1996-02

Pitch-based Streaming in Auditory Perception

PITCH-BASED STREAMING IN AUDITORY PERCEPTION

Stephen Grossberg

February 1996
Revised: June 1996
Revised: July 1996

Technical Report CAS/CNS-96-007

To appear in:
Niall Griffith and Peter Todd (Eds.)
Musical Networks: Parallel Distibuted Perception and Performance
Cambridge, MA: MIT Press, 1996

Boston University Center for Adaptive Systems and
Department of Cognitive and Neural Systems
677 Beacon Street
Boston, MA 02215

PITCH-BASED STREAMING IN AUDITORY PERCEPTION

Stephen Grossberg†
Department of Cognitive and Neural Systems
and
Center for Adaptive Systems
Boston University
677 Beacon Street
Boston, MA 02215

To appear in
Niall Griffith and Peter Todd (Editors)
Musical Networks: Parallel Distributed Perception and Performance
Cambridge, MA: MIT Press, 1996

February 1996
Revised: June 1996
Revised: July 1996

Technical Report CAS/CNS-TR-96-007
Boston, MA: Boston University

† Supported in part by the Air Force Office of Scientific Research (AFOSR F49620-92-J-0225), the Advanced Research Projects Agency (ONR N00014-92-J-4015), and the Office of Naval Research (ONR N00014-95-1-0409).

Acknowledgments: The author wishes to thank Cynthia E. Bradford and Diana Meyers for their valuable assistance in the preparation of this manuscript.

ABSTRACT

This chapter summarizes a neural model of how humans use pitch-based information to separate and attentively track multiple voices or instruments in distinct auditory streams, as in the cocktail party problem. The model incorporates concepts of top-down matching, attention, and resonance that have been used to analyse how humans can autonomously learn and stably remember large amounts of information in response to a rapidly changing environment. These Adaptive Resonance Theory, or ART, concepts are joined to a Spatial PItch NETwork, or SPINET, model to form an ARTSTREAM model for pitch-based streaming. The ARTSTREAM model suggests that a resonance between spectral and pitch representations is necessary for a conscious auditory percept to occur. Examples from auditory perception in noise and context-sensitive speech perception are discussed, such as the auditory continuity illusion and phonemic restoration. The Gjerdingen analysis of apparent motion in music is shown to have a natural embedding within the ARTSTREAM model.

Auditory Streaming, Pitch Perception, and Music Perception

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. How do we separate this jumbled mixture of sounds into distinct voices? This 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 voices into distinct sources, or auditory streams, then we could not hear the music as music, or intelligently recognize a speaker's sounds. A major cue for separating sounds into distinct sources is their pitch (Bregman, 1990). Thus, in order to understand how music is perceived, we need to understand how the pitch of a sound is determined and how different sources of sound are separated into distinct auditory streams.

A simple version of this competence is illustrated by the auditory continuity illusion (Miller and Licklider, 1950). This percept also calls attention to some remarkable properties of the events that lead to a conscious perception of music, or of any other sound. 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 Figure 1a. When this happens under appropriate conditions, 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 Figure 1b. Then the first tone does not seem to continue through the noise. It is perceived to stop before the noise. In Figure 1a, the second tone turns on only after the first tone and the subsequent noise turn off. How does the brain use the information about a future event, the second tone, to continue the first tone through the noise? 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?

The reality of this problem is emphasized by the third condition: If no noise occurs between two temporally disjoint tones, as in Figure 1c, then the tone is not heard across the silent interval. Instead, two temporally disjoint tones are heard. This fact raises the

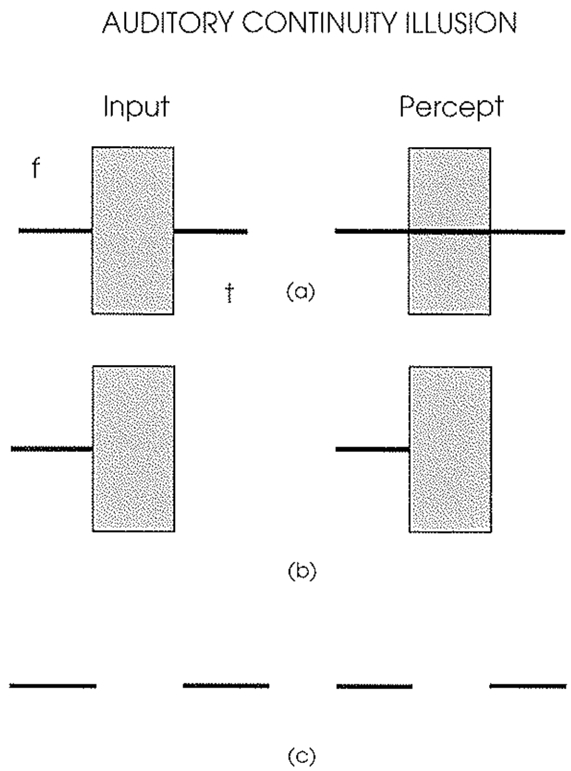

Figure 1. (a) Auditory continuity illusion: When a steady tone occurs both before and after a burst of noise, then under appropriate temporal and amplitude conditions, the tone is perceived to continue through the noise. (b) This does not occur if the noise is not followed by a tone. (c) Nor does it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

additional 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. I will argue that 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?

Phonemic Restoration, Attentive Matching, and Adaptive Resonance

I suggest that the neural mechanisms whereby auditory streaming is achieved are used, in specialized form, in other brain systems as well. Another example from the auditory system operates at a higher level of processing. It concerns how we understand speech. In this example, too, the process whereby conscious awareness occurs takes a long time, on the order of 100 milliseconds or more. The phenomenon in question is called phonemic restoration (Samuel, 1981; Warren, 1984; Warren and Sherman, 1974). Suppose that a listener hears a noise followed immediately by the words "eel is on the ...". If this string of words is followed by the word "orange", then "noise-eel" sounds like "peel". If the word "wagon" completes the sentence, then "noise-eel" sounds like "wheel". If the final word is "shoe", then "noise-eel" sounds like "heel".

This example vividly shows that the bottom-up occurrence of the noise is not sufficient for us to hear it. Somehow the sound that we *expect* to hear based upon our previous language experiences influences what we do hear. Such an expectation takes time to influence the speech that we consciously hear. As in the auditory continuity illusion, the brain works "backwards in time" to allow the meaning imparted by a later word to alter the sounds that we consciously perceive in an earlier word.

I suggest that this happens because, as the individual words occur, they are stored temporarily in a working memory. The working memory converts a temporal sequence of events into a spatial pattern of activation across the items that represent each word. A similar recoding enables musical phrases to be stored. As the items of the words are stored, they activate previously learned memories which attempt to categorize the stored sound stream into familiar language units at a higher processing level. Such learned categories encode abstract lists of items that may include the words themselves, their syllables, or even their phonemes (Cohen and Grossberg, 1986; Grossberg, 1984, 1987). Which list categories are chosen depends upon the temporal context in which all the sounds occur, whether they are the sounds of language or of music. The list category layer is designed to activate those groupings of working memory items that are most predictive in the context within which they appear.

The list categories, in turn, activate learned top-down expectations that are matched against the items stored in working memory to verify that the information expected from previous learning experiences is really there. This concept of bottom-up activation of learned categories by a working memory, followed by read-out of learned top-down expectations, is illustrated in Figure 2a.

What is the nature of this matching, or verification, process? Its properties have been clarified by experiments in which the spectral content of the noise was varied (Samuel, 1981).

