COGNITIVE SCIENCE **20,** 409–456 (1996)

Primitive Auditory Segregation Based on Oscillatory Correlation

DELIANG WANG

The Ohio State University

Auditory scene analysis is critical for complex auditory processing. We study auditory segregation from the neural network perspective, and develop a framework for primitive auditory scene analysis. The architecture is a laterally coupled two-dimensional network of relaxation oscillators with a global inhibitor. One dimension represents time and another one represents frequency. We show that this architecture, plus systematic delay lines, can in real time group auditory features into a stream by phase synchrony and segregate different streams by desynchronization. The network demonstrates a set of psychological phenomena regarding primitive auditory scene analysis, including dependency on frequency proximity and the rate of presentation, sequential capturing, and competition among different perceptual organizations. We offer a neurocomputational theory—shifting synchronization theory—for explaining how auditory segregation might be achieved in the brain, and the psychological phenomenon of stream segregation. Possible extensions of the model are discussed.

INTRODUCTION

A listener in a realistic auditory environment is always exposed to acoustic energy from many different sources. In order to recognize and understand the auditory environment, the listener must first disentangle the acoustic wave and capture each event. This process is referred to as *auditory scene analysis* (Bregman, 1990), or *auditory segregation.* Its task is to separate the sensory features of an auditory scene into multiple coherent streams, each having a high likelihood of coming from the same source. A *stream* in the auditory domain roughly corresponds to an object in the visual domain.

The author is grateful to M. Jones and E. Covey for their stimulating discussions and constructive suggestions, and M. Jones for critically reading an earlier draft. The author also wishes to thank two referees, D. Mellinger and R. Watrous, whose extensive comments have significantly improved the presentation of the article. The preparation of this article was supported in part by the ONR Grant N00014-93-1-0335, the NSF Grants IRI-9211419 and IRI-9423312, and the NSF Equipment Grant CDA-9413962.

Correspondence and requests for reprints should be sent to DeLiang Wang, Department of Computer and Information Science, The Ohio State University, Columbus, OH 43210-1277. Email: dwang@cis.ohiostate.edu.

Auditory scene analysis is a remarkable achievement of the auditory system, playing a fundamental role in auditory perception. It has much in common with segmentation of a visual scene into a set of objects, and figure-ground segregation, the perceptual processes much studied in Gestalt psychology (Koffka, 1935; Rock & Palmer, 1990). It should be clear that auditory segregation differs from the segmentation of a single auditory flow into different successive components, which corresponds to, for example, the process of separating a continuous utterance from a single speaker into successive phonemes or words.[1] Notice that successive segmentation is different from *sequential integration,* a term used by Bregman (1990) to refer to the grouping of a set of sequential tones that are interleaved in presentation with another set of sequential tones (discussed later). Auditory segregation, which is potentially a parallel process, should be a precursor to successive separation.

Auditory segregation can be divided into primitive auditory segregation and memory-based auditory segregation (Bregman, 1990). Primitive auditory segregation is an innate process, relying on the similarities of local qualities within the input scene itself, such as frequency, timing, or amplitude. Memory-based auditory segregation is based on prior knowledge stored in memory to segregate the auditory input. In this article, we deal with primitive auditory segregation only.

The remaining part of this section reviews the psychological background of auditory scene analysis and pertinent computational studies on the subject.

Psychological Evidence

The human ability to segregate multiple auditory sources was observed long ago (Helmholtz, 1863/1954). Auditory stream segregation was first systematically studied by Miller and Heise (1950), who noted that listeners split a signal with two alternating sine wave tones into two streams. Auditory segregation could be obtained with as little as a 15% difference in frequency and could be obtained throughout the frequency range from about 150 Hz to 7000 Hz. Bregman and his collaborators have carried out a series of studies on this subject (see Bregman, 1990, for an extensive review). In one of the early studies (Bregman & Campbell, 1971), participants were asked to report the temporal order of six tones in a sequence. Three of them were in a high-frequency range, and the other three in a low-frequency range. This situation is depicted in Figure 1. The results showed that at high rates of

[1] In this article, we reserve the word *segregation* to refer to the process of separating an auditory scene into multiple streams, and *segmentation* to refer to the process of marking boundaries in a continuous auditory flow, in line with the terminology used in auditory perception (Bregman, 1990; Handel, 1989). In visual processing, segmentation is widely used to mean separation of an image into multiple objects. Adding to the confusion, the word *segmentation* has been used to mean segregation in some previous computational studies on auditory scene analysis (von der Malsburg & Schneider, 1986; Wang, in press).

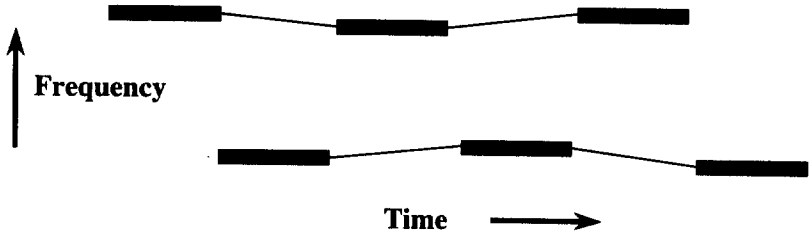


Figure 1. Six alternating pure tones as displayed in a spectrogram. Three of them are in a high-frequency range and the other three in a low-frequency range. When stream segregation occurs, the high-frequency tones form one stream and the low-frequency tones form another stream (indicated by thin lines).

presentation, participants perceived two separate sequences corresponding to high- and low-frequency tones respectively, and they were able to report only the temporal order of the tones within each sequence, but not across the two sequences. Bregman and Campbell called this phenomenon *stream segregation*. Furthermore, there is a trade-off between frequency separation and presentation rate, and the higher the presentation rate, the smaller the frequency separation is needed to generate stream segregation (Bregman & Campbell, 1971; van Noorden, 1975). The loss of order information across different streams was observed earlier by Warren, Obusek, Farmer, & Warren (1969).

This basic phenomenon of stream segregation was repeatedly verified in different contexts (see among others, Bregman, 1990, 1993; Bregman, Abramson, Doehring, & Darwing, 1985; Hartmann, 1988; Jones & Yee, 1993; Rasch, 1978; van Noorden, 1975). Although stream segregation is primarly about segregation of tones presented sequentially, the basic phenomenon also occurs in auditory stimuli with temporal overlap. In general, if the auditory patterns are displayed on a spectrogram, the results of stream segregation are analogous to Gestalt laws of grouping (the reverse process of segregation) that have been expressed in the visual domain (see also Handel, 1989). The following is a list of several important determinants of auditory scene analysis:

- Frequency (pitch) proximity. The closer the frequencies of two tones, the easier they are to be grouped into the same stream. Frequency proximity is considered the prime factor in auditory scene analysis (see Bregman, 1990).
- Presentation rate. The faster the presentation rate, the easier it is for each of interleaving tone sequences to be grouped into the same stream, or the harder are successive tones to be grouped into the same stream, where tones in each sequence have similar frequencies. This situation is illustrated in Figure 1. Because faster presentation of a tone corresponds

to shorter separation of the tones in each sequence, this property is consistent with the proximity principle as applied to the time domain of a spectrogram.

- Continuous/smooth transition. Tones tend to be grouped into the same stream if they form continuous transition (e.g., continuous frequency transition) or smooth but discontinuous transition (e.g., an ascending set of gliding tones), from one to another.
- Onset or offset synchrony. A person tends to perceive two or more tones as in the same stream if the tones have the same onset or offset time.
- Common modulation. Simultaneous tones that undergo the same kind of modulation at the same time tend to be bound together into a common stream. This principle applies to both frequency modulation and amplitude modulation.
- Prior knowledge. Prior knowledge can strongly influence auditory scene analysis, in the sense that components belonging to the same familiar pattern tend to be bound together.

Jones and her colleagues argued that a rhythmic structure in a sequence of tones can promote their grouping into the same stream (Jones, Jagacinski, Yee, Floyd, & Klapp, 1995; Jones, Kidd, & Wetzel, 1981). For example, they observed that if a sequence of tones forms a simple rhythm, for example, the duration between the onset times of the tones in the sequence is a constant, the sequence tends to be integrated into the same stream.

In speech perception, auditory segregation seems to contribute to a listener's ability to separate utterances from different speakers into different streams. One question arises: What prevents the undesirable effects of segregating the speech sound (like syllables) of the same speaker into competing streams, given that the rate of speech production is very high and *formants* (peaks in the spectrum of a speech signal) of the same speaker are spread widely? In general, the production of a syllable involves a set of transitions from several frequency partials associated with the consonant of the syllable to the same number of frequency partials associated with the vowel. One acoustic factor leading to the coherence of the speech sound of the same speaker is the relative degree of onset and offset synchrony of these frequency transitions (called *formant transitions*). A second factor leading to coherence is the continuous nature of some formant transitions. These continuous transitions are caused by the nature of articulation and coarticulation, the latter referring to the interactions between adjacent sounds due to overlapping articulatory movements of the vocal tract. These two factors are consistent with the general properties of auditory scene analysis (see Bregman, 1990; Handel, 1989).

Music perception is also subject to auditory segregation. As shown by Dowling and coworkers (Dowling, 1973; Dowling, Lund, & Herrbold,

1987), if the notes of two melodies whose pitch ranges do not overlap are interleaved in time so that adjacent tones come from the different melodies, the resulting sequence of tones is perceptually divided into two streams that correspond to the two melodies. If two melodies overlap in the pitch domain, they are no longer perceived as the same melodies by the listeners who are familiar with each of the melodies. In fact, composers of music have made purposeful use of pitch difference to permit the perception of interleaved melodies (Dowling, 1973). On the other hand, if listeners know in advance what melody to listen for, they can attend to it even if it is intermixed with other notes by making use of attentional expectancy, a "cocktail party effect" (Cherry, 1953) in music perception. These observations are also true for simultaneously presented melodies (see Chapter 5 in Bregman, 1990).

Computational Studies

The technology for auditory pattern recognition (sometimes called *temporal pattern recognition,* see Covey, Hawkins, McMullen, & Port, in press; Wang, 1995b), particularly speech recognition, has been rapidly advanced in recent years. It has been demonstrated that neural networks can make a significant contribution to this technology (Bourlard & Morgan, 1994; Wang, 1995b). However, segregation of interleaving or simultaneous auditory signals remains a tremendous challenge, one that has hardly been addressed at all. Almost all speech recognition systems assume presegregation, and many can perform well only if the input is from a single stream (Bourlard & Morgan, 1994; Rabiner, 1989). Obviously, auditory segregation is a necessary ability for any speech recognizer to work in a realistic environment. Thus, a successful method of auditory segregation would be a breakthrough in making speech recognition technology reach the real world.

Parsons (1976) developed a computer program to separate two speakers on the basis of different fundamental frequencies. It uses Fourier analysis as the front end to extract frequency partials. Based on the extracted frequency partials, the algorithm computes the fundamental frequency of the first speaker that can best account for these frequency partials. This fundamental frequency is later used to cancel those partials that fall in the harmonics of the fundamental. The remaining partials are then used to compute another fundamental frequency corresponding to the second speaker. The system is programmed to separate only two voices, and it cannot detect how many voices there are to be separated. Using a similar idea, Weintraub (1986) proposed another model for separating two simultaneous talkers. The speech to be separated was taken from a male speaker and a female speaker. The input signal was first processed by a cochlear model that extracts time and frequency information. The model uses a dynamic programming algorithm to compute the periods of two fundamentals presumably corresponding to the two talkers. Then a Markov model is used to identify whether the

voice of each talker is silent, periodic, or nonperiodic. The fundamental frequencies and the voice characteristics of each talker are used to compute a spectral estimate for each sound source. Because many other factors contributing to speech separation were not considered, the success of both models is quite limited even just for two input sources. It is not clear how the models could be extended to handle sound separation beyond two talkers speaking voiced sounds.

Beauvois and Meddis (1991) proposed a computational model to simulate stream segregation. The model uses a bank of bandpass filters to extract frequency partials of an auditory input, and assumes winner-take-all competition between different filter channels. The winning channel does not decrease its activity while other channels do. Their model is designed to simulate stream segregation of successive high-frequency and low-frequency tones. It is assumed that streaming occurs if the overall system consistently shows a higher response to one of the two alternating tones. The criterion of judging whether streaming occurs is somewhat arbitrary, and cannot explain the basic fact that an individual hears two streams when streaming occurs (Bregman & Campbell, 1971). Also, the model cannot address auditory segregation of temporally overlapping streams.

More recently, Brown and Cooke (1994) presented a multistage computational model for auditory scene analysis. Like Beauvois and Meddis (1991), their model starts with a bank of bandpass filters. After this stage, the model reveals auditory features, such as harmonicity, frequency transition, onset, and offset in the representation of auditory maps. Then a symbolic representation is extracted from the auditory maps to describe an auditory scene. Finally, a search algorithm is employed that groups the symbolic elements into streams. The grouping process is based on some known psychophysical principles. Another computational study by Mellinger (1992) also uses multistage processing. Similarly, Mellinger's model uses an algorithmic approach to derive the results of segregation. His model mainly uses the grouping cues of onset synchrony and common frequency modulation. Whereas the model of Brown and Cooke aims at general auditory signals, the model of Mellinger is oriented toward musical signals. Both models do not need to assume the number of streams to be segregated in advance, an advantage over those models already described. The performance of both models, however, is limited by the grouping cues integrated, and it remains to be seen how other cues can be effectively incorporated in the symbolic framework employed.

Sound localization seems to influence auditory segregation, for different sources have high probability of originating from different spatial locations. Models of sound localization are based primarily on analyzing interaural time difference, interaural intensity difference, and spectral cues. The

model of Lazzaro and Mead (1989) computes interaural time difference using layers of silicon chips, which can produce azimuth information in real time. It has been found that in mammals, due to the head and the pinna, the variation in stimulus level (intensity) with respect to frequency shows a unique relation with sound location. The models by Neti and Young (1992) and by Zakarauskas and Cynader (1993) use such spectral cues to compute the spatial location of sound. All these models concern localizing only a single sound, although the auditory system can identify multiple locations simultaneously (Blauert, 1983). It is not clear how the algorithms may be extended to deal with multiple simultaneous sounds, the very problem facing auditory segregation. Furthermore, it is not even clear whether sound localization aids auditory segregation or auditory segregation aids sound localization (Bregman, 1990), or both.

Perhaps the only neural network model that addresses the problem of auditory segregation was one proposed by von der Malsburg and Schneider (1986). They described the idea of using neural oscillations for expressing segregation, a form of the temporal correlation theory proposed earlier by von der Malsburg (1981). Because an oscillatory pattern has an extra degree of freedom, its *phase*, it can be used to elegantly represent synchronization and desynchronization among a group of oscillators. In this representation, a set of auditory features forms a stream if the corresponding oscillators oscillate in phase with no phase lag (synchronization). Oscillators representing different streams oscillate out of phase (desynchronization). Similar ideas were suggested in general contexts earlier by Milner (1974) and at about the same time by Abeles (1982). Using this idea, von der Malsburg and Schneider constructed a network of oscillators, each representing a specific auditory feature. Each oscillator connects to all the others in the network, and there is also a global inhibitory oscillator introduced to desynchronize different streams. With a mechanism of rapid modulation of connection strengths, they simulated segregation based on onset synchrony; that is, oscillators simultaneously triggered (by a stream) synchronize with each other, and these oscillators desynchronize with those representing another stream presented at a different time.

Although the idea of using oscillators for segregation has been around for quite some time, it has not led to successful solutions to the problem of auditory segregation. For example, the model of von der Malsburg and Schneider cannot reproduce the basic phenomenon of stream segregation shown in Figure 1. One major reason for this lack of success is that an auditory pattern in their model is represented as a set of features that have no geometrical (spectral) relationships. Stream segregation in contrast has a clear dependency on the distances among tones in the time and frequency domain. In other words, auditory segregation that is best explained by

Gestalt laws of grouping requires a representational framework that clearly exhibits geometrical relationships among different tones, such as proximity and continuity.

