Available online at www.sciencedirect.com
SCIENCE DIRECT®
Neural Networks 17 (2004) 511–536
Neural Networks
www.elsevier.com/locate/neunet

ARTSTREAM: a neural network model of auditory scene analysis and source segregation

Stephen Grossberg[a,*], Krishna K. Govindarajan[b], Lonce L. Wyse[c], Michael A. Cohen[a]

[a]*Department of Cognitive and Neural Systems, Center for Adaptive Systems, Boston University, 677 Beacon Street, Boston, MA 02215, USA*
[b]*SpeechWorks International, 695 Atlantic Avenue, Boston, MA 02111, USA*
[c]*Laboratories for Information Technology, 21 Heng Mui Keng Terrace, Kent Ridge, Singapore, Singapore 119613*

Received 3 June 2003; accepted 8 October 2003

Abstract

Multiple sound sources often contain harmonics that overlap and may be degraded by environmental noise. The auditory system is capable of teasing apart these sources into distinct mental objects, or streams. Such an 'auditory scene analysis' enables the brain to solve the cocktail party problem. A neural network model of auditory scene analysis, called the ARTSTREAM model, is presented to propose how the brain accomplishes this feat. The model clarifies how the frequency components that correspond to a given acoustic source may be coherently grouped together into a distinct stream based on pitch and spatial location cues. The model also clarifies how multiple streams may be distinguished and separated by the brain. Streams are formed as spectral-pitch resonances that emerge through feedback interactions between frequency-specific spectral representations of a sound source and its pitch. First, the model transforms a sound into a spatial pattern of frequency-specific activation across a spectral stream layer. The sound has multiple parallel representations at this layer. A sound's spectral representation activates a bottom-up filter that is sensitive to the harmonics of the sound's pitch. This filter activates a pitch category which, in turn, activates a top-down expectation that is also sensitive to the harmonics of the pitch. Resonance develops when the spectral and pitch representations mutually reinforce one another. Resonance provides the coherence that allows one voice or instrument to be tracked through a noisy multiple source environment. Spectral components are suppressed if they do not match harmonics of the top-down expectation that is read-out by the selected pitch, thereby allowing another stream to capture these components, as in the 'old-plus-new heuristic' of Bregman. Multiple simultaneously occurring spectral-pitch resonances can hereby emerge. These resonance and matching mechanisms are specialized versions of Adaptive Resonance Theory, or ART, which clarifies how pitch representations can self-organize during learning of harmonic bottom-up filters and top-down expectations. The model also clarifies how spatial location cues can help to disambiguate two sources with similar spectral cues. Data are simulated from psychophysical grouping experiments, such as how a tone sweeping upwards in frequency creates a bounce percept by grouping with a downward sweeping tone due to proximity in frequency, even if noise replaces the tones at their intersection point. Illusory auditory percepts are also simulated, such as the auditory continuity illusion of a tone continuing through a noise burst even if the tone is not present during the noise, and the scale illusion of Deutsch whereby downward and upward scales presented alternately to the two ears are regrouped based on frequency proximity, leading to a bounce percept. Since related sorts of resonances have been used to quantitatively simulate psychophysical data about speech perception, the model strengthens the hypothesis that ART-like mechanisms are used at multiple levels of the auditory system. Proposals for developing the model to explain more complex streaming data are also provided.

Keywords: Auditory scene analysis; Streaming; Cocktail party problem; Pitch perception; Spatial localization; Neural network; Resonance; Adaptive resonance theory; Spectral-pitch resonance

1. Introduction: cocktail party problem and auditory continuity illusion

When we talk to a friend in a crowded noisy room, we can usually keep track of our conversation above the hubbub, even though the sounds emitted by the friendly voice partially overlap the sounds emitted by other speakers and noise sources. How do we separate this jumbled mixture of sounds into distinct voices? This issue is often called the *cocktail party problem.* The same problem is solved whenever we listen to a symphony or other music wherein overlapping harmonic components are emitted by several instruments. If we could not separate the instruments or

* Corresponding author. Tel.: +1-617-353-7857; fax: +1-617-353-7755.
E-mail address: steve@bu.edu (S. Grossberg).

0893-6080/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2003.10.002

voices into distinct sources, or auditory streams, then we could not hear the music as music, or intelligently recognize a speaker's sounds. The ability to segregate these different signals has been generally termed *auditory scene analysis* (Bregman, 1990).

A simple version of this competence is illustrated by the *auditory continuity illusion* (Miller & Licklider, 1950). Suppose that a steady tone shuts off just as a broadband noise turns on. Suppose, moreover, that the noise shuts off just as the tone turns on once again; see Fig. 1a. When this happens under appropriate temporal constraints, the tone seems to continue right through the noise, which seems to occur in a separate auditory 'stream'. This example suggests that the auditory system can actively extract those components of the noise that are consistent with the tone and use them to track the 'voice' of the tone right through the hubbub of the noise.

In order to appreciate how remarkable this property is, let us compare it with what happens when the tone does not turn on again for a second time, as in Fig. 1b. Then the first tone does not seem to continue through the noise. It is perceived to stop before the noise ends. How does the brain know that the second tone will turn on after the noise shuts off, so that it can continue the tone through the noise, even though the tone is not perceived to persist through the noise if the second tone does not eventually occur? Does this not seem to require that the brain can operate 'backwards in time' to alter its decision as to whether or not to continue a past tone through the noise based on future events?

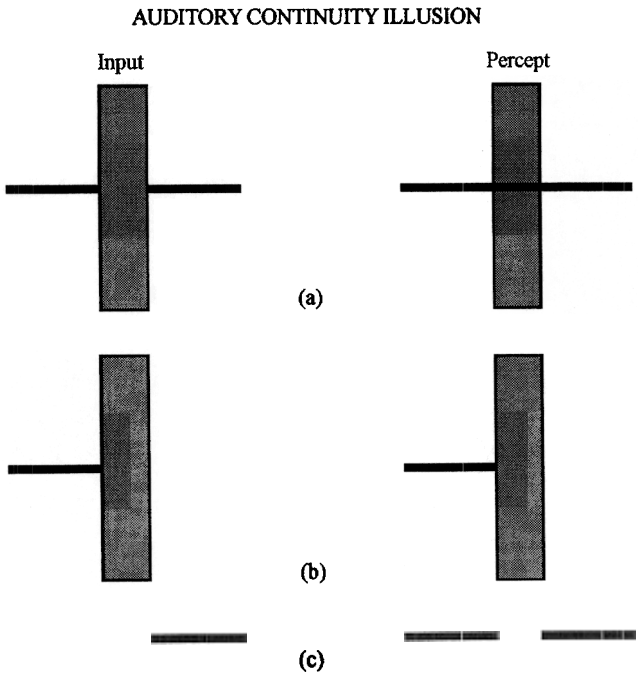

Fig. 1. (a) Auditory continuity illusion: when a steady tone occurs both before and after a burst of noise, then under appropriate temporal and amplitude conditions, the tone is perceived to continue through the noise. (b) This does not occur if the noise is not followed by a tone. (c) Nor does it occur if two tones are separated by silence.

Additional properties of this phenomenon are clarified by the third condition: If no noise occurs between two temporally disjoint tones, as in Fig. 1c, then the tone is not heard across the silent interval. Instead, two temporally disjoint tones are heard. This fact raises the additional question: how does the brain use the noise to continue the tone through it?

Many philosophers and scientists have puzzled about this sort of problem. This article clarifies how the process whereby we consciously hear the first tone takes some time to unfold, so that by the time we hear it, the second tone has an opportunity to influence it. To make this argument, we need to ask: Why does conscious audition take so long to occur after the actual sound energy reaches our brain? Just as important: why can the second tone influence the conscious percept so quickly, given that the first tone could not?

An analysis of the mechanisms of auditory scene analysis is important for understanding how the human auditory perceptual system operates, as well as for technological applications. While speech recognition systems have improved greatly within the last decade, they are still prone to noise and interference from other speakers.

1.1. Auditory scene analysis

The nomenclature associated with auditory scene analysis contains several keywords: source, stream, grouping and stream segregation. The source is a physical, external entity which produces sound; e.g. a speaker. The perceptual correlate of this source is a stream; i.e. it is what the brain takes to be a single sound. The stream is created by the perceptual grouping and segregation of acoustic properties that are thought to correspond to an acoustic object. Grouping and stream segregation, or streaming, assign appropriate combinations of frequency components to a stream through time. For an exhaustive review of auditory scene analysis, the reader is referred to Bregman (1990).

The scene analysis process can be thought of as two processes that interact: a simultaneous grouping process and a sequential grouping process. For example, in Fig. 2, the simultaneous grouping process tries to group B and C together if they have synchronous onsets and offsets, or if they are harmonically related. Similarly, the sequential grouping process tries to group A and B together based on their frequency and temporal proximity.

1.2. Grouping principles

In order to denote which acoustic attributes correspond to a stream, researchers, including Gestalt scientists and, more recently, Bregman (1990) and his colleagues, have suggested several grouping principles.

Proximity. The proximity grouping principle is shown in Fig. 2. If two tones are closer together in frequency and time, then it is more likely that they should be grouped

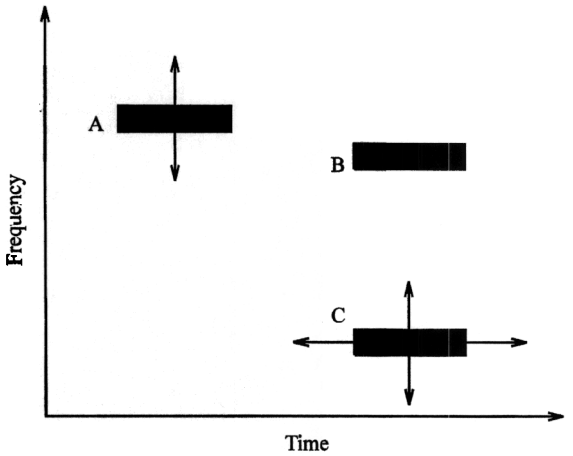

Fig. 2. A groups better with B if they are closer in frequency. However, simultaneous cues, such as common onsets, common offsets and harmonicity, can help group B and C. [Adapted with permission from Bregman and Pinker (1978).]

together, e.g. A and B should be grouped together if they are close enough.

Closure and belongingness. Closure and belongingness lead to percepts of continuity and completion. Closure is the perceptual phenomenon of completing streams when there is evidence for it. For example, listeners may hear a tone continuing through noise under certain conditions, even though the tone is not present during the noise, as in the auditory continuity illusion of Fig. 1a. Thus, the perceptual system completes the tone across the noise, given the evidence that the same frequency tone is present on either side of the noise.

Good continuation. Good continuation states that an object's sound does not make rapid jumps, but instead continues smoothly. For example, in Fig. 1a, the slope of the tone is the same on either side of the noise, and thus should be grouped together due to good continuity of the tone. However, if the post-noise tone was at a distant frequency, then the tone would not have good continuity and would not stream across the noise. Note that continuity is closely related to proximity.

Common fate. Common fate states that those attributes which are going through similar manifestations should be grouped together. For example, those frequency components which originate from the same spatial location share the same 'fate', and therefore, should correspond to the same object. Similarly, those frequency components which are being modulated (frequency or amplitude) at the same rate or have synchronous onsets and offsets should correspond to an object.

Principle of exclusive allocation. This principle states that attributes are assigned to one stream or another, but not both. While this principle seems to hold in sequential streaming, it can fail in simultaneous streaming, where harmonics of two streams can overlap.

1.3. Primitive versus schema-based segregation

Bregman (1990) noted that auditory stream segregation consists of a primitive, nonattentive, unlearned process and a schema-based, attentive, learned process. Bregman and Rudnicky (1975) found that tones in an unattended stream can capture tones from an attended stream. In addition, van Noorden (1975) presented a repetition of two alternating tones whose frequency and temporal spacing were manipulated to subjects. van Noorden obtained two curves: the temporal coherence boundary (TCB) and the fission boundary (FB). The TCB corresponds to the boundary where the frequency separation between the temporally adjacent tones was too large to hear one stream. The FB corresponds to the point where the two frequencies were too close in frequency to be heard as separate streams. The FB varied little as a function of the tone repetition rate, and was mainly a function of the frequency separation. On the other hand, the TCB showed that as the frequency separation between the tones increased, one needed to slow down the repetition rate in order to maintain one stream with both tones. Bregman (1990) argued that the FB corresponds to an attentional mechanism and the TCB corresponds to a nonattentional mechanism, and noted that the schema-based mechanisms can override the primitive mechanisms. The mechanism proposed here addresses the preattentive, primitive segregation mechanisms, but also proposes how automatic attentional mechanisms help to determine perceived streams.

2. Grouping cues

One can find acoustic attributes that correspond to the grouping principles. The attributes include temporal and frequency separation, harmonicity, spatial location, amplitude modulation, frequency modulation, and onsets and offsets.

2.1. Temporal and frequency separation

Bregman and Pinker (1978) showed that tones in a repeating sequence tend to group if they are closer in frequency, e.g. A and B in Fig. 2. In addition, faster presentation rates of alternating high and low frequency tones causes the two tones to be segregated into two streams (Bregman & Campbell, 1971). The effect of faster presentation rates is to narrow the temporal separation between adjacent instances of the high tone (and low tone), allowing the tones in each frequency region to form a separate stream. The Bregman and Rudnicky (1975) stimuli, which are shown in Fig. 3, show how tones that are part of one stream can be captured into a different stream by adding additional tones that are close in frequency. When A and B were presented by themselves, listeners could easily judge their temporal order. When A and B were flanked by tones F, listeners had a more difficult time. However, if the captor tones C surrounded the flankers, then F streamed with C,

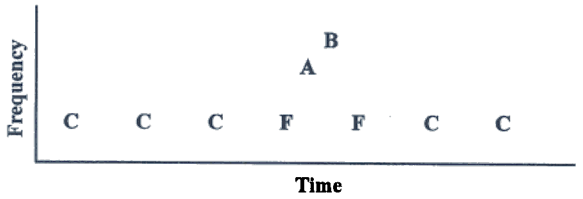

Fig. 3. When A and B are presented by themselves, listeners could easily judge the order of them. If A and B were flanked by tones F, then listeners had a more difficult time. However, if the captor tones C surrounded the flankers, then F streamed with C, leaving A–B to a different stream, allowing the listeners to hear the order once again. [Adapted with permission from Bregman and Rudnicky (1975).]

A–B split into a different stream, and the listeners could again hear the order of A–B. Thus, if A and B are in the middle of a stream, their order is more difficult to determine.

2.2. Continuity illusion

As mentioned above, proximity combined with closure leads to the auditory continuity illusion. In the continuity illusion, sound A seems to continue through sound B, even though sound A is not present during sound B. This illusion works for both tones and glides that are interrupted by brief bursts of noise.

An example involving glides is shown in Fig. 4. The top two figures show the two different stimuli that Steiger (1980) presented to listeners. In (b), broadband noise replaced the glide portion. However, for both the stimuli in (a) and (b), listeners heard the two streams shown in (c) and (d). Thus, in (b), the glide complex is completed, or continued, through the noise. Also in (b), a third stream is heard corresponding to the broadband noise bursts. This

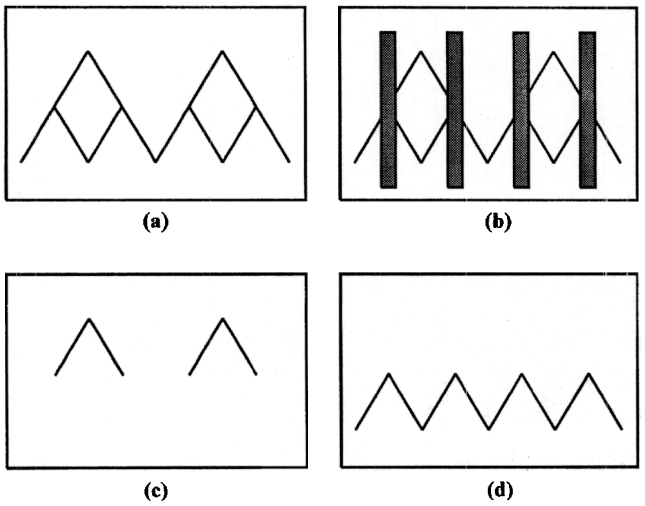

Fig. 4. Stimuli and percept of the experiment by Steiger (1980). (a) and (b) show the stimuli that were presented to the subjects. In (b), the noise is not added to the glides, but actually replaces the glide portions. For both the stimuli in (a) and (b), listeners hear the two streams shown in (c) and (d). In (b), a third stream is heard corresponding to the broadband noise bursts. [Adapted with permission from Steiger (1980).]

experiment is important because, in it, the principle of good continuation is overcome by frequency proximity.

2.3. Harmonicity and pitch

Periodic sources typically have frequency components, called harmonics, at integer multiples of the fundamental frequency, F_0. The subjective experience of F_0 is denoted as pitch, and is influenced by the harmonic content and other attributes of the signal. Consider a speaker producing a vowel at a particular fundamental frequency; e.g. 150 Hz. The vowel contains harmonics at integer multiples; e.g. 300, 450, 600, etc. and the relative amplitudes of these harmonics lead to a given vowel percept. Since a set of related harmonics will correspond to the same source, the pitch can be used to group these harmonic components.

A harmonic of a complex tone can be heard separate from the tone if it is mistuned by 1.5–3%, as well as causing the complex pitch to shift. If the mistuning is greater than 3%, the harmonic has little effect on the pitch, and is still heard as a second source (Moore, Glasberg, & Peters, 1985). Also, lower harmonics are easier to hear separately from a complex than higher harmonics, and harmonics are easier to capture out of a complex if the neighboring harmonics are removed (van Noorden, 1975). Partials spaced 14 semitones apart fuse better than ones that 16 semitones apart (Bregman, 1990). A semitone is the smallest pitch interval in Western music, and two tones separated by a semitone corresponds to tones at frequencies f and $(1.06)f$. These effects may be related to the resolution of the harmonics within the auditory channels (Cohen, Grossberg, & Wyse, 1995).