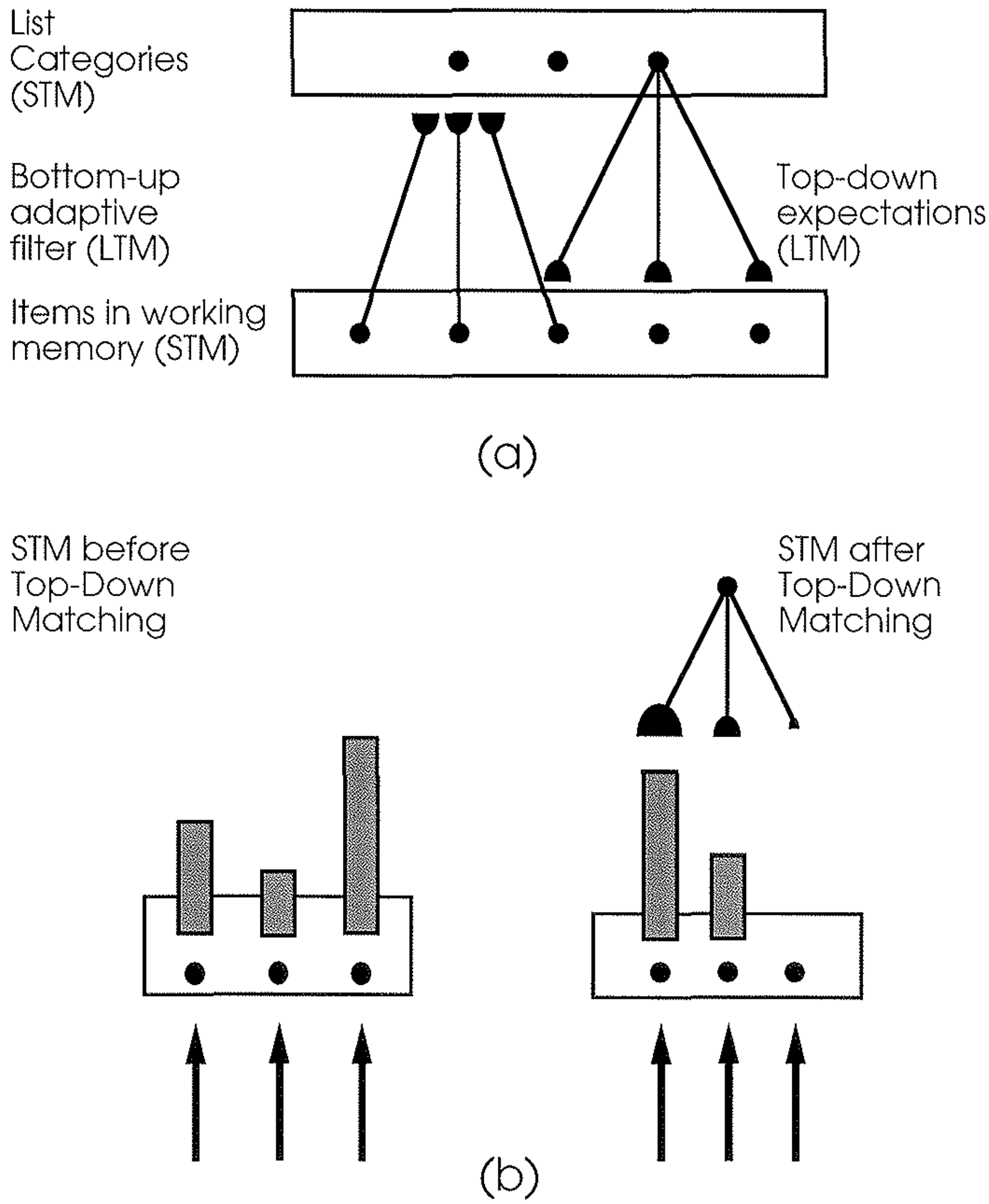

Figure 2. ART matching: (a) Auditory items activate short term memory (STM) traces in a working memory, which send bottom-up signals towards a level at which list categories, or chunks, are activated in STM. These bottom-up signals are multiplied by learned long term memory (LTM) traces which influence the competitive selection of the list categories that are stored in STM. The list categories, in turn, activate top-down expectation signals that are also read out of LTM. These expectations, or prototypes, are matched against the active STM pattern in working memory. (b) This matching process selects STM activations that are supported by contiguous LTM traces, and suppresses those that are not.

If the noise includes all the formants of the expected sound, then that is what the subject hears, and other spectral components of the noise are suppressed. If some formants of the expected sound are missing from the noise, then only a partial reconstruction is heard. If silence replaces the noise, then only silence is heard. The matching process thus cannot "create something out of nothing". It can select the expected features that are represented in the bottom-up signal and suppress the rest, as in Figure 2b.

The process whereby the top-down expectation selects some features while suppressing others helps to "focus attention" upon information that matches our momentary expectations. By filtering out the flood of irrelevant sensory signals, expectations prevent these signals from destabilizing previously learned memories (Carpenter and Grossberg, 1991; Grossberg, 1980, 1982).

What does all this have to do with our conscious percepts of speech and music? This can be seen by asking: If top-down expectations can select consistent bottom-up signals using an attentional focus, then what keeps the attended bottom-up signals from reactivating their top-down expectations in a continuing cycle of bottom-up and top-down feedback? Nothing does! In fact, this reciprocal feedback process takes awhile to equilibrate, and when it does, the bottom-up and top-down signals lock the activity patterns of the interacting levels into a resonant state that lasts much longer and is more energetic than any individual activation. I claim that only resonant states of the brain can achieve consciousness, and that the time needed for a bottom-up/top-down resonance to develop helps to explain why a conscious percept of an event takes so long to occur after its bottom-up input is delivered.

Adaptive Resonance Theory, or ART, is a cognitive theory that was introduced to explain how the brain continues to rapidly learn about the world throughout life without undergoing catastrophic forgetting (Carpenter and Grossberg, 1991, 1993; Grossberg, 1980, 1982, 1987). ART models how top-down expectations are learned and help to focus attention in the manner described above in order to ensure that learning can proceed in a stable fashion throughout life. A key result of ART is that only resonant states trigger the learning process - hence the name *adaptive* resonance – and that all conscious states are resonant states. Thus the properties of conscious audition that we are discussing may be viewed as special cases of how each brain can effectively learn about its world.

The same types of properties may now be seen to hold in the auditory continuity illusion. The first main point is that bottom-up activation by the tone is not immediately perceived.

A bottom-up/top-down resonance first needs to develop. This slower resonance time scale helps to explain why the tone continues to be heard even after the noise input begins. From here it is not hard to see how the second tone in Figure 1a can quickly access the already active tone resonance to keep it going through the percept of the noise, which also takes awhile to develop. All the percepts are hereby shifted in time relative to the onset times of their inputs.

The type of top-down matching in the auditory continuity illusion is also similar to that in phonemic restoration. An active categorical representation of the tone, as in Figure 1a, can use its top-down expectation to select those frequency components in the noise that are compatible with it and to suppress the rest. The selected frequency components can then resonate with their category until the percept of the tone becomes conscious.

This summary clarifies some properties of the auditory continuity percept but also raises new questions as it does so. For example, what is the "expectation" against which a sound, like the tone, is matched? How do we hear the noise as a separate perceptual stream from the tone? In a more general cocktail party or concert hall situation, how do we hear multiple voices or instruments? What are the rules whereby multiple streams of sound are simultaneously heard, even as each stream selectively suppresses the spectral components that do not belong to its source using top-down expectations?

Pitch Cues for Streaming

Perhaps the most important cue for perceptual streaming is the pitch of a sound. Naturally occurring periodic sources often have harmonic frequency components at integer multiples of the fundamental frequency, F_0. The subjective experience of F_0 describes the sound's pitch. For example, when a speaker produces a vowel at a particular fundamental frequency, (e.g., 150 Hz.), the vowel contains harmonics at integer multiples (e.g., 300, 450, 600 Hz., etc.), whose pattern of relative amplitudes corresponds to the vowel percept. Since such a set of related harmonics typically come from the same sound source, a categorical representation of pitch can be used to group the corresponding harmonic components together.

Pitch-based grouping is used by listeners in both speech and music perception. For example, 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 .5–2 semitones (Scheffers, 1983). When the two F_0 are chosen an octave apart, identification is poor (Brokx and Noteboom, 1982; Chalika and Bregman, 1989). Since an octave corresponds to a doubling of frequency, half the harmonics for the two vowels will overlap. In addition, a speech formant may become segregated from the vowel in which it occurs when the formant has a different F_0 (Broadbent and Ladefoged, 1957; Gardner, Gaskill, and Darwin, 1989) and speech stimuli with discontinuous pitch contours tend to segregate at the discontinuities (Darwin and Bethell-Fox, 1977).

A Neural Model of Pitch Perception and Auditory Streaming

The present chapter summarizes a model of how humans perceive pitch-based auditory streams. This model includes a specialized filter which inputs to a grouping network. The filter is a **S**patial **PI**tch **NET**work, or SPINET model, that models how the brain converts temporal streams of sound into spatial representations of pitch (Cohen, Grossberg, and Wyse, 1995). The grouping network is a specialized ART network that breaks sounds into separate streams based upon their pitch. The model wherein an ART streaming network is joined to a SPINET front end is called the ARTSTREAM model (Govindarajan, Grossberg, Wyse, and Cohen, 1994). This model was developed 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. Then a circuit for ART matching and resonance will be described and incorporated into the ARTSTREAM model, whose operation will be illustrated by a simulation of streaming. Finally it will be suggested how the Gjerdingen (1994) analysis of streaming percepts in music, which was based upon the motion perception model of Grossberg and Rudd (1989, 1992), can be incorporated into ARTSTREAM. Gjerdingen's analysis quantifies aspects of the analogy between visual and 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.