In this article, we study auditory segregation from a neurocomputational perspective, and present a neural network framework for primitive auditory segregation. The model is based on the idea of using synchronization of neural oscillations to represent one stream and desynchronization to represent different streams. The formation of synchrony is produced by lateral excitatory connections between relaxation oscillators, and the formation of desynchrony is produced by a global inhibitory mechanism. This mechanism of segregation is called *oscillatory correlation* (Terman & Wang, 1995). Both the building block—the single oscillator model—and the mechanism of reaching synchrony and desynchrony differ fundamentally from those used by von der Malsburg and Schneider (more comparisons are given in the Discussion section). Simulations show that the model is capable of replicating the basic phenomenon of stream segregation, and explaining a set of psychological observations concerning primitive auditory scene analysis. A neurocomputational theory, namely *shifting synchronization* theory, is provided for explaining the psychological effects of stream segregation, and the theory is argued to be consistent with auditory neurophysiology and neuroanatomy. The model proposed here promises to explain a varity of experimental data and to provide an effective computational approach to auditory segregation.

The rest of the article is organized as follows. The next section describes the computational elements of our auditory segregation network. The third section provides the simulation results of the network in performing stream segregation and a number of other auditory segregation tasks. On the basis of the simulation results, the following section presents the shifting synchronization theory. Some further discussions on the theory are provided following that. Finally, the last section concludes the article.

NEURAL ARCHITECTURE

We introduce a network of neural oscillators to model primitive auditory segregation, called *the segregation network*. As described previously, the idea is to achieve segregation by oscillatory correlation. More specifically, a set of auditory tones form a stream if their corresponding oscillators synchronize with each other. Different streams correspond to different groups whose oscillation desynchronize from each other. Synchrony and desynchrony are achieved by local excitation and global inhibition, respectively. In this representation, it is very easy to read out the results of segregation: Simple thresholding will reflect grouping and segregation of the involved oscillators. This is because the synchronized oscillators reach high and low activity simultaneously.

The building block of the segregation network, a single oscillator i, is defined in the simplest form as a feedback loop between an excitatory unit x_i and an inhibitory unit y_i (cf. Terman & Wang, 1995; Wang & Terman, 1995):

$$\frac{dx_i}{dt} = f(x_i, y_i) + I_i + S_i + \rho \tag{1a}$$

$$\frac{dy_i}{dt} = \epsilon\, g(x_i, y_i) \tag{1b}$$

where, in the following implementation, $f(x_i, y_i) = 3x_i - x_i^3 + 2 - yi$, and $g(x_i, y_i) = \gamma\,(1 + tanh(x_i/\beta)) - y_i$. I_i represents external stimulation to the oscillator, and S_i represents overall coupling from other oscillators in the network. The symbol ρ denotes the amplitude of a Gaussian noise term. The purpose of introducing the noise term is twofold. First, it can test the robustness of the system. Second and perhaps more importantly, it plays a role in helping desynchronize different input patterns that happen to start with very similar initial conditions (so-called symmetry breaking). The parameter ϵ is chosen to be a small positive number. In this case, Equation 1 without any coupling or noise, corresponds to a standard relaxation oscillator (Verhulst, 1990). The x-nullcline (namely $dx/dt = 0 = f(x, y) + I$) of Equation 1 is a cubic curve, while the y-nullcline ($dy/dt = 0 = g(x, y)$) is a sigmoid function with the parameter β. For $I > 0$, the two nullclines intersect only on the middle branch of the cubic, and Equation 1 gives rise to a stable periodic orbit for all values of ϵ sufficiently small (Figure 2a). The periodic solution alternates between a phase of relatively high values of x, called the *active phase* of the oscillator, and a phase of relatively low values of x, called the *silent phase* of the oscillator. Within these two phases, Equation 1 exhibits near steady state behavior. Compared to the behavior within the two phases, the transition between the phases takes place on a fast time scale. In this case, that is, $I > 0$, we call the oscillator *enabled*. For $I < 0$, the two nullclines intersect on the left branch of the cubic, and Equation 1 produces a stable fixed point (an equilibrium point) at a low value of x (Figure 2b). In this case, we call the oscillator *disabled*. The parameter γ is introduced to control the relative times an enabled oscillator spends in the two phases, and a larger γ yields a relatively short active phase.

The model of Equation 1 resembles the simple neuronal models of action potential generation by FitzHugh (1961) and Nagumo, Arimoto, and Yoshizawa (1962). However, the form of nonlinearity plus the parameter γ provides a dimension of flexibility that is missing from the FitzHugh–Nagumo equations. Thus, Equation 1 can be interpreted biologically as a model of action potential generation of a single neuron (see also Morris & Lecar, 1981). The oscillator model may also be interpreted as a mean field approximation to an interacting network of excitatory and inhibitory neurons (Buhmann, 1989; Sporns, Gally, Reeke, & Edelman, 1989).

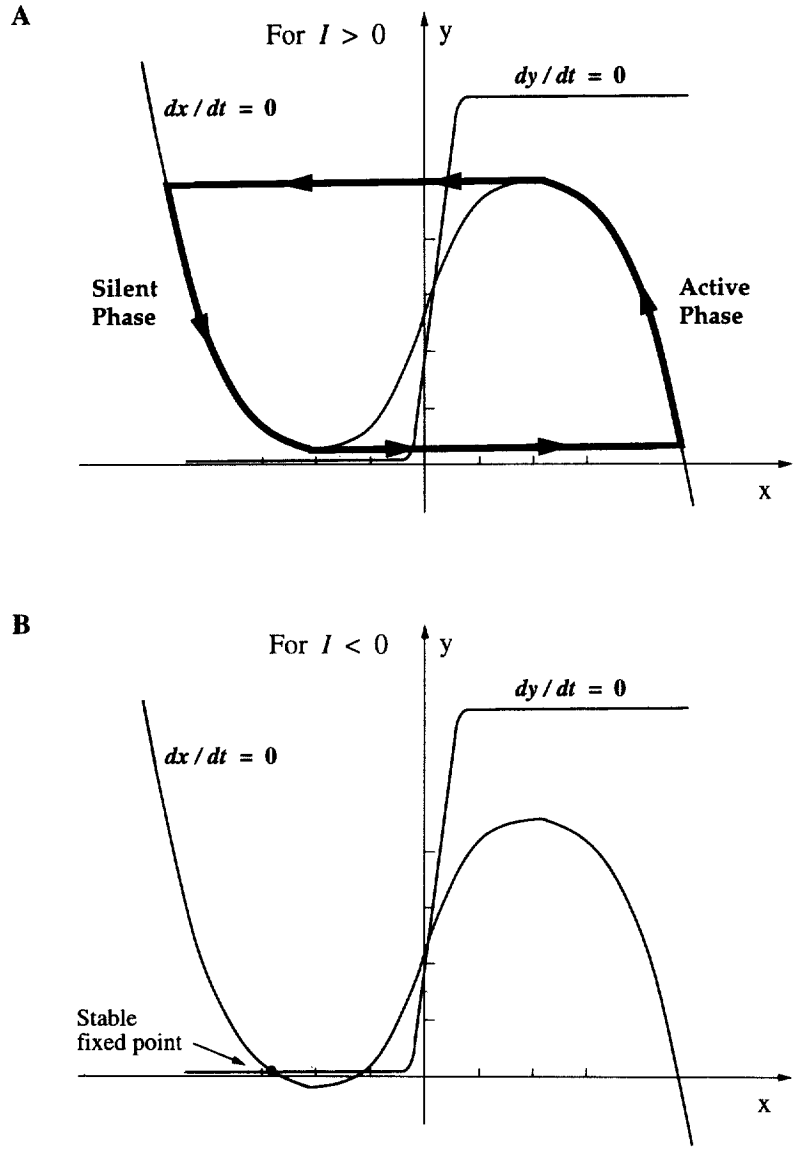


Figure 2. Nullclines and behavior for a single oscillator. (**A**) When $I > 0$, the system gives rise to a stable periodic (limit cycle) solution. The periodic orbit is shown with a bold curve. The orbit consists of an active phase and a silent phase of near steady state behavior (indicated by single arrowheads), and fast transitions between the two phases (indicated by two arrowheads on each transition). (**B**) When $I < 0$, there is no periodic solution. In this case, the system yields an asymptotically stable fixed point on the left branch of the cubic.

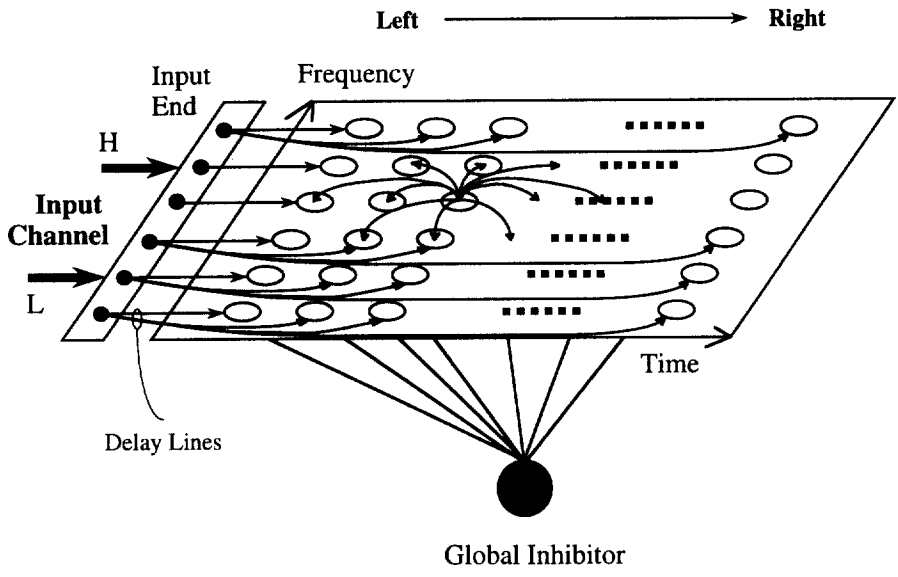


Figure 3. Diagrom of the segregation network. The main architecture is a two-dimensional time-frequency matrix. External input is applied to the input end that consists of a set of input channels. Each input channel detects input of a specific frequency and projects to a (horizontal) row of isofrequency oscillators with decreasing delays from left to right. The connections from a typical oscillator in the network are shown in the figure, and those from other oscillators are omitted for clarity. The global inhibitor receives an excitatory input from and sends an inhibitory output to every oscillator in the matrix. As in the following figures, symbol *H* indicates an auditory input with a high frequency and *L* indicates an auditory input with a low frequency.

The fact that auditory scene analysis is a time-varying process and depends on the rate of presentation requires a representation of time for any system that deals with auditory scene analysis. We treat time in this model as a separate dimension. To simplify the discussions, we consider only time and frequency; it should be straightforward to include other dimensions. The architecture of the network we study thus consists of a two-dimensional matrix of oscillators plus a global inhibitor. One dimension represents time, and another one represents frequency, as shown in Figure 3. Each oscillator in the matrix is laterally connected with its neighboring oscillators, and lateral connections are all excitatory. The global inhibitor receives excitation from each oscillator, and inhibits in turn each oscillator of the network. The network has an input end that consists of units representing distinct frequencies, called *input channels.* Each input channel connects to a corresponding oscillator row representing the same frequency, called a *frequency channel,* by delay lines with different delays. These delay lines are arranged so that delays decrease systematically from left to right. Thus,

each oscillator in the matrix is activated by an input of a specific frequency at a specific time relative to the present time.

Following Wang and Terman (1995) and Terman and Wang (1995), we use a pair of weights to describe the connection from one oscillator to another: one representing permanent weight and another dynamic weight. Specifically, we use T_{ij} and J_{ij} to represent the permanent connection weight and the dynamic connection weight from oscillator j to oscillator i, respectively. Permanent connections reflect the hardwired structure of a network (see Figure 3), whereas dynamic connections quickly change their strengths from time to time, depending on the current state of the network. For neural computation, only dynamic connections formed on the basis of permanent connections play an effective role. The idea of using two types of connection weights was first proposed by von der Malsburg (1981; von der Malsburg & Schneider, 1986). Wang (1993, 1995a) later formulated the idea to the dynamic normalization mechanism for synchronizing a population of locally coupled neural oscillators. Dynamic normalization is adopted in the present model as the modification rule of dynamic links J_{ij}, which combines a Hebbian rule (Hebb, 1949) that emphasizes coactivation of oscillators i and j and normalization of all incoming connections to an oscillator. More specifically, it is a two-step procedure: First, update dynamic connections on the basis of permanent connections and then normalize dynamic connections:

$$\Delta J_{ij} = \eta\, T_{ij}\, h(x_i)\, h(x_j) \tag{2a}$$

$$J_{ij} = \frac{W_T\,(J_{ij} + \Delta J_{ij})}{[c + \sum_k (J_{ik} + \Delta J_{ik})]} \tag{2b}$$

where the parameter η controls the speed of dynamic modification, and W_T specifies the overall strength of dynamic connections converging on a single enabled oscillator at any given time; that is, every enabled oscillator receives the same amount of dynamic connections. The small constant c is introduced to prevent dividing by 0. The function $h(x_i)$ measures whether oscillator i is enabled. It is implemented as $h(x) = 1$ if $<x>$ is greater than a specified constant (0.05 in the following simulations) and $h(x) = 0$ otherwise. The angular bracket $<x>$ stands for temporal averaging of the activity x over a most recent time period roughly corresponding to the period of the oscillation as defined in Equation 1. All J_{ij}'s are initialized to 0. Thus, if any given oscillator i is disabled, that is, $h(x_i) = 0$, then $J_{ij} = 0$ for every j. Normalization of all incoming dynamic weights to a single unit is commonly used in various neural models (see among others, Goodhill & Barrow, 1994; von der Malsburg, 1973; Wang & Arbib, 1990).

The permanent connectivity pattern between the oscillators in the segregation network, except for the self-connection, is assumed to take on a two-

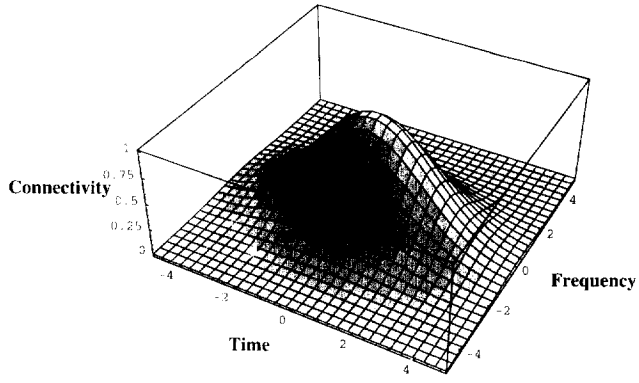


Figure 4. Two-dimensional Gaussian distribution that describes the strengths of the permanent lateral connections of the center oscillator. The distribution is generated according to Equation 3 in the text. The parameter values are $t_i=0$, $f_i=0$, $\sigma_t=8.0$, and $\sigma_f=5.0$.

dimensional Gaussian distribution: The permanent connection strength between any pair of oscillators on the matrix of Figure 3 falls off exponentially with the distance between the two oscillators. Let the two-dimensional indices of oscillator i be (t_i, f_i), representing the time and frequency coordinates of the oscillator respectively. Then, oscillator i connects to oscillator j with strength

$$T_{ij} = exp[-(\frac{(t_j - t_i)^2}{\sigma_t^2} + \frac{(f_j - f_i)^2}{\sigma_f^2})] \tag{3}$$

where the parameters σ_t and σ_f determine the widths of the Gaussian distribution along the time axis and the frequency axis, respectively. It is easy to see that $T_{ij} = T_{ji}$, that is, the permanent weights are symmetrical. Figure 4 illustrates a two-dimensional Gaussian distribution, which shows the permanent connection strengths of the center oscillator. The oscillator receives connections from the oscillators with higher (indicated by a positive number) or lower (indicated by a negative number) frequency coordinates, and higher or lower time coordinates. The self connectivity T_{ii} is set to 0. Such a Gaussian distribution is often used to describe the lateral connection pattern in the brain (Hubel, 1988; von der Malsburg, 1973). Once T_{ij}'s are defined, J_{ij}'s are updated according to Equation 2.

