractive because considerable computation is needed to analyze real auditory scenes, and analog VLSI technology is known for its high speed and compact size, both desired for real-time implementation.

To conclude, we have studied ASA from a neurocomputational perspective and have proposed a multistage model for segregating speech from interfering sounds, where grouping and segregation are performed by a two-layer oscillator network. The lateral connections within the network embody proximity in frequency and time, and harmonicity. The network forms auditory segments first, which correspond to connected acoustic components that are atomic and perceptually relevant elements for further analysis. Streams then emerge from the network that groups harmonically related segments. The model is founded on auditory neurobiology, and has been systematically evaluated using a corpus of voiced speech mixed with a variety of interfering sounds.

References

[1] P. F. Assmann and Q. Summerfield, "Modeling the perception of concurrent vowels: Vowels with different fundamental frequencies," *J. Acoust. Soc. Am.*, vol. 88, pp. 680–697, 1990.

[2] D. S. Barth and K. D. MacDonald, "Thalamic modulation of high-frequency oscillating potentials in auditory cortex," *Nature*, vol. 383, pp. 78–81, 1996.

[3] M. W. Beauvois and R. Meddis, "Computer simulation of auditory stream segregation in alternating-tone sequences," *J. Acoust. Soc. Am.*, vol. 99, pp. 2270–2280, 1996.

[4] S. W. Beet, "Automatic speech recognition using a reduced auditory representation and position-tolerant discrimination," *Computer Speech and Language*, vol. 4, pp. 17–33, 1990.

[5] J. P. L. Brokx and S. G. Nooteboom, "Intonation and the perceptual separation of simultaneous voices," *J. Phonetics*, vol. 10, pp. 23–36, 1982.

[6] A. S. Bregman, *Auditory Scene Analysis*, Cambridge, MA: MIT Press, 1990.

[7] G. J. Brown and M. P. Cooke, "Computational auditory scene analysis," *Computer Speech and Language*, vol. 8, pp. 297–336, 1994.

[8] G. J. Brown and M. P. Cooke, "Perceptual grouping of musical sounds: A computational model," *J. New Music Research*, vol. 23, pp. 107–132, 1994.

[9] G. J. Brown and M. P. Cooke, "Temporal synchronization in a neural oscillator model of primitive auditory stream segregation," in *Computational Auditory Scene Analysis*, D. F. Rosenthal and H. Okuno (Eds.), Mahwah, NJ: Lawrence Erlbaum, pp. 87–103, 1998.

[10] G. J. Brown and D. L. Wang, "Modeling the perceptual segregation of double vowels with a network of neural oscillators," *Neural Networks*, vol. 10, pp. 1547–1558, 1997.

[11] L. H. Carney, "Sensitivities of cells in the anteroventral cochlear nucleus of cat to spatiotemporal discharge patterns across primary afferents," *J. Neurophysiol.*, vol. 64, pp. 437–456, 1990.

[12] M. P. Cooke, *Modeling Auditory Processing and Organization*, Cambridge, U.K.: Cambridge University Press, 1993.

[13] M. P. Cooke and G. J. Brown, "Separating simultaneous sound sources: Issues, challenges and models," in *Speech Recognition and Speech Synthesis*, E. Keller (Ed.), London: John Wiley and Sons, 1994.

[14] E. de Boer and H. D. de Jongh, "On cochlear encoding: Potentialities and limitations of the reverse correlation technique," *J. Acoust. Soc. Am.*, vol. 63, pp. 115–135, 1978.

[15] R. C. deCharms and M. M. Merzenich, "Primary cortical representation of sounds by the coordination of action-potential timing," *Nature*, vol. 381, pp. 610–613, 1996.

[16] D. P. W. Ellis, "Prediction-driven computational auditory scene analysis," *Ph.D. Dissertation*, MIT Department of Electrical Engineering and Computer Science, 1996.

[17] D. P. W. Ellis and D. Rosenthal, "Mid-level representations for computational auditory scene analysis: The weft element," in *Computational Auditory Scene Analysis*, D. F. Rosenthal and H. Okuno (Eds.), Mahwah, NJ: Lawrence Erlbaum, pp. 257–272, 1998.