The SPINET Model

The SPINET model (Cohen, Grossberg, and Wyse, 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. Figure 3 shows the main processing stages of the SPINET model. A key goal of the model is to transform a spectral representation of an acoustic source into

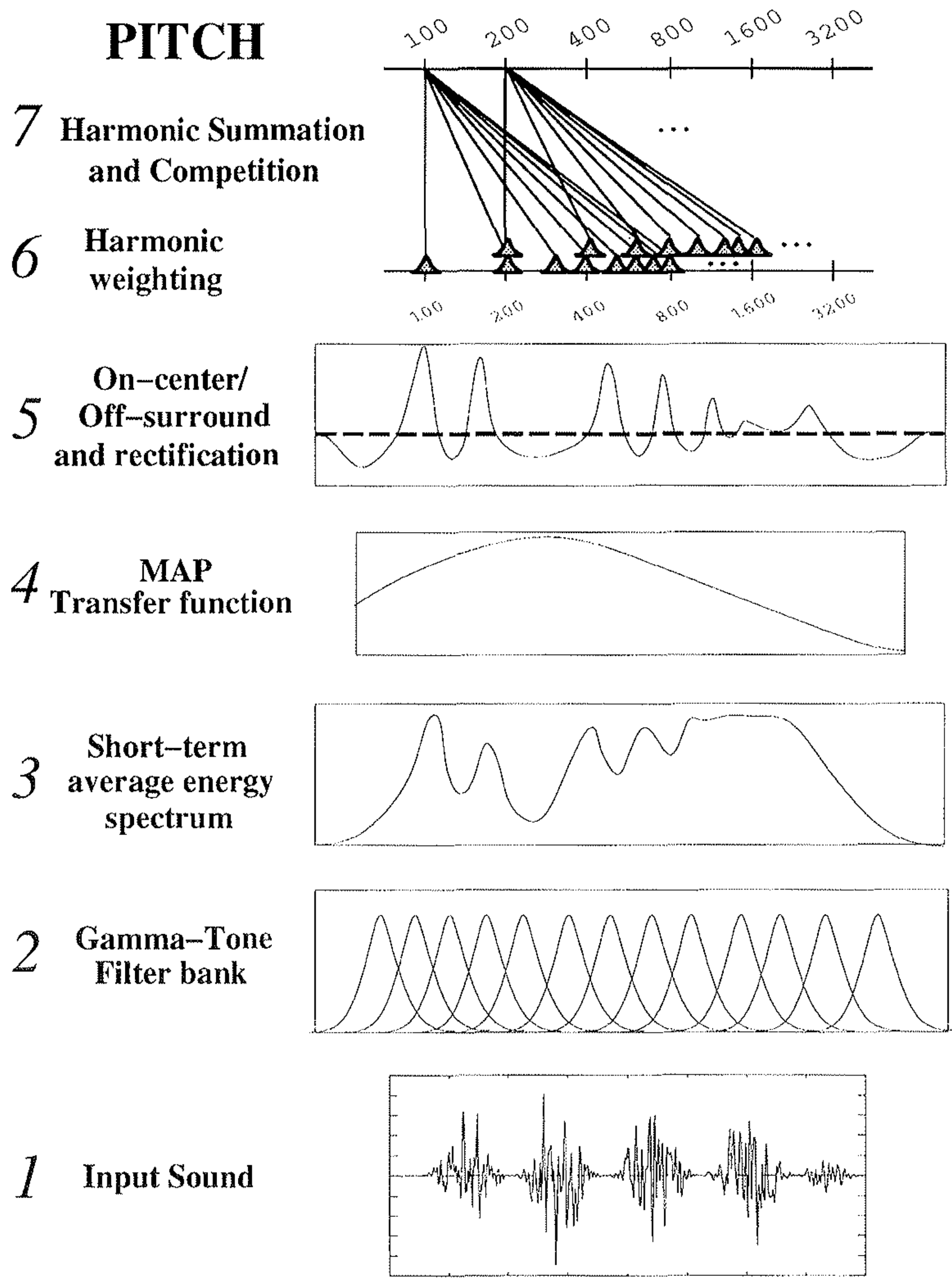

Figure 3. SPINET model processing stages. See the Appendix for more details. [Reprinted with permission from Cohen, Grossberg, and Wyse (1995).]

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 SPINET model preprocesses sounds at Levels 1-5 in order to generate a spectral representation of sound across a spatial array of frequency-tuned cells at Level 6. The spatial interactions from the spectral representation of Level 6 to the pitch representation of Level 7 are critical for our analysis of pitch perception and streaming. These interactions show that SPINET is a type of pattern matching model, a class that also includes the pitch models of Goldstein (1973) and Wightman(1973). Each possible pitch samples regions of the spectrum with a sampling period equal to the pitch frequency. That is, a region around nf_0, for integers n and fundamental frequency f_0, contributes to the strength of the pitch percept at frequency f_0. The weighting function for the region is Gaussian and symmetric in log frequency space (Figure 3), causing the resolution of the filter to scale with frequency.

In all, these interactions define a weighted "harmonic sieve" whereby the strength of activation of a given pitch depends upon a weighted sum of narrow regions around the harmonics of the nominal pitch value, and higher harmonics contribute less to a pitch than lower ones (Duifhuis, Willems, and Sluyter, 1982; Goldstein, 1973; Scheffers, 1983; Terhardt, 1972). Suitably chosen harmonic weighting functions support computer simulations of pitch perception data involving mistuned components (Moore *et al.*, 1985), shifted harmonics (Patterson and Wightman, 1976; Schouten, Ritsma, and Cardozo, 1962), and various types of continuous spectra including rippled noise (Bilsen and Ritsma, 1970; Yost, Hill, and Perez-Falcon, 1978). The weighting functions also produce the dominance region (Plomp, 1967; Ritsma, 1967), octave shifts of pitch in response to ambiguous stimuli (Patterson and Wightman, 1976; Schouten, Ritsma, and Cardozo, 1962), and how they lead to a pitch region in response to the octave-spaced Shepard tone complexes and Deutsch tritones (Deutsch, 1992a, 1992b; Shepard, 1964) without the use of attentional mechanisms to limit pitch choices. The on-center off-surround network in the model (Level 5) helps to produce noise suppression, partial masking and edge pitch (von Békésy, 1963). The model's peripheral filtering and short term energy measurements (Levels 2-4) produce pitch estimates that are sensitive to certain component phase relationships (Ritsma and Engel, 1964; Moore, 1977).

Figure 4b compares an illustrative computer simulation with pitch data in Figure 4a 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

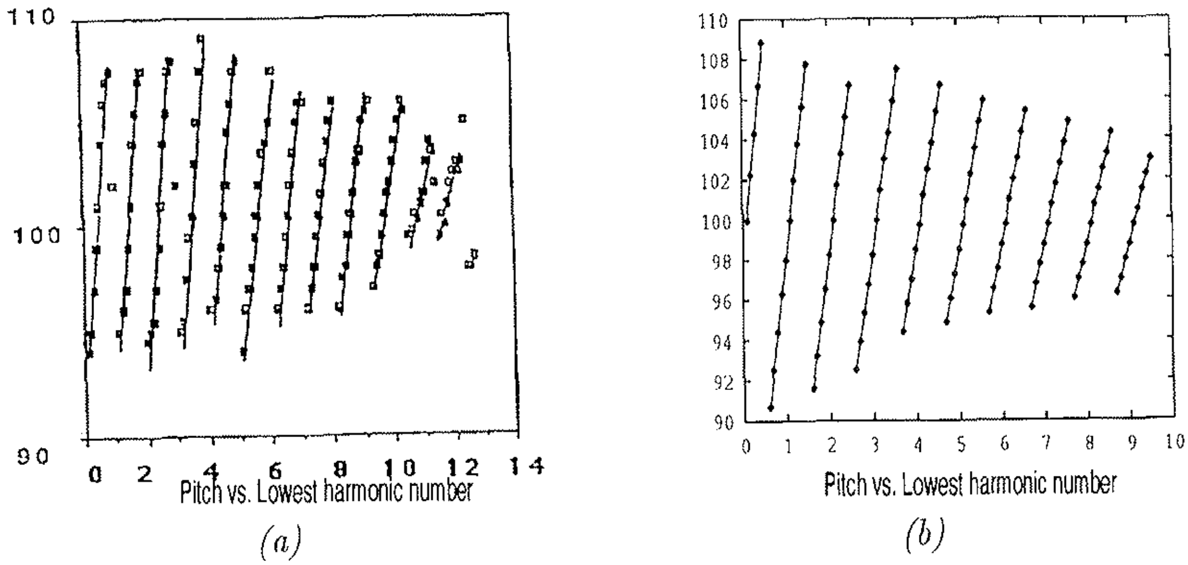

Figure 4. Pitch shift in response to a complex of 6 components spaced by 100 Hz, as a function of the lowest component's harmonic number. (a) Data from Patterson and Wightman (1976). (b) Maximally activated pitch produced by the network model. See text for details. [Reprinted with permission from Cohen, Grossberg, and Wyse (1995).]

frequency so that they maintain their spacing of f_0, ($f_n = nf_0 + \Delta, n = 1, ...$), the pitch shift in linear frequency is slower than that of the components (Patterson and Wightman, 1976; Schouten, Ritsma, and Cardozo, 1962). The data exhibit an ambiguous pitch region at shift values of $\Delta = lf_0$, $l = .5, 1.5, 2.5, ...$ where the most commonly perceived pitch jumps down to below the value of f_0. Figure 4a shows the pitch of components spaced by $f_0 = 100Hz$ as a function of the lowest component's harmonic number, l. When the shift value Δ is near a harmonic of f_0 ($\Delta = lf_0$, $l = 0, 1, 2, ...$), then the pitch is unambiguous and near 100 Hz.