The coupling term S_i (see Equation 1a) from the oscillator network to oscillator i is given by

$$S_i = \sum_j J_{ij} S_\infty(x_k, \theta_x) - W_1 S_\infty(z_1, \theta_1) - W_2 S_\infty(z_2, \theta_2) \tag{4}$$

$$S_\infty(x, \theta) = \frac{1}{1 + exp[-\kappa(x - \theta)]} \tag{5}$$

The first term of the right-hand side of Equation 4 describes the lateral excitatory connections to i, and the second term and the third term describe the inhibition from the global inhibitor. Note that dynamic weights J_{ij}'s, not permanent weights T_{ij}'s, are used in Equation 4. The parameter θ_x is a threshold above which an oscillator can influence the oscillators it connects to (see the sigmoid function of Equation 5). W_1 and W_2 are the weights of inhibition from the global inhibitor, which we denote by a pair of the units z_1 and z_2. The activity of the global inhibitor is defined as

$$\frac{dz_1}{dt} = \phi \cdot (\sigma_\infty - z_1) \tag{6a}$$

$$\frac{dz_2}{dt} = \phi \cdot (\sigma_0 / N_t - z_2) \tag{6b}$$

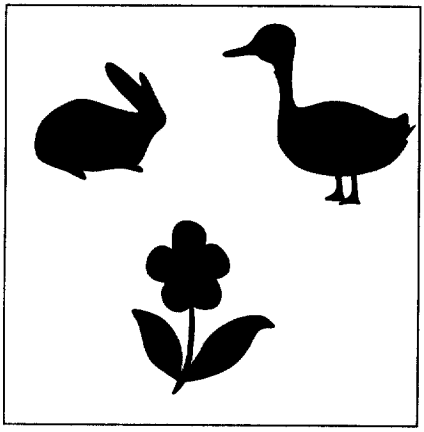
where $\sigma_\infty = 0$ if $x_i < \theta_z$ for every oscillator i and $\sigma_\infty = 1$ if $x_i \geq \theta_z$ for at least one oscillator i, and σ_0 equals the number of oscillators whose x activities are greater than or equal to θ_z. Hence, θ_z represents a threshold. If the x activity of every oscillator is below θ_z, both z_1 and z_2 approach 0, and the oscillators on the network receive no inhibition. On the other hand, if the x activity of at least one oscillator is above the threshold, the global inhibitor will receive input. In this case, z_1 approaches 1, and z_2 is proportional to the number of oscillators that exceed the threshold. Thus, every oscillator on the network will sense inhibition when z_1 is above θ_1 or z_2 is above θ_2. N_t in Equation 6b is the total number of the oscillators in a row of the segregation network (Figure 3), and θ_2 is set to $1/(2N_t)$. Finally, ϕ determines the rate at which the inhibitor reacts to such stimulation.

The system of Equations 1 to 6 in a slightly simplified version has recently been analyzed in general contexts by Terman and Wang (1995; for a much abbreviated version see Wang & Terman, 1995). The analysis concerns the properties of synchronization and desynchronization of Locally Excitatory Globally Inhibitory Oscillator Networks (LEGION; see Wang & Terman, 1995). Because their studies lay down the computational foundation of the present investigation, let us briefly summarize the main results. Their analysis is conducted with a general two-dimensional laterally connected matrix of relaxation oscillators with a global inhibitor, the same structure shown in Figure 3 after the delay lines are excluded. In particular, they have analyzed nearest neighbor connections on the network and the global inhibitor composed of only z_1. Let a *pattern* be a connected region, and a *block* be a subset of oscillators stimulated by a given pattern. For the case $\epsilon > 0$ sufficiently small, the network exhibits a mechanism of *selective gating,* whereby an enabled oscillator jumping up to the active phase rapidly recruits the oscillators stimulated by the same pattern, while preventing other oscillators from jumping up. With the selective gating mechanism, the

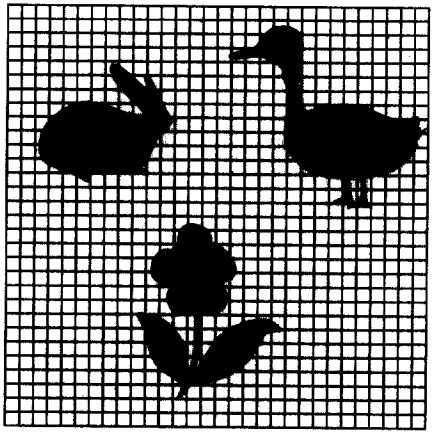
network rapidly achieves both synchronization within each block and desynchronization between different blocks. Here, desynchronization between two blocks means that they never stay in the active phase at the same time. Starting with random initial conditions, the overall time the system takes to achieve both synchronization and desynchronization is no greater than m cycles of oscillations, where m is the number of patterns simultaneously presented to the network. These results are true with an arbitrary number of oscillators, and can extend to lateral connections beyond nearest neighbor coupling.

The selective gating mechanism can be understood as a result of the interaction between local cooperation through excitatory lateral coupling and global inhibition through the global inhibitor. As described in Equations 6a and 4, once an oscillator jumps up to the active phase, it triggers the global inhibitor, which then inhibits the entire network. On the other hand, an oscillator in the active phase will spread its activation to its nearest neighbors (see Equation 4), and from them to its further neighbors. To illustrate how selective gating can be applied to scene segregation, Figure 5 shows a simulation where the input image is composed of three objects: a "rabbit," a "duck," and a "flower." These patterns are simultaneously presented to the system as shown in Figure 5A. Notice that this way of presenting the entire input image *at once,* although justified in the visual domain, cannot apply to the auditory domain. Human subjects can easily separate them visually, as the three objects are not connected in space. In the simulation, the input image was presented to a 30×30 oscillator grid as shown in Figure 5B. In terms of permanent connections, each oscillator on the grid connects only to its four nearest neighbors, except on the boundary where no wrap-around is assumed. All permanent connections have the same weight, whose precise value does not matter for computation. All the oscillators stimulated (mostly covered, see Figure 5B) by the objects received an external input $I = 0.2$, whereas the others had $I = -0.02$. Thus the oscillators under stimulation become enabled, whereas the others are disabled. For each oscillator, $\rho = 0.02$ (see Equation 1a). Namely, compared to the external input, a 10% noise is included in every oscillator. The phases of all the oscillators on the matrix were randomly initialized. Figure 5C–5F shows the instantaneous activity (snapshot) of the network at various stages of dynamic evolution. The diameter of each black circle represents the x activity of the corresponding oscillator. That is, if the range of x values of all the oscillators is given by x_{min} and x_{max}, then the diameter of the black circle corresponding to an oscillator is proportional to $(x - x_{min})/(x_{max} - x_{min})$. Figure 5C shows a snapshot of the network at the beginning of the simulation. The activities of the network were random at this time. Figure 5D shows a snapshot after the system had evolved for a short time period. One can clearly see the effect of grouping and segregation: All the oscillators

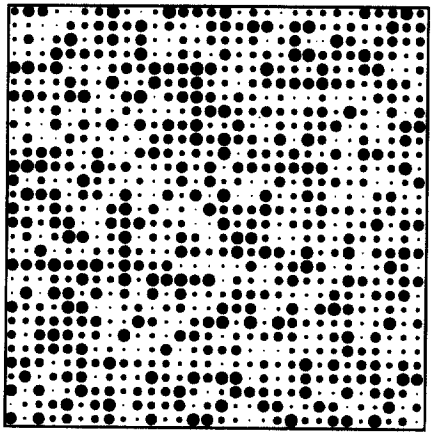
A



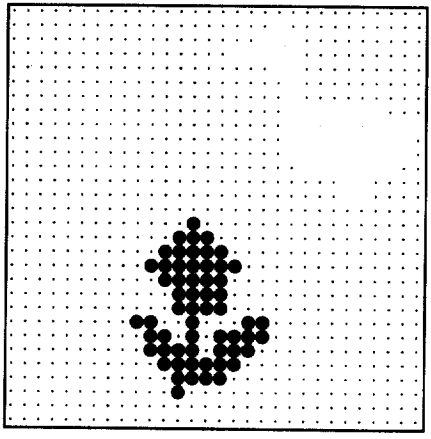
B



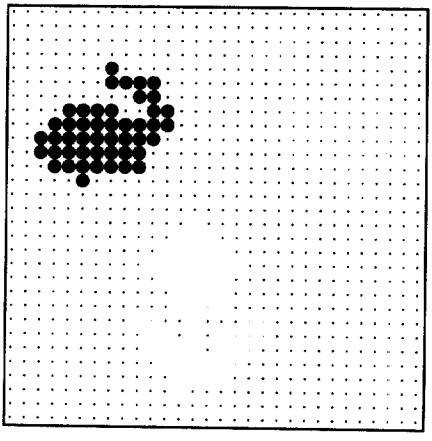
C



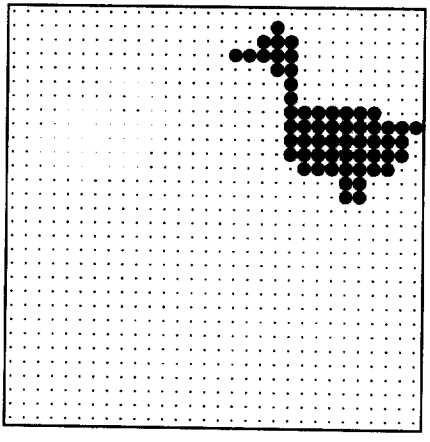
D



E



F



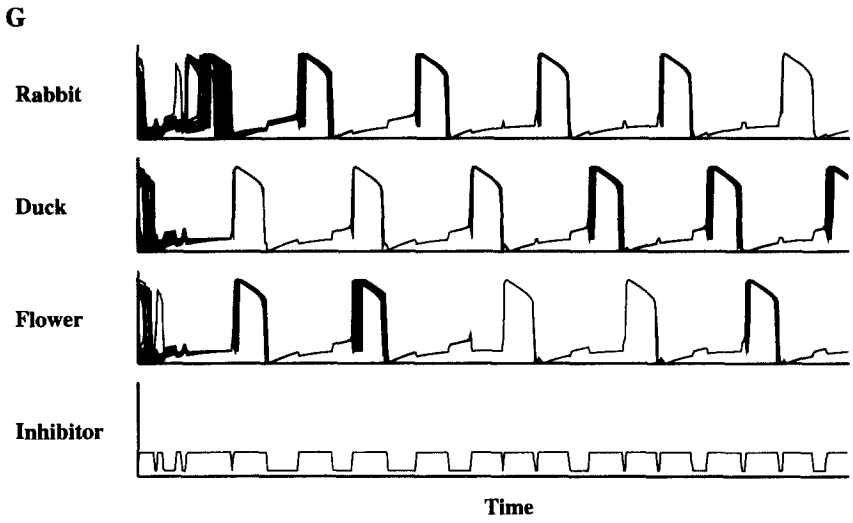


Figure 5. Scene segmentation. (**A**) An input image with three patterns, each being a connected region (Thanks to Ping Bai for drawing the illustration). (**B**) The input image as sampled by a 30×30 grid of oscillators. An oscillator receives an external input $I=0.2$ if the corresponding box is covered by the image at least by half. Otherwise $I=-0.02$ for an oscillator. (**C**) A snapshot (instantaneous recording) of the activities of the oscillator grid at the beginning of dynamic evolution. (**D**) A snapshot taken shortly after the beginning. (**E**) Another snapshot taken shortly after D. (**F**) Another snapshot taken shortly after E. (**G**) Temporal activities of the oscillator grid. The upper three traces show the combined temporal activities of the oscillator blocks representing the patterns indicated by the symbols to their left, respectively. The bottom trace shows the temporal activity of the global inhibitor. The parameters: $\epsilon=0.02$, $\phi=3.0$, $\gamma=6.0$, $\beta=0.1$, $\kappa=50$, $\eta=10.0$, $W_T=6.0$, $W_1=1.5$, $\theta_X=-0.5$, $\theta_Z=\theta_1=0.1$. The simulation took 8,000 integration steps.

representing the "flower" were entrained and had large activities. At the same time, the oscillators stimulated by the other two patterns had very small activities. A short time later, as shown in Figure 5E, the oscillators stimulated by the "rabbit" reached high values and were separated from the rest of the input. Finally in Figure 5F, the oscillators representing the "duck" were highly active and the rest of the input remained inactive. The almost empty regions in Figure 5D–5F reflect the fact that when an active block of oscillators jumps down to the silent phase, these oscillators reach their minimum x values (see Figure 5G). This successive "pop-out" of the objects continued in a stable periodic fashion. To provide a complete picture, Figure 5G shows the temporal evolution of each oscillator. Because the oscillators receiving no external input were disabled during the entire simulation process, they were excluded from the display. The activities of the oscillators stimulated by each object are combined together in the figure. Thus, if they are synchronized, they appear like a single oscillator. In Figure 5G, the three upper traces represent the activities of the three oscillator blocks, and the bottom trace represents the activity of the global

inhibitor. Although the network started with random phases, the synchronization within each block and the desynchronization between different blocks are clearly shown within just two oscillation cycles of dynamic evolution.

Emergent synchrony on the basis of local connections as illustrated in Figure 5 (see also Somers & Kopell, 1993; Wang, 1993, 1995a) eliminates the use of either all-to-all connections (Sompolinsky, Golomb, & Kleinfeld, 1991; von der Malsburg & Schneider, 1986) or a global phase coordinator (Kammen, Holmes, & Koch, 1989), the two mechanisms commonly used for reaching global phase synchrony. The failure to achieve emergent synchrony with local coupling is largely due to the use of sinusoidal (or harmonic) oscillators and linear coupling, which have been shown to differ qualitatively from relaxation oscillators and nonlinear coupling as used in the preceding definitions (Somers & Kopell, 1993; Terman & Wang, 1995; Wang, 1993). Note that the global inhibitor in Figure 3 is used here for producing desynchrony, and that rapid synchrony occurs regardless of the global inhibitor (Terman & Wang, 1995). From the perspective of scene analysis, this is significant because global mechanisms of synchronization would lead to indiscriminate synchrony. For example, the task of segmenting the three objects in Figure 5A could not be performed by an oscillator network that relies on global mechanisms. This is because critical information of geometrical (spatial) relations between the objects is lost in a globally connected network, and thus the oscillator block representing the "rabbit" and the block representing the "duck," for example, would be connected in the same way as the oscillators within the same block.

The precise definitions of the cubic and the sigmoid in $f(x, y)$ and $g(x, y)$ do not matter for following simulations (see Equation 1). Terman and Wang (1995) gave more discussions about the generality of their mathematical analysis of LEGION. What is important is the network ability to rapidly achieve both synchrony with local connections and desynchrony among different oscillator groups. To the best of our knowledge, LEGION is the only model that can achieve this network ability. Furthermore, this behavior of LEGION is established by mathematical analysis. The rapidity of synchrony and desynchrony is particularly important for the auditory domain, because auditory signals to be segregated vary quickly from time to time.

What complicates auditory segregation is *time.* An auditory scene must be unfolded over a certain time, and the auditory scene keeps changing. In addition, temporal relationships unique to audition must be captured. Before we present detailed simulations, let us see how the oscillator network of Figure 3 functions in general terms. From the study of the general oscillator network (Terman & Wang, 1995; Wang & Terman, 1995), we know that a pair of oscillators synchronize if they both are enabled and their mutual dynamic excitation is sufficiently strong to overcome the global inhibition. Otherwise, the two oscillators would be desynchronized.

According to Equation 2, in order to form an effective dynamic connection between two oscillators, they both must be enabled by external stimulation and there must be nonzero permanent connections between them. If a dynamic connection can be formed, its strength will be proportional to the strength of the corresponding permanent connection. The previous results (Terman & Wang, 1995) also suggest that synchronization is transitive: If separately oscillator i can synchronize with oscillator j and oscillator j can synchronize with oscillator k, then all three are synchronized. The delay lines in the segregation network serve to provide some form of short-term memory (STM) that can make simultaneously available a recent history of external stimulation. Thus, the transitivity of synchronization plus the Gaussian distribution of permanent connections (Figure 4) should promote grouping of a sequence of interleaved tones that have similar frequencies (proximity in frequency) and/or high presentation rates (proximity in time). At the same time, tones that cannot be grouped will be segregated due to the global inhibition. We now present simulation experiments using the segregation network with respect to a set of psychological phenomena regarding primitive auditory scene analysis.

SIMULATION RESULTS

The segregation network (Figure 3) with the detailed quantitative description defined in the previous section has been simulated with respect to auditory segregation. To reduce numerical computations involved in integrating a large number of the differential equations of Equations 1 and 6, a computer algorithm has been extracted based on these equations. The algorithm follows every major step in the numerical simulation of the equations, preserving the essential properties of relaxation oscillators, such as two time scales (fast and slow) and the properties of synchrony and desynchrony. More specifically, the following approximations have been made:

1. When no oscillator is in the active phase (see Figure 2), the one closest to the jumping point among all enabled oscillators is selected to jump up to the active phase.
2. An oscillator jumps up to the active phase immediately if it receives an excitatory input from its neighbors and the net input it receives from external input, neighboring oscillators, and the global inhibitor is positive.
3. The alternation between the active phase and the silent phase of a single oscillator takes one time step only.
4. All of the oscillators in the active phase jump down if no more oscillators can jump up. This situation occurs when the oscillators stimulated by the same stream have all jumped up.