Segregation based on harmonicity is used by listeners in speech perception. It has been shown that listeners can use F_0 to segregate multiple voices. Listeners' identification of two concurrent vowels increases as the difference in the two F_0 increases, and plateaus between 0.5 and 2 semitones (Scheffers, 1983). When F_0 was an octave apart, identification is also very poor (Brokx & Noteboom, 1982; Chalika & Bregman, 1989). Since an octave corresponds to a doubling of frequency, half the harmonics for the two vowels will overlap. It should be noted that listeners can identify concurrent vowels with the same F_0 with greater than chance accuracy, implying that listeners can also use schema-based segregation. In addition, a formant (frequencies with greater energy that correspond to vowel identity) of a single vowel may become segregated when the formant has a differing F_0 under certain conditions (Broadbent & Ladefoged, 1957; Gardner, Gaskill, & Darwin, 1989). Finally, speech stimuli with discontinuous pitch contours tend to segregate at the discontinuities (Darwin & Bethell-Fox, 1977).

2.4. Bounce and cross percept in crossing glide complexes

While the harmonicity cues can cause components to group, they can also compete with frequency proximity

cues, leading to a bounce or a cross percept in the perception of crossing glides. The influence of harmonicity is seen in the experiments of Bregman and Doehring (1984), who showed that a glide can be captured into a stream if two partials form a harmonic frame around the glide. While harmonicity can cause streaming, glides which cross sometimes produce a bounce percept, presumably due to frequency proximity at the crossing point (Halpern, 1977; Tougas & Bregman, 1990). A bounce percept corresponds to hearing two streams, one with a ' ∪ ' shaped percept and another with a ' ∩ ' shaped percept, due to the crossing of glides. The cross percept corresponds to hearing two streams, each stream containing one of the glides. Halpern (1977) presented the six different one second glide stimuli shown in Fig. 5 to subjects and asked them to rate how well they produced a bounce percept. The numbers below each figure corresponds to the preference of hearing a bounce or a cross: numbers greater than 2.5 correspond to a bounce percept, and numbers below 2.5 correspond to a cross percept. The numbers next to the glides correspond to the harmonic number of an underlying F_0. The stimuli in (a) and (d) produced a bounce percept, while the others produced a cross percept. This experiment shows that the harmonic structure in (b) and (c) help to overcome the ambiguity at the crossing point that occurs in (a) and promotes a cross percept.

Tougas and Bregman (1990) performed an experiment very similar to that of Halpern. Tougas and Bregman had four different harmonic stimuli: rich crossing, rich bouncing, all pure, and all rich (Fig. 6). All but the rich crossing condition produced a bounce percept, even when the interval I was filled with silence, noise, or just the glides. The bounce percept was greatest for rich bouncing, then all pure, and then all rich, for all three-interval conditions. An implication of this experiment is that regardless of noise, silence, or glide during the crossing point, one gets the same percept.

2.5. Spatial location

While spatial location seems to be a strong principle for grouping, the auditory system does not treat it as a dominant cue. The principle that frequency components arising from the same spatial location should belong to the same object seems reasonable, but the pliable nature of sound confounds the unambiguous implementation of this idea. Since sounds can travel around objects or corners, one object's sound can

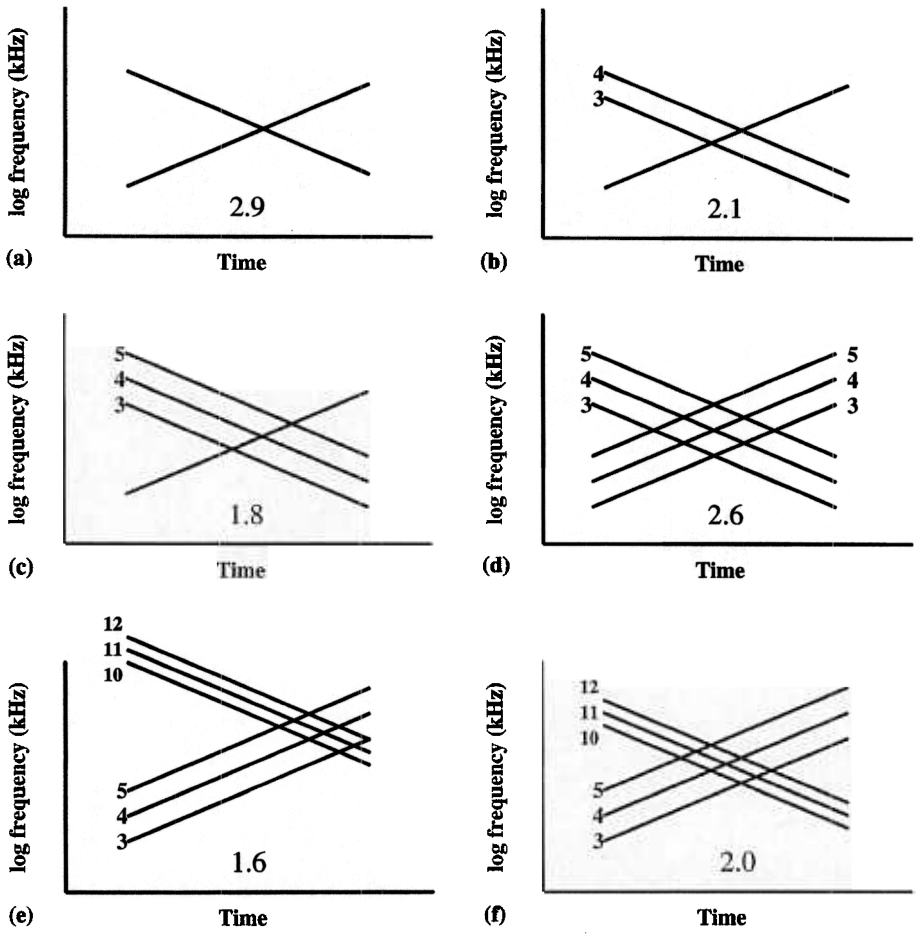

Fig. 5. Stimuli and listeners' responses in Halpern (1977) for different harmonic conditions. The complex glides were all 1 second long, and the numbers next to a glide is its harmonic number. The numbers below each figure corresponds to the preference of hearing a bounce or a cross: numbers greater than 2.5 correspond to a bounce percept, and numbers below 2.5 correspond to a cross percept. [Adapted with permission from Halpern (1977).]

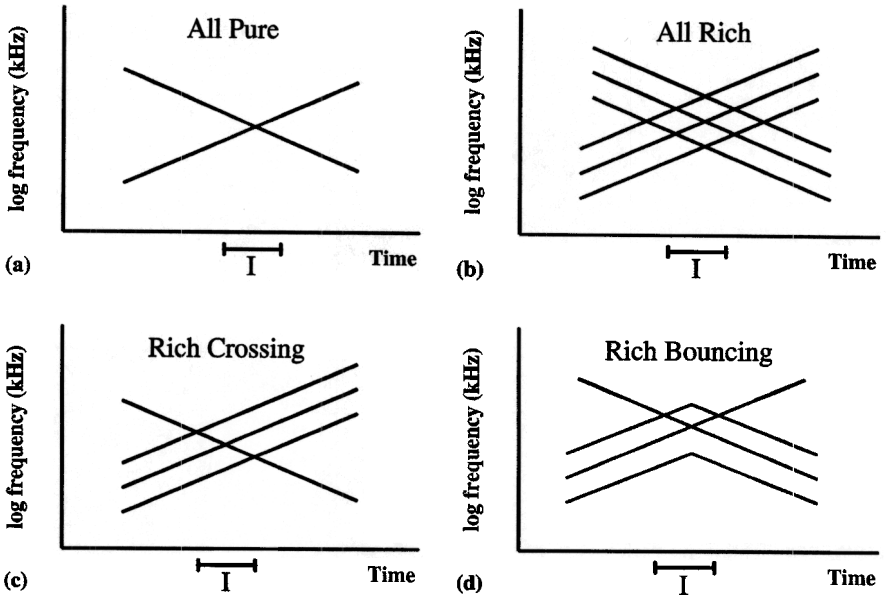

Fig. 6. Stimuli of Tougas and Bregman (1990) for four different harmonic conditions. All but the rich crossing condition produced a bounce percept, even when the interval I was filled with silence, noise, or just the glides. The order, from greatest to the least, of bounciness was rich bouncing, all pure, and all rich. [Adapted with permission from Tougas and Bregman (1990).]

travel through another object's sound. Moreover, two sounds can arise from the same location, e.g. two talkers on a monophonic radio, which listeners can easily segregate. Thus, spatial cues alone are not sufficient to separate streams. Shackleton, Meddis, and Hewitt (1994) presented two different concurrent vowels to listeners and varied the spatial and pitch separation of the two vowels. They found no improvement in identification of both vowels by introducing a spatial difference, while keeping the pitch the same for both vowels. However, by introducing a pitch difference and no spatial cue, performance improved by 35.8%. With both a pitch difference and a spatial difference, the performance improved by 45.5%.

Grouping can also affect perceived location. If a tone located in the medial plane is captured by a left ear tone (due to frequency proximity), as opposed to a right ear tone, then the central tone will be perceived to come from the left side (Bregman & Steiger, 1980). The scale illusion of Deutsch (1975) also illustrates this point (Fig. 7a). In this illusion, a downward and an upward scale are played at the same time, except that every other tone in a given scale is presented to the opposite ear. In the figure, the ear presentation is shown as an L or R for left and right ear. The result is that listeners grouped the sounds based on frequency proximity, and heard the two streams A and B shown in Fig. 7b. In addition, right-handed listeners stated that they heard the higher tones (A) in the right ear, and the lower tones (B) in the left ear.

Overall, it seems that spatial cues are secondary cues, and the perceptual system relies more on harmonicity and proximity cues. Section 6 describes how the model integrates both pitch and spatial position cues to offer an explanation of the scale illusion.

2.6. Amplitude modulation (AM)

Amplitude modulation (AM) can be a possible cue if the perceptual system groups those frequency components which have correlated amplitude fluctuations. One effect of AM is that the perception of a tone, which is masked by a noise band centered on the tone, can become easier to perceive if another band of noise is modulated with the centered noise (Hall & Grose, 1988). The release of the tone from masking is known as comodulation masking release. Despite this effect, an experiment by Summerfield and Culling (1992) showed that, at slow AM rates (2.5 Hz), segregation of two vowels did not improved due to AM. So, the influence of AM on segregation of multiple voices of seems unlikely.

2.7. Frequency modulation (FM)

Frequency modulation (FM) could act as a streaming cue if the auditory system could detect correlated frequency

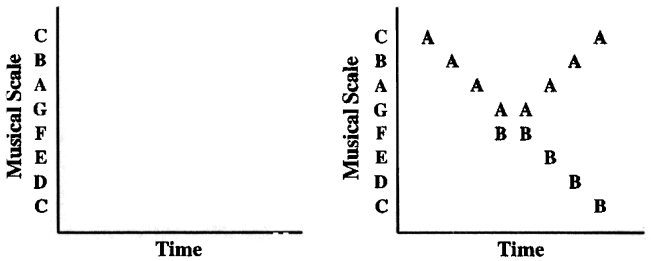

Fig. 7. (a) Scale illusion in which a downward and an upward scale are being played at the same time, except that every other tone in a given scale is presented to the opposite ear, corresponding to an L or R for left and right ear. (b) The result is that listeners group based on frequency proximity, and heard the two streams A and B. [Adapted with permission from Deutsch (1975).]

changes among spectral components. One needs to distinguish coherent FM from incoherent FM. In coherent FM, all partials (a harmonic or inharmonic component of a complex tone) are modulated at the same rate. In incoherent FM, the partials are modulated independently. Changes in F_0 correspond to coherent FM since all the harmonics are being changed by a proportionate amount. Thus, segregation based on coherent FM could be a result of changes in F_0.

Several psychophysical experiments seem to imply that segregation based on FM is not used. Carlyon (1991) found that with inharmonic complex tone pairs, listeners could not distinguish between coherent and incoherent FM, per se. Extending this, Carlyon (1992) found that if listeners did discriminate between coherent and incoherent FM, it was due to mistuning a harmonic and not to FM explicitly. Moreover, McAdams (1989) showed that by adding vibrato and jitter to different components of a three vowel mixture, the components did not segregate. Summerfield (1992) found that identification of a vowel presented with another vowel did not improve when a difference in FM was used, and all the harmonics had been randomly shifted. However, there was some benefit if the components of one vowel in a two vowel presentation was frequency modulated while the other was not (Summerfield & Culling, 1992). This result could, however, be due to pitch difference cues. Thus, for the most part, it seems that FM is not used as cue for segregation.

2.8. Onsets and offsets

Common onset and offset cause grouping, even over sequential grouping (Bregman & Pinker, 1978; Dannenbring & Bregman, 1978). Bregman and Pinker (1978) presented the stimulus shown in Fig. 1 as a repeating sequence. They found that as A and B were further separated in frequency, onset and offset synchrony grouped B and C together. However, as B and C became asynchronous, A and B grouped together to form a stream.

The interaction between harmonicity and onset asynchrony was investigated by Darwin and Ciocca (1992). They found that if a harmonic started 160 ms before rest of a complex tone, then it had a diminished influence on pitch of the complex tone. Moreover, if it started 300 ms before the complex, then it has no influence on the pitch. Finally, Bregman and Rudnicky (1975) found that two 250 ms tones that have 88% overlap fuse into one stream.

While not as strong as onset asynchrony, offset asynchrony influences grouping. A harmonic which has an offset asynchrony of 30 ms with respect to a vowel complex contributes less to its identity than one with a synchronous offset (Darwin, 1984; Darwin & Sutherland, 1984).

3. Existing models of segregation

Meddis and Hewitt (1992) presented a static model that segregated concurrent vowels based on pitch. The pitch was derived using an autocorrelation. However, the model did not handle temporally varying stimuli. Brown (1992) and Cooke (1991) have presented models which perform segregation of temporally varying stimuli. These models use pitch cues derived from autocorrelation methods to perform segregation. However, these models use time-frequency kernels to achieve segregation. In other words, they treat the stimuli as a static pattern, a spectrogram, and then perform dynamic programming and spatio-temporal processing, which treats time as another spatial dimension. None of these models has tried to model the process dynamically.

4. ARTSTREAM model of auditory streaming

4.1. From SPINET and ART to ARTSTREAM

The ARTSTREAM model developed in this article suggests how harmonicity and frequency proximity interact in the brain. The model, which is shown in Fig. 8, consists of several stages. The model includes a specialized filter which inputs to a network that groups frequency components based on pitch. The filter is a Spatial PItch NETwork, or SPINET model, that has been developed in order to simulate psychophysical data concerning how the brain converts sound streams into frequency spectra that activate spatial representations of pitch (Cohen et al., 1995). The grouping network is the type of circuit that arises in Adaptive Resonance Theory, or ART. ART proposes how the brain rapidly learns to recognize and categorize vast amounts of information by using learned top-down expectations and attentional focusing to help stabilize the learning process (Carpenter & Grossberg, 1991, 1993; Grossberg, 1976, 1980, 1999b). A specialized version of such an ART grouping network has been joined to a SPINET front end in the ARTSTREAM model of auditory scene analysis, in order to simulate psychophysical data concerning how the brain achieves pitch-based separation and streaming of multiple acoustic sources.

First, the SPINET model will be introduced and its operations illustrated by a simulation of pitch perception. Next, some general ART principles will be reviewed. Finally the ARTSTREAM model will be described and illustrative streaming simulations presented. In Section 8, ARTSTREAM will be compared with the Gjerdingen (1994) analysis of streaming percepts in music, which was based upon the motion perception model of Grossberg and Rudd (1989, 1992). Gjerdingen's analysis quantifies an analogy between visual motion perception and auditory streaming that several authors have noted; see Bregman (1990) for a review. Other extensions of the ARTSTREAM model will also be discussed.