The model simulates these data in Figure 4b 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 Figure 3). 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 Figure 4.

As indicated above, 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 Figure 3 filters each frequency spectrum through a harmonic sieve that transforms logarithmically scaled and Gaussian-weighted harmonic components into activations of pitch cells 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.

ART Matching and Resonance in ARTSTREAM

In particular, the ARTSTREAM model incorporates all the stages of the SPINET model, as shown in Figure 5, but also elaborates them into multiple spectral and pitch layers that are capable of representing multiple streams of sound. As in the SPINET model, each of the bottom-up filters from spectral to pitch layers forms a harmonic sieve. In addition, there are top-down filters that also form harmonic sieves and satisfy ART matching constraints. This is how top-down signals can select those spectral components that are compatible with the chosen pitch node, while suppressing all other frequencies that may have initially activated that spectral stream layer.

These ART matching circuits satisfy the following constraints:

Bottom-Up Automatic Activation: A cell, or node, 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 become sensitized, or subliminally active, without generating output signals, if it receives only a large top-down expectation input. Such

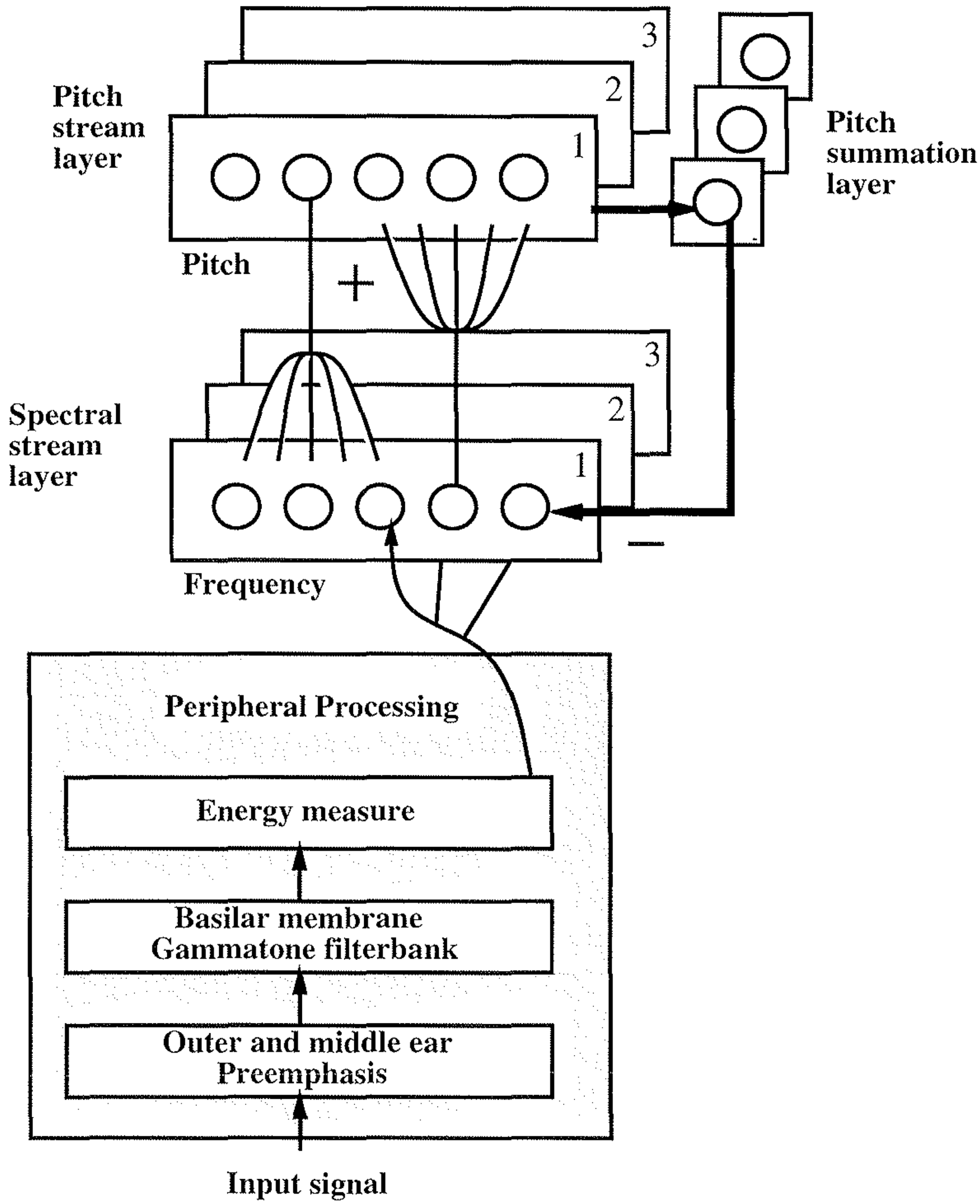

Figure 5. Block diagram of the ARTSTREAM auditory streaming model. Note the nonspecific top-down inhibitory signals from the pitch level to the spectral level that realize ART matching within the network. [Reprinted with permission from Govindarajan *et al.* (1994).]

a top-down signal prepares, or primes, 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 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.

Figure 6 illustrates perhaps the simplest way that the ART matching rule can be realized. Figure 5 embeds this circuit into multiple copies of the spectral and pitch layers. By this scheme, bottom-up signals to the spectral stream level can excite their frequency-sensitive cells if top-down signals are not active. Top-down signals try to excite those spectral nodes that are consistent with the pitch node that activates them. By themselves, top-down signals fail to activate spectral nodes because the pitch node also activates a pitch summation layer that nonspecifically inhibits all spectral nodes in its stream. The nonspecific top-down inhibition hereby prevents the specific top-down excitation from supraliminally activating any spectral nodes. On the other hand, when excitatory bottom-up and top-down signals occur together, then those spectral nodes that receive both types of signals can be fully activated. All other nodes in that stream are inhibited, including spectral nodes that were previously activated by bottom-up signals but received no subsequent top-down pitch support. Attention hereby selectively activates consistent nodes while nonselectively inhibiting all other nodes in a stream. Because the top-down signals form a (fuzzy) harmonic sieve, only spectral components that are (nearly) harmonically related to the active pitch node can survive a top-down match.

Resonant Dynamics During Auditory Streaming

Resonant processing is used in the ARTSTREAM model to help explain separation of distinct voices or instruments into auditory streams, as in the auditory continuity illusion of Figure 1. In the ARTSTREAM model, sounds are grouped into streams at the spectral and pitch stream levels, as in Figure 5. After the auditory signals are preprocessed by SPINET mechanisms, the active spectral, or frequency, components are redundantly represented in