In these steps, 1 and 4 are particularly effective in saving simulation time, because these two steps dramatically shorten the time a stream stays in the active phase and the silent phase, the two relatively stable and time-consuming stages in dynamical evolution (see Figure 2 and Figure 5G). Despite these approximations, it should be straightforward to see that the behavior of the segregation network produced in the following simulations will be exhibited in corresponding dynamical systems defined in Equations 1 and 6.

Simulated auditory tones are presented to the segregation network in simulated real time by triggering appropriate input channels. An oscillator in the network is activated if the stimulus has the same frequency as represented by the oscillator and an appropriate delay has elapsed since the onset of the stimulus. In order to relate to real time, we assume that the basic delay interval, that is, the difference in delay between two neighboring isofrequency oscillators is 40 ms. Assuming the same length of silences between succesive tones, the rate of presentation of a sequence of tones is inversely proportional to the duration of each tone in the sequence. The duration then corresponds to the number of enabled oscillators occupied by the tone along the time axis. When an oscillator is triggered, a random phase is generated for it. Additionally, the phases of all enabled oscillators are randomized after every delay step (40 ms). This is a reasonable assumption because an oscillator senses its input channel after every delay step (cf. Figure 3). The simulation results are presented in the following four subsections.

Auditory Stream Segregation

We first study segregation of sequential tones. Some preliminary simulation results of stream segregation were previously presented with a different oscillator system (Wang, 1994, in press). As illustrated in Figure 1, a sequence of six alternating tones *HLHLHL* is used as input in this simulation. All *L* tones are assumed to trigger the same frequency channel (F_a), and so are all *H* tones (F_b). The distance between F_a and F_b was first set to eight rows, corresponding to the condition of large frequency separation. The distance was later set to four rows, corresponding to the condition of small frequency separation. The sequence was repeatedly presented to the network, as in the psychological experiments. We conducted the simulation with durations of 160 ms and 320 ms per tone, corresponding to fast and slow presentations, respectively. Thus, for fast presentation, each tone occupies four oscillators, and for slow presentation eight oscillators.

A network of 15×30 oscillators was simulated for fast presentation. We present the results with large frequency separation first, and with small frequency separation next. Figure 6 shows the diagram of the stimulus condition with large frequency separation. The figure depicts the stimulus pattern as mapped to the oscillator network at one time, which in a sense reflects a spectrogram. Notice that the silence of one delay step (40 ms) was included

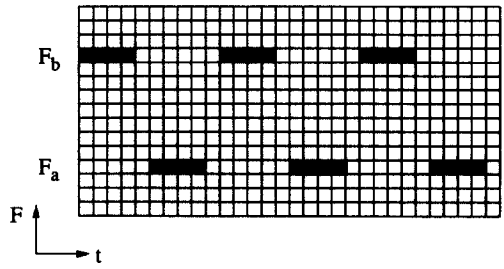


Figure 6. A stimulus pattern as mapped to the segregation network at a specific time. This condition corresponds to fast presentation and large frequency separation. The network consists of a matrix of 15×30 oscillators. The stimulus pattern is a sequence of six tones. Each tone has a frequency of either F_a or F_b, representing an *L* or an *H* tone, respectively.

between successive tones for better correspondence with experimental conditions. The presence of such silence does not affect the qualitative response of the network. Figure 7 displays the complete response of the two frequency channels triggered by the tones. Because all of the other frequency channels were not stimulated, the oscillators in those channels were disabled, hence omitted in the display. Each trace displays the activity of the excitatory unit of one oscillator. The top 30 traces represent the 30 isofrequency oscillators with progressively increasing response delays (latencies) in the F_b channel (*H* tone). Similarly, the bottom 30 traces represent the 30 oscillators in the F_a channel (*L* tone). A total of 40 delay steps were simulated, corresponding to 1,600 ms, while the complete sequence of the six tones corresponds to 30 delay steps or 1,200 ms (see Figure 6). The vertical lines were included in the figure to help compare the phases of different oscillators. As can be seen from the figure, except for a beginning period, all enabled oscillators of the F_b channel rapidly reached synchronization, and so did the oscillators of the F_a channel. Furthermore, the oscillators of one channel became desynchronized with those of the other channel. Taken together, all *H* tones are grouped into one stream, and all *L* tones are grouped into another stream. Relating to the experiments, stream segregation occurred in this simulation for fast presentation with large frequency separation, and two streams were segregated apart in real time.

Compared to Figure 5 where only connected regions are synchronized, here tones that are disconnected in the time/frequency domain can still be synchronized. This is because of the pattern of lateral connectivity. In Figure 5, only nearest neighbors are connected excitatorily. In this simulation, however, a more general scheme of lateral excitatory connectivity is adopted: a two-dimensional Gaussian distribution (see Figure 4).

Figure 8 shows the combined x values (see Equation 1) of all oscillators of each of the frequency channels for one typical delay interval after the presentation of a full sequence of six tones was completed. The top and the middle panels show the combined activities of the F_b and F_a frequency

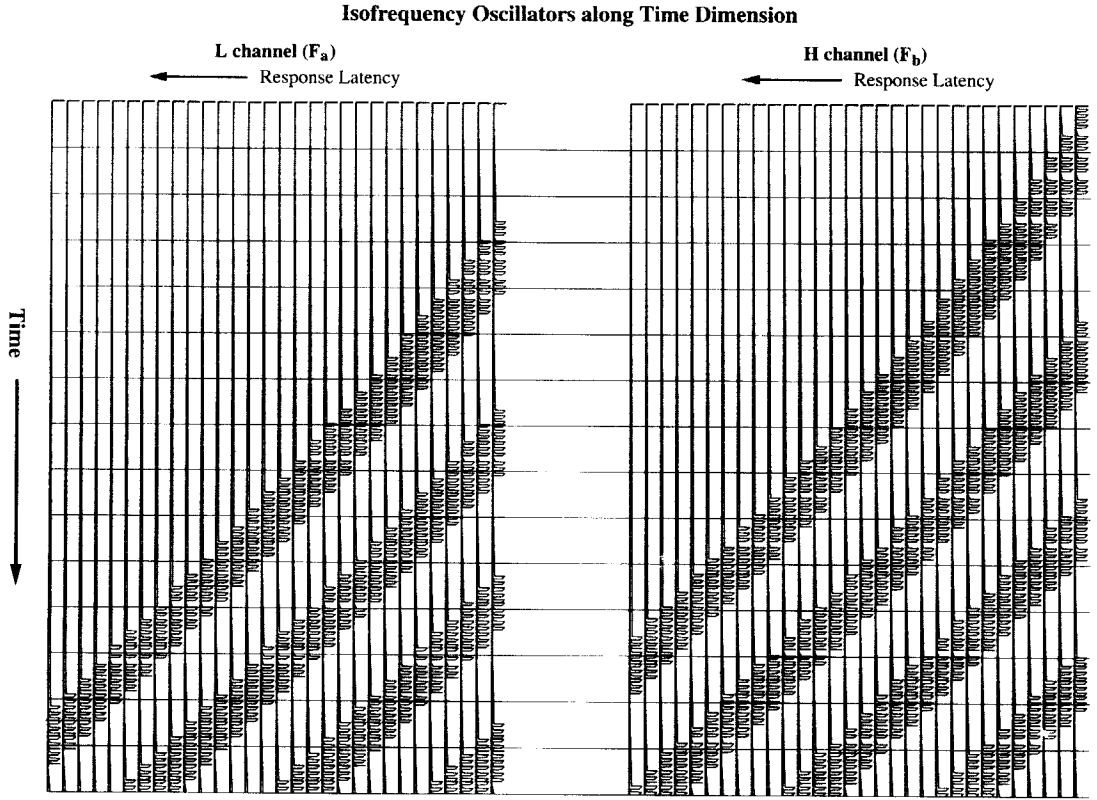


Figure 7. Response of the two corresponding frequency channels to fast presentation of alternating *H* and *L* tones with large frequency separation (8 rows). Each isofrequency channel (*H* or *L*) consists of 30 oscillators with increasing delays downward. Each activity trace represents the normalized value of the excitatory unit of an oscillator. The vertical lines are included to facilitate comparing the phases of different oscillators. The parameter values used: $\varkappa=50$, $\eta=10.0$, $W_T=6.0$, $\theta_X=-0.5$, $\theta_Z=0.1$ $\theta_1=0.5$, $W_1=0.5$, $W_2=1.0$, $\sigma_f=8.0$, $\sigma_f=5.0$, $I_i=0.2$ if oscillator i is externally stimulated, and $I_i=-0.02$ otherwise. The algorithm took 1,500 steps.

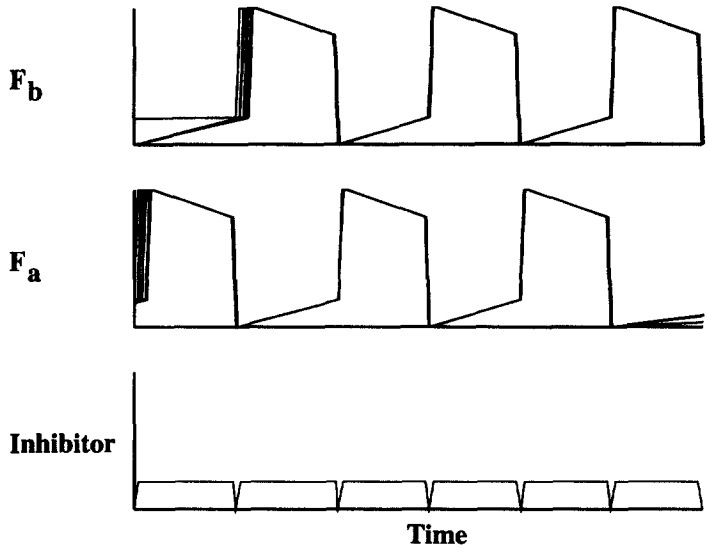


Figure 8. The combined activities of all the enabled oscillators during one delay interval in the high- (F_b) or the low-frequency (F_a) channels, respectively. The bottom trace shows the activity of the global inhibitor during the same period. The ordinates indicate the normalized x values of the oscillators and the value of the global inhibitor.

channels respectively, and the bottom panel shows the activity of the global inhibitor during the same time period. For each frequency channel, only the enabled oscillators are included in the display because disabled oscillators are always silent. As is clear in the figure, the quality of synchronization within the same frequency channel improved after the first cycle of oscillations. The frequency of the global inhibitor is double that of an enabled oscillator in the segregation network, because the inhibitor is activated by both streams. Rather rigid oscillations shown in Figure 8 as compared to Figure 5G result from algorithmic implementation of the system of Equations 1 to 6. But synchronization within the same frequency channels and desynchronization across the two channels are clearly captured in the format of illustration of Figure 8, which will be used for all the following simulations for the sake of simplicity.

It should be clear that synchrony and desynchrony in these simulations are the emergent properties of the segregation network, not caused by input. The input that a particular input channel received due to stimulation was a constant. The rhythm that happens to exist between the sequence of H (or L) tones (Figure 6), has nothing to do with the periodic oscillations generated in the oscillator network. As shown in Figure 7, the stimulus rhythm has a very different frequency than that of each oscillator. Also in the simulations, emergent synchrony can still be generated even if no rhythmic

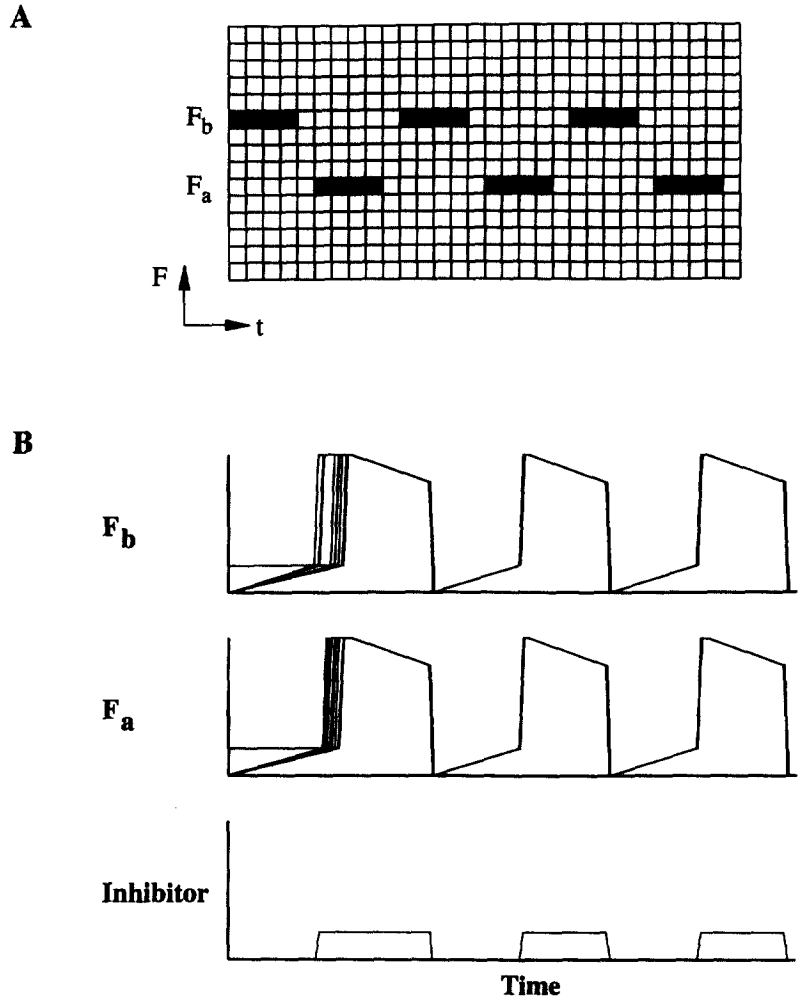


Figure 9. Response of two corresponding frequency channels to fast presentation of alternating *H* and *L* tones, with small frequency separation (4 rows). (**A**) The stimulus pattern as mapped to the segregation network at a specific time. The stimulus is a sequence of six tones, and each tone has a frequency of either F_a or F_b, representing an *L* tone or an *H* tone respectively. (**B**) The combined activities of all of the enabled oscillators in the high- or the low-frequency channels respectively, plus the activity of the global inhibitor. Only shown is one delay interval after the full sequence of six tones was presented. The parameter values are the same as in Figure 7. The algorithm took 1,500 steps.

structure exists in stimulus tones. Jones has suggested that a rhythm in a tone sequence helps the grouping of the sequence into a single stream (see the Introduction), but it is not a required condition for producing stream segregation (Bregman, 1990).

Figure 9 shows the simulation results with the same presentation rate but small frequency separation. The stimulus condition is illustrated in Figure 9A, where F_a is separated from F_b by four rows. Figure 9B displays the combined oscillatory activities for the F_a and the F_b channels as well as the activity of the global inhibitor, for a typical delay step after the full sequence of six tones was presented. The figure clearly shows that in this stimulus condition all enabled oscillators across both frequency channels were synchronized. Thus all tones are grouped into a single stream. In other words, no stream segregation occurred when frequency separation was small.

In another test, we kept large frequency separation as in Figure 6, but slowed down the presentation rate to eight oscillators per tone. In this case, a network of 15 $\times$ 54 oscillators was simulated, so that the entire sequence of six tones could be represented in the simulation. Figure 10A depicts the stimulation pattern, and Figure 10B provides the simulation results. Different from either of these cases, there was no phase synchronization at all between different tones, even within the same frequency channel. Rather, each tone formed its own stream, and the entire "scene" was segregated into six streams. This can also be seen from the activity of the global inhibitor, whose frequency is now six times that of the each stream.

Another logical condition, which is small frequency separation (four rows) combined with slow presentation (eight oscillators per tone), has also been tested. As expected, all six tones are grouped into the same stream as similarly shown in Figure 9. We have also tested an intermediate duration for each tone—six oscillators per tone—with large frequency separation. Consistent with an earlier observation (Wang, in press), the results are somewhere between Figure 8 and Figure 10B. That is, partial stream segregation occurred for this medium presentation rate.

From all the given simulations, we conclude that tones can be grouped together based on their proximity in frequency, and stream segregation critically depends on the rate of presentation. Stream segregation is best for high rates of presentation and absent for low rates. These simulation results are consistent with classical psychological findings on stream segregation as reviewed in the Introduction.