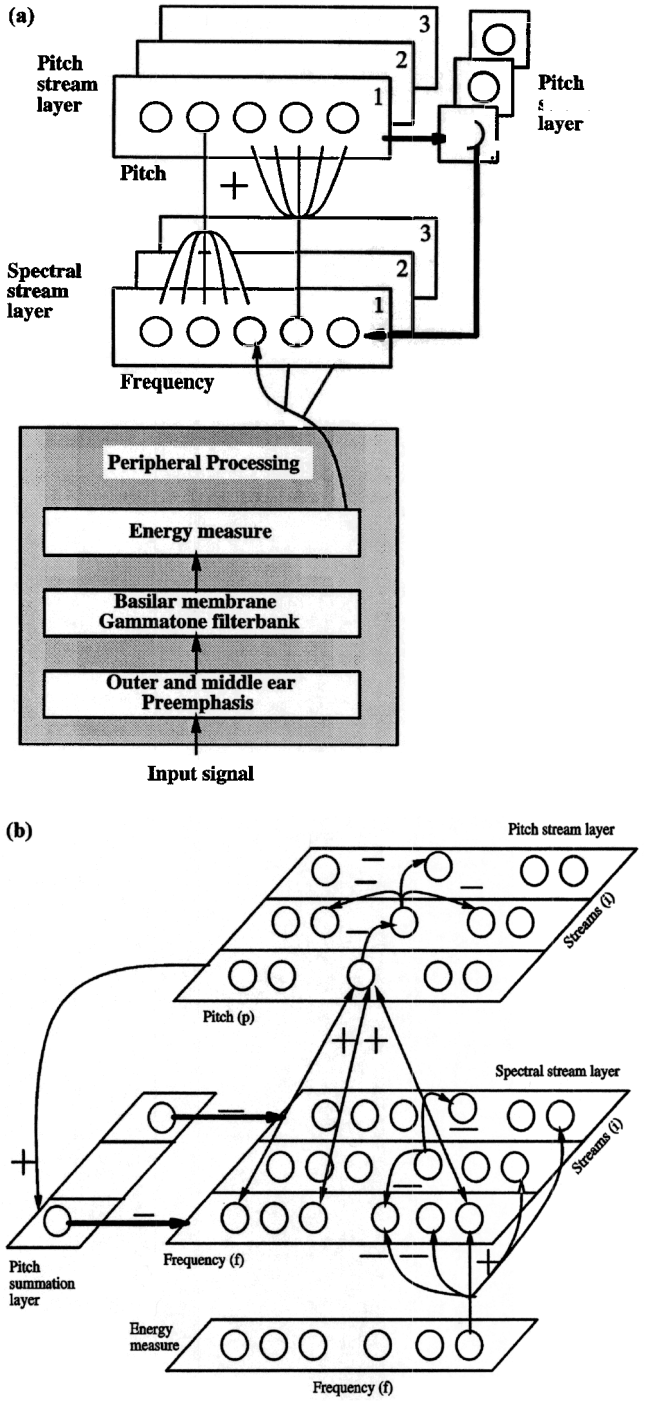

Fig. 8. (a) Block diagram of the ARTSTREAM auditory streaming model. See text for further details. (b) Interaction between the energy measure, the spectral stream layer, the pitch stream layer, and the pitch summation layer. The energy measure layer is fed forward in a frequency-specific one-to-many manner to each frequency-specific stream node in the spectral stream layer. This feed-forward activation is contrast-enhanced. Competition occurs within the spectral stream layer across streams for each frequency so that a component is allocated to only one stream at a time. Each stream in the spectral stream layer activates its corresponding pitch stream in the pitch stream layer. Each pitch neuron receives excitation from its harmonics in the corresponding spectral stream. Since each pitch stream is a winner-take-all network, only one pitch can be active at any given time. Across streams in the pitch stream layer, asymmetric competition occurs for each pitch so that one stream is biased to win and the same pitch cannot be represented in another stream. The winning pitch neuron feeds back excitation to its harmonics in the corresponding spectral stream. The stream also receives nonspecific inhibition from the pitch summation layer, which sums up the activity at the pitch stream layer for that stream. This nonspecific inhibition helps to suppress those components that are not supported by the top-down excitation, which plays the role of a priming stimulus or expectation. [Reprinted with permission from Grossberg (1999b).]

4.2. The SPINET model

The SPINET model (Cohen et al., 1995) was developed in order to neurally instantiate ideas from the spectral pitch modeling literature and join them to neural network signal processing designs to simulate a broader range of perceptual pitch data than previous spectral models. A key goal of SPINET is to transform a spectral representation of an acoustic source into a spatial distribution of pitch strengths that could be incorporated into a larger network architecture, such as ARTSTREAM, for separating multiple sound sources in the environment. The first several stages of SPINET are based on a model of the physiology and psychophysics of the auditory periphery (Cohen et al., 1995). The peripheral processing preemphasizes the signal, or boosts the amplitude of higher frequencies, which emulates the outer and middle ears. Next, the preemphasized signal is filtered by a bank of bandpass filters, which emulates the cochlea. Finally, an energy measure is obtained at the output of these filters. This energy measure inputs to a spatial representation of the frequencies in the sound. These frequencies pass through a filter to activate pitch category cells. This filter converts spectral frequency activations into pitch category activations by using a weighted *harmonic sieve* whereby the strength of activation of a given pitch category is derived from activations by a weighted sum of narrow regions around the frequency harmonics of that pitch at the spectral layer, with higher harmonics contributing less to a pitch than lower ones.

Suitably chosen harmonic weighting functions enabled computer simulations of pitch perception data involving mistuned components (Moore et al., 1985), shifted harmonics (Patterson & Wightman, 1976; Schouten, Ritsma, & Cardozo, 1962), and various types of continuous spectra including rippled noise (Bilsen & Ritsma, 1970; Yost, Hill, & Perez-Falcon, 1978). It was shown how the weighting functions produce the dominance region (Plomp, 1967; Ritsma, 1967), how they lead to octave shifts of pitch in response to ambiguous stimuli (Patterson & Wightman, 1976; Schouten, Ritsma, & Cardozo, 1962), and how they lead to a pitch region in response to the octave-spaced Shepard tone complexes and Deutsch tritones (Deutsch, 1992a,b; Shepard, 1964) without the use of attentional mechanisms to limit pitch choices. An on-center off-surround network in the model helped to produce noise suppression, partial masking and edge pitch (von Bekesy,

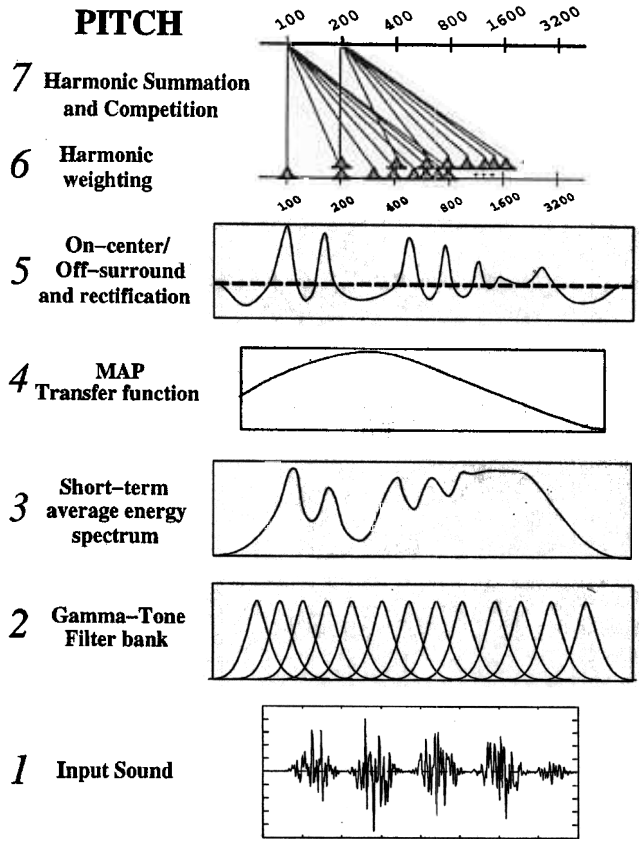

Fig. 9. Graphical representation of the SPINET model processing stages. [Reprinted with permission from Cohen, Grossberg, and Wyse (1995).]

1963; Small & Daniloff, 1967). Finally, it was shown how peripheral filtering and short term energy measurements produced a model pitch estimate that is sensitive to certain component phase relationships (Moore, 1977; Ritsma & Engel, 1964).

Fig. 9 shows the main processing stages of the SPINET model. Fig. 10b compares an illustrative computer simulation of pitch data in Fig. 10a concerning pitch shifts as a function of shifts in component harmonics. In particular, when harmonic components ($f_n = nf_0$, $n = 1,\ldots$) are all shifted by a constant amount, Δ, in frequency so that they maintain their spacing of f_0, ($f_n = nf_0 + \Delta$, $n = 1,\ldots$), the pitch shift in linear frequency is slower than that of the components (Patterson & Wightman, 1976; Schouten, Ritsma, & Cardozo, 1962). The data exhibit an ambiguous pitch region at shift values of $\Delta = lf_0$, $l = 0.5, 1.5, 2.5,\ldots$ where the most commonly perceived pitch jumps down to below the value of f_0. Fig. 10 shows the pitch of components spaced by $f_0 = 100$ Hz as a function of the lowest component's harmonic number, l. When the shift value Δ is near a harmonic of f_0 ($\Delta = lf_0$, $l = 0, 1, 2,\ldots$), then the pitch is unambiguous and near 100 Hz.

The model explains these data, as in Fig. 10b, in terms of the gradual reduction in the contribution a component makes to a pitch as it is mistuned, combined with the effect of filters whose widths are approximately constant in log coordinates for high frequencies (see Level 6 in Fig. 9). As the components shift together in linear frequency away from harmonicity, the higher components move into the shallow skirts of the filters centered at harmonics of the original nominal pitch frequency much more slowly than do the lower components, thereby slowing the shift away from the original pitch. Moreover, as the lowest stimulus component increases in harmonic number, all components are moving through broader filters, so the slopes of the pitch shift become less steep, as can be seen in both the data and the model output in Fig. 10.

Various other pitch data explanations of the SPINET model depend for their explanation upon properties of other model processing levels. The full array of simulated data makes use of all these levels. A key hypothesis of the model in all these explanations is that the harmonic summation at Level 7 of Fig. 9 filters each frequency spectrum through a *harmonic sieve* (Duifhuis, Willems, & Sluyter, 1982; Goldstein, 1973; Scheffers, 1983; Terhardt, 1972) that transforms logarithmically scaled and Gaussianly weighted harmonic components into activations of pitch nodes (or cell

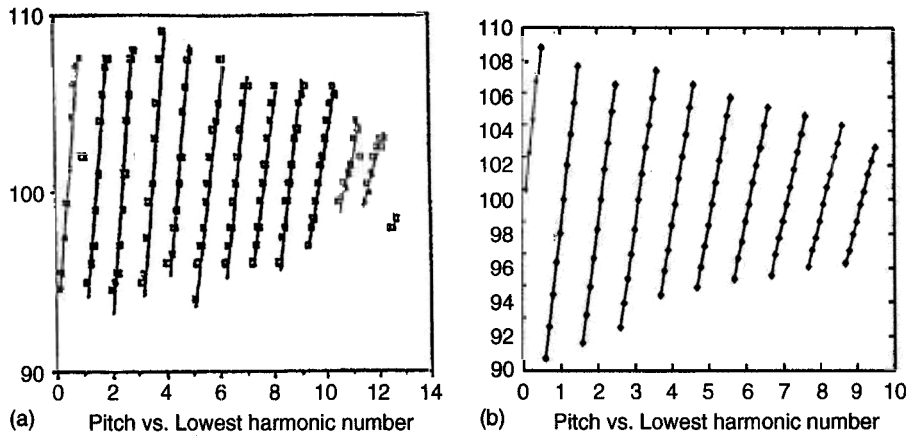

Fig. 10. Pitch shift in response to a complex of 6 components spaced by 100 Hz, as a function of the lowest component's harmonic number. (a) Data from Patterson and Wightman (1976). (b) Maximally activated pitch produced by the network model. [Reprinted with permission from Cohen, Grossberg, and Wyse (1995).]

populations) at the model's final layer. The harmonic sieve prevents spectral components that are not harmonically related to a prescribed pitch from activating the corresponding pitch node. It is assumed that the harmonic sieve gets adaptively tuned during development in response to harmonic preprocessing by peripheral acoustic mechanisms. This learning process is not explicitly modeled in SPINET, but the use of ART matching and resonance mechanisms in the ARTSTREAM model clarify how this learning process could occur.

4.3. The ARTSTREAM model

Accordingly, the final two spectral (Level 6) and pitch (Level 7) layers of the SPINET model in Fig. 9, including the harmonic sieve, are embedded in the ARTSTREAM model of Fig. 8, where they are elaborated into multiple spectral and pitch stream layers that interact via excitatory and inhibitory pathways. In particular, instead of there being just one spectral or pitch representation, ARTSTREAM contains multiple copies of the spectral and pitch representations (Fig. 8), each one providing a spatial substrate for a different stream. Said in another way, each frequency can activate a *band* of cells in the spectral representation. The cells in a given frequency band lie at spatial positions that are perpendicular to, or at least different from, the positions at which different frequencies are represented. The spatial organization of excitatory and inhibitory interactions converts these bands of cells into different perceptual streams.

For example, as in the SPINET model, each of the bottom-up filters from spectral to pitch layers forms a harmonic sieve. In addition, the top-down filters also form harmonic sieves. As clarified below, these top-down signals select those spectral components that are harmonically related to a chosen pitch category, while suppressing all other frequencies that may have initially activated that spectral stream layer. The ARTSTREAM model incorporates general ART principles which clarify how the bottom-up and top-down harmonic sieves are learned, and then used to generate percepts of distinct auditory streams.

4.4. ART: fast learning and stable memory in a changing world

Humans are able to rapidly learn enormous amounts of new information throughout life. For example, after seeing and hearing an exciting movie, we can tell our friends many details about it later on, even though the individual scenes flashed by very quickly. More generally, we can quickly learn about new environments, even if no one tells us how the rules of each environment differ. To a surprising degree, new facts can be learned without forcing rapid forgetting of what we already know.

The brain hereby solves a very hard problem that many current approaches to technology have not solved: It is a self-organizing system that is capable of rapid yet stable autonomous learning of huge amounts of data in a nonstationary environment. Discovering the brain's solution to this key problem is as important for understanding ourselves as it is for developing new pattern recognition and prediction applications in technology.

The problem whereby the brain learns quickly and stably without catastrophically forgetting its past knowledge has been called the *stability–plasticity dilemma* (Grossberg, 1980). The stability–plasticity dilemma must be solved by every brain system that needs to rapidly and adaptively respond to the flood of signals that subserves even the most ordinary experiences. If the brain's design is parsimonious, then similar design principles should operate in all the brain systems that can stably learn an accumulating knowledge base in response to changing conditions throughout life. The discovery of such principles should clarify how the brain unifies diverse sources of information into coherent moments of conscious experience.

4.5. ART matching and resonance: the link between attention, intention, and consciousness

Adaptive resonance theory claims that, in order to solve the stability–plasticity dilemma, resonant states, such as the ones mentioned above, can drive new learning. That is why the theory is called *adaptive* resonance theory. How this works is more completely explained in Carpenter and Grossberg (1991) and Grossberg (1999b). Some implications of ART principles are as follows.

The first implication provides a new answer to why, as philosophers have asked for many years, humans are 'intentional' beings who are always anticipating or planning their next behaviors and their expected consequences. ART suggests that 'stability implies intentionality'. That is, stable learning requires that we have expectations about the world that are continually matched against world data. In the special case of the ARTSTREAM model, these expectations are top-down harmonic sieves that are activated by pitch categories. The second implication is that 'intention implies attention and consciousness'. That is, expectations start to focus attention on data worthy of learning, and these attentional foci are confirmed when the system as a whole incorporates them into resonant states that are predicted to include conscious states of mind. In the ARTSTREAM model, these attentional foci are harmonics of a selected pitch category.

Implicit in the concept of intentionality is the idea that one can get *ready* to experience an expected event so that, when it finally occurs, it can be reacted to it more quickly and vigorously, and until it occurs, we are able to ignore other, less desired, events. This property is an example of *priming*. It shows that, when a top-down expectation is read-out in the absence of a bottom-up input, it can modulate, or subliminally select, the cells that would ordinarily respond to the bottom-up input, but not vigorously fire them, while it

suppresses cells whose activity is not expected. Correspondingly, the ART matching rule computationally realizes the following properties at any processing level where bottom-up and top-down signals are matched.

Bottom-up automatic activation. A cell, or cell population, can become active enough to generate output signals if it receives a large enough bottom-up input, other things being equal.

Top-down priming. A cell can be sensitized, modulated, or subliminally activated, but cannot generate large output signals, if it receives only a large top-down expectation input. Such a top-down priming signal prepares a cell to react more quickly and vigorously to subsequent bottom-up input that matches the top-down prime.

Match. A cell can become active if it receives large convergent bottom-up and top-down inputs. Such a matching process can generate enhanced activation and synchronization with other primed cells as resonance takes hold.

Mismatch. A cell is suppressed even if it receives a large bottom-up input if it also receives only a small, or zero, top-down expectation input.

This ART matching rule and the resonance rule that it implies have been mathematically proved necessary to solve the stability–plasticity dilemma (Carpenter & Grossberg, 1991). In particular, where they are violated, examples have been constructed wherein learning is unstable through time. These examples illustrate how we can continue to learn rapidly and stably about new experiences throughout life by matching bottom-up signal patterns from more peripheral to more central brain processing stages against top-down signal patterns from more central to more peripheral processing stages. The top-down signals represent the brain's learned expectations of what the bottom-up signal patterns should be based upon past experience. The matching process is designed to confirm those combinations of features in the bottom-up pattern that are consistent with the top-down expectations, and to suppress those features that are inconsistent. This top-down matching step initiates the process whereby the brain selectively pays attention to experiences that it expects, binds them into coherent and synchronous internal representations through resonant states, and incorporates them through learning into its knowledge about the world.

ART predicted (Carpenter & Grossberg, 1987; Grossberg, 1999b) that the brain uses the simplest possible circuit to realize the ART matching rule; namely, a *modulatory top-down on-center off-surround network.* In such a network, excitation and inhibition are approximately balanced within the on-center, so that top-down attentive priming can sensitize but not fire target cells, yet matched bottom-up and top-down signals can fire and even gain-amplify the activities of cells to which attention is paid. The off-surround can vigorously suppress mismatched cells. Many psychophysical and neurobiological experiments have by now supported this predicted link between attention, competition, and matching, and circuits have been identified that are proposed to realize it within the laminar architecture of neocortex. See Grossberg (1999a, 2003b) and Raizada and Grossberg (2003) for reviews.