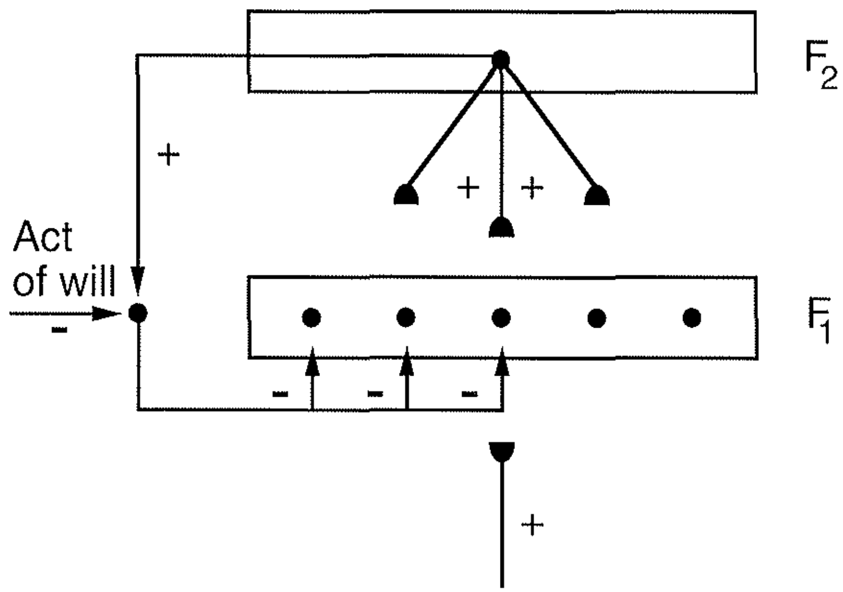

Figure 6. One way to realize the ART matching rule between two successive processing levels (a feature level F_1 and a category level F_2) using top-down activation of nonspecific inhibitory interneurons, as in Figure 5. In Figure 5, the feature level codes spectra and the category level codes pitches. In this circuit, an "act of will" can shut off top-down inhibition, thereby enabling internally generated fantasy activities, such as hearing a familiar tune in your head, to occur. Several mathematically possible alternative ways are suggested in the Appendix of Carpenter and Grossberg (1987). See Grossberg (1995) for other applications of this rule in auditory and visual perception.

multiple spectral streams. These streams are then filtered by bottom-up harmonic sieve signals that activate multiple representations of the sound's pitch at the pitch stream level. These pitch representations compete across streams to select a winner, which inhibits the redundant representations of the same pitch across streams. The winning pitch node also sends matching signals through its top-down harmonic sieve 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 binds together the frequency components that correspond to a prescribed auditory source. The selected frequency components inhibit redundant representations of the same frequency across steams, thereby achieving a type of exclusive allocation (Bregman, 1990). In addition, all the frequency components that are suppressed by ART matching in the resonant spectral stream are freed to activate and resonate with a different pitch in a different stream, thereby realizing a type of old-plus-new heuristic (Bregman, 1990). The net result is multiple resonances, each selectively grouping together into pitches those frequencies that correspond to distinct auditory sources.

Figure 7 depicts in greater detail the balance of excitatory and inhibitory interactions within and between the spectral and pitch stream layers that enables multiple streams to capture their own frequency components and inhibit their redundant activation within other streams, while freeing other components to resonate in these streams.

Using the ARTSTREAM model, Govindarajan *et al.* (1994) have simulated a number of basic streaming percepts, including those in Figure 8. The percept summarized in Figure 8c is the auditory continuity illusion. It occurs, I contend, because the spectral-stream resonance takes awhile to develop that is commensurate to the duration of the subsequent noise. Once the tone resonance does develop, the second tone can quickly act through bottom-up signaling to support and maintain it throughout the duration of the noise. ART matching selects the tone from the noise and the interstream competitive interactions enable the noise to be captured by different streams. Of course, for this to make sense, one needs to accept the fact that the tone resonance does not start to get consciously heard until well after the noise begins.

Simulation of the Auditory Continuity Illusion

Model dynamics are illustrated by a computer simulation of the auditory continuity illusion, whereby continuation of a tone occurs in noise, even though the tone is not physically present in the noise (Miller and Licklider, 1950). In addition, for a tone-silence-tone stimulus (Figure 8b), the tone should not continue across the silence, but should stop near the onset of silence. Figure 9 shows the simulated spectrogram and the resulting spectral layer and pitch layer activities for the tone-silence-tone stimulus for the selected stream (numbered 1). and for an unselected stream (numbered 2). The figures show that the first stream captures the tone, but does not remain active in the silent interval. An acoustic percept is assumed to occur when there is a spectral-pitch resonance that supports activity in the spectral stream. Thus the tone is not perceived within the model to fill the silent interval. The same stream then captures the tone after the silence as well. The second stream is not active since there

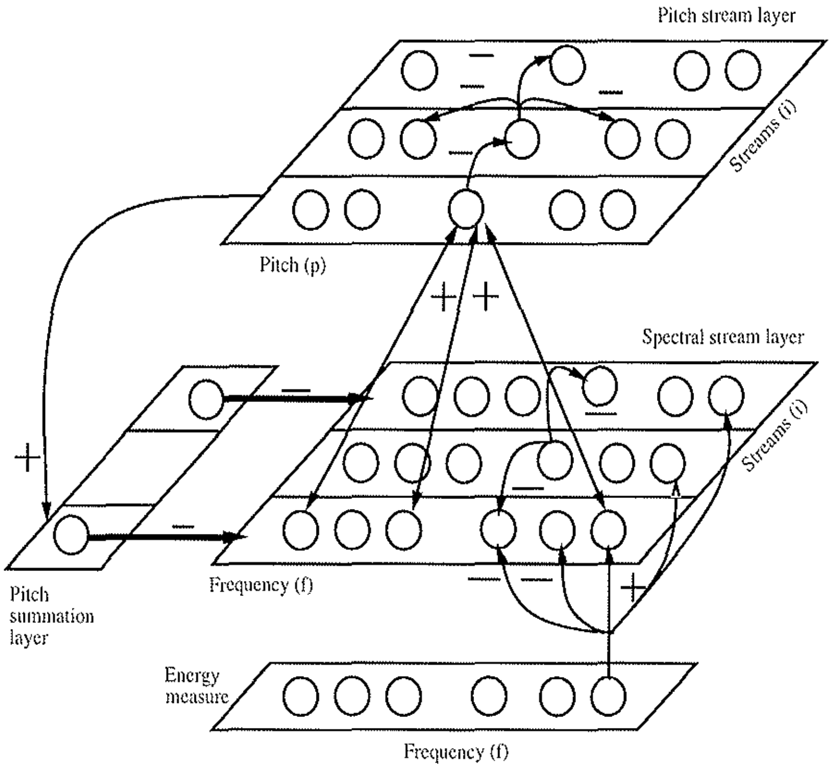

Figure 7. Interaction between the energy measure, the spectral stream layer, the pitch stream layer, and the pitch summation layer. The energy measure layer is fed forward in a frequency-specific one-to-many manner to each frequency-specific stream node in the spectral stream layer. This feed-forward 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 can not 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 Govindarajan *et al.* (1994).]

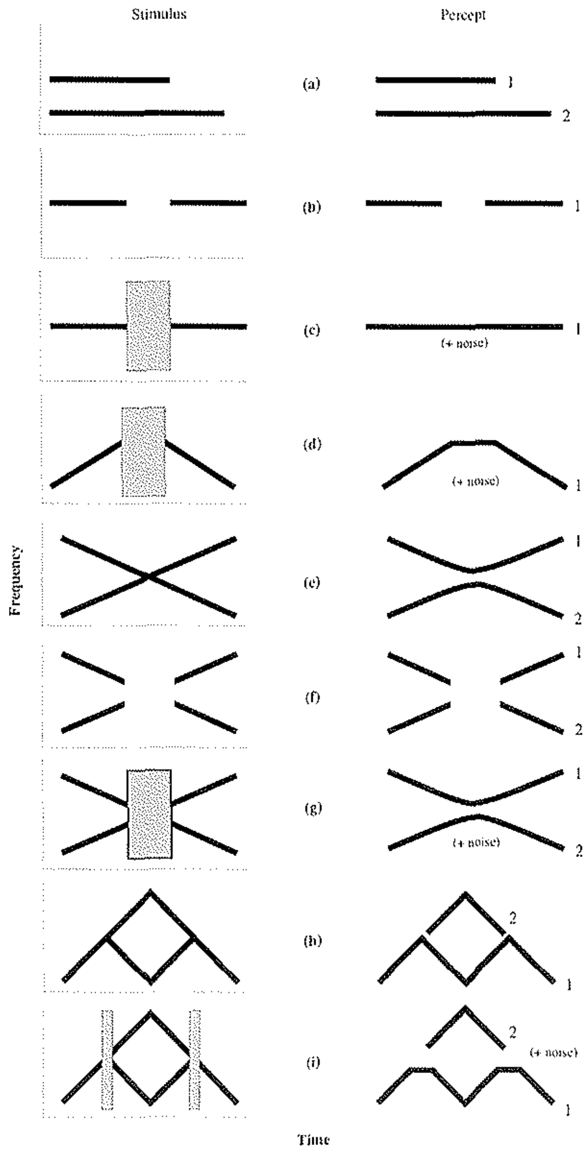

Figure 8. Illustrative stimuli and the listeners' percepts that ARTSTREAM model simulations emulate. The hashed boxes represent broadband noise. The stimuli consist of: (a) two inharmonic tones, (b) tone-silence-tone, (c) tone-noise-tone, (d) a ramp or glide-noise-glide, (e) crossing glides, (f) crossing glides where the intersection point has been replaced by silence, (g) crossing glides where the intersection point has been replaced by noise, (h) Steiger diamond stimulus, and (i) Steiger diamond stimulus where bifurcation points have been replaced by noise. [Reprinted with permission from Govindarajan *et al.* (1994).]