Frequency Modulation

The phenomenon of stream segregation is established similarly for frequency-modulated (FM) tones, such as glides with gradually increasing or decreasing frequencies (Bregman, 1990, pp. 58–65; Steiger & Bregman, 1981). The simulations presented here concern only pure tones. The basic simulation results of stream segregation, however, extend to FM tones as well. This is demonstrated in the following simulation. As shown in Figure 11A, the stimulus pattern consists of eight glides that fall into two frequency regions, R_1 and R_2. The two regions are separated in frequency by six rows.

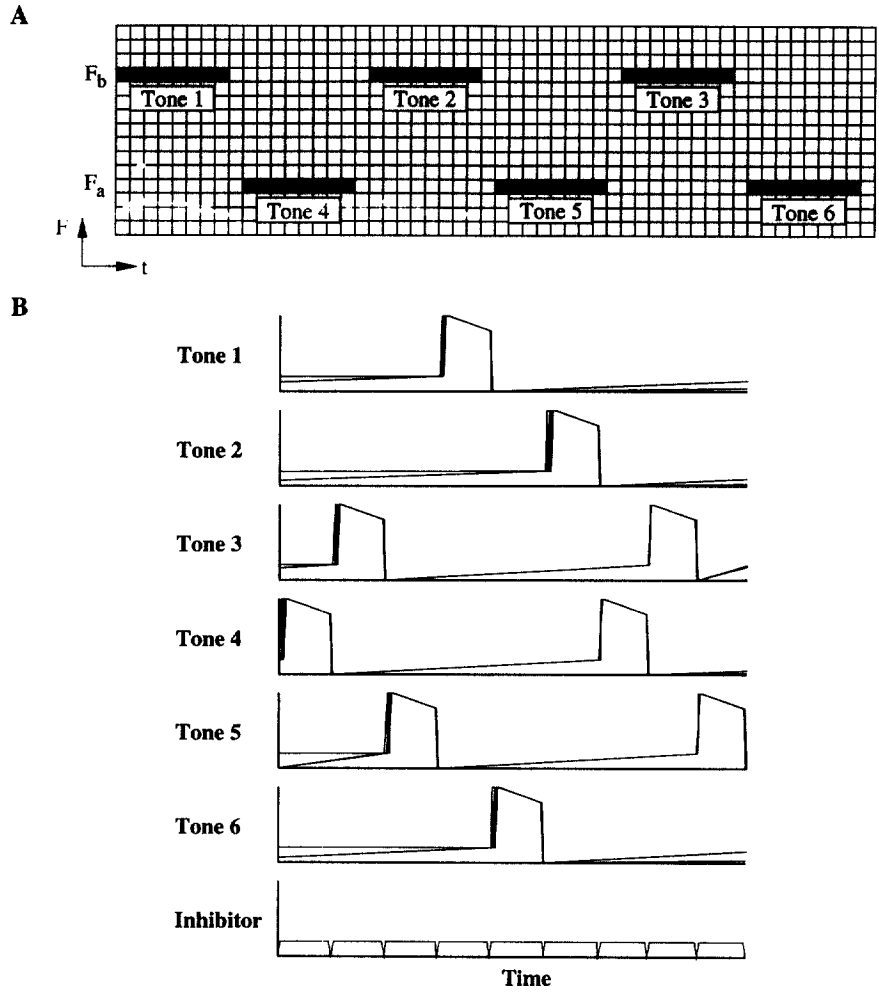


Figure 10. Response of two corresponding frequency channels to slow presentation of alternating *H* and *L* tones, with large frequency separation (8 rows). (**A**) The stimulus as mapped to the segregation network at a specific time. The stimulus is a sequence of six tones. Each tone has a frequency of either F_a or F_b, representing an *L* tone or an *H* tone respectively. The network consists of a matrix of 15×54 oscillators. (**B**) The combined activities of all of the enabled oscillators representing each single tone, plus the activity of the global inhibitor. Only shown is one delay interval after the full sequence of six tones was presented. The parameter values are the same as in Figure 7. The algorithm took 2,800 steps.

A network of 15×40 oscillators was simulated. As clearly shown in Figure 11B, all of the enabled oscillators in R_1 were synchronized, and so were the enabled oscillators in R_2. But the oscillations in the two regions were desynchronized from each other. The simulation results demonstrate that stream segregation occurred and the tones are grouped into two different streams.

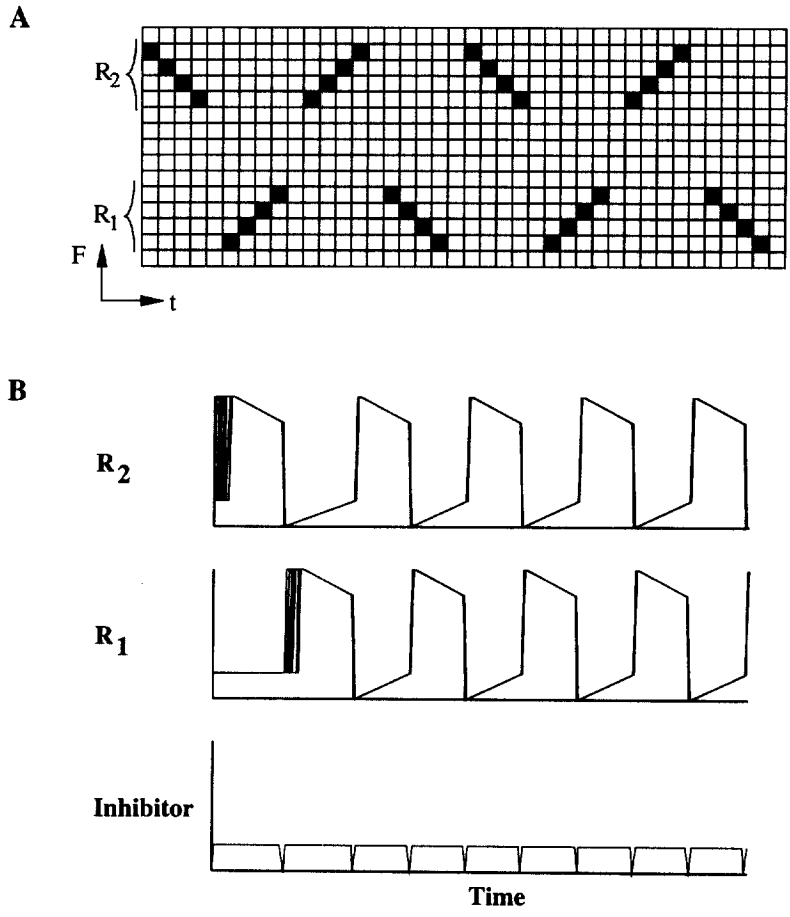


Figure 11. Response of two corresponding frequency regions, separated by 6 rows, to fast presentation of a sequence of 8 FM tones. (**A**) The stimulus pattern as mapped to the segregation network at a specific time. The network consists of a matrix of 15×40 oscillators. (**B**) The combined activities of all of the enabled oscillators in the two frequency regions respectively, plus the activity of the global inhibitor. Only shown is one delay interval after the full sequence of six tones was presented. The parameter values are the same as in Figure 7. The algorithm took 2,400 steps.

Here, frequency proximity plays the dominant role in stream segregation because relatively strong connections between oscillators of similar frequency but dissimilar time (see Figure 4) are responsible for grouping the tones in each frequency region together. The simulation results conform with the findings of Steiger and Bregman (1981), which show that frequency relations are of primary importance for segregation of FM tones. To be clear, this simulation does not deal with common FM modulation, which is known to be a factor of grouping (see Introduction). A discussion of how to incorporate FM modulation into this model is given in the Discussion section.

Before we conduct further simulation experiments, let us understand why the previous behavior of the segregation network occurs. It can be explained by general principles of competition and cooperation in the segregation network. Oscillators that are adjacent in both the frequency and the time domains will always synchronize because they are strongly coupled with each other (see Figure 4). For the fast presentation, interleaved tones of similar frequencies are separated only by six oscillators (Figures 6, 9, and 11). Thus, with strong coupling, all tones of similar frequencies are grouped into the same stream. With large frequency separation (Figures 6 and 11), tones of dissimilar frequencies cannot be grouped together. Furthermore, because of global inhibition exerted by the global inhibitor, tones that cannot be grouped will be segregated, thus producing stream segregation. With small frequency separation (Figure 9), tones of different frequencies can still be grouped together and no streaming occurs. On the other hand, slowing down the rate of presentation increases the distance between interleaved tones of similar frequencies, and thus reduces the probability of grouping these tones. This explains the results in Figure 10, where no grouping occurs at all. The global inhibitor plays a critical role in segregation. Without it, all enabled oscillators in the network would synchronize regardless of which tones they represent, as these oscillators would form a locally coupled population by the Gaussian connectivity pattern (Figure 4), which we have shown will reach global synchrony (see Terman & Wang, 1995; Wang & Terman, 1995).

Sequential Capturing

One of the well-known phenomena in auditory scene analysis is so-called sequential capturing. It was first reported by Bregman and Pinker (1978; see also Bregman, 1990), who tested a repeating sequence formed by a pure tone T_3 and a complex tone composed of two pure tone components T_1 and T_2 (Figure 12A). By varying the frequencies of T_2 and T_3 and the onset time differences between T_1 and T_2, they show that T_3 may capture T_2 from the complex T_1/T_2 to form a new stream T_2/T_3. Sequential capturing is promoted by decreasing frequency distance between T_2 and T_3, increasing frequency distance between T_1 and T_2, or increasing onset time differences between T_1 and T_2.

We have simulated the phenomenon of sequential capturing. To capture the frequency relations between different tones, we held T_1 and T_3 tones at constant frequencies and set T_2 to two different frequencies. As in the experiments, a repeating sequence of three simulated pure tones were used, as shown in Figure 12A. Relating to the experiments of Bregman and Pinker, F_a tones correspond to T_1 tones, F_c to T_2, and F_d to T_3. The tones with frequencies F_a and F_c constitute a complex tone, with substantial overlapping in time (we avoided full onset synchrony between them). We

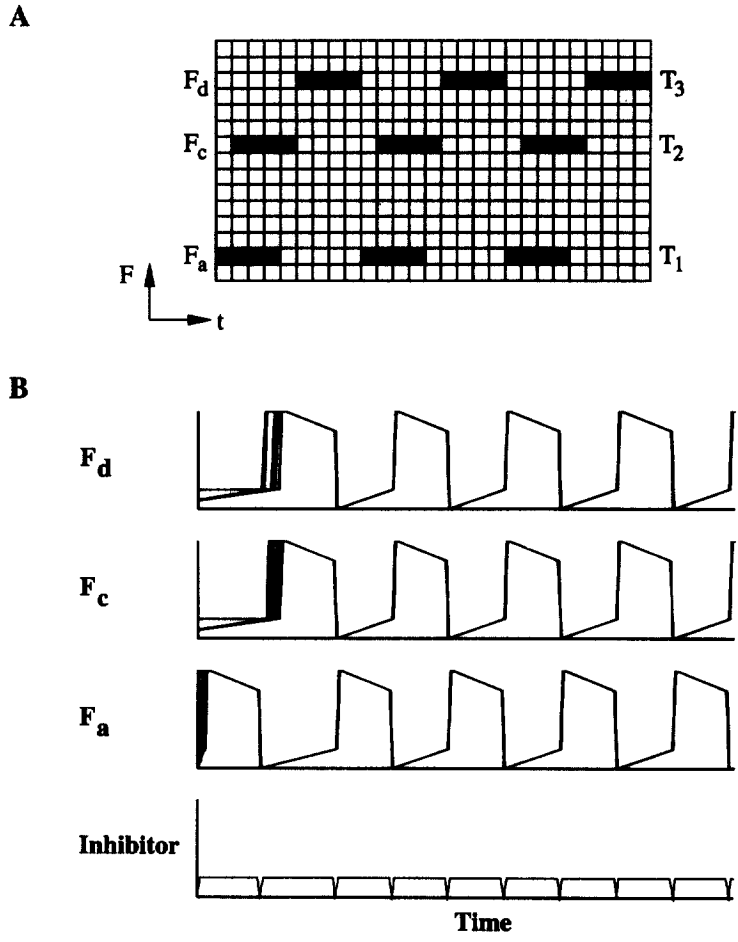


Figure 12. Sequential capturing: Case 1. (**A**) The stimulus pattern to the segregation network at a specific time. The stimulus is a sequence of nine tones, triggering three frequency channels: F_a, F_c, or F_d. Overlapping F_a and F_c tones compose the complex tone and F_d tones are the captor. The separation between F_a and F_c is 7 rows, and between F_c and F_d is 4 rows. The network consists of a matrix of 15×27 oscillators. (**B**) The combined activities of all of the enabled oscillators in each of three frequency channels, plus the activity of the global inhibitor. Only shown is one delay interval after the full sequence of six tones was presented. The parameter values are the same as in Figure 7 except that $W_1 = 0.3$ and $W_2 = 1.2$. The algorithm took 1,600 steps.

first tested the case that the distance between T_2 and T_3 is closer than that between T_1 and T_2. A network of 15×27 oscillators was simulated so that it could represent a sequence of three repetitions of the stimuli. Figure 12B provides the simulation results within a typical delay step after the whole sequence was presented. As is clear in the figure, the enabled oscillators in the

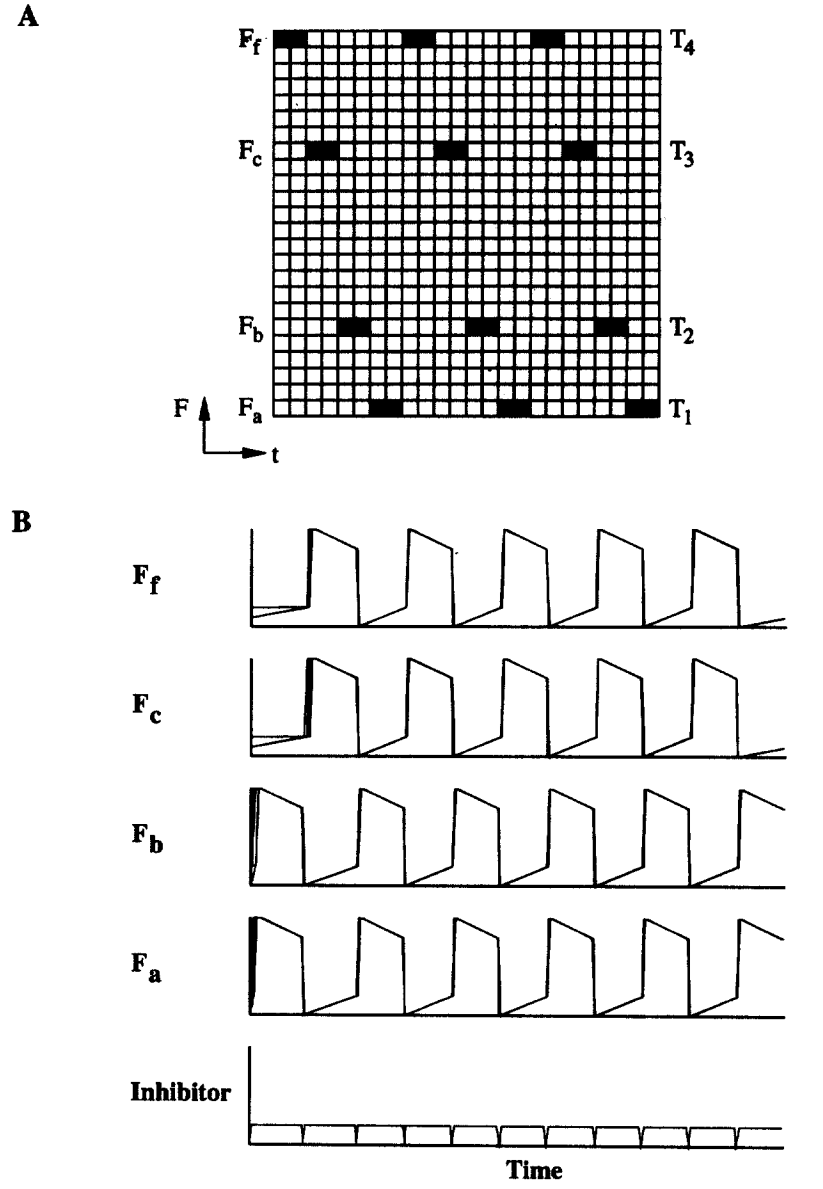


Figure 14. Competition among different organizations: (**A**) The stimulus pattern as mapped to the segregation network at a specific time. The stimulus is a sequence of 12 tones, triggering four frequency channels: F_a, F_b, F_c or F_f. The network consists of a matrix of 24×24 oscillators. (**B**) The combined activities of all of the enabled oscillators in each of four frequency channels, plus the activity of the global inhibitor. Only shown is one delay interval after the full sequence of six tones was presented. The parameter values are the same as in Figure 7 except that $W_1 = 0.1$ and $W_2 = 1.8$. The algorithm took 1,600 steps.