[18] W. J. Freeman, "Nonlinear dynamics in olfactory information processing," in *Olfaction*, J. L. Davis and H. Eichenbaum (Eds.), Cambridge, MA: MIT Press, pp. 225–249, 1991.

[19] R. D. Frisina, R. L. Smith, and S. C. Chamberlain, "Encoding of amplitude-modulation in the gerbil cochlear nucleus. 1. A hierarchy of enhancement," *Hearing Research*, vol. 44, pp. 99–122, 1990.

[20] R. Galambos, S. Makeig, and P. J. Talmachoff, "A 40-Hz auditory potential recorded from the human scalp," in *Proc. Natl. Acad. Sci.*, USA, 1981, vol. 78, pp. 2643–2647.

[21] B. R. Glasberg and B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing Research*, vol. 47, pp. 103–138, 1990.

[22] D. J. Godsmark and G. J. Brown, "Context-sensitive selection of competing auditory organizations: A blackboard model," in *Computational Auditory Scene Analysis*, D. F. Rosenthal and H. Okuno (Eds.), Mahwah, NJ: Lawrence Erlbaum, pp. 139–155, 1998.

[23] S. Grossberg, "Pitch-based streaming in auditory perception," in *Creative Networks*, N. Griffith and P. Todd (Eds.), Cambridge, MA: MIT Press, 1998.

[24] ISO, *Normal Equal-Loudness Level Contours for Pure Tones Under Free-Field Listening Conditions (ISO 226)*, International Standards Organization.

[25] M. Joliot, U. Ribary, and R. Llinas, "Human oscillatory brain activity near to 40 Hz coexists with cognitive temporal binding," in *Proc. Natl. Acad. Sci.*, USA, 1994, vol. 91, pp. 11748–11751.

[26] K. Kashino, K. Nakadai, T. Kinoshita and H. Tanaka, "Application of the Bayesian probability network to music scene analysis," in *Computational Auditory Scene Analysis*, D. F. Rosenthal and H. Okuno (Eds.), Mahwah, NJ: Lawrence Erlbaum, pp. 115–137, 1998.

[27] F. Klassner, V. Lesser, and S. H. Nawab, "The IPUS blackboard architecture as a framework for computational auditory scene analysis," in *Computational Auditory Scene Analysis*, D. F. Rosenthal and H. Okuno (Eds.), Mahwah, NJ: Lawrence Erlbaum, pp. 105–114, 1998.

[28] P. S. Linsay and D. L. Wang, "Fast numerical integration of relaxation oscillator networks based on singular limit solutions," *IEEE Trans. Neural Networks*, vol. 9, pp. 523–532, 1998.

[29] R. Llinás and U. Ribary, "Coherent 40-Hz oscillation characterizes dream state in humans," in *Proc. Natl. Acad. Sci.*, USA, 1993, vol. 90, pp. 2078–2082.

[30] P. E. Maldonado and G. L. Gerstein, "Neuronal assembly dynamics in the rat auditory cortex during reorganization induced by intracortical microstimulation," *Exp. Brain Res.*, vol. 112, pp. 431–441, 1996.

[31] S. L. McCabe and M. J. Denham, "A model of auditory streaming," *J. Acoust. Soc. Am.*, vol. 101, pp. 1611–1621, 1997.

[32] R. Meddis, "Simulation of auditory-neural transduction: Further studies," *J. Acoust. Soc. Am.*, vol. 83, pp. 1056–1063, 1988.

[33] R. Meddis and L. O'Mard, "A unitary model of pitch perception," *J. Acoust. Soc. Am.*, vol. 102, pp. 1811–1820, 1997.

[34] R. Meddis and M. Hewitt, "Modeling the identification of concurrent vowels with different fundamental frequencies," *J. Acoust. Soc. Am.*, vol. 91, pp. 233–245, 1992.

[35] D. K. Mellinger, "Event formation and separation in musical sound," Ph.D. dissertation, Department of Computer Science, Stanford University, 1992.

[36] B. C. J. Moore, *An Introduction to the Psychology of Hearing*, 4th Edition, San Diego, CA: Academic, 1997.