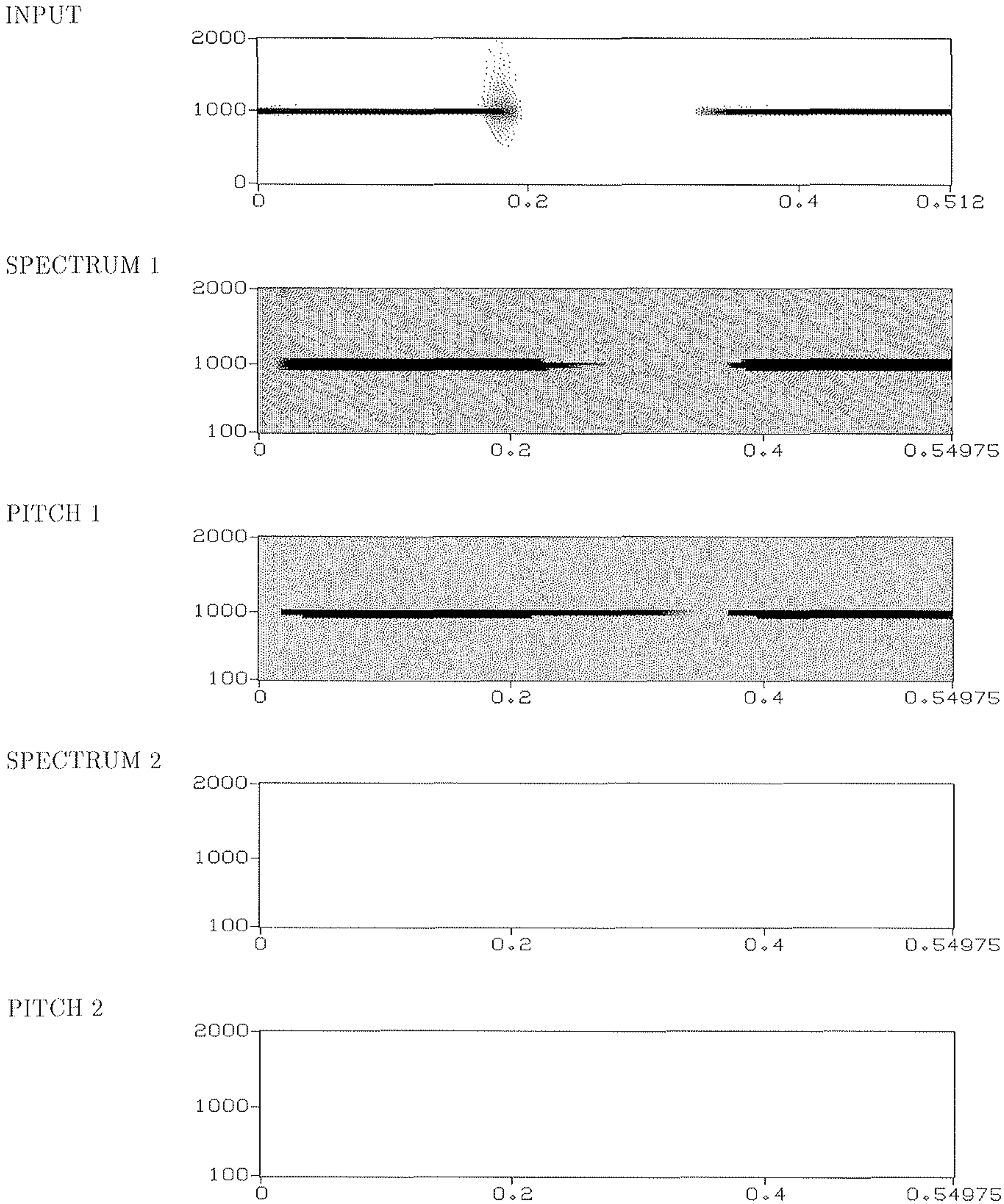

Figure 9. Computer simulation of the tone-silence-tone stimulus and percept. [Adapted with permission from Govindarajan *et al.* (1994).]

are no other components to capture.

The simulation of the case where the silent interval is replaced by noise is illustrated by Figure 10, which shows the spectrogram and the resulting spectral and pitch layer activations of two streams. The first stream here captures the tone and the resonance between the spectral and pitch layers continues through and past the noise interval. The noise is captured by the second stream. The use of more streams could possibly break up the noise into smaller groupings.

An extension of the ARTSTREAM model to include interactions between pitch cues and spatial location cues clarifies how acoustic sources that are placed at different angles with respect to the head can be separated into streams more easily than sources which are not. This interaction has also been used to suggest an explanation of the scale illusion of Deutsch (1975). In this percept, a downward scale 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. Listeners group the scales based upon frequency proximity, so that the alternating ear of origin is not heard. Moreover, at the point where the scales intersect, a bounce percept is heard, so that each ear hears a rising and descending sequence of tones in one ear, and a descending and rising sequence in the other, rather than a complete scale. Thus, as in the Steiger (1980) percept of Figure 8h, grouping is dominated by frequency proximity. In all these cases, the stream resonances provide the coherence that allows distinct voices or instruments to be separated and tracked through a multiple source environment.

Apparent Motion in Music?

The model is being further developed in order to emulate other streaming phenomena. For example, the existing model does not yet contain 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 and Dannenberg, 1977). Another set of data that need further investigation demonstrate how the addition of harmonics can help overcome grouping by proximity. In particular, the addition of harmonics to one glide in a stimulus that consists of crossing ascending and descending glides can lead to a percept of crossing

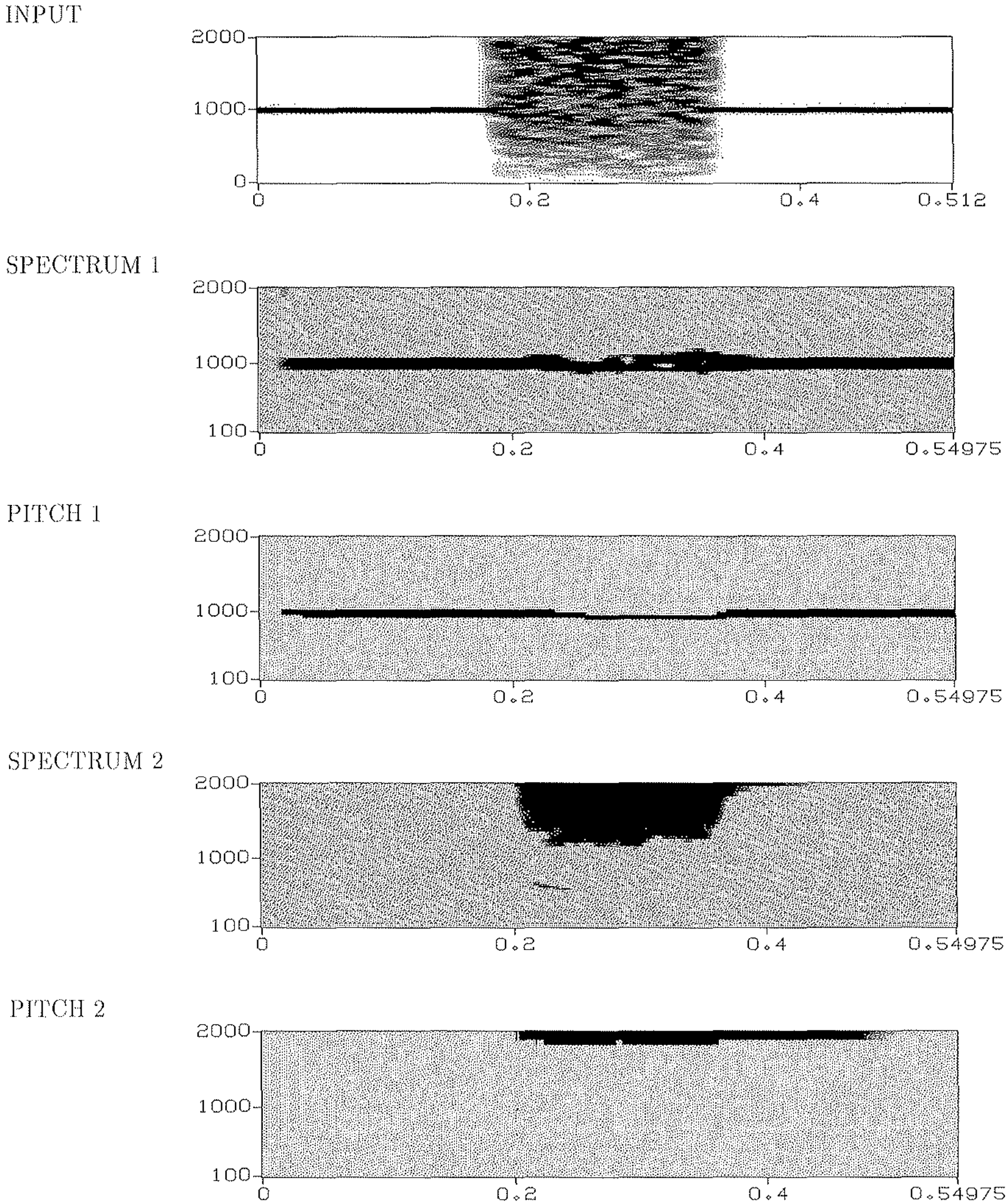

Figure 10. Computer simulation of the tone-noise-tone stimulus and percept. [Adapted with permission from Govindarajan *et al.* (1994).]