The simulation results in Figures 12 and 13 resemble those of the corresponding psychological experiments (Bregman, 1990; Bregman & Pinker, 1978). Different from stream segregation, these simulations of sequential capturing used tones with time overlaps. As shown in the results, unlike other models of auditory segregation (e.g., Beauvois & Meddis, 1991), our model handles simultaneous tones equally well as sequential tones. We did not test the case of full onset synchrony between T_1 and T_2 because the simple segregation network of Figure 3 lacks the ability of detecting stimulus onsets, which should be incorporated in a more comprehensive model (see the Discussion section).

Competition Among Alternative Organizations

Whether a set of tones is grouped into the same stream depends not only on the arrangement within the set but also on the context of which the set is part. Grouping takes place as though different streams competed for the belonging of a specific tone (Bregman, 1978, 1990, pp. 165–172; Idson & Massaro, 1976; McNally & Handel, 1977). To test whether our model exhibits appropriate competition among rival organizations, we have simulated the experiment of Bregman (1978). In the experiment, two tones T_1 and T_2 with a fixed frequency separation formed the first pair and were presented successively. At the same time, another two tones T_3 and T_4, also with fixed frequency separation, formed the second pair and were also presented successively. The presentation of the two pairs was interleaved in time. When the frequency separation between the pairs was large, each pair formed its own stream. But when the two pairs were brought into the same frequency region so that the frequency distance within each pair was greater than the distances between T_1 and T_3 and between T_2 and T_4 across the pairs, then T_1 of the first pair formed a stream with T_3 of the second pair and T_2 formed a stream with T_4. The experiment demonstrates that whether the sequence of the first pair can form its own stream, even if the time/and frequency relations within the sequence are kept fixed, depends on the presence of other tones in the auditory scene.

To simulate the experiment of Bregman (1978), a network of 24 $\times$ 24 oscillators was used so that each tone could be repeated three times. Following the experiment, each tone was assumed to be brief. More specifically, it occupied only two oscillators. We first tested the case with large frequency separation between the two pairs. The stimulus condition is described in Figure 14A, where the first pair of T_1/T_2 stimulates the frequency channels of F_a and F_b respectively, and the second pair of T_3/T_4 stimulates the frequency channels of F_c and F_f respectively. The frequency distance between F_a and F_b is 5 rows, between F_c and F_f is 7 rows, and between F_b and F_c is 11 rows. Figure 14B displays the simulation results. The figure shows that the sequence of T_1/T_2 tones (the first pair) was grouped into the same stream by

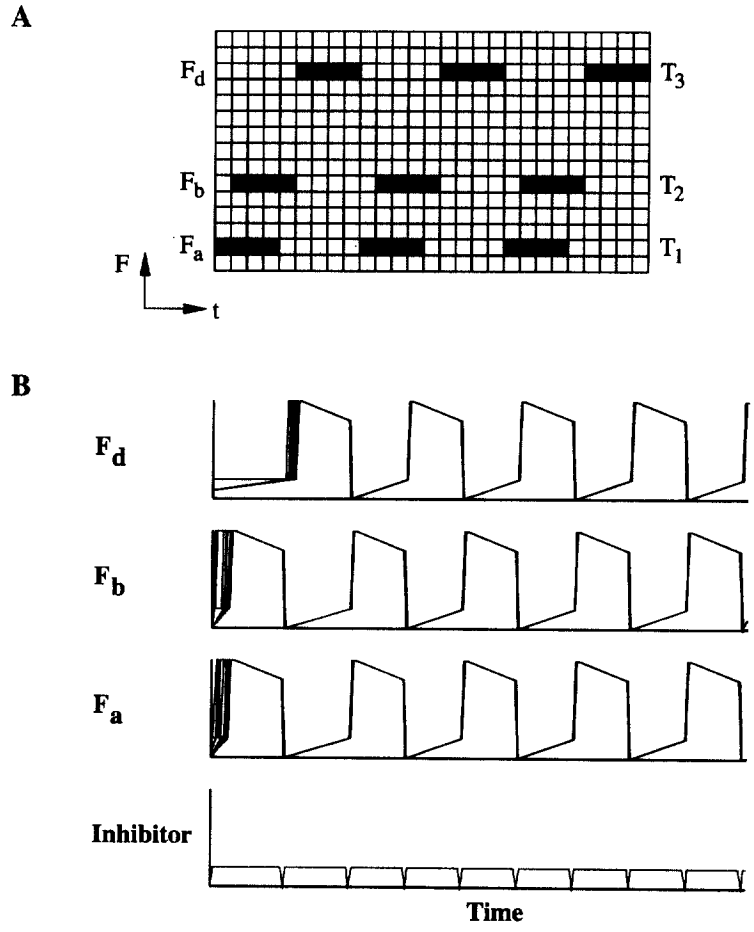


Figure 13. Sequential capturing: Case 2. The only difference from Figure 12 is that T_2 tones in this case have frequency F_b instead of F_c.

F_c and F_d frequency channels were synchronized, and their oscillations were desynchronized from those in the F_a channel. In this case, the T_3 tones captured the T_2 tones from the T_1/T_2 complex to form a new stream T_3/T_2.

Next we tested the case where the frequency separation between T_2 and T_3 is larger than that between T_1 and T_2, as shown in Figure 13A. Different from the previous case, F_b tones correspond to T_2 tones. The simulation results are shown in Figure 13B, clearly demonstrating that the complex tone of T_1/T_2 is kept together and it is separated from the captor T_3. That is, no capturing was exhibited in the simulation. As in Figure 12B, the global inhibitor in Figure 13B oscillated with a frequency double that of an enabled oscillator in the segregation network, signifying auditory segregation.

phase synchronization, and the sequence of T_3/T_4 tones (the second pair) was grouped into a different stream. We later moved the first pair up to the same frequency region as the second pair, keeping the frequency distance within each pair unchanged. In this case, the first pair of T_1/T_2 stimulated the frequency channels of F_d and F_e respectively, as shown in Figure 15A. The simulation results are given in Figure 15B. In this case, T_1 tones were grouped with T_3 tones, instead of with T_2 tones. T_2 tones were grouped with T_4 tones. The two pairs were broken down into different streams.

The simulation results well resemble the experimental findings of Bregman (1978), and capture the essential properties of competition among alternative organizations. Three aspects of the segregation network are responsible for the successful simulation: (a) lateral excitatory coupling that falls off with distance (see Figure 4); (b) dynamic normalization of Equation 2 in which only relative connection strengths are important; and (c) global inhibition. When the first pair was brought into the same frequency region as the second pair (Figure 15A), dynamic normalization led to much weakened relative coupling between frequency channels F_c and channel F_f, because each had much stronger coupling with the first pair as compared with the situation in Figure 14A. The weakened coupling within the second pair could not overcome global inhibition. Thus the pair was segregated into two streams, each of which then grouped with one tone sequence of the first pair instead.

4. A NEUROCOMPUTATIONAL THEORY OF AUDITORY SEGREGATION

Based on extensive empirical studies, Bregman put forward a theory to explain experimental phenomena of auditory scene analysis (Bregman, 1990). Bregman's theory extends Gestalt principles of grouping that result from visual observations to the auditory domain. Time is viewed as a separate dimension, and the theory builds on a number of rules of perceptual organization, such as proximity, similarity, good continuation, common fate, and so on (Rock & Palmer, 1990). Bregman's theory can explain a variety of experimental data in terms of these rules. As for primitive auditory scene analysis, his theory claims that it is a bottom-up, data-driven process, and that it is immediate and innate. It involves no attention or learning, both of which, however, are assumed to happen in the process of schema-driven (memory-based) auditory scene analysis. On the other hand, Jones' theory of rhythmic attention emphasizes the unique properties of auditory perception (Jones, 1976; Jones et al., 1981). According to the rhythmic attention theory, listeners group tones that fall into a certain rhythmic structure and anticipate the onset of a next tone using the rhythm.

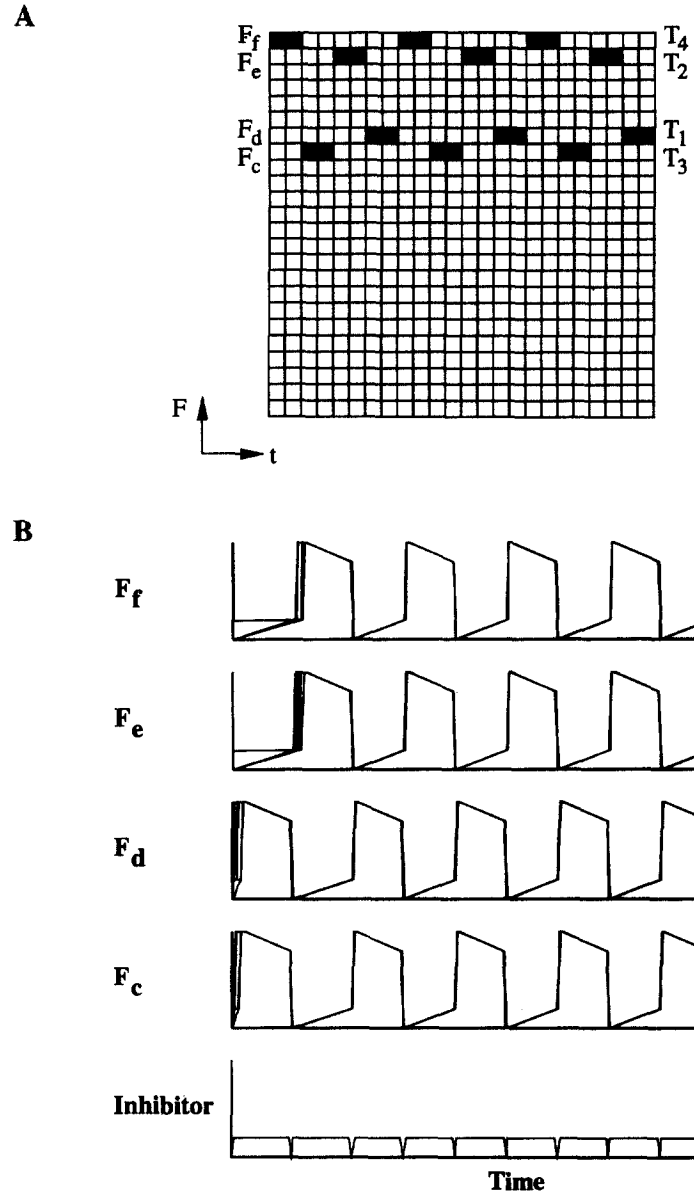


Figure 15. Competition among different organizations: Case 2: The only difference from Figure 14 is that in this simulation T_1 and T_2 tones trigger F_d and F_e frequency channels, respectively.

Contrasting Bregman, for Jones, attention is intimately involved even in what is regarded as primitive auditory scene analysis by Bregman. Also, Jones emphasizes anticipation, which involves a top-down process. The theory is supported by the experiments that vary the rhythmic structure in a sequence of sequential tones (see the first section). Furthermore, the rhythmic attention theory asserts the existence of an internal biological mechanism that is also rhythmic. This internal rhythmic mechanism entrains itself with the rhythms in a sequence of input tones, and generates active anticipation (Jones, 1976).

Neither of these theories, however, makes specific reference to the underlying neural mechanisms. Bregman's theory makes no attempts to link to biological principles. Although Jones makes an appeal to a possible biological mechanism, her theory does not offer a plausible neural model.

We present the following neural theory to explain primitive auditory segregation. The underlying structure of our theory is the oscillator network of Figure 3. Relaxation oscillators, the building blocks of the network, exhibit nonstationary behavior. This limit-cycle dynamics differs fundamentally from the commonly used limit-point dynamics, which always approaches stationary behavior (Guckenheimer & Holmes, 1983). Oscillators in the network are enabled by auditory stimuli. Because current auditory stimuli always change and the input end projects to the oscillator network through systematic delay lines, the ensemble of the oscillators thus stimulated shifts on the network constantly (see Figure 3). The basic hypothesis is that a set of tones, presented either sequentially or simultaneously, forms the same stream if the underlying oscillators enabled by the tones reach synchrony because of lateral excitatory connections. Moreover, different streams correspond to oscillator groups that desynchronize from each other because of global inhibition. Such a theory is referred to as shifting synchronization (see the Introduction).[2] More elements of the theory are given in the following.

A fundamental effect accompanying stream segregation is the loss of information about the order of successively presented tones. Take the classical example of a sequence of alternating pure tones *HLHLHL*. When stream segregation occurs, the individual perceives two streams, one for *H* tones and another for *L* tones. In the meantime, the individual cannot identify relative order between tones across the two streams, for example,

[2] The word *shifting* is used to indicate that the ensemble of the oscillators that is enabled by auditory stimuli shifts on the oscillator network. As a result, which oscillators form a synchronized group and which oscillator groups desynchronize from each other also shift on the oscillator network as the stimuli change in time. The word is also consistent with our model description that attention shifts back and forth (alternates) between different segments when stream segregation occurs (see the following paragraph).

the second tone (L) in the sequence and the third one (H). The shifting synchronization theory can explain the phenomena on the basis of the following two assumptions. First, we assume that *attention is paid to a stream when its constituent oscillators reach their active phases*. This is a natural assumption for auditory segregation on the basis of oscillatory correlation. So in our computational model, attention quickly alternates between the two different streams, and attention can be paid to multiple interleaved tones that belong to the same stream. Second, we assume that *the temporal order between two items (tones) can be perceived only if (a) they have distinct onset times, and (b) they are in the same stream, or they constitute two different streams, one of which comes before another.* The first requirement is natural, because the order would be undefined otherwise. The second one merits further discussion. If the two items are in the same stream, they will be attended to simultaneously according to the first assumption. If they do not overlap in onset times, the individual should be able to tell their order of occurrence on the basis of recency (which may be coded in STM). If, on the other hand, one item constitutes a stream that has clear temporal order with respect to the stream constituted by another item (i.e., the two streams do not interleave in time), then the sequential relation of the two streams enables the individual to perceive the order. With the two assumptions, the shifting synchronization theory explains the loss of temporal order in stream segregation as follows.

1. For fast presentation, all H tones are grouped into a single stream, and all L tones are grouped into a different stream (see Figure 8). According to the second assumption, the relative order between the tones within each stream is perceived because these tones are attended to simultaneously and have a clear ordering in their presentations. But the relative order between tones across the two streams is not perceived because they are in two streams that interleave in time.

2. For slow presentation, a number of situations are possible. If the frequency separation between the tones is not large, all tones may form a single stream (see Figure 9). In this case, the temporal order between the tones is perceived according to the assumptions. If the frequency separation is large, each tone may form its own stream (see Figure 10). In this case, the order can still be revealed because these streams do not interleave in time. For intermediate frequency separation, neighboring pairs of H and L tones may form a single stream. In this case, the order of tones is again perceived, because streams do not overlap in time. In sum, the order is fully perceived for slow presentation.

These explanations of the theory are consistent with typical psychological data of stream segregation as reviewed in the Introduction. What happens for intermediate rates of presentation? As observed in our simulations,

pairs of interleaved tones with similar frequencies may form a single stream, together with streams that are made of adjacent *H* and *L* tones. Because this situation is somewhere between the two previous cases, our theory predicts that the individual should be able to tell the order across different frequencies to a certain extent. Thus, the performance on recalling the ordering of tones should be between total confusion and full perception of the order. This effect of intermediate stream segregation seems to happen in the experiments (Bregman, 1990, pp. 143–165).

The main point of the shifting synchronization theory is that attention is directed to streams, each of which is formed by phase synchrony of its constituent oscillators, and the overlapping of different streams in time is responsible for the individual's inability to report the temporal order of tones across different streams. The latter prediction can be verified by psychological experiments. As an example, in the experiments of Bregman and Pinker (1978), sequential capturing occurs when the frequency of the captor tone T_3 is near that of tone T_2 of the complex T_1/T_2. Our theory predicts that when sequential capturing occurs, the individual cannot perceive the relative order across the sequences of T_1 tones and the T_3 tones (see Figure 12B). Because the original study of Bregman and Pinker did not discuss the loss of order information, this prediction remains to be experimentally tested.