In the ARTSTREAM model (Fig. 8), the top-down excitatory harmonic sieve is balanced by inhibition from the pitch summation layer to realize these properties. As a result, feedback from the pitch stream layer to the spectral stream layer activates a matching process that reinforces consistent spectral components and suppresses inconsistent components. The inconsistent spectral components are then freed to be captured by other streams, as in the 'old-plus-new heuristic' of Bregman (1990). Competition between streams for each frequency component (Fig. 8b) presents a frequency from being simultaneously allocated to two streams; hence, a frequency is uniquely assigned to a pitch whose top-down harmonic filter succeeds in selecting it. Reciprocal excitatory interactions between active pitch stream neurons and their consistent spectral components may continue until they give rise to a nonlinear resonance across both layers. The listener's conscious percept is hypothesized to correspond to the activity at the spectral stream layer when there is resonance between it and the pitch stream layer. In other words, a conscious streaming percept is predicted to arise from a *spectral-pitch resonance.*

4.6. Resonant dynamics explain the auditory continuity illusion

Resonant processing in the ARTSTREAM model helps to explain cocktail party separation of distinct voices into auditory streams, as in the auditory continuity illusion of Fig. 1, as follows. As noted above, after the auditory signals are preprocessed by SPINET mechanisms, the active spectral, or frequency, components are redundantly represented in multiple spectral streams. These streams are then filtered by bottom-up signals that activate multiple representations of the sound's pitch at the pitch stream level. These pitch representations compete to select a winner, which inhibits the redundant representations of the same pitch across streams, while also sending top-down matching signals back to the spectral stream level. By the ART matching rule, the frequency components that are consistent with the winning pitch node are selected, and all others are suppressed. The selected frequency components reactivate their pitch node which, in turn, reads out selective top-down signals. In this way, a spectral-pitch resonance develops within the stream of the winning pitch node. The pitch layer hereby coherently binds together the frequency components that correspond to a prescribed auditory source. All the frequency components that are suppressed by ART matching in this stream are freed to activate and resonate with a different pitch in a different stream. The net result is multiple resonances, each selectively grouping together into pitches those frequencies that correspond to distinct auditory sources.

The fact that noise is needed to continue the tone in Fig. 1a is due to the fact that top-down expectations in ART can *select* active bottom-up signals, but cannot create suprathreshold activation in their absence, which also explains the property in Fig. 1c. The fact that a future tone can help the resonance persist through the noise is traced to the fact that it takes a relatively long time for a spectral-pitch resonance to become suprathreshold and conscious, but a much shorter time for a consistent bottom-up signal to maintain such a resonance after it begins. Similar properties help to explain a lot of data about speech perception, including classical percepts like phonetic restoration (Grossberg, 1999b, 2003b; Grossberg, Boardman, & Cohen, 1997; Grossberg & Myers, 2000).

5. ARTSTREAM model

The ARTSTREAM model is mathematically defined in this section. Readers can skip to Section 6 for model simulations before studying the model equations.

5.1. Auditory peripheral processing

5.1.1. Outer and middle ear

The outer and middle ear act as a broad bandpass filter, linearly boosting frequencies between 100 and 5000 Hz. An approximation to this is to preemphasize the signal using a simple difference

$$y(t) = x(t) - Ax(t - \Delta f), \tag{1}$$

where A is the preemphasis parameter, and Δt is the sampling interval. In the simulations, A was set to 0.95, and $\Delta t = 0.125$ ms, corresponding to a sampling frequency of 8 kHz.

5.1.2. Cochlear filterbank

The overall effect of the basilar membrane is to act as a filterbank, where the response at a particular location on the basilar membrane acts like a bandpass filter. This bandpass characteristic has been modeled as a fourth order gammatone (de Boer & de Jongh, 1978; Cohen et al., 1995) filter

$$g_{f_0}(t) = \begin{cases} t^{n-1}\, \mathrm{e}^{-2\pi t b(f_0)} \cos(2\pi f_0 t + \phi), & \text{if } t > 0, \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

and its frequency response is

$$G_{f_0}(f) = [1 + \mathrm{j}(f - f_0)/b(f_0)]^n, \tag{3}$$

where n is the order of the filter; f_0, the center frequency of the filter; ϕ, a phase factor; $b(f)$ is the gammatone filter's bandwidth parameter, corresponding to:

$$b(f) = 1.02\,\mathrm{ERB}(f). \tag{4}$$

The equivalent rectangular bandwidth (ERB) of a gammatone filter is the equivalent bandwidth that a rectangular filter would have if it passed the same power:

$$\mathrm{ERB}(f) = 6.23\,\mathrm{e}^{-6} f^2 + 93.39\,\mathrm{e}^{-3} f + 28.52. \tag{5}$$

Sixty gammatone filters, which were equally spaced in ERB, were used to cover the range 100–2000 Hz. The output of each gammatone filter was converted into an energy measure.

5.1.3. Energy measure

The energy measures a short-time energy spectra (Cohen et al., 1995)

$$e_f(t) = \frac{\Delta t}{W} \sum_{k=0}^{W/\Delta t} |g_f(t - k\Delta t)|^2 \mathrm{e}^{-\alpha \Delta t k}, \tag{6}$$

where $e_f(t)$ is the energy measure output of the gammatone filter $g_f(t)$ centered at frequency f at time t; W is the time window over which the energy measure is computed; and α represents the decay of the exponential window. In the simulations, $\alpha = 0.995$, and $W = 5$ ms. The output of the energy measure feeds identically to the multiple fields in the spectral stream layer.

5.2. Spectral stream layer

Segregation based on harmonicity is achieved by having objects compete for frequency channels, which are excited by their pitch counterparts and supported by the bottom-up input (Fig. 8b). As noted above, the spectral stream layer is a plane with one axis representing frequency, and the other axis representing frequency bands that can be allocated to different auditory streams.

Each frequency channel in the energy measure, e_f, feeds up to the corresponding frequency channel in the spectral stream layer S_f in a one-to-many manner, so that all streams in the spectral stream layer receive equal bottom-up excitation. After the spectral stream layer becomes activated, the different streams activate their corresponding pitch streams in the pitch stream layer. When a pitch is selected in a given stream, it feeds back excitation to its spectral harmonics, and inhibits that pitch value in other streams in the pitch stream layer. An asymmetric gradient of inhibition across streams prevents a deadlock in the selection of a stream. In addition, nonspecific inhibition, mediated by the pitch summation layer, helps to suppress those spectral components that do not belong to the given pitch within its stream, and thereby realizes the ART matching rule.

The following equation describes the dynamics of the spectral stream layer:

$$\dot{S}_{if} = -AS_{if} + [B - S_{if}]E_{if} - [C + S_{if}]I_{if}, \tag{7}$$

$$E_{if} = \sum_g D_{fg} s(e_g) + F \sum_p \sum_k M_{f,kp} g(P_{ip}) h(k) \tag{8}$$

and

$$I_{if} = \sum_{g \neq f} E_{fg} s(e_g) + J \sum_{k \neq i} \sum_g N_{fg} [S_{kg}]^+ + \mathrm{LT}_i, \tag{9}$$

where S_{if} is the activity of the spectral stream layer neuron corresponding to the ith stream and frequency f. Term $-AS_{if}$ in Eq. (7) is the spontaneous decay. Term $D_{fg}s(e_g)$ in Eq. (8) is the excitation from the energy measure, which has been passed through a sigmoid $s(x)$ to compress the dynamic range:

$$s(x) = \begin{cases} x^2/(N_s + x^2), & \text{if } x > 0, \\ 0, & \text{otherwise} \end{cases} \tag{10}$$

Similarly, $E_{fg}s(e_g)$ in Eq. (9) is the inhibition from the energy measure, which has been passed through a sigmoid $s(x)$. Thus, with both $D_{fg}s(e_g)$ and $E_{fg}s(e_g)$, each spectral stream layer receives a contrast-enhanced version of the energy measure. Both D_{fg} and E_{fg} are Gaussians which are centered at frequency f, and have standard deviation parameters, σ_D and σ_E, and scaling parameters D and E, respectively; namely

$$D_{fg} = DG(f, \sigma_D) = D\frac{1}{\sigma_D\sqrt{2\pi}}e^{-0.5(f-g)^2/\sigma_D^2} \tag{11}$$

and

$$E_{fg} = EG(f, \sigma_F) = E\frac{1}{\sigma_E\sqrt{2\pi}}e^{-0.5(f-g)^2/\sigma_E^2}. \tag{12}$$

In addition, the term $F\Sigma_p\Sigma_k M_{f,kp}g(P_{ip})h(k)$ in Eq. (8) is the sum of all the pitches p which have a harmonic kp near frequency f in the pitch stream layer corresponding to stream i. In Eq. (8), $g(x)$ is a sigmoid function

$$g(x) = \begin{cases} x^2/(N_g + x^2), & \text{if } x > 0, \\ 0, & \text{otherwise} \end{cases} \tag{13}$$

where $h(k)$ is the harmonic weighting function, which weights the lower harmonics more heavily than higher harmonics:

$$h(k) = \begin{cases} 1 - M_h \log_2(k), & \text{if } 0 < M_h \log_2(k) < 1, \\ 0, & \text{otherwise} \end{cases} \tag{14}$$

and $M_{f,kp}$ is a normalized Gaussian, so that if a harmonic is slightly mistuned it will still be within the Gaussian and thus get partially reinforced. The width of the Gaussian dictates the tolerance for mistuning. Kernel $M_{f,kp}$ is centered at frequency f and has a standard deviation parameter, σ_M :

$$M_{f,kp} = G(f, \sigma_M) = \frac{1}{\sigma_M\sqrt{2\pi}}e^{-0.5(f-kp)^2/\sigma_M^2}. \tag{15}$$

The term $J\sum_{k\neq i}\sum_g N_{fg}[S_{kg}]^+$ in Eq. (9) represents the competition across streams for a component, so that a harmonic will belong to only one object. This inhibition embodies the principle of 'exclusive allocation.' Since a harmonic can be mistuned slightly, a Gaussian window N_{fg} exists within which the competition takes place. Kernel N_{fg} is centered at frequency f and has a standard deviation parameter, σ_N

$$N_{fg} = G(f, \sigma_N) = \frac{1}{\sigma_N\sqrt{2\pi}}e^{-0.5(f-g)^2/\sigma_N^2}. \tag{16}$$

Term LT_i in Eq. (9) is the inhibition from the pitch summation layer, which nonspecifically inhibits all components in stream i. The effect of this is to subtract out those nonharmonic components which are not reinforced by the top-down excitation from the pitch unit in the pitch stream layer. This is akin to the matching process used in Adaptive Resonance Theory (Carpenter & Grossberg, 1991, 1993; Grossberg, 1980). As realizes the ART matching rule, so that a spectral stream layer neuron can become

- Active if only an energy input is present (bottom-up automatic activation),
- Partially, or subliminally, active if only a pitch input is present (top-down priming),
- Active if both energy and pitch inputs are present (bottom-up and top-down consistency),
- Inactive if both energy and pitch inputs are present, but the spectral component is not a harmonic of pitch (bottom-up and top-down inconsistency).

The first constraint allows bottom-up activation to initiate the segregation process. So, if there is no pitch unit that is active, then there is no inhibition from the pitch stream layer, via the pitch summation layer. Thus, the spectral stream layer will become active. The second constraint makes sure that the pitch units do not activate spurious spectral units by themselves, but only in conjunction with an input. This is accomplished by letting the inhibition from the pitch summation layer be no smaller than the excitation from the pitch units. The third and fourth constraints state that only harmonics of the particular pitch that are present in the input are excited. This is accomplished by setting the combined excitation from the input and pitch stream unit to be greater than the inhibition from the pitch summation layer. If a spectral unit is a harmonic of a pitch P and it has an input at that frequency, then the spectral unit will remain active. However, if the unit is not a harmonic (or a slightly mistuned harmonic), then the inhibition from the pitch summation layer will be greater than only the bottom-up input. In all the simulations, the parameters were set to: $A = 1$, $B = 1$, $C = 1$, $D = 500$, $E = 450$, $F = 3$, $J = 1000$, $L = 5$, $M_h = 0.3$, $N = 0.01$, $N_s = 10\,000$, $N_g = 0.01$, $\sigma_D = 0.2$, $\sigma_E = 4$, $\sigma_M = 0.2$, and $\sigma_N = 1$.

5.3. Pitch summation layer

The pitch summation layer sums up the pitch activity at stream i, and provides nonspecific inhibition LT_i to stream i's spectral stream layer in Eqs. (7)–(9) so that only those harmonic components that correspond to the selected pitch

remain active:

$$\dot{T}_i = AT_i + [B - T_i] \sum_p g(P_{ip}), \tag{17}$$

where $g(x)$ is the sigmoid function described in Eq. (13). In the simulations, $A = 100$, $B = 100$.

5.4. Pitch stream layer

The original SPINET model had two components: the spectral layer and a pitch layer. The spectral and pitch representations in ARTSTREAM enable multiple streams to compete between pitch units within and across streams (Fig. 8b). The modified pitch strength activation is

$$\dot{P}_{ip} = -AP_{ip} + [B - P_{ip}]E_{ip} - [C + P_{ip}]I_{ip}, \tag{18}$$

where

$$E_{ip} = E \sum_k \sum_f M_{f,kp}[S_{if} - \Gamma]^+ h(k) \tag{19}$$

and

$$I_{ip} = J \sum_{p \neq q} H_{pq} g(P_{iq}) + L \sum_{k>i} g(P_{kp}), \tag{20}$$

where P_{ip} is the pth pitch unit of object i. The term $E\Sigma_k\Sigma_f M_{f,kp}[S_{if} - \Gamma]^+ h(k)$ in Eq. (19) corresponds to the Gaussian excitation $M_{f,kp}$ from the spectral layer which has suprathreshold components near a harmonic kp of pitch p, which is weighted by the harmonic weighting function $h(k)$. The harmonic weighting function $h(k)$ and the Gaussian $M_{f,kp}$ are same as in the spectral layer (Eqs. (14) and (15), respectively). The term $J\Sigma_{p \neq q} H_{pq} g(P_{iq})$ in Eq. (20) represents the symmetric off-surround inhibition across pitches within a stream. The off-surround competition across pitches within a stream makes the layer act as a winner-take-all net so that only one pitch tends to be active within a stream. In addition, H_{pq} is defined to be one within a neighborhood around pitch unit j and zero otherwise, so that a stream can maintain a pitch even if the pitch fluctuates:

$$H_{pq} = \begin{cases} 1, & \text{if } |p - q| > \sigma_H, \\ 0, & \text{otherwise} \end{cases} \tag{21}$$

The term $L\Sigma_{k>i} g(P_{kp})$ in Eq. (20) represents asymmetric inhibition across streams for a given pitch, so that only one stream will activate a given pitch. This asymmetry across streams also provides a systematic choice of streams, and prevents deadlock between two streams for a given pitch, since all pitch streams receive equal bottom-up excitation from the spectral layer initially. In all the simulations, the parameters were set to: $A = 100$, $B = 1$, $C = 10$, $E = 5000$, $J = 300$, $L = 2$, $\sigma_H = 0.2$, and $\Gamma = 0.005$.

6. Streaming simulations

The model qualitatively emulates bounce percepts for crossing glides, as well as several variants of the continuity illusion. Fig. 11 shows the stimuli and the listeners' percepts that the model emulates. It should be reiterated that the percept that a listener would hear corresponds to the *resonant* activity in the spectral layer.

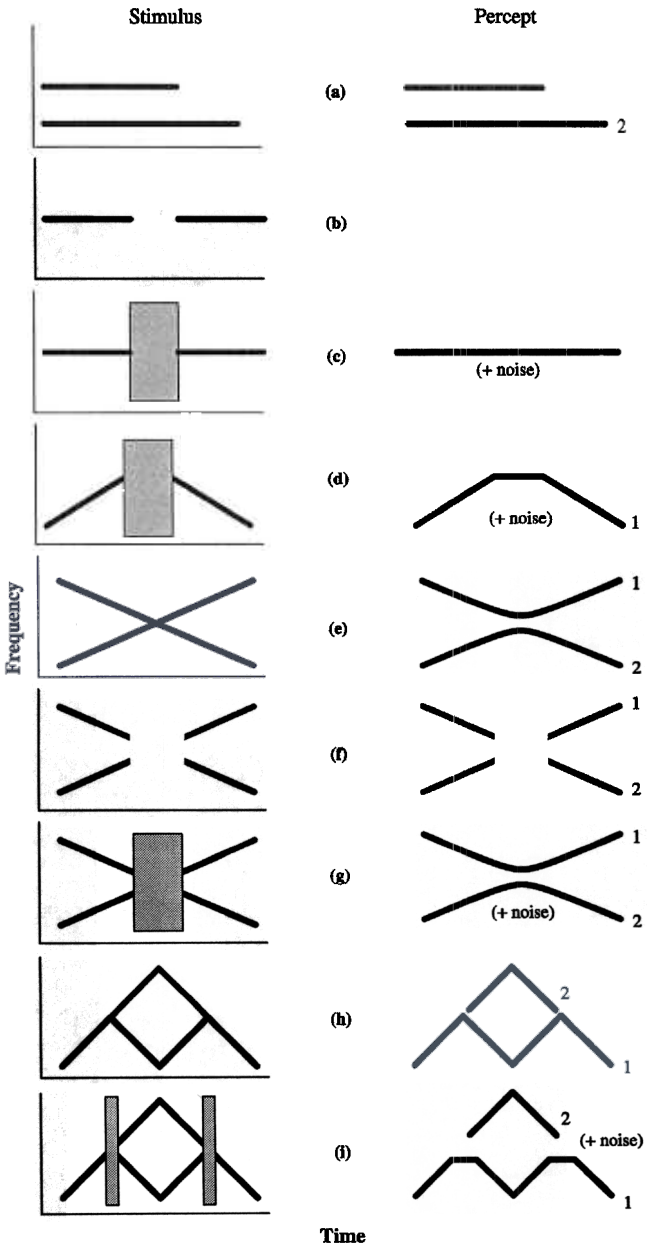

Fig. 11. Stimuli and the listeners' percepts that model simulations emulate. The hashed boxes represent broadband noise. The stimuli consist of: (a) two inharmonic tones, (b) tone–silence–tone, (c) tone–noise–tone, (d) a ramp or glide–noise–glide, (e) crossing glides, (f) crossing glides where the intersection point has been replaced by silence; (g) crossing glides where the intersection point has been replaced by noise, (h) Steiger (1980) diamond stimulus, and (i) Steiger (1980) diamond stimulus where bifurcation points have been replaced by noise.