glides rather than of a bounce that separates them into V and inverted V percepts of pitch streaming (Bregman, 1990). Using analog, rather than winner-take-all, activations of pitch stream neurons helps to explain these cases by making the activity of pitch nodes covary with the number of harmonics that activate them.

Gjerdingen (1994, this book) has exploited the similarities between apparent motion in vision and streaming in audition by applying the Grossberg and Rudd (1989, 1992) motion model to simulate a variety of streaming percepts that are found in music perception. His analysis takes as a point of departure the realization 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" (Gjerdingen, 1994, p. 335). Using the Grossberg-Rudd model, 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.

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 (Figure 11a). 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 (Figure 11b). 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 com-

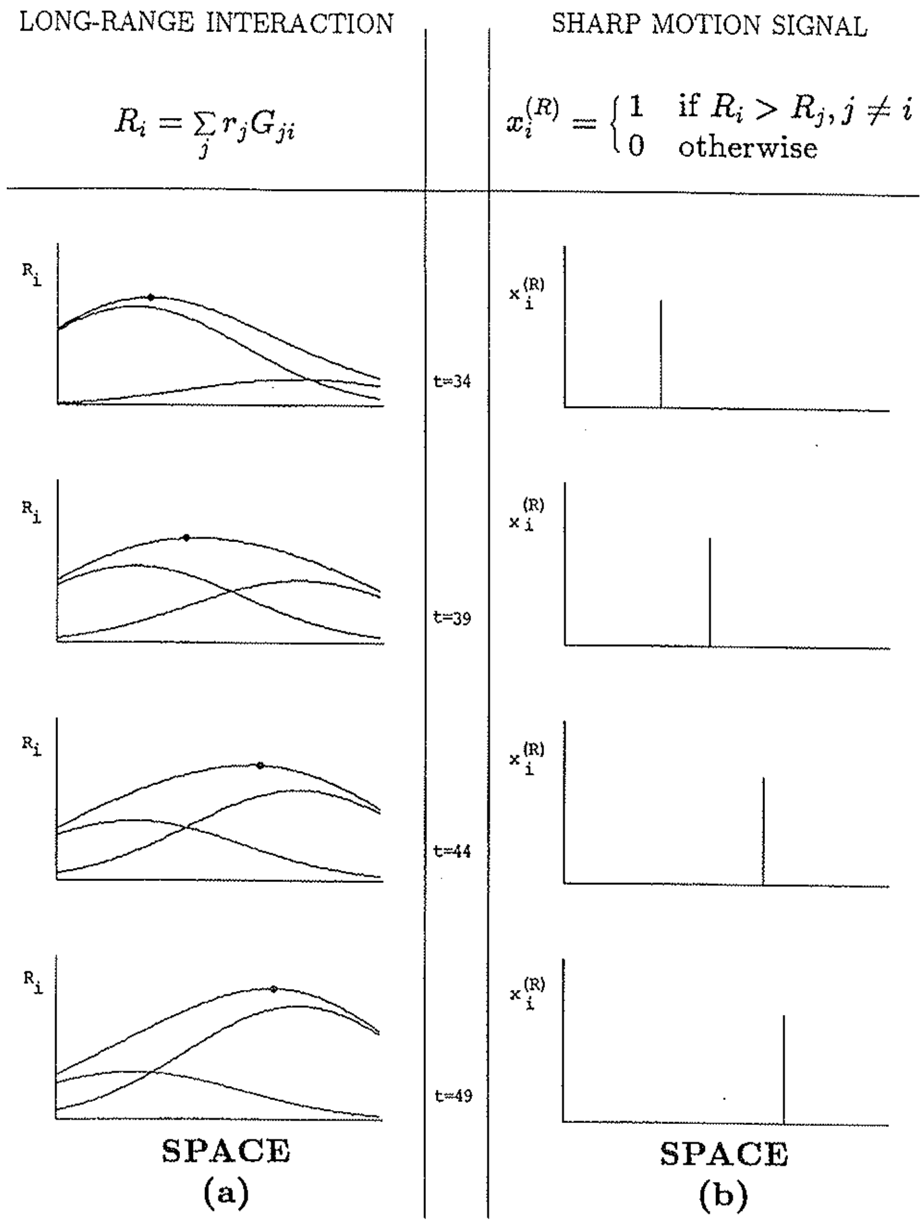

Figure 11. Simulation of an apparent motion G-wave. Each successively lower row depicts a later time. In (a), the two lower curves in each row depict the waning (leftward) and waxing (rightward) Gaussians through time. The upper curve depicts the sum of these Gaussians. Its maximum moves continuously from the location (on the left) of the first flash to the location (on the right) of the second flash. In (b), this maximum is plotted at successive times after the contrast-enhancing competition selects the node that receives the maximum total input. [Reprinted with permission from Grossberg and Rudd (1989).]

petition. 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 equations (18)–(20) in the Appendix. Term E_{ip} there describes the Gaussian-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.

Michiro Negishi and I are now working to further develop the ARTSTREAM model using the visual motion model of Chey, Grossberg, and Mingolla (1994, 1995) that builds upon the Grossberg-Rudd model. The Chey *et al.* model uses transient cells and multiple spatial scales to simulate human psychophysical data concerning the perceived speed and direction of moving objects. Analogous mechanisms in the ARTSTREAM model are helping to explain directionally selective auditory streaming percepts (e.g., Bregman, 1990; Steiger and Bregman, 1981) as well as properties of directionally-sensitive auditory neurons (e.g., Wagner and Takahashi, 1992). All the properties simulated by Gjerdingen (1994) should also be achievable within this version of the ARTSTREAM model when the Gaussians, transient cells, and multiple scales are combined. These several developments should enable the ARTSTREAM model to simulate a broader variety of phenomena about musical phrasing and separation into multiple voices.

Finally, no learning presently exists in the ARTSTREAM model. An exploration of how an organism can learn during development to adaptively tune the harmonic sieves that abut its pitch stream representations remains to be developed. Previous analyses of learning by ART networks should provide helpful guideposts for these future studies, which may ultimately shed light on cultural differences in music perception.

APPENDIX

The mathematical equations that define the ARTSTREAM model, including its embedded SPINET mechanisms, are now summarized to clarify how pitch-based streaming is achieved by it.

Outer and Middle Ear

The outer and middle ear act as a broad bandpass filter that linearly boost frequencies between 100 to 5000 Hz. This is approximated by preemphasizing the signal using a difference equation:

$$y(t) = x(t) - A * x(t - \Delta t), \tag{1}$$

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.

Cochlear Filterbank

The basilar membrane acts like a filterbank whose responses at a particular location act like a bandpass filter. This bandpass characteristic was modeled as a fourth order gammatone (de Boer and de Jongh, 1978; Cohen, Grossberg, and Wyse, 1995) filter:

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

Its frequency response is:

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

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

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

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:

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

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

Energy Measure

The energy measure measures a short-time energy spectrum

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

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 sends the same signal pattern to the multiple fields in the spectral stream layer.

Spectral Stream Layer

The spectral stream layer is a plane with one axis representing frequency, and the other axis representing different auditory streams. Each frequency channel inputs the energy measure e_f in (6) to each spectral stream layer 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 in other streams of the pitch stream layer. In addition, nonspecific inhibition, via the pitch summation layer, helps to realize ART matching and to thereby suppress those spectral components that do not belong to the given pitch within its stream.

