

1994-12

A Neural Network Model of Auditory Scene Analysis and Source Segregation

<https://hdl.handle.net/2144/2177>

"Downloaded from OpenBU. Boston University's institutional repository."

**A neural network model of auditory scene analysis
and source segregation**

**Krishna Govindarajan, Stephen Grossberg, Lonce Wyse,
and Michael Cohen**

December, 1994

Technical Report CAS/CNS-1994-039

Permission to copy without fee all or part of this material is granted provided that: 1. The copies are not made or distributed for direct commercial advantage; 2. the report title, author, document number, and release date appear, and notice is given that copying is by permission of the BOSTON UNIVERSITY CENTER FOR ADAPTIVE SYSTEMS AND DEPARTMENT OF COGNITIVE AND NEURAL SYSTEMS. To copy otherwise, or to republish, requires a fee and / or special permission.

Copyright © 1994

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

A NEURAL NETWORK MODEL
OF AUDITORY SCENE ANALYSIS
AND SOURCE SEGREGATION

by

Krishna K. Govindarajan¹, Stephen Grossberg², Lonce L. Wyse^{*3}, and Michael A. Cohen².

Department of Cognitive and Neural Systems
and
Center for Adaptive Systems
Boston University
111 Cummington Street
Boston, MA 02215 USA
email: RLL@cns.bu.edu

*Institute for Systems Science
National University of Singapore
Heng Mui Keng Terrace
Kent Ridge, Singapore 0511
email: lwyse@iss.nus.sg

December 1994

Please address all reprint inquiries to Professor Stephen Grossberg.

Technical Report CAS/CNS-TR-94-039

¹Supported 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).

²Supported in part by the Air Force Office of Scientific Research (AFOSR F49620-92-J-0225).

³Supported in part by the American Society for Engineering Education and by the Air Force Office of Scientific Research (AFOSR F49620-92-J-0225)

The authors wish to thank Carol Y. Jefferson and Robin L. Locke for their valuable assistance in the preparation of the manuscript.

ABSTRACT

In environments with multiple sound sources, the auditory system is capable of teasing apart the impinging jumbled signal into different mental objects, or streams, as in its ability to solve the cocktail party problem. A neural network model of auditory scene analysis, called the ARTSTREAM model, is presented that groups different frequency components based on pitch and spatial location cues, and selectively allocates the components to different streams. The grouping is accomplished through a resonance that develops between a given object's pitch, its harmonic spectral components, and (to a lesser extent) its spatial location. Those spectral components that are not reinforced by being matched with the top-down prototype read-out by the selected object's pitch representation are suppressed, thereby allowing another stream to capture these components, as in the "old-plus-new heuristic" of Bregman. These resonance and matching mechanisms are specialized versions of Adaptive Resonance Theory, or ART, mechanisms. The model is used to simulate data 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. The model also simulates illusory auditory percepts 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. The stream resonances provide the coherence that allows one voice or instrument to be tracked through a multiple source environment.

Key words: auditory scene analysis, streaming, cocktail party problem, neural network, resonance, adaptive resonance theory, ART.

1 Introduction

The ability of a listener to pay attention to a particular speaker in a noisy room or in a room with other speakers, e.g. at a cocktail party, attests to the robustness of the auditory perceptual system. Even though the harmonics of various sources are mixed together to produce one signal at the listener's ear, the auditory system is capable of teasing apart this jumbled signal to recognize different mental objects for the different sound sources. The ability to segregate these different signals has been termed auditory scene analysis (Bregman, 1990). The scene analysis corresponds to the mechanisms by which the auditory system selectively groups certain acoustic features, while excluding others, to form internal representations of auditory objects.

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 Figure 1, 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.

(Figure 1)

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 Figure 1. If two tones are closer together in frequency and time, then it is more likely that they should be grouped 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 (Figure 2), even though the tone is not present during the noise (Miller and Licklider, 1950). 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. This is also known as the auditory continuity illusion.

- Good continuation

Good continuation states that an object's sound does not make rapid jumps, but instead continues smoothly. For example, in Figure 2 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.

(Figure 2)

1.3 Primitive versus schema-based segregation

Bregman (1990) noted that auditory stream segregation consists of a primitive, non-attentive, unlearned process and a schema-based, attentive, learned process. Bregman and Rudnick (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 fre-

quency 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 non-attentional mechanism, and noted that the schema-based mechanisms can override the primitive mechanisms. The mechanism proposed here addresses the pre-attentive, primitive segregation mechanisms.