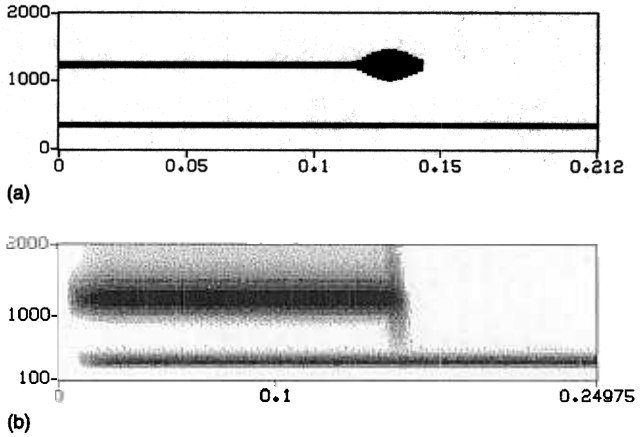

Fig. 12. (a) Spectrogram and (b) result of energy measure for the two tone stimulus.

6.1. Inharmonic simple tones

If two inharmonic tones are presented, then they should segregate into two different streams since they do not have a common pitch (Moore et al., 1985). Fig. 11a shows the stimulus and the listeners' percept for two inharmonic tones. Fig. 12a shows the spectrogram for two inharmonic tones, whose frequencies are 358 and 1233 Hz. Fig. 12b shows the result after peripheral processing; i.e. the result after the energy measure. Fig. 13 shows the resulting spectral and pitch layers for the two tone stimulus for two different streams. Fig. 13c shows how the streams initially compete for the tones, but the first stream, which is inherently biased in the pitch stream layer, wins the higher frequency component, allowing the second stream to capture the lower frequency tone.

Fig. 14 shows a schematic of how the grouping process works for the two inharmonic tones. After the two tones are processed by the peripheral processing, the higher frequency tone has a larger activity due to the preemphasis. The preprocessed activities feed into the spectral stream layers at time $t = 0$. Since there is no top-down activity at the spectral stream layers, the two spectral layers are equally active. Next, at time $t = t_1$, the pitch stream layer receives activation from the spectral stream layer. Since stream 1's pitch layer is inherently biased over stream 2's pitch layer, and since the higher frequency tone has a larger activity, the 1233 Hz tone is chosen by stream 1's pitch layer. Since the pitch layer is a winner-take-all network, only one pitch can be active within a pitch stream layer. Once the 1233 Hz tone is chosen by stream 1, the corresponding frequency in stream 2's pitch layer is inhibited by the stream 1's winning pitch neuron, allowing the 358 Hz tone to be captured by stream 2's pitch layer. Next, at time $t = t_2$, the winning pitch neurons excite their corresponding harmonic components in the spectral layer. In addition, the nonspecific inhibition (shown as the darker arrow) inhibits all

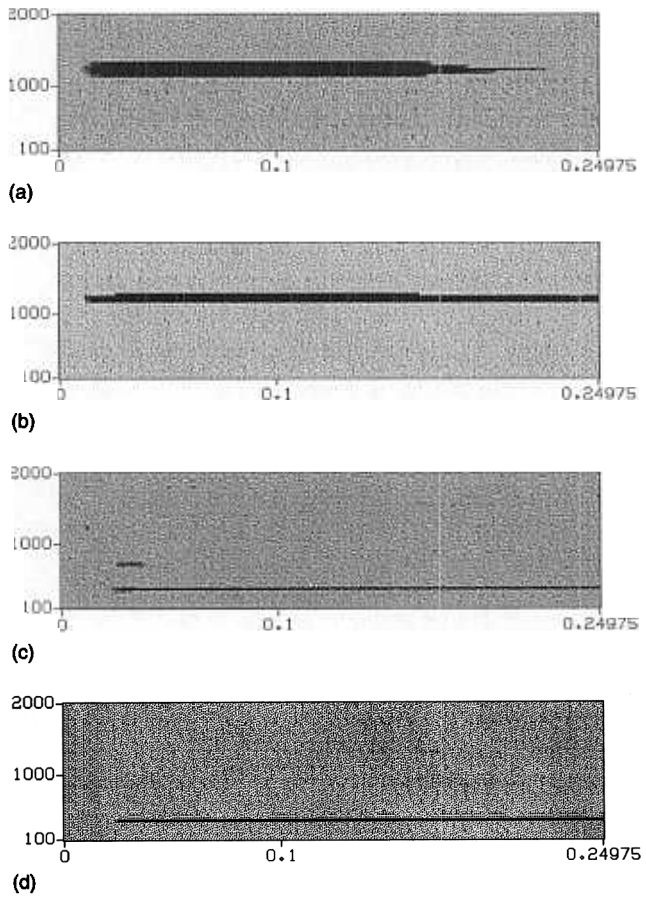

Fig. 13. Model results for the two tone stimulus. (a) spectral stream layer and (b) pitch stream layer for stream 1; and (c) spectral stream layer and (d) pitch stream layer for stream 2.

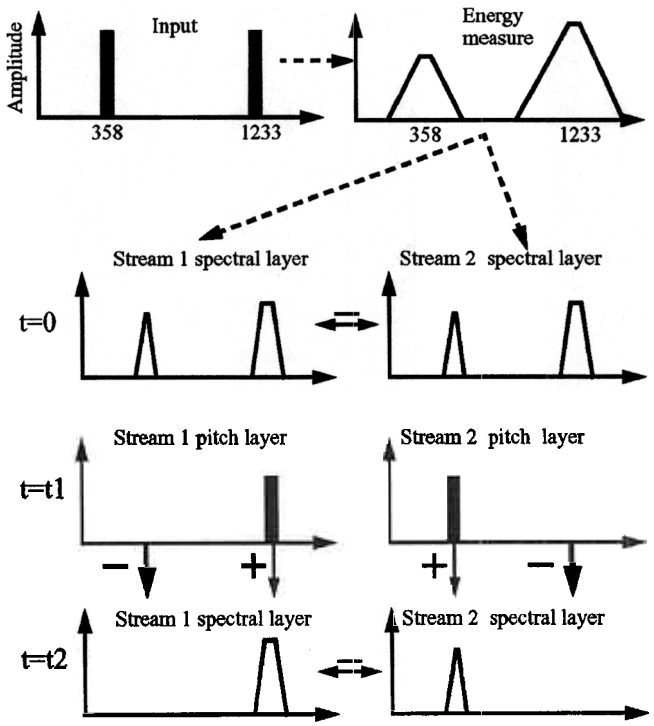

Fig. 14. Schematic of how the model segregates the two inharmonic tones into two different streams. See text for explanation.

components in the spectral layer. Therefore, those components that are not specifically excited by the pitch layer are suppressed. For example, the 358 Hz tone is suppressed in stream 1 since it is receiving top-down nonspecific inhibition and no top-down specific excitation, whereas the 1233 Hz tone receives top-down excitation allowing it to remain active.

6.2. Auditory continuity illusion

The model is capable of simulating continuation of a tone in noise, even though the tone is not physically present in the noise (Miller & Licklider, 1950). In order to appreciate the result for tone–noise–tone condition, one should consider the result of the model for a tone–silence–tone stimulus (Figs. 1c and 11b). For this stimulus, the tone should not continue across the silence, but should stop before penetrating the noise. Fig. 15 shows the spectrogram and the result after the peripheral processing for the tone–silence–tone stimulus. Fig. 16 shows the resulting spectral and pitch layers for the tone–silence–tone stimulus for two different streams. The figures show that the first stream captures the tone, which decays into to the silent interval but does not remain active throughout the silent interval. Since the model does not yet have any onset/offset mechanisms, the spectral stream activity slowly decays into the silent interval. The percept does not, however, persist this long because the pitch layer activity decays more quickly, thereby aborting the spectral-pitch resonance. The same stream then captures the tone after the silence as well. The second stream is not active since there are no extraneous components to capture.

Now consider the case where the silent interval is replaced by noise; i.e. the tone–noise–tone stimulus. For appropriate signal levels in the tone and noise, the tone percept should continue across the noise, even though the tone is not physically present during the noise interval. Fig. 17 shows the spectrogram and the result after the peripheral processing for the tone–noise–tone stimulus.

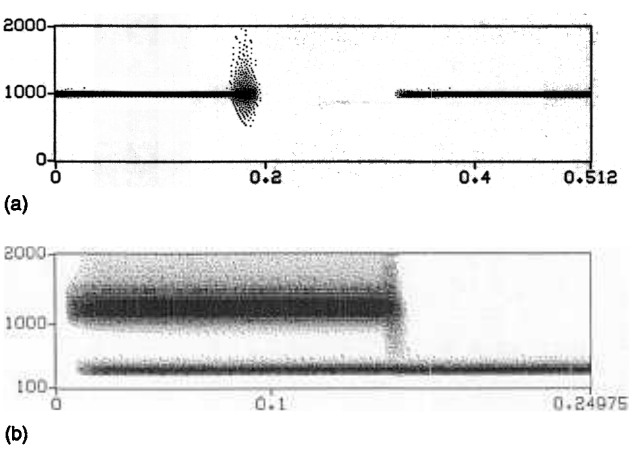

Fig. 15. (a) Spectrogram and (b) result of energy measure for the tone–silence–tone stimulus.

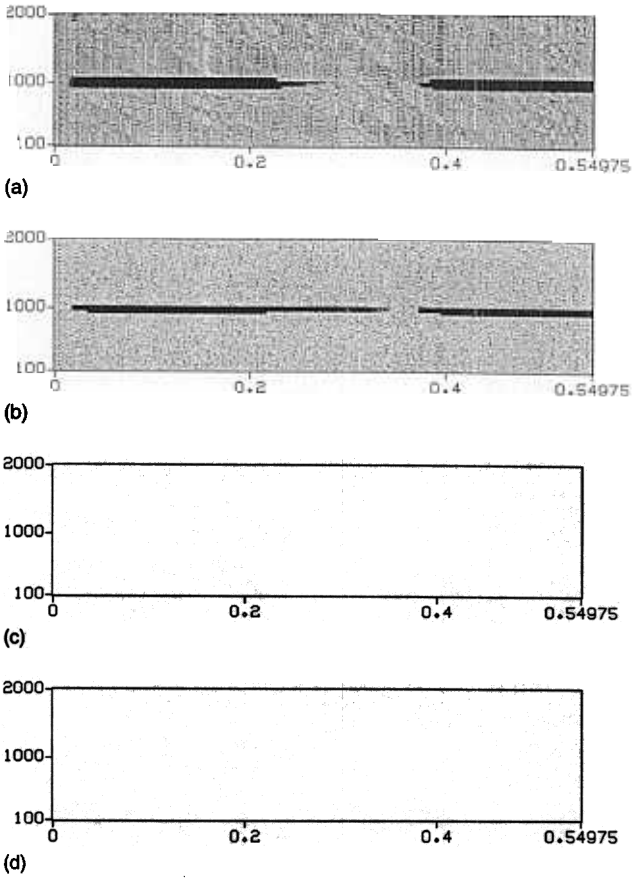

Fig. 16. Model results for the tone–silence–tone stimulus. (a) spectral stream layer and (b) pitch stream layer for stream 1; and (c) spectral stream layer and (d) pitch stream layer for stream 2.

Fig. 18 shows the resulting spectral and pitch layers for the stimulus for the first two streams, and Fig. 19 shows a third stream. The figures show that the first stream captures the tone, and that the resonance between the spectral and pitch layers continues through and past the noise interval.

The second and third streams contain the noise. The reason that the second stream captures the high frequency noise as opposed to the low frequency noise is due to

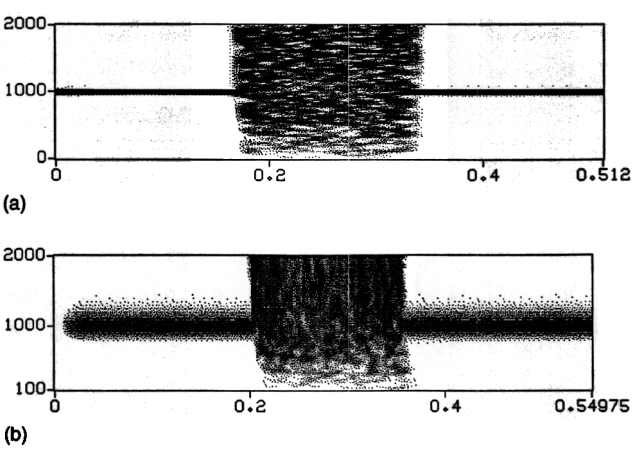

Fig. 17. (a) Spectrogram and (b) result of energy measure for the tone–noise–tone stimulus.

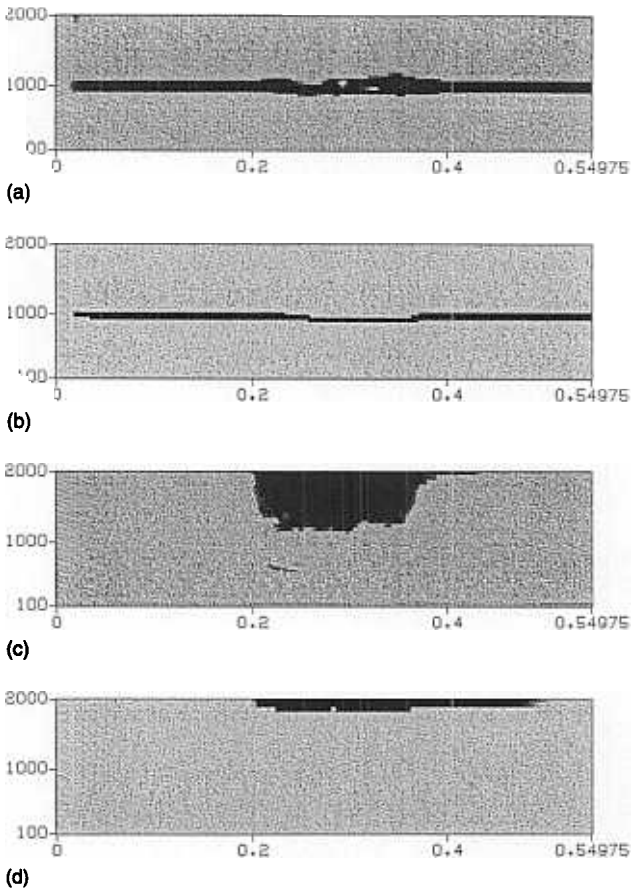

Fig. 18. Model results for the tone–noise–tone stimulus. (a) spectral stream layer and (b) pitch stream layer for stream 1; and (c) spectral stream layer and (d) pitch stream layer for stream 2.

preemphasis: the noise at the highest frequency is most active, and so it is captured by the second stream. If more streams were present in the model, then they would capture finer subsets of noise components.

The model is also capable of producing the continuity illusion for the ramped stimulus shown in Fig. 11d. Fig. 20 shows the spectrogram and the result after the peripheral processing. Fig. 21 shows the resulting spectral and pitch

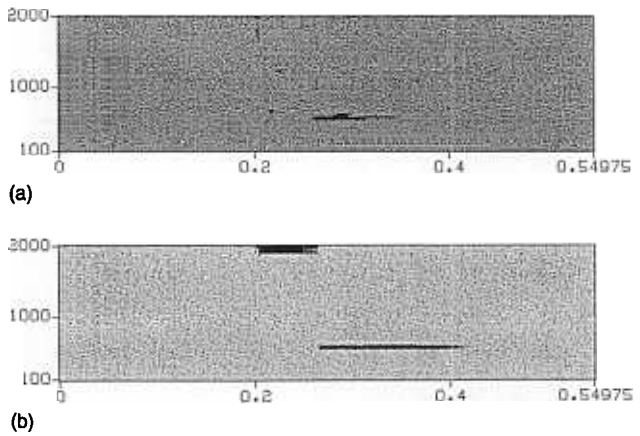

Fig. 19. The (a) spectral and (b) pitch stream layers for stream 3 for the tone–noise–tone stimulus.

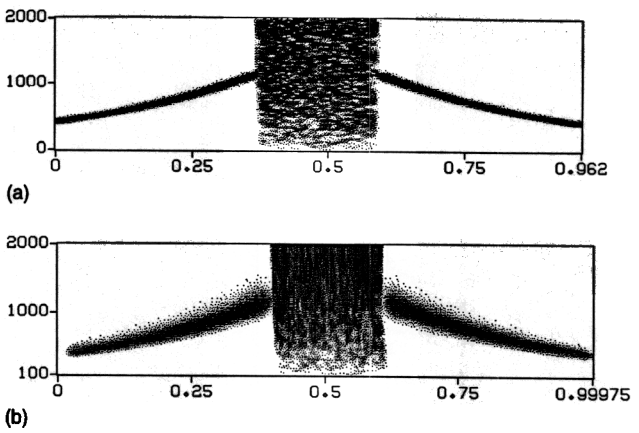

Fig. 20. (a) Spectrogram and (b) result of energy measure for the ramp stimulus.

layers for the stimulus for the two different streams. The figures show that the first stream captures the upward glide, which then continues through the noise interval. After the noise interval, the same stream captures the downward glide, leading to the ramp percept. The reason that the ramp completes across the noise is due to the same reason that the tone completes across the noise in the tone–noise–tone stimulus; namely, the temporal averaging at the spectral

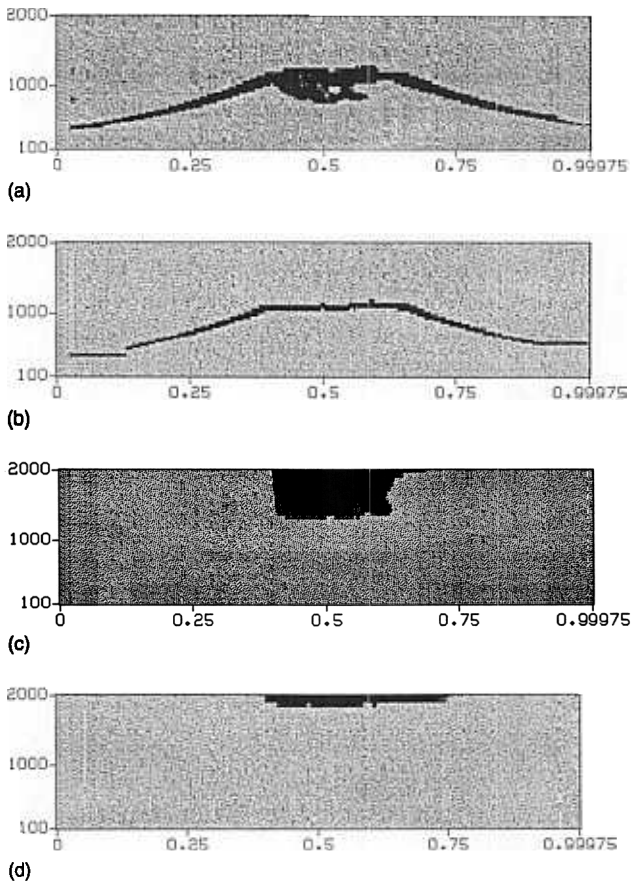

Fig. 21. Model results for the ramp stimulus. (a) Spectral stream layer and (b) pitch stream layer for stream 1; and (c) spectral stream layer and (d) pitch stream layer for stream 2.

stream layer is reinforced by top-down excitation from the pitch stream layer. Also, during the noise interval, some noise adjacent to the plateau is active since the top-down inhibition is not strong enough to suppress this activity. Meanwhile, the second stream contains the extraneous noise. If other streams were present, they might also capture some noise components.