[37] T. Nakatani, H. Okuno, M. Goto, and T. Ito, "Multiagent based binaural sound stream segregation," in *Computational Auditory Scene Analysis*, D. F. Rosenthal and H. Okuno (Eds.), Mahwah, NJ: Lawrence Erlbaum, pp. 195–214, 1998.

[38] D. L. Oliver and D. K. Morest, "The central nucleus of the inferior colliculus in the cat," *J. Comp. Neurol.*, vol. 222, pp. 237–264, 1984.

[39] A. R. Palmer, "Physiology of the cochlear nerve and cochlear nucleus," in *Hearing*, M. P. Haggard and E. F. Evans, Eds., Edinburgh: Churchill Livingstone, 1987.

[40] T. W. Parsons, "Separation of speech from interfering speech by means of harmonic selection," *J. Acoust. Soc. Am.*, vol. 60, no. 4, pp. 911–918, 1976.

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

[42] U. Ribary *et al.*, "Magnetic field tomography of coherent thalamocortical 40-Hz oscillations in humans," in *Proc. Natl. Acad. Sci.*, USA, 1991, vol. 88, pp. 11037–11041.

[43] M. T. M. Scheffers, "Sifting vowels: Auditory pitch analysis and sound segregation," Ph.D. dissertation, University of Gröningen, 1983.

[44] C. E. Schreiner and G. Langner, "Periodicity coding in the inferior colliculus of the cat. II. Topographical organization," *J. Neurophysiology*, vol. 60, pp. 1823–1840, 1988.

[45] S. A. Shamma, "Speech processing in the auditory system I: The representation of speech sounds in the responses of the auditory nerve," *J. Acoust. Soc. Am.*, vol. 78, pp. 1613–1621, 1985.

[46] W. Singer and C. M. Gray, "Visual feature integration and the temporal correlation hypothesis," *Ann. Rev. Neurosci.*, vol. 18, pp. 555–586, 1995.

[47] M. Slaney and R. F. Lyon, "A perceptual pitch detector," in *Proc. ICASSP*, 1990, pp. 357–360.

[48] R. J. Stubbs and Q. Summerfield, "Algorithms for separating the speech of interfering talkers: Evaluations with voiced sentences, and normal-hearing and hearing-impaired listeners," *J. Acoust. Soc. Am.*, vol. 87, pp. 359–372, 1990.

[49] D. Terman and D. L. Wang, "Global competition and local cooperation in a network of neural oscillators," *Physica D*, vol. 81, pp. 148–176, 1995.

[50] B. van der Pol, "On relaxation oscillations," *Philosophical Mag.*, vol. 2, no. 11, pp. 978–992, 1926.

[51] C. von der Malsburg, "The correlation theory of brain function," *Internal Report 81-2*, Max-Planck-Institute for Biophysical Chemistry, 1981.

[52] C. von der Malsburg and W. Schneider, "A neural cocktail-party processor," *Biol. Cybern.*, vol. 54, pp. 29–40, 1986.

[53] D. L. Wang, "Auditory stream segregation based on oscillatory correlation," in *Proc. IEEE Workshop on Neural Networks for Signal Processing*, IEEE, 1994, pp. 624–632.

[54] D. L. Wang, "Object selection based on oscillatory correlation," *Tech. Rep. 96-67*, OSU Department of Computer and Information Science, 1996.

[55] D. L. Wang, "Primitive auditory segregation based on oscillatory correlation," *Cognit. Sci.*, vol. 20, pp. 409–456, 1996.

[56] D. L. Wang and D. Terman, "Image segmentation based on oscillatory correlation," *Neural Comp.*, vol. 9, pp. 805–836, 1997 (for errata see *Neural Comp.*, vol. 9, pp. 1623–1626, 1997).

[57] M. Weintraub, "A computational model for separating two simultaneous talkers," in *Proc. IEEE ICASSP*, 1986, pp. 81–84.

DeLiang L. Wang (A'95), for a biography, see this issue, p. 572.

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

He is currently a lecturer in computer science at the University of Sheffield. He has studied computational models of auditory perception since 1989, and also has research interests in hearing impairment, computer-assisted learning, and music technology.