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 Figure 1. In addition, faster presentation rates of alternating high and low frequency tones causes the two tones to be segregated into 2 streams (Bregman and 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 Rudnick (1975) stimuli, which are shown in Figure 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 the temporal order. When A and B were flanked by tones F, listeners had a more difficult time. However, if the capture tones C surrounded the flankers, then F streamed with C, 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.

(Figure 3)

2.2 Continuity illusion

As mentioned above, proximity combined with closure has led 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 (Figure 2).

A more complex example is shown in Figure 4. The top two figures show the two different stimuli that Steiger (1980) presented to listeners. In (b), the 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). In (b), a third stream was also heard corresponding to the broadband noise bursts. Thus, the glide complex had been completed, or continued, through the noise. This experiment is important in that the principle of “good continuation” has been overcome by frequency proximity.

(Figure 4)

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 to 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, and 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). Partially 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, 1994).

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 .5-2 semitones (Scheffers, 1983). When F_0 was an octave apart, identification is also very poor (Brokx and Neteboom, 1982; Chalika and 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 and Ladefoged, 1957; Gardner, Gaskill, and Darwin, 1989). Finally, speech stimuli with discontinuous pitch contours tend to segregate at the discontinuities (Darwin and 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 and Bregman, 1990). A bounce percept corresponds to hearing two streams, one with a “U” 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 Figure 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.

(Figure 5)

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 (Figure 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. The consequence of this experiment is that regardless of noise, silence, or glide during the crossing point, one gets the same percept.

(Figure 6)

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 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 and Steiger, 1980). The scale illusion of Deutsch (1975) also illustrates this point (Figure 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 Figure 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.

(Figure 7)

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 and Grose, 1988). The release of the tone from masking is known as comodulation masking release (CMR). Despite this effect, a recent experiment by Summerfield and Culling (1992) showed that at slow AM rates (2.5Hz), 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 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 .

The results from recent 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 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 and Culling, 1992). This result could be due to pitch difference cues though. 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 and Pinker, 1978; Dannenbring and Bregman, 1978). Bregman and Pinker (1978) presented the stimulus shown in Figure 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 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 and 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

The neural model developed in this article suggests how harmonicity and frequency proximity interact in the brain. The model, which is shown in Figure 8, consists of several stages. The model first preprocesses the incoming signal in the peripheral processing modules. The preprocessed signal is then used to group frequency components based on pitch.

(Figure 8)

The first several stages are based on a model of the physiology and psychophysics of the auditory periphery (Cohen, Grossberg, and Wyse, 1992, 1994). 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 feeds into the different cell arrays, or fields, in the spectral stream layer, where different fields correspond to different streams. There is competition between these streams for each frequency component. No component can be simultaneously allocated to two streams after the competition acts. In addition, this competition causes a component that is not harmonically related to the other components in a given stream to “pop out” of the spectrum assigned to that stream and become active in another stream.

The spectral stream layer has reciprocal connections with the pitch stream layer to determine which spectral components belong to a given pitch. Thus, a pitch is associated with each active stream via a bottom-up filter. The feedback from the pitch stream layer

to the spectral stream layer activates a matching process that reinforces consistent spectral components and suppresses inconsistent components, as in Adaptive Resonance Theory, or ART (Carpenter and Grossberg, 1991; Grossberg, 1980). The inconsistent spectral components are then freed to be captured by other streams, as in the “old-plus-new heuristic” of Bregman (1990). The reciprocal 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 percept is hypothesized to correspond to the activity at the spectral stream layer when there is resonance between it and the pitch stream layer. The fact that the core of the streaming model utilizes ART matching and resonance mechanisms has led us to call the model the ARTSTREAM model. The mathematical operations of the ARTSTREAM model are defined as follows.

4.1 Auditory peripheral processing

4.1.1 Outer and middle ear

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

$$y(t) = x(t) - A * x(t - \Delta t), \quad (1)$$

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.

4.1.2 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 and de Jongh, 1978; Cohen, Grossberg, and Wyse, 1994) filter:

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

and its frequency response is:

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

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:

$$b(f) = 1.02ERB(f). \quad (4)$$

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:

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

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.

4.1.3 Energy measure

The energy measures a short-time energy spectra (Cohen, Grossberg, and Wyse, 1992, 1994):

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

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.