The shifting synchronization theory explains the phenomena of competition among alternative organizations in terms of general principles of local cooperation and global competition, characteristic of the segregation network (see the previous section). These experimental phenomena (Bregman, 1978; Idson & Massaro, 1976; McNally & Handel, 1977) appear to pose thorny problems to both Bregman's theory and Jones' rhythmic attention theory. For Bregman's theory, no obvious Gestalt principle seems to readily provide a solution. For the rhythmic attention theory, the rhythmic structure does not change when the first pair (T_1/T_2) is moved to the frequency range of the second pair (T_3/T_4). The competition among different organizations seems to call for a global control mechanism for auditory segregation, which in our model is mediated by the global inhibitor.

Relating back to the previous theories of primitive auditory scene analysis, our theory is consistent with many aspects of Bregman's theory. For example, our segregation network represents an innate neural structure, involving no learning. Primitive auditory segregation in our model is also an immediate process, which is achieved in real time. Gestalt principles of grouping are embodied in the network architecture (Figure 3 is a much simplified version), such as proximity and continuation. Linking to the rhythmic attention theory, our theory also assumes the involvement of attention that rapidly switches between different streams. However, it should be clear that

attention in the shifting synchronization theory does not influence which tones should be grouped or segregated, as assumed in the rhythmic attention theory.

Is the segregation network of Figure 3 neurally plausible? As mentioned earlier, an oscillator may be interpreted either as a single neuron, as a neuron generates action potentials from time to time, or as the collective behavior of a local group of neurons. In the latter case, an oscillator corresponds to the average activity of an excitatory cell group that is mutually connected with an inhibitory cell group. The cells in the excitatory cell group have recurrent connections among themselves and excite the inhibitory cells. The inhibitory cells send inhibition to the excitatory cell group. It has been shown that such a network of excitatory and inhibitory binary neurons exhibits emergent neural oscillations (Buhmann, 1989; Sporns et al., 1989).

There is ample evidence that suggests the existence of neural oscillations in the brain (Buzsáki, Llinás, Singer, Berthoz, & Christen, 1994). It was first observed that the central olfactory system, namely the olfactory bulb and the olfactory cortex, yields oscillatory activity in response to olfactory stimulation (for a review, see Freeman, 1991). More recently, it has been reported that local field potentials in the visual cortex and the sensorimotor cortex show oscillations that are initiated by appropriate visual stimuli (Eckhorn et al., 1988; Gray, König, Engel, & Singer, 1989; Murthy & Fetz, 1992). The frequencies of these oscillations are generally between 20 and 80 Hz, often referred to as 40-Hz oscillations. In addition, neural oscillations seem to exhibit temporal coherence (synchronization) across remote regions of the visual system when stimulated by a coherent stimulus pattern. In auditory processing, 40-Hz oscillations have also been observed. Galambos, Makeig, & Talmachoff (1981) first reported that auditory evoked potentials in humans by a tone show 40 Hz oscillations, which can last for several cycles after the stimulus presentation is over. These oscillations of auditory evoked potentials were later confirmed by Madler and Pöppel (1987), who further found that these characteristic oscillations were absent from the patients under deep anesthesia. The work by Mäkelä and Hari (1987) also confirmed the observation of Galambos et al., and further suggested that the auditory cortex gives rise to the oscillations.

The work by Ribary et al. (1991) lends perhaps the most direct support to the shifting synchronization theory. Using a noninvasive imaging technique called magnetic field tomography, they recorded three-dimensional human brain activity during auditory processing. Their results show 40-Hz activity in localized brain regions both at the cortical level and at the thalamic level in the auditory system. These oscillations are synchronized over considerable cortical areas, and the synchronized oscillations can be elicited by both rhythmic and transient sound stimuli. Llinás and Ribary (1993) in a later report described 40-Hz oscillations triggered by frequency modulated tones.

An important part of the architecture of the model is an array of delay lines (see Figure 3). An array of delay lines has been argued to be neurally plausible (Hopfield & Tank, 1989), and it has been used as a basis for temporal pattern recognition (see Tank & Hopfield, 1987; Waibel, Hanazawa, Hinton, Shikano, & Lang, 1989). Latencies of neuronal responses to auditory stimuli have been found at various levels of the auditory pathway, and the range of delays increases greatly in higher auditory structures (Popper & Fay, 1992). For instance, electrophysiological recordings in the cat auditory cortex identify up to 1.6 s delays in response to a tone sequence (Hocherman & Gilat, 1981; McKenna, Weinberger, & Diamond, 1989). In echolocating bats, Dear, Simmons, & Fritz (1993) discovered that cells in the auditory cortex show systematic response latencies. Furthermore, these cortical cells with different response latencies encode multiple objects located at different distances so that the echoes arriving at different times can concurrently trigger their corresponding cells. The use of systematic delay lines is also consistent with the proposal of shifter circuits by Anderson and van Essen (1987), who argued that such neural circuits exist at many levels in the visual system. Our model uses delay lines to provide some form of STM that can simultaneously make available a recent history of external stimulation. In an electrophysiological experiment, neurons that exhibit STM characteristics have been found to constitute most of the recorded cells in the auditory cortex of the monkey during a delay task (Gottlieb, Vaadia, & Abeles, 1989).

Other structural characteristics of the neural architecture of Figure 3 include tonotopic organization and lateral connections (see Figure 4). Tonotopic organization is a characteristic structural principle of the auditory system, including the auditory cortex (Popper & Fay, 1992). Lateral connections within the auditory cortex have been observed (for a review, see Winer, 1992), and the local projections of cortical cells can span up to 3 mm in the auditory cortex. The global inhibitor in our model exerts control to the entire oscillator network, and its role is to segregate simultaneously active multiple streams. The shifting synchronization theory regards attention as shifting between different streams; thus it is reasonable to assume that the global inhibitor is involved in some form of attentional control. Crick (1984, 1994) suggested that part of the thalamus, the thalamic reticular complex in particular, may be involved in selective attention. The thalamus is located in a unique place in the structure of the brain: It sends projections to and receives input from almost the entire cortex. In light of Crick's suggestion and the structural properties of the thalamus, the global inhibitor may be speculated to correspond to a neuronal group in the thalamus. The activity of the global inhibitor should be taken to represent the collective behavior of the neuronal group.

On the basis of these discussions, it is tempting, although preliminary, to suggest that the two-dimensional oscillator network, as shown in Figure 3,

describes an aspect of the structure of the auditory cortex. The auditory cortex is bound to be more complex than the model we have specified. On the other hand, this suggestion, in the coarse scale, seems to be consistent with the neurobiology of the auditory cortex. Thus, as a working hypothesis, we suggest that auditory segregation may be achieved in the auditory cortex. Here a parallel can be made with the visual system where neurobiological evidence suggests that motion-based visual segmentation may occur in the primary visual cortex (Lamme, van Diijk, & Spekreijse, 1993; Stoner & Albright, 1992).

It is interesting to compare the neural pathways underlying vision and audition. Each human retina contains roughly 130 million light receptors, which converge onto roughly 1 million (optic) fibers of the visual nerve. Optic fibers in turn innervate about 100 million neurons in each side of the visual cortex. In contrast, each human cochlea contains rougly 12,000 receptors (inner and outer hair cells), which give rise to roughly 30,000 fibers of the auditory nerve (Kandel, Schwartz, & Jessell, 1991). The connections from the auditory nerve to the auditory cortex are relayed by many auditory nuclei, each having an increased number of neurons so that there are about 100 million cells in each side of the auditory cortex (Handel, 1989; Kandel et al., 1991). In other words, the number of visual and the number of auditory cortical cells are roughly identical. It is a striking phenomenon that the number of auditory cells significantly increases along the auditory pathway. The time delay network of Figure 3 provides an explanation of this phenomenon as follows. Approximately speaking, the representation of auditory stimuli in the cochlea is in real time; that is, the representation is instantaneous. The numerous relay stations along the auditory pathway progressively expand the time dimension so that by the level of the auditory cortex, a broad range of latencies (delays) are present. The significant increase in the cell populations along the ascending stations of the auditory pathway may provide the neural substrate for the increased ranges of latencies. However, ever increased latencies cannot stretch endlessly. The limit on the ranges of latencies may reflect the nature of STM, which provides a sliding representation that is shortlived.

We call the shifting synchronization theory *neurocomputational* (the title of this section) in order to emphasize that its underlying elements and architecture are neurally plausible.

DISCUSSION

The global inhibitor in the segregation network plays the role of breaking groups of oscillators that have weak mutual coupling into different streams by phase desynchronization. In other words, it adds a dimension of competition into the oscillator network so that only relatively strong coupling

leads to phase entrainment. The segregation network incorporates both competition (global inhibition) and cooperation (local excitation; see Arbib, 1989, for a general discussion of the roles of competition and cooperation in the brain). As a result, the network does not segregate tones based on their absolute distances in the frequency domain or the time domain, which would have been the case without global competition. Instead, the context of the tones plays an important role in the overall outcome of segregation. The influence of such contexts is responsible for competition among different organizations (Figures 14 and 15, for example). Because of this, the network also predicts an interaction between frequency proximity and presentation rate. Such "trading" between frequency and time seems to be present in human stream segregation (Jones, 1976).

Although structural similar to the comparator model in Kammen et al. (1989), the global inhibitor of our model serves an entirely different role. In the comparator model, the comparator receives input from every oscillator of a population of uncoupled oscillators, and feeds back a function of the average phase of the population to every oscillator. The comparator is used to synchronize all the enabled oscillators in the oscillator population, whereas the global inhibitor is to desynchronize oscillator groups, each of which is synchronized. Llinás and his colleagues (Llinás & Ribary, 1993; Ribary et al., 1991) also suggested that synchronization in the brain is achieved by a mechanism similar to the comparator model. More precisely, they suggest that the thalamus, through its mutual connections with the cortex, plays the role of synchronizing cortical oscillations. In our model, synchrony within each oscillator group is led to by lateral excitatory connections within the network, consistent with the suggestions of Singer and colleagues (Engel, König, Kreiter, & Singer, 1991a, 1991b; Singer, 1993; Singer & Gray, 1995) that coherent oscillations in the visual cortex result from lateral interactions within the cortex. Because both the comparator and the global inhibitor are implicated to be located in the thalamus, which provides the necessary organizational structure, the disputes may be settled by the following experiment. Assume that there are only two streams. As noted previously, our model predicts that the global inhibitor oscillates with a frequency double that of the oscillators on the network when stream segregation occurs (see Figure 8). The prediction implies that if the auditory cortex shows 40-Hz oscillations then the thalamus oscillates with a frequency of 80 Hz. In contrast, the comparator model assumes that the feedback loop between the thalamus and the cortex produces synchrony, and thus would predict that the thalamus oscillates with the same frequency as the auditory cortex. The occurrence of stream segregation is a critical condition. Otherwise, both of these models predict the same frequency of the oscillations in the auditory cortex and the thalamus. We recognize that this physiological experiment would be difficult to conduct because stream segregation is demonstrated so far only in human psychological experiments. However, animal experi-

ments in the visual domain may be able to help verify the contrasting models, as both models have also been implicated in visual segmentation.

In terms of representing multiple streams, our shifting synchronization theory can be regarded as a special form of the temporal correlation theory of von der Malsburg (1981; von der Malsburg & Schneider, 1986). Generally speaking, temporal correlation is not necessarily committed to the use of neural oscillators. Besides, the idea of using a global inhibitor to segregate multiple streams is common to both models. However, the two theories differ fundamentally in underlying *computational* models. More specifically, the single oscillator model in von der Malsburg and Schneider (1986) is constructed by ad hoc definitions that make further analysis very hard, if not impossible. Our single oscillator is a standard relaxation oscillator with well-understood mathematical properties. As mentioned in the Introduction, their network relies on full connectivity to achieve synchrony, whereas ours is based on local connectivity. As described in the first two sections of this article, the difference between full connectivity and local connectivity in generating synchrony is critical for general segregation, including auditory segregation. Although a global inhibitor is used in both models, the definition of the global inhibitor and the mechanism to achieve segregation are much different. To summarize, our model is a LEGION network that has been shown to exhibit the property of selective gating, which in turn gives rise to rapid synchrony and rapid desynchrony. These computational properties, plus systematic delay lines that represent time, which are responsible for the simulation results presented in this article, are missing from their model. As a result, their simulation of auditory segregation is limited to illustrating their idea of temporal correlation. In addition, our shifting synchronization theory provides an explanation of how auditory scene analysis may arise from an interacting neural structure (see the previous section).

The simulations in the Simulation Results section represent only a start at providing neurocomputational explanations of an important aspect of auditory perception. Many psychological data remain to be explained at this stage of model development. But, we believe that the approach outlined here—lateral connections providing the basis for encoding similarities and yielding synchronization—holds significant promise to explain a variety of experimental phenomena. Take, for example, onset synchrony, which has been used by von der Malsburg and Schneider (1986) as the determinant for producing synchronization. Our model can be extended to include another layer for detecting stimulus onset (various onset-detecting neurons have been identified in the auditory system; see Pickles, 1988; Popper & Fay, 1992). For example, Smith (1994) proposed to use different filters to detect onsets of sound, such as the difference of Gaussians. The determinant of onset synchrony can be embodied by strengthened connections between onset detectors that are activated more or less at the same time. These extra

links derived from the onset detection layer should suffice to promote grouping by synchronization among tones that exhibit onset synchrony. This is attributed to transitivity of synchronization as discussed in the earlier section on neural architecture, which is a unique property of our segregation network. Also, the permanent connectivity pattern of a Gaussian distribution (Figure 4) strongly biases the network toward the grouping of sounds that have continuous frequency transitions, which is consistent with the analysis of speech perception (Bregman, 1990; Handel, 1989).

Although much can be accomplished with the basic architecture of Figure 3 as demonstrated earlier, to achieve more realistic auditory pattern segregation, the architecture must be extended to incorporate other qualities of auditory stimuli, such as amplitude, rhythm, harmonics, timbre, and so on. Grouping based on common amplitude/frequency modulation may be handled in a way similar to onset synchrony. Grouping of multiple frequency partials that form harmonics of a fundamental frequency may be incorporated based on two cues. The first is onset/offset synchrony among these partials. The second is spectral relations among the partials. There are many models for pitch perception on the basis of harmonic relations (Moore, 1989). Most of them are not built on neurobiology. A topic of future research is how to combine the detection of harmonic relations with the present model of auditory segregation.

Our discussion so far has been concerned with only primitive auditory segregation. A good deal of psychological evidence suggests that segregation is also influenced by prior knowledge possessed by an individual (Bregman, 1990). Thus, knowledge effects also must be considered in future research. We previously proposed a model of oscillatory associative memory (Wang, Buhmann, & von der Malsburg, 1990). This associative memory model can separate an input that is a mixture of multiple patterns, based on the patterns stored in the memory. In the future, these two types of segregation must be integrated into a coherent model. Such a model should be able to address the data on melody segregation observed by Dowling (1973) and Dowling et al. (1987), which showed the influence of previously acquired melodies. The model should also be able to simulate the experiments of Tougas and Bregman (1985), which involve extracting a pattern held in memory from a mixture of tones. Their experiments tested participants' ability to follow a tone sequence, and they found that grouping based on frequency proximity dominates over grouping based on a smooth trajectory.

The shifting synchronization theory proposes a novel approach for tackling automatic auditory scene analysis, which is largely an unsolved engineering problem. Automatic auditory segregation is a critical part of auditory signal processing, real-time speech recognition, and music transcription in natural environments. Compared to existing computer algorithms for auditory input separation (Brown & Cooke, 1994; Mellinger, 1992; Parsons, 1976;

Weintraub, 1986), the oscillatory correlation approach offers many unique advantages. Due to oscillatory dynamics, no single stream can dominate and suppress the perception of the rest of the auditory scene for a long time. The processing is inherently parallel. The organizational simplicity renders the segregation network particularly feasible for VLSI implementation. Also, rapid continuous time dynamics allows real-time processing.