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

$$\frac{d}{dt} S_{if} = -AS_{if} + [B - S_{if}]E_{if} - [C + S_{if}]I_{if}, \tag{7}$$

where

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

and

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

In (7), S_{if} is the activity of the spectral stream layer neuron corresponding to the ith stream and frequency f. Equation (7) is a membrane, or shunting equation, with passive decay ($-AS_{if}$), excitation ($[B - S_{if}]E_{if}$), and inhibition ($-[C + S_{if}]I_{if}$) terms. The total excitatory input is E_{if} and the total inhibitory input is I_{if}. The excitatory term $D_{fg}s(e_g)$ in (8) is the bottom-up excitatory input from the energy measure, which has been passed through a sigmoid signal function $s(x)$ to contrast-enhance its signal and to compress its 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 (9) is the bottom-up inhibitory input from the energy measure, which has also been passed through a sigmoid $s(x)$. Both $D_{fg}s(e_g)$ and $E_{fg}s(e_g)$ thus input to each spectral stream layer a contrast-enhanced version of the energy measure. Signal $s(e_g)$ is distributed across frequencies by the kernels D_{fg} and E_{fg}, which are Gaussians that are centered at frequency f, and have standard deviation parameters, σ_D and σ_E, and scaling parameters D and E, respectively:

$$D_{fg} = DG(f, \sigma_D) = D\frac{1}{\sigma_D\sqrt{2\pi}}e^{-.5(f-g)^2/\sigma_D^2} \tag{11}$$

$$E_{fg} = EG(f, \sigma_F) = E\frac{1}{\sigma_E\sqrt{2\pi}}e^{-.5(f-g)^2/\sigma_E^2} \tag{12}$$

The on-center $D_{fg}s(e_g)$ and off-surround $E_{fg}s(e_g)$ inputs balance each other so that the spectral stream layer can respond sensitively, without saturation, to the pattern of $s(e_g)$ signals across frequency (Grossberg, 1973, 1982).

Term $F\sum_p\sum_k M_{f,kp}g(P_{ip})h(k)$ in (8) is the top-down harmonic sieve signal. It sums all the pitches p which have a harmonic kp near frequency f in the pitch stream layer that corresponds to stream i. In (8), P_{ip} is the activity that represents pitch p in stream i, and $g(x)$ is a sigmoid function:

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

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

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

and $M_{f,kp}$ is a normalized Gaussian that represents the top-down harmonic sieve. 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:

$$M_{f,kp} = G(f, \sigma_M) = \frac{1}{\sigma_M \sqrt{2\pi}} e^{-.5(f-kp)^2/\sigma_M^2}. \tag{15}$$

Term $J \sum_{k \neq i} \sum_g N_{fg}[S_{kg}]^+$ in (9) represents the competition across streams for a component, so that a harmonic will resonate within only one stream. This inhibition embodies the principle of "exclusive allocation" (Bregman, 1990). 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:

$$N_{fg} = G(f, \sigma_N) = \frac{1}{\sigma_N \sqrt{2\pi}} e^{-.5(f-g)^2/\sigma_N^2}. \tag{16}$$

Term LT_i in (9) is the top-down inhibition from the pitch summation layer that nonspecifically inhibits all components in stream i. It hereby suppresses those non-harmonic components that are not reinforced by the top-down harmonic sieve excitation from the active pitch unit in the ith pitch stream layer. This is akin to the matching process that is used in ART.

In all the simulations of Govindarajan *et al.* (1994), the parameters were set to: $A = 1, B = 1, C = 1, D = 500, E = 450, F = 3, J = 1000, L = 5, M_h = .3, N = .01, N_s = 10000, N_g = .01, \sigma_D = .2, \sigma_E = 4, \sigma_M = .2$, and $\sigma_N = 1$.

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 (7)–(9) so that only those harmonic components that correspond to the selected pitch node remain active. The activity T_i of the ith pitch summation layer obeys:

$$\frac{d}{dt} T_i = -AT_i + [B - T_i] \sum_p g(P_{ip}), \tag{17}$$

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

Pitch Stream Layer

In the ARTSTREAM model, the spectral and pitch representations of the SPINET model are modified to allow multiple streams to cooperate and compete between pitch units within and across streams. The pitch strength activation P_{ip} of pitch p in stream i obeys a membrane equation:

$$\frac{d}{dt}P_{ip} = -AP_{ip} + [B - P_{ip}]E_{ip} - [C + P_{ip}]I_{ip}, \tag{18}$$

where

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

and

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

Term $E\sum_k \sum_f M_{f,kp}[S_{if} - \Gamma]^+ h(k)$ in (19) corresponds to the bottom-up harmonic sieve input. Kernel $M_{f,kp}$ in (19) Gaussianly filters signals from the spectral layer that have suprathreshold components near a harmonic kp of pitch p. This Gaussian kernel is further weighted by the harmonic weighting function $h(k)$. The harmonic weighting function $h(k)$ and the Gaussian $M_{f,kp}$ are the same as in the spectral layer (equations (14) and (15), respectively). Term $J\sum_{p\neq q} H_{pq}g(P_{iq})$ in (20) allows pitches to compete within a stream. This off-surround competition across pitches within a stream converts each pitch stream into a winner-take-all network (Grossberg, 1973, 1982) wherein only one pitch tends to be active within each stream. For simplicity, kernel 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:

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

Term $L\sum_{k>i} g(P_{kp})$ in (20) represents inhibition across streams for a given pitch, so that only one stream can activate a given pitch. This is a form of asymmetric inhibition, from higher to lower pitches, that prevents deadlock from occurring between two streams that are competing for a given pitch. This inhibition breaks the symmetry that arises from the fact that all pitch streams initially receive equal bottom-up excitation from the spectral layer. In all the simulations, the parameters were set to: $A = 100, B = 1, C = 10, E = 5000, J = 300, L = 2, \sigma_H = .2$, and $\Gamma = .005$.

REFERENCES

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

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

Bregman, A.S. and Dannenberg, G. (1977). Auditory continuity and amplitude edges. *Journal of Psychology*, *31*, 151–159.

Broadbent, D.E. and 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. and Noteboom, S.G. (1982). Intonation and the perceptual separation of simultaneous voices. *Journal of Phonetics*, *10*, 23–26.

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

Carpenter, G.A. and 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. and 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., and Mingolla, E. (1994). Neural dynamics of motion processing and speed discrimination. *Technical Report CAS/CNS-TR-94-030*, Boston University. Submitted for publication.

Chey, J., Grossberg, S., and Mingolla, E. (1995). Neural dynamics of motion speed and directional grouping: From aperture ambiguity to plaid coherence. *Technical Report CAS/CNS-TR-95-031*, Boston University. Submitted for publication.

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

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

de Boer, E. and de Jongh, H.R. (1978). On cochlear encoding: Potentialities and limitations of the reverse correlation technique. *Journal of the Acoustical Society of America*, *63*, 115–135.

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., and 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., and 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.

Govindarajan, K.K., Grossberg, S., Wyse, L.L., and Cohen, M.A. (1994). A neural network model of auditory scene analysis and source segregation. *Technical Report CAS/CNS-TR-94-039*, Boston University. Submitted for publication.

Grossberg, S. (1973). Contour enhancement, short-term memory, and constancies in reverberating neural networks. *Studies in Applied Mathematics*, *52*, 217–257.

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

Grossberg, S. (1982). *Studies of mind and brain*. Boston: Reidel Press.

Grossberg, S. (1984). Unitization, automaticity, temporal order, and word recognition. *Cognition and Brain Theory*, *7*, 263–283. Reprinted in G.A. Carpenter and S. Grossberg (Eds.), *Neural networks for vision and image processing*. Cambridge, MA: MIT Press, 1992.

Grossberg, S. (1987). Competitive learning: From interactive activation to adaptive resonance. *Cognitive Science*, *11*, 23–63.

Grossberg, S. (1995). The attentive brain. *American Scientist*, *83*, 438–449.

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

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

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

Miller, G.A. and 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 and J. Wilson (Eds.), *Psychophysics and physiology of hearing*. New York: Academic Press.

Moore, B.C.J., Glasberg, B.R., and 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.

Patterson, R. and 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.

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

Samuel, A.G. (1981). Phonemic restoration: Insights from a new methodology. *Journal of*

Experimental Psychology: General, *110*, 474–494.

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

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

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

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

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

Terhardt, E. (1972). Zur Tonhöhenwahrnehmung von Klängen. *Acustica*, *26*, 173–199.

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

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

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

Warren, R.M. (1984). Perceptual restoration of obliterated sounds. *Psychological Bulletin*, *96*, 371–383.

Warren, R.M. and Sherman, G.L. (1974). Phonemic restoration based on subsequent context. *Perception and Psychophysics*, *16*, 150–156.

Wightman, F.L. (1973). The pattern-transformation model of pitch. *Journal of the Acoustical Society of America*, *54*, 407–416.

Yost, W., Hill, R., and 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.