4.2 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 (Figure 9). The spectral stream layer is a plane with one axis representing frequency, and the other axis representing different auditory streams.

(Figure 9)

Each frequency channel in the energy measure, e_f , feeds up to each stream's 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. In addition, nonspecific inhibition, via the pitch summation layer, helps to 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:

$$\dot{S}_{if} = -AS_{if} + [B - S_{if}]\mathcal{E}_{if} - [C + S_{if}]\mathcal{I}_{if} \quad (7)$$

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

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

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

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

Similarly, $E_{fg}s(e_g)$ in (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:

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

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

In addition, the term $F \sum_p \sum_k M_{f, kp} g(P_{ip}) h(k)$ in (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 (8), $g(x)$ is a sigmoid function:

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

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

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

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 :

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

The 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 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 :

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

Term LT_i in (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 non-harmonic 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 and Grossberg, 1991, 1993; Grossberg, 1980). As a result of this matching process, 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: $A = 1, B = 1, C = 1, D = 500, E = 450, F = 3, J = 1000, L = 5, M_h = .3, N = .01, N_s = 10000, N_g = .01, \sigma_D = .2, \sigma_E = 4, \sigma_M = .2$, and $\sigma_N = 1$.

4.3 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 (7)-(9) so that only those harmonic components that correspond to the selected pitch remain active:

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

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

4.4 Pitch stream layer

To determine the pitch, the neural pitch model of Cohen, Grossberg, and Wyse (1992, 1994), called the SPINET model, was used. The original pitch model had two components: the spectral layer and a pitch layer. The spectral and pitch representations have been modified so that there are multiple streams such that competition occurs between pitch units within and across streams. The modified pitch strength activation is:

$$\dot{P}_{ip} = -AP_{ip} + [B - P_{ip}]\mathcal{E}_{ip} - [C + P_{ip}]\mathcal{I}_{ip} \quad (18)$$

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

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

where P_{ip} is the p th pitch unit of object i . The term $E \sum_k \sum_f M_{f,kp} [S_{if} - \Gamma]^+ h(k)$ in (19) corresponds to the Gaussian excitation $M_{f,kp}$ from the spectral layer which have 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 (equations (14) and (15), respectively). The term $J \sum_{p \neq q} H_{pq} g(P_{iq})$ in (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 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.

$$H_{pq} = \begin{cases} 1, & \text{if } |p - q| > \sigma_H \\ 0, & \text{else} \end{cases} \quad (21)$$

The term $L \sum_{k > i} g(P_{kp})$ in (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: $A = 100, B = 1, C = 10, E = 5000, J = 300, L = 2, \sigma_H = .2$, and $\Gamma = .005$.

5 Simulation results of model

The model is here shown to qualitatively emulate bounce percepts for crossing glides, as well as several variants of the continuity illusion. Figure 10 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.

(Figure 10)

5.1 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, Glasberg, and Peters, 1985). Figure 10a shows the stimulus and the listeners' percept for two inharmonic tones. Figure 11a shows the spectrogram for two inharmonic tones, whose frequencies are 358 Hz and 1233 Hz. Figure 11b shows the result after peripheral processing, i.e. the result after the energy measure. Figure 12 shows the resulting spectral and pitch layers for the two tone stimulus for two different streams. Figure 12C illustrates 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.

(Figure 11)

(Figure 12)

Figure 13 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 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.

(Figure 13)

5.2 Continuity illusion

The model is capable of producing the continuity illusion: continuation of a tone in noise, even though the tone is not physically present in the noise (Miller and 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 (Figure 10b). For this stimulus, the tone should not continue across the silence, but should stop at the onset of silence. Figure 14 shows the spectrogram and the result after the peripheral processing for the tone-silence-tone stimulus. Figure 15 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 in 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.

(Figure 14)

(Figure 15)

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. Figure 16 shows the spectrogram and the result after the peripheral processing for the tone-noise-tone stimulus. Figure 17 shows the resulting spectral and pitch layers for the

stimulus for the first two streams, and Figure 18 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.