6.3. Bounce percepts for crossing glides

The model is capable of qualitatively replicating the Halpern (1977) and the Tougas and Bregman (1990) data. For these stimuli, one obtains bounce percepts for crossing glides (Fig. 11e), even if the crossing interval is replaced by silence (Fig. 11f) or noise (Fig. 11g). Fig. 22 shows the spectrogram and the result after the energy measure for the standard crossing glide stimulus; and Fig. 23 shows the resulting spectral and pitch activity for the two streams. As one can see, one stream supports a ' ∪ ' percept, while the other stream has a ' ∩ ' percept. The ARTSTREAM explanation for the bounce percept in response to the standard crossing glide stimulus is as follows: initially, the higher frequency glide is captured by the first stream since it has a larger activation, and thus the lower frequency glide is captured by the second stream. The glides are maintained within their streams as they approach the intersection point. At the intersection point, the glides activate multiple, adjacent channels at the spectral layer. These adjacent channels can belong to the two different streams such that the larger frequency channel belongs to the first stream, and thus groups with the upper glide; and the lower adjacent frequency channel belongs to the second stream, and thus groups with the lower glide.

Fig. 24 shows the crossing glide stimulus for the silent-center condition and the result of the energy measure. Fig. 25 shows the spectral and pitch layers for two different streams. The result corresponds to a bounce percept, which does not continue across the silent interval. The reason one obtains the grouping of the upper glides is as follows. The first stream

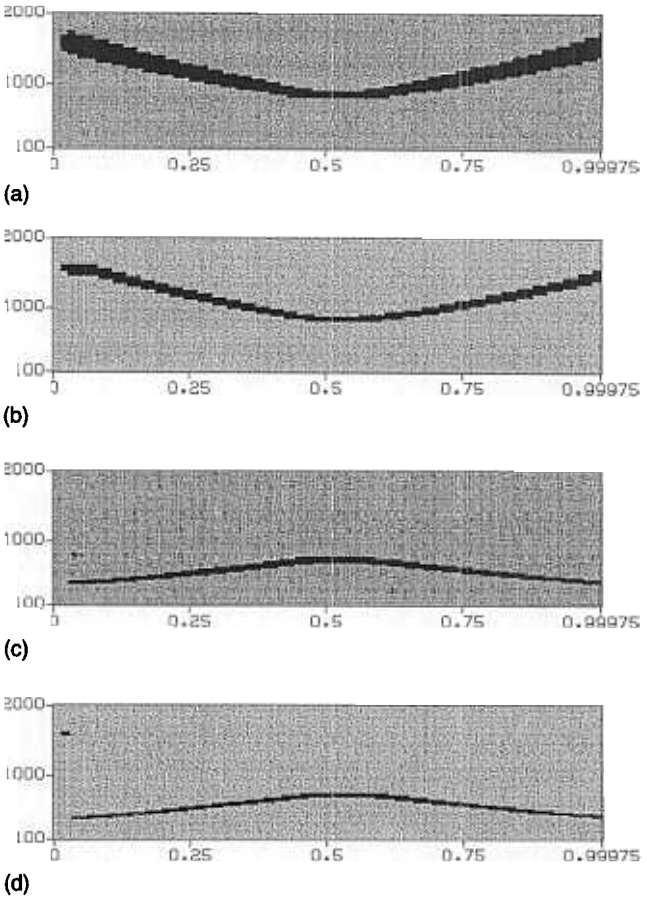

Fig. 23. Model results for the crossing glide stimulus. (a) Spectral stream layer and (b) pitch stream layer for stream 1; and (c) spectral stream layer and (d) pitch stream layer for stream 2.

captures the higher frequency glide at the onset of the stimulus and after the silent interval since these components have a larger activity than the lower frequency glides due to preemphasis. Since these components have a larger activity, the first stream will choose these components, leading to the grouping of the upper glides by stream 1, and the lower glides by stream 2; i.e. a bounce percept.

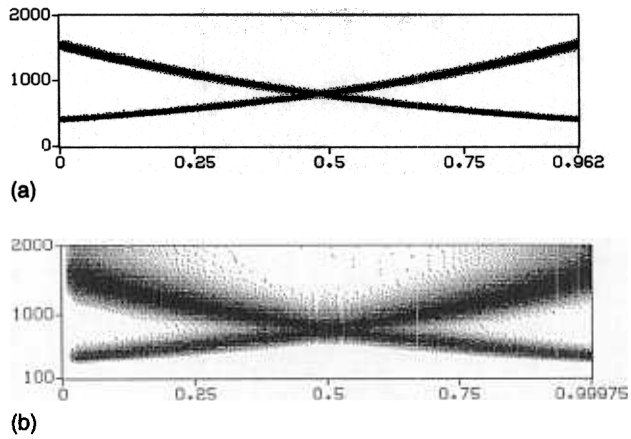

Fig. 22. (a) Spectrogram and (b) result of energy measure for the crossing glide stimulus.

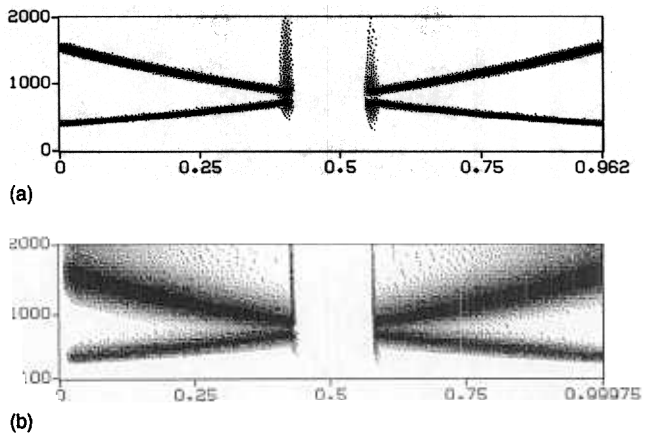

Fig. 24. (a) Spectrogram and (b) result of energy measure for the crossing glide stimulus with silence replacing the intersection point.

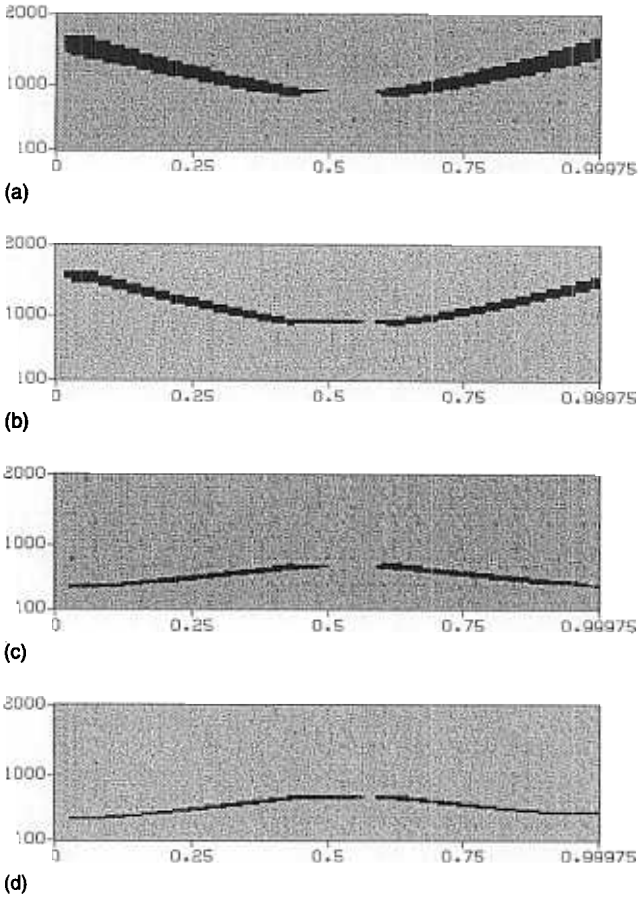

Fig. 25. Model results for the crossing glide stimulus with silence replacing the intersection point. (a) Spectral stream layer and (b) pitch stream layer for stream 1; and (c) spectral stream layer and (d) pitch stream layer for stream 2.

Fig. 11g shows the crossing glide stimulus where the intersection point has been replaced by noise, and the subjects' percepts of a bounce that is completed across the noise interval. Fig. 26 shows the spectrogram and the result of the energy measure for the crossing glide with noise-center stimulus, and Fig. 27 shows the spectral and pitch layers for two different streams. Once again, the bounce percept is evident, but there is continuity of the bounce through the noise interval. Stream 2 shows some noise activity that 'leaks' through, which is due to not enough top-down inhibition. The reason that the model produces the bounce phenomenon can be understood from the results on the auditory continuity illusion and the standard crossing glide stimulus. Initially, the upper frequency glide is chosen by stream 1, and the lower frequency glide is chosen by stream 2, just as in the standard crossing glide stimulus. The continuity illusion explanation clarifies how top-down activity from the pitch layer helps maintain the tone across the noise interval at the same frequency as the offset of the glide. In addition, the temporal averaging of the noise at the spectral stream layer provides uniform activity over time that aids the resonance between

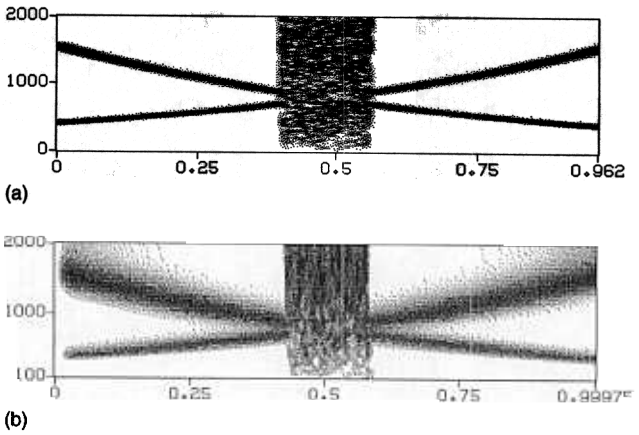

Fig. 26. (a) Spectrogram and (b) result of energy measure for the crossing glide stimulus with noise replacing the intersection point.

the spectral and pitch layers, and thus, maintaining the tone across the noise interval. At the offset of the noise, the glides are at approximately the same frequency as the tones that were continuing through the noise. Thus, these glides are grouped with the stream that has a tone close to its frequency. As a result, one obtains a bounce percept, where the bounce completes across the noise interval.

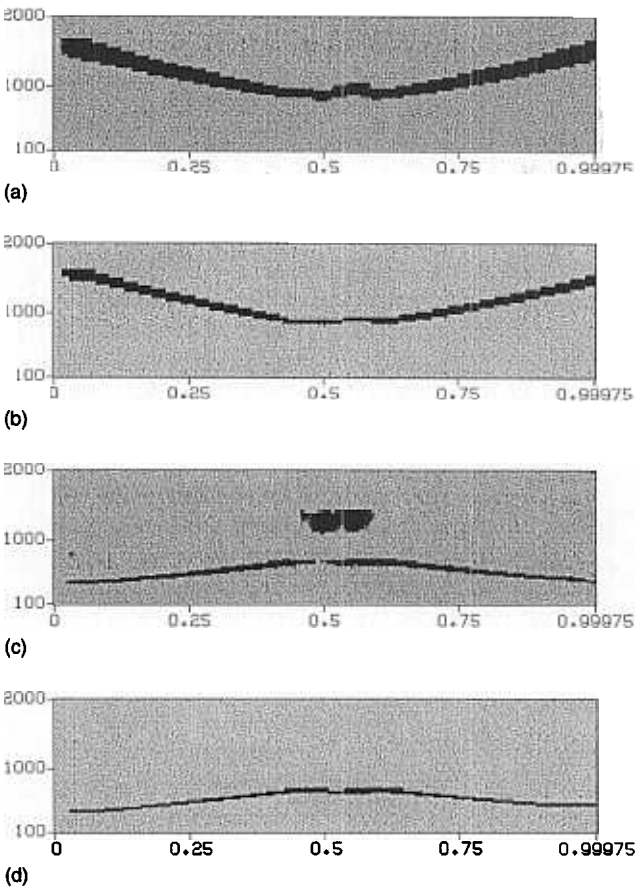

Fig. 27. Model results for the crossing glide stimulus with noise replacing the intersection point. (a) Spectral stream layer and (b) pitch stream layer for stream 1; and (c) spectral stream layer and (d) pitch stream layer for stream 2.

6.4. Steiger (1980) diamond stimulus

For the Steiger (1980) diamond stimulus (Fig. 11h), the percept consists of two streams, an 'M' stream and an inverted 'V' stream. This percept shows that the principle of continuity can be overcome by frequency proximity. Fig. 28 shows the Steiger (1980) stimulus and the result after the peripheral processing. Fig. 29 shows the spectral and pitch layer for two different streams. As one can see, the lower 'M' shaped component falls into one stream, while the inverted 'V' is in the other stream, which qualitatively emulates the percept. The reason the model emulates the Steiger data is similar to the explanation for the bounce percept for the standard crossing glide explanation. Initially, stream 1 is active with the lower frequency glide and stream 2 is inactive, since there is only one component present in the stimulus. At the bifurcation point, stream 1 continues with the lower frequency glide since this frequency component was previously active in stream 1. In other words, due to the temporal averaging of the spectral layer activity and resonance with the pitch layer, the frequency component that was activated immediately prior to the bifurcation point will remain active and group with the same frequency component immediately after the bifurcation point. Since the first stream groups the lower frequency glides together, the second stream is capable of capturing the higher frequency glides. Thus, stream 1 contains the 'M' percept, while stream 2 contains the inverted 'V' percept.

Fig. 30 shows the spectrogram and the result of the energy measure for the Steiger (1980) stimulus where the bifurcation points have been replaced by noise. Fig. 31 shows the spectral and pitch layers for the two streams for the Steiger (1980) stimulus when the bifurcation points have been replaced by noise. The figures show that the 'M' and the inverted 'V' segregate into two different streams, and the 'M' continues across the noise interval. The noise activates other streams, which are not shown. The reason the model emulates this percept derives from the explanation of the Steiger (1980) diamond stimulus and the continuity illusion; e.g. the ramp stimulus of Fig. 11d. Stream 1 initially captures the increasing glide, while stream 2 is inactive, just as in Steiger (1980) diamond stimulus. During the noise interval, stream 1 completes across the noise interval just as in the ramp stimulus, allowing stream 2 to capture the inverted 'V' component.

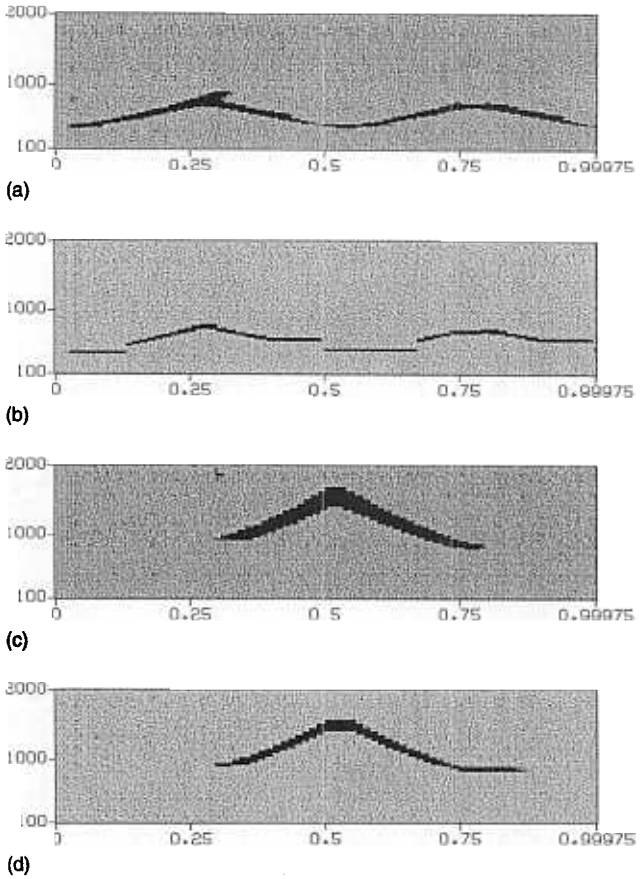

Fig. 29. Model results for the Steiger (1980) diamond stimulus. (a) Spectral stream layer and (b) pitch stream layer for stream 1; and (c) spectral stream layer and (d) pitch stream layer for stream 2.