CONCLUSION

Auditory segregation is critical for complex auditory pattern processing. This article presents a novel neurocomputational theory, namely the shifting synchronization theory, of how primitive auditory segregation might be achieved in the brain. The architecture of the model is a laterally coupled two-dimensional network of neural oscillators with a global inhibitor, one dimension representing time and the other representing frequency. Computer simulation of the segregation network exhibits a set of psychological phenomena of primitive auditory segregation, including stream segregation. We have argued that the model is neurally plausible and may provide an effective computational approach to automatic auditory segregation.

REFERENCES

Abeles, M. (1982). *Local cortical circuits*. New York: Springer.

Anderson, C.H., & van Essen, D.C. (1987). Shifter circuits: A computational strategy for dynamic aspects of visual processing. *Proceedings of the National Academy of Sciences of USA, 84,* 6297-6301.

Arbib, M.A. (1989). *The metaphorical brain 2: Neural networks and beyond.* New York: Wiley Interscience.

Beauvois, M.W., & Meddis, R. (1991). A computer model of auditory stream segregation. *Quarterly Journal of Experimental Psychology, 43 A,* 517-541.

Blauert, J. (1983). *Spatial hearing: The psychophysics of human sound localization* (J.S. Allen, Trans). Cambridge, MA: MIT Press.

Bourlard, H.A., & Morgan, N. (1994). *Connectionist speech recognition: A hybrid approach.* Norwell, MA: Kluwer Academic.

Bregman, A.S. (1978). Auditory streaming: Competition among alternative organizations. *Perception and Psychophysics, 23,* 391-398.

Bregman, A.S. (1990). *Auditory scene analysis.* Cambridge, MA: MIT Press.

Bregman, A.S. (1993). Auditory scene analysis: Hearing in complex environments. In S. McAdams & E. Bigand (Eds.), *Thinking in sound.* Oxford, England: Clarendon.

Bregman, A.S., Abramson, J., Doehring, P., & Darwing, C.J. (1985). Spectral integration based on common amplitude modulation. *Perception & Psychophysics, 37,* 483-493.

Bregman, A.S., & Campbell, J. (1971). Primary auditory stream segregation and perception of order in rapid sequences of tones. *Journal of Experimental Psychology, 89,* 244-249.

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

Brown, G., & Cooke, M. (1994). Computational auditory scene analysis. *Computer Speech and Language, 8,* 297-336.

Buhmann, J. (1989). Oscillations and low firing rates in associative memory neural networks. *Physics Review A, 40,* 4145–4148.

Buzsáki, G., Llinás, R., Singer, W., Berthoz, A., & Christen, Y. (Eds.) (1994). *Temporal coding in the brain*. Berlin: Springer-Verlag.

Cherry, E.C. (1953). Some experiments on the recognition of speech, with one and with two ears. *Journal of the Acoustical Society of America, 25,* 975–979.

Covey, E., Hawkins, H., McMullen, T., & Port, R. (Eds. in press). *Neural representation of temporal Patterns*. New York: Plenum.

Crick, F. (1984). Function of the thalamic reticular complex: The searchlight hypothesis. *Proceedings of the National Academy of Sciences of USA, 81,* 4586–4590.

Crick, F. (1994). *The astonishing hypothesis*. New York: Scribner.

Dear, S.P., Simmons, J.A., & Fritz, J. (1993). A possible neuronal basis for representation of acoustic scenes in auditory cortex of the big brown bat. *Nature, 364,* 620–623.

Dowling, W.J. (1973). The perception of interleaved melodies. *Cognitive Psychology, 5,* 322–337.

Dowling, W.J., Lund, K.M-T., & Herrbold, S. (1987). Aiming attention in pitch and time in the perception of interleaved melodies. *Perception & Psychophysics, 41,* 642–656.

Eckhorn, R., Bauer, R., Jordan, W., Brosch, M., Kruse, W., Munk, M., & Reitboeck, H.J. (1988). Coherent oscillations: A mechanism of feature linking in the visual cortex? *Biological Cybernetics, 60,* 121–130.

Engel, A.K., König, P., Kreiter, A.K., & Singer, W. (1991a). Interhemispheric synchronization of oscillatory neuronal responses in cat visual cortex. *Science, 252,* 1177–1179.

Engle, A.K., König, P., Kreiter, A.K., & Singer, W. (1991b). Synchronization of oscillatory neuronal responses between striate and extrastriate visual cortical areas of the cat. *Proceedings of the National Academy of Sciences of USA, 88,* 6048–6052.

FitzHugh, R. (1961). Impulses and physiological states in models of nerve membrane. *Biophysical Journal, 1,* 445–466.

Freeman, W.J. (1991). Nonlinear dynamics in olfactory information processing. In J.L. Davis & H. Eichenbaum (Eds.), *Olfaction*. Cambridge, MA: MIT Press.

Galambos, R., Makeig, S., & Talmachoff, P.J. (1981). A 40-Hz auditory potential recorded from the human scalp. *Proceedings of the National Academy of Sciences of USA, 78,* 2643–2647.

Goodhill, G.J., & Barrow, H.G. (1994). The role of weight normalization in competitive learning. *Neural Computation, 6,* 255–269.

Gottlieb, Y., Vaadia, E., & Abeles, M. (1989). Single unit activity in the auditory cortex of a monkey performing a short term memory task. *Experimental Brain Research, 74,* 139–148.

Gray, C.M., König, P., Engel, A.K., & Singer, W. (1989). Oscillatory responses in cat visual cortex exhibit inter-columnar synchronization which reflects global stimulus properties. *Nature, 338,* 334–337.

Guckenheimer, J., & Holmes, P. (1983). *Nonlinear oscillations, dynamical systems and bifurcations of vector fields*. New York: Springer-Verlag.

Handel, S. (1989). *Listening: An introduction to the perception of auditory events*. Cambridge, MA: MIT Press.

Hartmann, W.M. (1988). Pitch perception and the segregation and integration of auditory entities. In G.M. Edelman, W.E. Gall, & W.M. Cowan (Eds.), *Auditory function: Neurobiological bases of hearing*. New York: Wiley.

Hebb, D.O. (1949). *The organization of behavior*. New York: Wiley.

Helmholtz, H. (1954). *On the sensation of tone* (A.J. Ellis, Trans.). Braunschweig: Vieweg & Son. (Original work published 1863)

Hocherman, S., & Gilat, E. (1981). Dependence of auditory cortex evoked unit activity on interstimulus interval in the cat. *Journal of Neurophysiology, 45,* 987–999.

Hopfield, J.J., & Tank, D.W. (1989). Neural architecture and biophysics for sequence recognition. In J.H. Byrne & W.O. Berry (Eds.), *Neural models of plasticity*. San Diego, CA: Academic Press.

Hubel, D.H. (1988). *Eye, brain, and vision*. New York: Freeman.

Idson, W.L., & Massaro, D.W. (1976). Cross-octave masking of single tones and musical sequences: The effects of structure on auditory recognition. *Perception & Psychophysics, 19,* 155–175.

Jones, M.R. (1976). Time, our lost dimension: Toward a new theory of perception, attention, and memory. *Psychological Review, 83,* 323–355.

Jones, M.R., Jagacinski, R.J., Yee, W., Floyd, R.L., & Klapp, S.T. (1995). Tests of attentional flexibility in listening to polyrhythmic patterns. *Journal of Experimental Psychology: Human Perception and Performance, 21,* 293–307.

Jones, M.R., Kidd, G., & Wetzel, R. (1981). Evidence for rhythmic attention. *Journal of Experimental Psychology: Human Perception and Performance, 7,* 1059–1073.

Jones, M.R., & Yee, W. (1993). Attending to auditory events: The role of temporal organization. In S. McAdams & E. Bigand (Eds.), *Thinking in sound*. Oxford, England: Clarendon.

Kammen, D.M., Holmes, P.J., & Koch, C. (1989). Origin of oscillations in visual cortex: Feedback versus local coupling. In R.M.J. Cotterill (Ed.), *Models of brain functions*. Cambridge, England: Cambridge University Press.

Kandel, E.R., Schwartz, J.H., & Jessell, T.M. (1991). *Principles of neural science* (3rd ed.). New York: Elsevier.

Koffka, K. (1935). *Principles of Gestalt psychology*. New York: Harcourt.

Lamme, V.A.F., van Dijk, B.W., & Spekreijse, H. (1993). Coutour from motion processing occurs in primary visual cortex. *Nature, 363,* 541–543.

Lazzaro, J., & Mead, C. (1989). Silicon modeling of pitch perception. *Proceedings of the National Academy of Sciences of USA, 86,* 9597–9601.

Llinás, R., & Ribary, U. (1993). Coherent 40-Hz oscillation characterizes dream state in humans. *Proceedings of the National Academy of Sciences of USA, 90,* 2078–2082.

Madler, C., & Pöppel, E. (1987). Auditory evoked potentials indicate the loss of neuronal oscillations during general anesthesia. *Naturwissenschaften, 74,* 42–43.

Mäkelä, J.P., & Hari, R. (1987). Evidence for cortical origin of the 40 Hz auditory evoked response in man. *Electroencephalographical and Clinical Neurophysiology, 66,* 539–546.

McKenna, T.M., Weinberger, N.M., & Diamond, D.M. (1989). Responses of single auditory cortical neurons to tone sequences. *Brain Research, 481,* 142–153.

McNally, K.A., & Handel, S. (1977). Effect of element composition on streaming and the ordering of repeating sequences. *Journal of Experimental Psychology: Human Perception and Performance, 3,* 451–460.

Mellinger, D.K. (1992). *Event formation and separation in musical sound*. Unpublished doctoral dissertation, Department of Computer Science, Stanford University, Stanford, CA.

Miller, G.A., & Heise, G.A. (1950). The trill threshold. *Journal of Acoustical Society of America, 22,* 637–638.

Milner, P.M. (1974). A model for visual shape recognition. *Psychological Review, 81,* 521–535.

Moore, B.C.J. (1989). *An introduction to the psychology of hearing*. San Diego, CA: Academic Press.

Morris, C., & Lecar, H. (1981). Voltage oscillations in the barnacle giant muscle fiber. *Biophysical Journal, 35,* 193–213.

Murthy, V.N., & Fetz, E.E. (1992). Coherent 25- to 35-Hz oscillations in the sensorimotor cortex of awake behaving monkeys. *Proceedings of the National Academy of Sciences of USA, 89,* 5670–5674.

Nagumo, J., Arimoto, S., & Yoshizawa, S. (1962). An active pulse transmission line simulating nerve axon. *Proceedings of the Institute of Radio Engineers, 50,* 2061–2070.

Neti, C., & Young, E.D. (1992). Neural network models of sound localization based on directional filtering by the pinna. *Journal of Acoustical Society of America, 92,* 3140–3156.

Parsons, T.W. (1976). Separation of speech from interfering speech by means of harmonic selection. *Journal of Acoustical Society of America, 60,* 911–918.

Pickles, J.O. (1988). *An introduction to the physiology of hearing* (2nd ed.). London: Academic Press.

Popper, A.N., & Fay, R.R. (Eds.) (1992). *The mammalian auditory pathway: Neurophysiology*. New York: Springer-Verlag.

Rabiner, L.R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the Institute of Electrical and Electronics Engineers, 77,* 257–286.

Rasch, R.A. (1978). The perception of simultaneous notes such as in polyphonic music. *Acustica, 40,* 22–33.

Ribary, U., Ioannides, A.A., Singh, K.D., Hasson, R., Bolton, J.P.R., Lado, F., Mogilner, A., & Llinás, R. (1991). Magnetic field tomography of coherent thalamocortical 40-Hz oscillations in humans. *Proceedings of the National Academy of Sciences of USA, 88,* 11037–11041.

Rock, I., & Palmer, S. (1990, December). The legacy of Gestalt psychology. *Scientific American, 263,* 84–90.

Singer, W. (1993). Synchronization of cortical activity and its putative role in information processing and learning. *Annual Review of Physiology, 55,* 349–374.

Singer, W., & Gray, C.M. (1995). Visual feature integration and the temporal correlation hypothesis. *Annual Review of Neuroscience, 18,* 555–586.

Smith, L.S. (1994). Sound segmentation using onsets and offsets. *Journal of New Music Research, 23,* 11–23.

Somers, D., & Kopell, N. (1993). Rapid synchronization through fast threshold modulation. *Biological Cybernetics, 68,* 393–407.

Sompolinsky, H., Golomb, D., & Kleinfeld, D. (1991). Cooperative dynamics in visual processing. *Physics Review A, 43,* 6990–7011.

Sporns, O., Gally, J.A., Reeke, G.N., Jr., & Edelman, G.M. (1989). Reentrant signaling among simulated neuronal groups leads to coherency in their oscillatory activity. *Proceedings of the National Academy of Sciences of USA, 86,* 7265–7269.

Steiger, H., & Bregman, A.S. (1981). Capturing frequency components of glided tones: Frequency separation, orientation and alignment. *Perception & Psychophysics, 30,* 425–435.

Stoner, G.R., & Albright, T.D. (1992). Neural correlates of perceptual motion coherence. *Nature, 358,* 412–414.

Tank, D.W., & Hopfield, J.J. (1987). Neural computation by concentrating information in time. *Proceedings of the National Academy of Sciences of USA, 84,* 1896–1900.

Terman, D., & Wang, D.L. (1995). Global competition and local cooperation in a network of neural oscillators. *Physica D, 81,* 148–176.

Tougas, Y., & Bregman, A.S. (1985). Crossing of auditory streams. *Journal of Experimental Psychology: Human Perception and Performance, 11,* 788–798.

van Noorden, L.P.A.S. (1975). *Temporal coherence in the perception of tone sequences.* unpublished doctoral dissertation, The Institute of Perception Research, Eindhoven, The Netherlands.

Verhulst, F. (1990). *Nonlinear differential equations and dynamical systems.* Berlin: Springer-Verlag.

von der Malsburg, C. (1973). Self-organization of orientation sensitive cells in the striate cortex. *Kybernetik, 14,* 85–100.

von der Malsburg, C. (1981). *The correlation theory of brain function* (Internal Report No. 81-2). Max-Planck-Institut for Biophysical Chemistry, Göttingen, Germany.

von der Malsburg, C., & Schneider, W. (1986). A neural cocktail-party processor. *Biological Cybernetics, 54,* 29–40.

Waibel, A., Hanazawa, T., Hinton, G.E., Shikano, K., & Lang, K.J. (1989). Phoneme recognition using time-delay neural networks. *IEEE Transactions on Acoustics, Speech, and Signal Processing, 37,* 328–339.

Wang, D.L. (1993). Modeling global synchrony in the visual cortex by locally coupled neural oscillators. In *Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.

Wang, D.L. (1994). Auditory stream segregation based on oscillatory correlation. In *Proceedings of the IEEE 1994 Workshop on Neural Networks for Signal Processing*. Piscataway, NJ: IEEE.

Wang, D.L. (1995a). Emergent synchrony in locally coupled neural oscillators. *IEEE Transactions on Neural Networks, 6,* 941–948.

Wang, D.L. (1995b). Temporal pattern processing. In M.A. Arbib (Ed.), *Handbook of brain theory and neural networks*. Cambridge, MA: MIT Press.

Wang, D.L. (in press). An oscillatory correlation theory of temporal pattern segmentation. In E. Covey, H. Hawkins, T. McMullen, & R. Port, (Eds.), *Neural representation of temporal Patterns*. New York: Plenum.

Wang, D.L., & Arbib, M.A. (1990). Complex temporal sequence learning based on short-term memory. *Proceedings of the Institute of Electrical and Electronics Engineers, 78,* 1536–1543.

Wang, D.L., Buhmann, J., & von der Malsburg, C. (1990). Pattern segmentation in associative memory. *Neural Computation, 2,* 95–107.

Wang, D.L., & Terman, D. (1995). Locally excitatory globally inhibitory oscillator networks. *IEEE Transactions on Neural Networks, 6,* 283–286.

Warren, R.M., Obusek C.J., Farmer, R.M., & Warren, R.P. (1969). Auditory sequence: Confusion of patterns other than speech or music. *Science, 164,* 586–587.

Weintraub, M. (1986). A computational model for separating two simultaneous talkers. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*. Tokyo: IEEE.

Winer, J.A. (1992). The functional architecture of the medial geniculate body and the primary auditory cortex. In D.B. Webster, A.N. Popper, & R.R. Fay (Eds.), *The mammalian auditory pathway: Neuroanatomy*. New York: Springer-Verlag.

Zakarauskas, P., & Cynader, M.S. (1993). A computational theory of spectral cue localization. *Journal of Acoustical Society of America, 94,* 1323–1331.