(Figure 16)

(Figure 17)

(Figure 18)

The reason that the tone continues through the noise derives from two factors. The first factor is that the spectral layer slowly integrates the input, and so, the noise is temporally averaged, or smoothed over time. Due to this smoothing, if there is no top-down activity, the noise is relatively constant over time. The second factor is that the top-down activity from the pitch layer remains active at the onset of the noise due to the prior tone. Due to both of these factors, the noise at the same frequency as the tone is reinforced by the top-down activity, while the other frequency components are inhibited, allowing the “tone” to complete across the noise. The second and third streams contain the other spurious noise. The reason that the second stream captures the high frequency noise as opposed to the low frequency noise is due to 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.

(Figure 19)

The model is also capable of producing the continuity illusion for the ramped stimulus shown in Figure 10d. Figure 19 shows the spectrogram and the result after the peripheral processing. Figure 20 shows the resulting spectral and pitch 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 stream layer and the

prior 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.

(Figure 20)

5.3 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 (Figure 10e), even if the crossing interval is replaced by silence (Figure 10f) or noise (Figure 10g). Figure 21 shows the spectrogram and the result after the energy measure for the standard crossing glide stimulus; and Figure 22 shows the resulting spectral and pitch activity for the two streams. As one can see, one stream contains the “U” percept, while the other stream has a “O” percept. The reason one obtains the bounce percept for the standard crossing glide stimulus is due to the following. 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, grouped with the upper glide; and the lower adjacent frequency channel belongs to the second stream, and thus, grouped with the lower glide.

(Figure 21)

(Figure 22)

Figure 23 shows the crossing glide stimulus for the silent-center condition and the result of the energy measure. Figure 24 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 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.

(Figure 23)

(Figure 24)

Figure 10g 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. Figure 25 shows the spectrogram and the result of the energy measure for the crossing glide with noise-center stimulus, and Figure 26 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 seen from the results on the 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, e.g. the for ramp stimulus of Figure 10d, applies during the noise interval. At the onset of the noise, the 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 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.

(Figure 25)

(Figure 26)

5.4 Steiger (1980) diamond stimulus

For the Steiger (1980) diamond stimulus (Figure 10h), the percept consists of two streams, a “M” stream and an inverted “V” stream. This percept shows that the principle of continuity can be overcome by frequency proximity. Figure 27 shows the Steiger (1980) stimulus and the result after the peripheral processing. Figure 28 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.

(Figure 27)

(Figure 28)

Figure 29 shows the spectrogram and the result of the energy measure for the Steiger (1980) stimulus where the bifurcation points have been replaced by noise. Figure 30 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 Figure 10d. Stream 1 initially captures the increasing glide, while stream 2 is inactive, just as in the 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.

(Figure 29)

(Figure 30)

6 Model 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 Section 4. 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.

6.1 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. Figure 31 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.

(Figure 31)

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 some 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 Figure 32. 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.

(Figure 32)

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 . τ represents radial direction, taking on values from -600 to $600 \mu s$. 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. Figure 33 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.

(Figure 33)

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 $L \sum_{k>i} g(P_{kp})$ in equation (20), exists only at the central location; non-central 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 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 the Deutsch (1975) scale illusion (Figure 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, hi 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 hi 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 hi 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.

7 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 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 the 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 and Grossberg, 1991, 1993; Grossberg, 1980; Grossberg and Stone, 1986). 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 subliminally 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 spectral components that are consistent with the pitch, but to inhibit 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 and Grossberg, 1987, 1991).

Bregman (1990) distinguishes primitive segregation mechanisms from higher-order processes that he calls schema-based segregation. Grossberg, Boardman, and Cohen (1994) 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. These examples provide convergent evidence that similar ART matching and resonance processes operate on multiple levels of the auditory system. These results extend the analyses of Grossberg (1978, 1986) of a variety of speech and word recognition data properties using ART mechanisms; also see Cohen and Grossberg (1986) and Grossberg and Stone (1986).

While the present model of primitive segregation is capable of qualitatively producing correct responses for the key streaming stimuli mentioned above, the model needs to be further developed in order to emulate other 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 Dannenbring, 1977). Another set of data that need further investigation demonstrate how the addition of harmonics can help overcome grouping by proximity; e.g., how the addition of harmonics to one glide in a crossing glide stimulus leads to a cross percept and not a bounce percept in Figure 5c. Using analog, rather than winner-take-all, activations of pitch stream neurons should handle these cases. Finally, no learning exists in the model, and thus an exploration of how an organism can learn to adaptively tune its pitch stream representations for primitive auditory scene analysis remains to be developed. Previous analyses of learning by ART networks will provide helpful guideposts for these future studies.