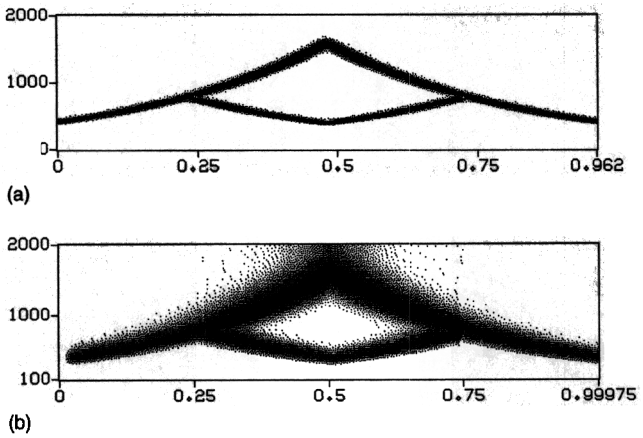

Fig. 28. (a) Spectrogram and (b) result of energy measure for the Steiger (1980) diamond stimulus.

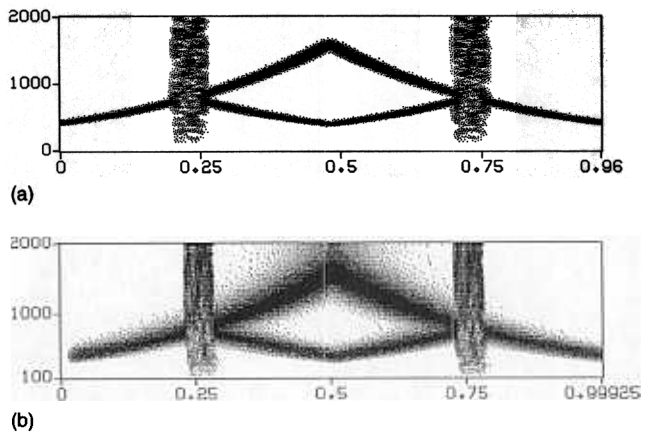

Fig. 30. (a) Spectrogram and (b) result of energy measure for the Steiger (1980) diamond stimulus with noise bursts replacing the bifurcation points.

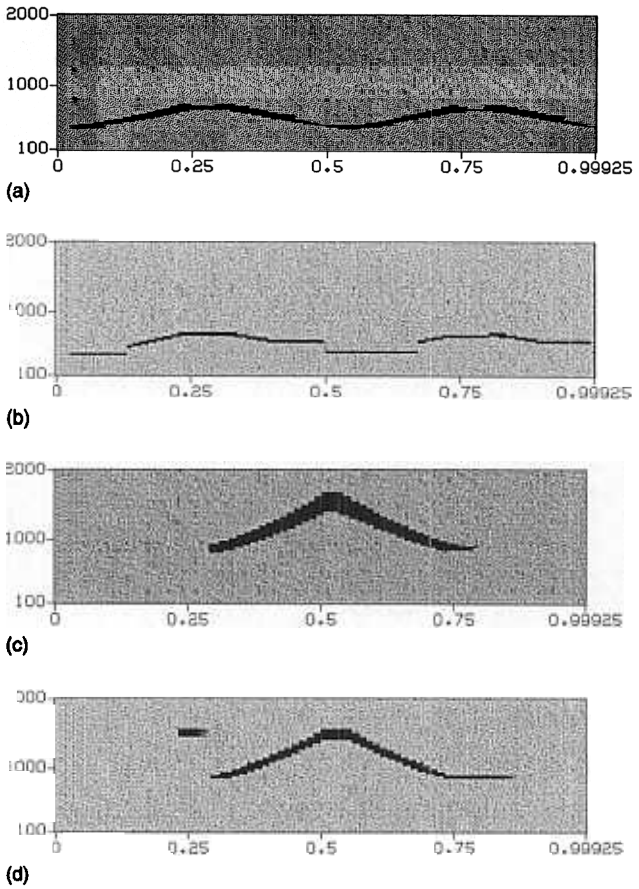

Fig. 31. Model results for the Steiger (1980) diamond stimulus with noise bursts replacing the bifurcation points. (a) Spectral stream layer and (b) pitch stream layer for stream 1; and (c) spectral stream layer and (d) pitch stream layer for stream 2.

7. Interactions between pitch and spatial location cues

This section outlines how spatial location cues can be incorporated into the model to aid the segregation process. The spatial location cues indirectly influence grouping by assisting grouping based on pitch. Spatial cues by themselves cannot group objects, but require a pitch difference to exist, in keeping with the data from Shackleton, Meddis, and Hewitt (1994). The model is extended using the same types of ART matching and resonance circuits that have been used to achieve grouping based on pitch in the previous sections. The extended model shows how spatial location cues can prime the pitch stream layer, and how the system can generate resonances that consistently incorporate all the pitch and spatial location cues that are available.

7.1. Influence of spatial location cues on streaming

The auditory system localizes sounds using two different mechanisms: interaural time differences (ITD) and interaural intensity differences (IID). The concept behind both ITD and IID is that the listener is comparing the signal between the two ears (interaural) and making a judgment on the sound's location (Handel, 1989).

ITD, which operates at low frequencies (less than 5 kHz), corresponds to comparing the arrival time of a signal to the two ears. If a signal is to the left, it will arrive at the left ear some microseconds before it arrives at the right ear. Thus at 0 ITD, the source is centralized, and at other ITDs the source is more lateral. However, ITDs only work for low frequency, where the wavelength is long compared to the size of the head. Fig. 32 shows a schematic representation of an object that is lateralized to the right. As the object emits a sound, it will arrive at the right ear first, and then at the left ear τ microseconds later, corresponding to the extra path distance d that the source has to travel.

At high frequencies, the head 'shadows' a sound lateralized to one side, causing an IID, or intensity difference. For example, if a high frequency sound is located to the left, the intensity of the sound to the right ear is diminished compared to the left ear. Thus, one can localize the sound by a computation based on the intensity difference at the two ears. The extended model presented here incorporates only ITDs in the segregation process.

The proposed model extension is schematized in Fig. 33. The model first preprocesses the incoming signal in the peripheral processing modules. This preprocessed signal is then used to determine spatial locations for the frequency components, and at the same time to group frequency components based on pitch using the spectral and pitch stream layers from the original model. Segregation of components is accomplished in the pitch and spectral stream layers; the spatial locations nonspecifically prime their corresponding pitch stream layer to bias them towards grouping components. Next, those components which have been grouped by pitch are reinforced based on their spatial locations.

The peripheral preprocessing is identical for both the left and right 'ears', and consists of the same module as in the original model. The output of this peripheral processing is fed to the $f-\tau$ plane (Colburn, 1973, 1977), where individual frequencies f are assigned to a spatial location τ.

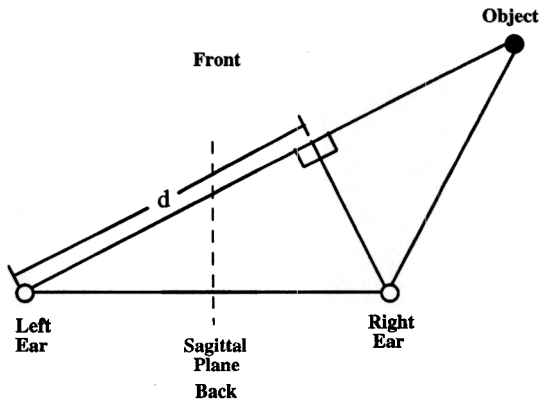

Fig. 32. Geometric representation of spatial lateralization using interaural timing differences (ITD).

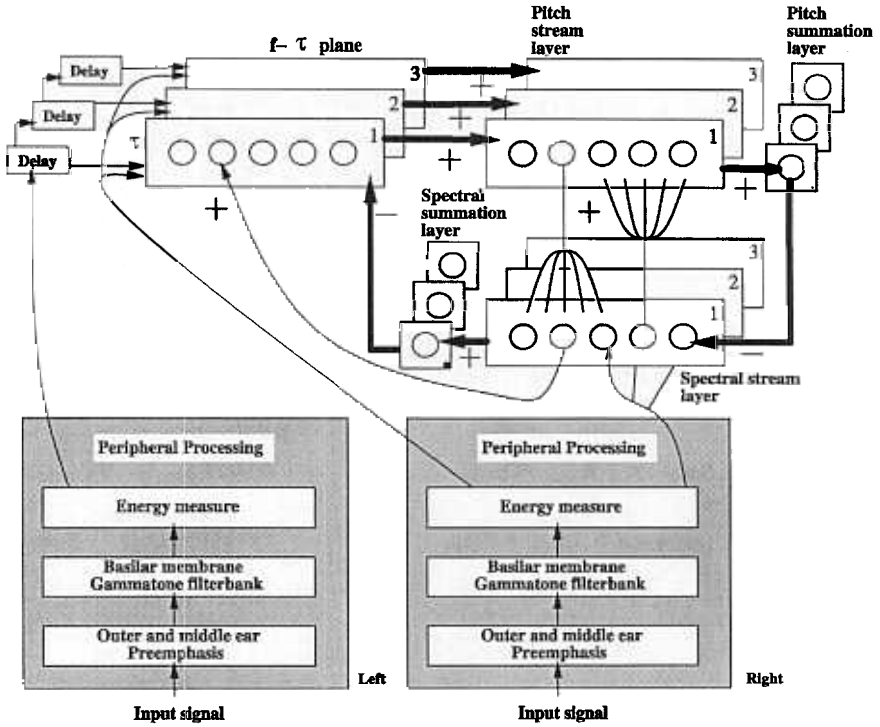

Fig. 33. Block diagram of an ARTSTREAM model that incorporates both pitch and spatial location cues.

Variable τ represents radial direction, taking on values from -600 to 600 μs. The value $\tau = 0$ corresponds to the central location, which is a location centered between the 'ears' and in front of the listener; $\tau = -600$ corresponds to a location that is directly to the left of the listener; and $\tau = 600$ corresponds to a location that is directly to the right of the listener. It is assumed that τ maps to radial direction in a linear fashion. It is also assumed that only one stream can occupy one spatial location, except at the central 'head-centered' location, where multiple streams can be represented, as when a symphony is heard through a pair of balanced monaural microphones. This scheme realizes a type of 'acoustic fovea' which donates more representational space to centered sounds than to peripheral sounds. Once components have been assigned to a given location, the location nonspecifically primes all the neurons in its corresponding pitch stream layer. Fig. 34 depicts how the spatial locations nonspecifically prime the pitch stream layers, and how a frequency component at a given spatial location in the $f-\tau$ is reinforced by its corresponding frequency component in the spectral stream layer.

The output of the right channel also feeds into the different streams of the spectral stream layer. The spectral stream layers are the same as in the original model. The pitch stream layer is modified so that all neurons within a stream become active if there are any components present at that given location. Thus, a pitch stream layer will be biased to win over another pitch stream layer if there are components present at that location. At the central location, the N streams are all excited. In addition, the asymmetric competition across streams, term $L\Sigma_{k>i}g(P_{kp})$ in Eq. (20), exists only at the central location; noncentral streams equally inhibit each other.

In addition, there is feedback from the spectral stream layer back to the $f-\tau$ plane. The feedback consists of

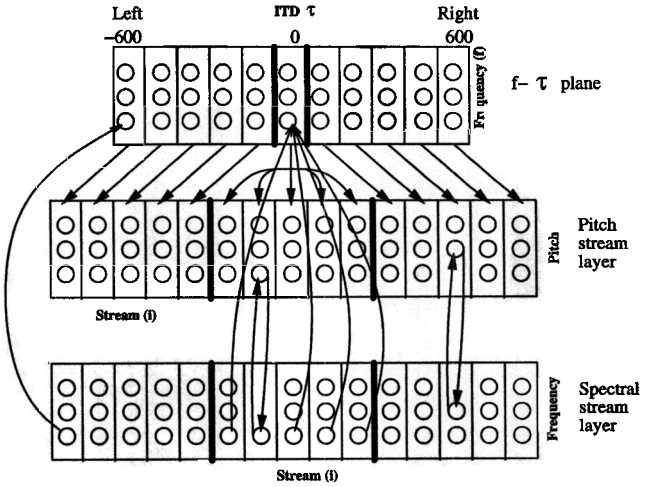

Fig. 34. Interaction between spatial locations in the $f-\tau$ field, pitch stream layer, and the spectral stream layer. The nonspecific inhibitory neurons are not shown. Only one stream can occupy one spatial location, except at the central 'head-centered' location $\tau = 0$, where multiple streams can be represented. Once a spatial location has been derived, the spatial location nonspecifically primes all the neurons in its corresponding pitch stream layer. At the central location, the N streams are all primed. Once components have been grouped based on pitch, the neurons in a spectral stream layer specifically excite the components at their corresponding spatial location. At the central location, the spectral neurons, corresponding to a given frequency, from all N streams excite the corresponding neuron at $\tau = 0$.

a specific excitatory feedback and a nonspecific inhibitory feedback, akin to the connectivity from the pitch stream layer to the spectral stream layer. The specific feedback excites those harmonic components existing at a given location where a pitch has been determined. At the central location, the spectral neurons, corresponding to a given frequency, from all N streams excite the corresponding neuron at $\tau = 0$. The spectral summation layer provides nonspecific inhibitory feedback to suppress those (inharmonic) frequency components that do not belong to that pitch, allowing other spatial locations to capture that frequency component, and in turn, leading to complete resonance within the model.

The extended model is capable of replicating Deutsch (1975) scale illusion (Fig. 7), where a downward and an upward scale are played at the same time, except that every other tone in a given scale is presented to the opposite ear. The result is that listeners group based on frequency proximity, and hear a bounce percept. In order to understand qualitatively how the model can explain this phenomenon, one needs to recall that the model does not group based on spatial location, but instead, spatial location only primes the grouping based on the pitch process. For the first two simultaneous tones, high C presented to the left ear and a low C presented to the right ear, the left and right spatial locations become active, priming their corresponding pitch stream layers. This in turn causes the left stream to capture the high C tone and the right stream to capture the low C tone. For the next two simultaneous tones, a B presented to the right ear and a D presented to the left ear, both the left and right channels are still equally active, which causes both the left and right pitch stream layers to remain equally primed. Now due to frequency proximity in the spectral stream layer, the B will be grouped with the high C tone, and the D will be grouped with the low C tone. Thus, due to equal activation of the left and right spatial locations, grouping based on frequency proximity overcomes grouping based on spatial location. Similarly, the rest of the tones in the sequence will be grouped based on proximity, leading to the bounce percept.

8. Discussion

This paper neurally models aspects of the process that Bregman (1990) calls primitive auditory scene analysis. The model suggests how the brain segregates overlapping auditory components using pitch cues to create different coherent mental objects, or streams. The model is shown to qualitatively replicate listeners' percepts of hearing two streams for two inharmonic tones, variants of the auditory continuity illusion, bounce percepts for crossing glides even if the intersection point is replaced by silence or noise, and the 'M' and inverted 'V' percept for Steiger (1980) diamond stimulus even if the bifurcation points are replaced by noise.

The model is called an ARTSTREAM model because the core mechanisms that control the streaming process are specializations of Adaptive Resonance Theory, or ART, mechanisms (Carpenter & Grossberg, 1991; Grossberg, 1980, 1999b, 2003a; Grossberg & Stone, 1986; Raizada & Grossberg, 2003). These include the matching process which enables bottom-up energy inputs to activate spectral stream components in the absence of top-down pitch-activated inputs, top-down inputs to prime consistent spectral components in the absence of bottom-up energy inputs, and a confluence of bottom-up and top-down inputs to selectively amplify those harmonic spectral components that are consistent with the pitch, while inhibiting inconsistent spectral components. Rejected components are then freed to be represented by other streams, as in the 'old-plus-new heuristic' of Bregman (1990). After matching selects consistent components, the continued reciprocal action of bottom-up and top-down inputs generates a resonance that is hypothesized to give rise to an auditory percept. In many applications of ART, this resonance also creates the dynamical substrate for triggering adaptive tuning of the weights in the bottom-up and top-down pathways; hence the name *adaptive* resonance theory. The ART matching and resonance mechanisms have been proved to be capable of stabilizing this learning process in response to dynamically changing input patterns (Carpenter & Grossberg, 1987, 1991).

Bregman (1990) distinguishes primitive segregation mechanisms from higher-order processes that he calls schema-based segregation. Grossberg et al. (1997) and Grossberg and Myers (2000) have shown that psychophysical data about such a schema-based process, namely variable-rate speech categorization, can also be quantitatively modeled using ART matching and resonance rules; see Grossberg (2003b) for a review. On the other hand, auditory streaming and phonetic processes seem to have distinguishable properties. For example, streaming includes the setting up of spectral-pitch resonances, whereas phonetic processing generates (working memory)–(list chunk) resonances in a different part of the brain. Due to harmonic bottom-up and top-down filters that bind spectral components to pitch categories during auditory streaming (Figs. 8 and 9), the role of harmonics is more important during auditory streaming than during phonetic perception, as has been experimentally demonstrated by Remez, Pardo, Piorkowski, and Rubin (2001) and Remez, Rubin, Berns, Pardo, and Lang (1994). These examples provide convergent evidence that similar ART matching and resonance processes operate on multiple levels of the auditory system. These results extend other ART explanations of a variety of speech and word recognition data (Cohen & Grossberg, 1986 and Grossberg & Stone, 1986).

While the present model of primitive segregation is capable of qualitatively producing correct responses for the streaming stimuli mentioned above, the model needs to be further developed in order to emulate other streaming

phenomena. For example, the present version of ARTSTREAM does not contain transient onset or offset mechanisms to help create more sharply synchronized resonant onsets and offsets. As a result, the spectral layer decays slowly at the offset of a tone. In addition, onset and offset cues can influence the segregation process itself. For example, the continuity illusion of hearing a tone in noise can be destroyed by decreasing or increasing the amplitude of the tone at the onset or offset of the noise (Bregman, 1990; Bregman & Dannenbring, 1977). Another set of data that need further investigation demonstrate how the addition of harmonics can help overcome grouping by proximity. In particular, as in Fig. 5c, the addition of harmonics to one glide in a stimulus that consists of crossing ascending and descending glides can lead to a cross percept and not a bounce percept (Bregman, 1990). Using analog, rather than binary, winner-take-all, activations of pitch stream neurons should handle these cases by making the activity of pitch nodes covary with the number of harmonics that activate them.

Streaming percepts in music perception have been simulated by Gjerdingen (1994), who has exploited similarities between apparent motion in vision and streaming in audition. Gjerdingen notes that "a great deal of the motion perceived in music is apparent rather than real. On the piano, for example, no continuous movement in frequency occurs between two sequentially sounded tones. Though a listener may perceive a movement from the first tone to the second, each tone merely begins and ends at its stationary position on the frequency continuum" (p. 335). By applying Grossberg and Rudd (1989, 1992) model of visual apparent motion, Gjerdingen has simulated properties of the van Noorden (1975) melodic-fission/temporal-coherence boundary, various Gestalt effects involving musical phrasing and rhythm, aspects of dynamic attending, and the Narmour (1990) categorical distinction between those musical intervals that imply a continuation and those that imply a reversal of direction.

Why is visual apparent motion relevant to auditory streaming? In an apparent motion display, two successive flashes of light at different locations can cause a percept of continuous motion from the first flash to the second flash if their time delay and spatial separation fall within certain bounds (Kolers, 1972). A key mechanism that helps to simulate this percept in the Grossberg–Rudd model is Gaussian filtering of visual inputs across space followed by contrast-enhancing competition. If the input (flash) to one Gaussian wanes through time as the input (flash) to another waxes, then the sum of the Gaussian outputs has a maximum that moves continuously between the input locations if the Gaussians overlap sufficiently. In other words, a traveling wave of activity moves continuously from one location to the other. The contrast-enhancing competition spatially localizes the maximum activity as it moves across space. This Gaussian wave, or G-wave, has properties of apparent motion percepts in response to a variety of stimulus conditions.

In the acoustic domain, visual flashes are replaced by acoustic tones. Gaussian filtering of visual inputs across space followed by contrast-enhancing competition is replaced by Gaussian filtering of acoustic inputs across frequency followed by contrast-enhancing competition. For example, although an arpeggio is composed of temporally discrete tones, it leads to the perception of a continuous musical phrase, which Gjerdingen (1994) has compared with the properties of a G-wave. Such properties include the key fact that a G-wave can continuously link distinct tones whose relative timing is uniform but whose frequency separation is variable.

How do the Gaussian and contrast-enhancing properties needed to generate G-waves compare with properties of the ARTSTREAM model? Remarkably, these properties are already part of the spectral and pitch stream layers of the ARTSTREAM model; see Eqs. (18)–(20). Term E_{ip} describes the Gaussianly distributed kernel $M_{f,kp}$ across frequency. Term I_{ip} describes contrast-enhancing competition. Thus, the ARTSTREAM model, in its original form, already incorporates the key mechanisms for causing 'apparent motion' between successive tones. Within ARTSTREAM, these mechanisms are a manifestation of the need for harmonic grouping of frequency spectra into winning pitch representations.

Other relevant properties of the Grossberg–Rudd model are the use of transient cells that are sensitive to input onsets and offsets, and multiple spatial scales to cope with objects that move across space at variable speeds. In the acoustic domain, a movement across space at variable speeds is replaced by movement across frequencies with variable speed or spacing.

Chey, Grossberg, and Mingolla (1997, 1998) and Grossberg, Mingolla, and Viswanathan (2001) have built upon the Grossberg–Rudd model to explain more data about visual motion perception. The motion BCS model uses transient cells and multiple spatial scales to simulate human psychophysical data concerning the perceived speed and direction of moving objects. Analogous mechanisms can be naturally integrated into the ARTSTREAM model to explain directionally selective auditory streaming percepts (e.g. Bregman, 1990; Steiger & Bregman, 1981) as well as properties of directionally sensitive auditory neurons (Wagner & Takahashi, 1992). All the properties simulated by Gjerdingen (1994) should also be achievable with such an extended ARTSTREAM model when the Gaussians, transient cells, and multiple scales are combined.

Finally, no learning occurs presently within the ARTSTREAM model. Simulations of how an animal can learn during development to adaptively tune the harmonic sieves that abut its pitch stream representations remain to be carried out. Previous analyses of learning by ART networks provide helpful hints for how these bottom-up and top-down

learning processes may be regulated by resonant states of the brain.

Acknowledgements

Stephen Grossberg was supported in part by the Air Force Office of Scientific Research (AFOSR F49620-01-1-0397 and AFOSR F49620-92-J-0225) and the Office of Naval Research (ONR N00014-01-1-0624). Krishna K. Govindarajan was supported in part by the Advanced Research Projects Agency (ONR N00014-92-J-4015), the Air Force Office of Scientific Research (AFOSR F49620-92-J-0225), British Petroleum (BP 89A-1204), and the National Science Foundation NSF IRI-90-00530). Lonce L. Wyse was supported in part by the American Society for Engineering Education and by the Air Force Office of Scientific Research (AFOSR F49620-92-J-0225). Michael A. Cohen was supported in part by the Air Force Office of Scientific Research (AFOSR F49620-92-J-0225). The authors wish to thank Susanne Daley and Carol Y. Jefferson for their valuable assistance in the preparation of the manuscript.

References

von Bekesy, G. (1963). Hearing theories and complex sound. *Journal of the Acoustical Society of America, 35*, 588–601.

Bilsen, F., & Ritsma, R. (1970). Some parameters influencing the perceptibility of pitch. *Journal of the Acoustical Society of America, 47*, 469–475.

deBoer, E., & deJongh, H. R. (1978). On cochlear encoding: potentialities and limitations of the reverse correlation technique. *Journal of the Acoustical Society of America, 63*, 115–135.

Bregman, A. S. (1990). Auditory scene analysis: the perceptual organization of sound. Cambridge, MA: MIT Press.

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., & Dannenbring, G. (1977). Auditory continuity and amplitude edges. *Journal of Psychology, 31*, 151–159.

Bregman, A. S., & Doehring, P. (1984). Fusion of simultaneous tonal glides: the role of parallelness and simple frequency relations. *Perception and Psychophysics, 36*, 251–256.

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

Bregman, A. S., & Rudnicky, A. (1975). Auditory segregation: stream or streams? *Journal of Experimental Psychology: Human Perception and Performance, 1*, 263–267.

Bregman, A. S., & Steiger, H. (1980). Auditory streaming and vertical localization: interdependence of 'what' and 'where' decisions in audition. *Perception and Psychophysics, 28*, 539–546.

Broadbent, D. E., & Ladefoged, P. (1957). On the fusion of sounds reaching different sense organs. *Journal of the Acoustical Society of America, 29*, 708–710.

Brokx, J. P. L., & Noteboom, S. G. (1982). Intonation and the perceptual separation of simultaneous voices. *Journal of Phonetics, 10*, 23–26.

Brown, G. J (1992). *Computational auditory scene analysis: A representational approach.* PhD thesis, University of Sheffield.

Carlyon, R. P. (1991). Discriminating between coherent and incoherent frequency modulation of complex tones. *Journal of the Acoustical Society of America, 89*, 329–340.

Carlyon, R. P. (1992). The psychophysics of concurrent sound segregation. In R. P. Carlyon, C. J. Darwin, & I. J. Russell (Eds.), *Processing of complex sounds by the auditory system.* Oxford: Clarendon Press.

Carpenter, G. A., & Grossberg, S. (1987). A massively parallel architecture for a self-organizing neural pattern recognition machine. *Computer Vision, Graphics, and Image Processing, 37*, 54–115.

Carpenter, G. A., & Grossberg, S. (1991). Pattern recognition by self-organizing neural networks. Cambridge, MA: MIT Press.

Carpenter, G. A., & Grossberg, S. (1993). Normal and amnesic learning, recognition, and memory by a neural model of cortico-hippocampal interactions. *Trends in Neurosciences, 16*, 131–137.

Chalika, M. H., & Bregman, A. S. (1989). The perceptual segregation of simultaneous auditory signals: pulse train segregation and vowel segregation. *Perception and Psychophysics, 46*, 487–497.

Chey, J., Grossberg, S., & Mingolla, E. (1997). Neural dynamics of motion grouping: from aperture ambiguity to object speed and direction. *Journal of the Optical Society of America A, 14*, 2570–2594.

Chey, J., Grossberg, S., & Mingolla, E. (1998). Neural dynamics of motion processing and speed discrimination. *Vision Research, 38*, 2769–2786.

Cohen, M. A., & Grossberg, S. (1986). Neural dynamics of speech and language coding: developmental programs, perceptual grouping, and competition for short term memory. *Human Neurobiology, 5*, 1–22.

Cohen, M. A., Grossberg, S., & Wyse, L. (1995). A spectral network model of pitch perception. *Journal of the Acoustical Society of America, 98*, 862–879.

Colburn, H. S. (1973). Theory of binaural interaction based on auditory-nerve data. I. General strategy and preliminary results on interaural discrimination. *Journal of the Acoustical Society of America, 54*, 1458–1470.

Colburn, H. S. (1977). Theory of binaural interaction based on auditory-nerve data. II. Detection of tones in noise. *Journal of the Acoustical Society of America, 61*, 525–533.

Cooke, M. P (1991). *Modelling auditory processing and organisation.* PhD thesis, University of Sheffield.

Dannenbring, G. L., & Bregman, A. S. (1978). Streaming vs. fusion of sinusoidal components of complex waves. *Perception and Psychophysics, 24*, 369–376.

Darwin, C. J. (1984). Perceiving vowels in the presence of another sound: constraints on formant perception. *Journal of the Acoustical Society of America, 76*, 1636–1647.

Darwin, C. J., & Bethell-Fox, C. E. (1977). Pitch continuity and speech source attribution. *Journal of Experimental Psychology: Human Perception and Performance, 3*, 665–672.

Darwin, C. J., & Ciocca, V. (1992). Grouping in pitch perception: effects of onset asynchrony and ear of presentation of a mistuned component. *Journal of the Acoustical Society of America, 91*, 3381–3390.

Darwin, C. J., & Sutherland, N. S. (1984). Grouping frequency components of vowels: when is a harmonic not a harmonic? *The Quarterly Journal of Experimental Psychology, 36A*, 193–208.

Deutsch, D. (1975). Two-channel listening to musical scales. *Journal of the Acoustical Society of America, 57*, 1156–1160.

Deutsch, D. (1992a). Paradoxes of musical pitch. *Scientific American, 264*, 88–95.

Deutsch, D. (1992b). Some new pitch paradoxes and their implications. *Philosophical Transactions of the Royal Society of London, 336*, 391–397.

Duifhuis, H., Willems, L. F., & Sluyter, R. (1982). Measurement of pitch in speech: an implementation of Goldstein's theory of pitch perception. *Journal of the Acoustical Society of America, 71*, 1568–1580.

Gardner, R. B., Gaskill, S. A., & Darwin, C. J. (1989). Perceptual grouping of formants with static and dynamic differences in fundamental frequency. *Journal of the Acoustical Society of America, 85*, 1329–1337.

Gjerdingen, R. O. (1994). Apparent motion in music? *Music Perception, 11*, 335–370.

Goldstein, J. (1973). An optimum processor theory for the central formation of the pitch of complex tones. *Journal of the Acoustical Society of America, 54*, 1496–1515.

Grossberg, S. (1976). Adaptive pattern classification and universal recoding. II. Feedback, expectation, olfaction, and illusions. *Biological Cybernetics, 23*, 187–202.

Grossberg, S. (1980). How does a brain build a cognitive code? *Psychological Review, 87*, 1–51.

Grossberg, S. (1999a). How does the cerebral cortex work? Learning, attention, and grouping by the laminar circuits of visual cortex. *Spatial Vision, 12*, 163–185.

Grossberg, S. (1999b). The link between brain learning, attention, and consciousness. *Consciousness and Cognition, 8*, 1–44.

Grossberg, S. (2003a). How does the cerebral cortex work? Development, learning, attention, and 3D vision by laminar circuits of visual cortex. *Behavioral and Cognitive Neuroscience Reviews*, in press.

Grossberg, S. (2003b). Resonant neural dynamics of speech perception. *Journal of Phonetics*, in press.

Grossberg, S., Boardman, I., & Cohen, M. A. (1997). Neural dynamics of variable-rate speech categorization. *Journal of Experimental Psychology: Human Perception and Performance, 23*, 481–503.

Grossberg, S., Mingolla, E., & Viswanathan, L. (2001). Neural dynamics of motion integration and segmentation within and across apertures. *Vision Research, 41*, 2521–2553.

Grossberg, S., & Myers, C. (2000). The resonant dynamics of conscious speech: interword integration and duration-dependent backward effects. *Psychological Review, 107*, 735–767.

Grossberg, S., & Rudd, M. E. (1989). A neural architecture for visual motion perception: group and element apparent motion. *Neural Networks, 2*, 421–450.

Grossberg, S., & Rudd, M. E. (1992). Cortical dynamics of visual motion perception: short-range and long-range apparent motion. *Psychological Review, 99*, 78–121.

Grossberg, S., & Stone, G. O. (1986). Neural dynamics of word recognition and recall: attentional priming, learning, and resonance. *Psychological Review, 93*, 46–74.

Hall, J. W., & Grose, J. H. (1988). Comodulation masking release: evidence for multiple cues. *Journal of the Acoustical Society of America, 84*, 1669–1675.

Halpern, L (1977). *The effect of harmonic ratio relationships on auditory stream segregation*. Technical report, Department of Psychology, McGill University.

Handel, S. (1989). Listening. Cambridge, MA: MIT Press.

Kolers, P. A. (1972). Aspects of motion perception. Elmsford, NY: Pergamon Press.

McAdams, S. (1989). Segregation of concurrent sounds. I. Effects of frequency modulation coherence. *Journal of the Acoustical Society of America, 86*, 2148–2159.

Meddis, R., & Hewitt, M. J. (1992). Modelling the identification of concurrent vowels with different fundamental frequencies. *Journal of the Acoustical Society of America, 91*, 233–245.

Miller, G. A., & Licklider, J. C. R. (1950). Intelligibility of interrupted speech. *Journal of the Acoustical Society of America, 22*, 167–173.

Moore, B. C. J. (1977). Effects of relative phase of the components on the pitch of three-component complex tones. In E. Evans, & J. Wilson (Eds.), *Psychophysics and physiology of hearing*. New York: Academic Press.

Moore, B. C. J., Glasberg, B. R., & Peters, R. W. (1985). Relative dominance of individual partials in determining the pitch of complex tones. *Journal of the Acoustical Society of America, 77*, 1853–1860.

Narmour, E. (1990). *The analysis and cognition of basic melodic structures: The implication–realization model*. Chicago: University of Chicago Press.

van Noorden, L. P. A. S (1975). *Temporal coherence in the perception of tone sequences*. PhD thesis, Eindhoven University of Technology.

Patterson, R., & Wightman, F. (1976). Residue pitch as a function of component spacing. *Journal of the Acoustical Society of America, 59*, 1450–1459.

Plomp, R. (1967). Pitch of complex tones. *Journal of the Acoustical Society of America, 41*, 1526–1533.

Raizada, R. D. S., & Grossberg, S. (2003). Towards a theory of the laminar architecture of cerebral cortex: computational clues from the visual system. *Cerebral Cortex, 13*, 100–113.

Remez, R. E., Pardo, J. S., Piorkowski, R. L., & Rubin, P. E. (2001). On the bistability of sine wave analogues of speech. *Psychological Science, 12*, 24–29.

Remez, R. E., Rubin, P. E., Berns, S. M., Pardo, J. S., & Lang, J. M. (1994). On the perceptual organization of speech. *Psychological Review, 101*, 129–156.

Ritsma, R. (1967). Frequencies dominant in the perception of the pitch of complex sounds. *Journal of the Acoustical Society of America, 42*, 191–198.

Ritsma, R., & Engel, F. (1964). Pitch of frequency-modulated signals. *Journal of the Acoustical Society of America, 36*, 1637–1644.

Scheffers, M. T. M (1983). *Sifting vowels: Auditory pitch analysis and sound segregation*. PhD thesis, Groningen University.

Schouten, J., Ritsma, R., & Cardozo, B. (1962). Pitch of the residue. *Journal of the Acoustical Society of America, 34*, 1418–1424.

Shackleton, T. M., Meddis, R., & Hewitt, M. J. (1994). The role of binaural and fundamental frequency difference cues in the identification of concurrently presented vowels. *The Quarterly Journal of Experimental Psychology, 47A*, 545–563.

Shepard, R. (1964). Circularity in judgments of relative pitch. *Journal of the Acoustical Society of America, 36*, 2346–2353.

Small, A. M., & Daniloff, R. G. (1967). Pitch of noise bands. *Journal of the Acoustical Society of America, 41*, 506–512.

Steiger, H (1980). *Some informal observations concerning the perceptual organization of patterns containing frequency glides*. Technical report, McGill University, Montreal.

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

Summerfield, Q. (1992). Roles of harmonicity and coherent frequency modulation in auditory grouping. In M. E. H. Schouten (Ed.), *Audition, speech, and language*. Berlin: Mouton.

Summerfield, Q., & Culling, J. F. (1992). Auditory segregation of competing voices: Absence of effects of fm or am coherence. In R. P. Carlyon, C. J. Darwin, & I. J. Russel (Eds.), *Processing of complex sounds by the auditory system*. Oxford: Clarendon Press.

Terhardt, E. (1972). Zur Tonhohenwahrnehmung von Klangen. *Acustica, 26*, 173–199.

Tougas, Y., & Bregman, A. S. (1990). The crossing of auditory streams. *Journal of Experimental Psychology: Human Perception and Performance, 11*, 788–798.

Wagner, H., & Takahashi, T. (1992). Influence of temporal cues on acoustic motion–direction sensitivity of auditory neurons in the owl. *Journal of Neurophysiology, 68*, 2063–2076.

Yost, W., Hill, R., & Perez-Falcon, T. (1978). Pitch and pitch discrimination of broadband signals with rippled power spectra. *Journal of the Acoustical Society of America, 63*, 1166–1173.