References

- Bregman, A. S. (1990). **Auditory scene analysis: The perceptual organization of sound**. Cambridge, MA: MIT Press.
- Bregman, A. S. and 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. and Dannenbring, G. (1977). Auditory continuity and amplitude edges. *Journal of Psychology*, **31**, 151–159.
- Bregman, A. S. and 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. and Pinker, S. (1978). Auditory streaming and the building of timbre. *Canadian Journal of Psychology*, **32**, 19–31.
- Bregman, A. S. and Rudnický, A. (1975). Auditory segregation: Stream or streams? *Journal of Experimental Psychology: Human Perception and Performance*, **1**, 263–267.
- Bregman, A. S. and 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. 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 Neteboom, 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. Ph.D. 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, and I. J. Russell (Eds.), **Processing of complex sounds by the auditory system**. Oxford: Clarendon Press.
- 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.
- 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. (1992). A neural network for synthesizing the pitch of an acoustic source. **Proceedings of the international joint conference on neural networks, IV**, 414–419. Piscataway, NJ: IEEE Service Center.
- Cohen, M. A., Grossberg, S., and Wyse, L. (1994). A neural network spectral model of pitch detection and representation. Technical Report CAS/CNS-TR-92-024, Boston, MA: Boston University. *Journal of the Acoustical Society of America*, in press.

- 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. Ph.D. Thesis, University of Sheffield.
- Dannenbring, G. L. and 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. and 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. and 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. and Sutherland, N. S. (1984). Grouping frequency components of vowels: When is a harmonic not a harmonic? *Q. Journal of Experimental Psychology*, **36A**, 193–208.
- 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.

- 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.
- Grossberg, S. (1978). A theory of human memory: Self-organization and performance of sensory-motor codes, maps, and plans. In R. Rosen and F. Snell (Eds.), **Progress in theoretical biology, volume 5**. New York: Academic Press, pp. 233–374. Reprinted in Grossberg, S. (1982). **Studies of mind and brain: Neural principles of learning, perception, development, cognition, and motor control**. Boston, MA: Reidel Press.
- Grossberg, S. (1980). How does a brain build a cognitive code? *Psychological Review*, **87**, 1–51.
- Grossberg, S. (1986). The adaptive self-organization of serial order in behavior: Speech, language and motor control. In E. C. Schwab and H. C. Nusbaum (Eds.), **Pattern recognition by humans and machines, Volume 1: Speech perception**. New York: Academic Press, pp. 187–294.
- Grossberg, S., Boardman, I., and Cohen, M.A. (1994). Neural dynamics of variable-rate speech categorization. Technical Report CAS/CNS-TR-94-038. Boston, MA: Boston University.
- Grossberg, S. and Stone, G. O. (1986). Neural dynamics of word recognition and recall: Attentional priming, learning, and resonance. *Psychological Review*, **93**, 46–74.
- Hall, J. W. and 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, McGill University, Psychology Department.

- Handel, S. (1989). **Listening**. Cambridge, MA: MIT 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. and 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. and Licklider, J. C. R. (1950). Intelligibility of interrupted speech. *Journal of the Acoustical Society of America*, **22**, 167–173.
- 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.
- Scheffers, M. T. M. (1983). Sifting vowels: Auditory pitch analysis and sound segregation. Ph.D. Thesis, Groningen University.
- Shackleton, T. M., Meddis, R., and Hewitt, M. J. (1994). The role of binaural and fundamental frequency difference cues in the identification of concurrently presented vowels. Technical Report, University of Technology, England. (submitted).
- Steiger, H. (1980). Some informal observations concerning the perceptual organization of patterns containing frequency glides. Technical Report, McGill University, Montreal.
- 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. and Culling, J. F. (1992). Auditory segregation of competing voices: Absence of effects of fm or am coherence. In R. P. Carlyon, C. J. Darwin, and I. J.

Russell (Eds.), **Processing of complex sounds by the auditory system**. Oxford: Clarendon Press.

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

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

Figure Captions

Figure 1: 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 from Bregman and Pinker (1978).

Figure 2: Stimulus and percept of the continuity illusion. (a) shows the stimulus that is presented to listeners, and (b) represents the percept. Note that in the stimulus, the tone does not continue through the noise, but stops at the onset of the noise, and continues at the offset of the noise, but the percept is that the tone continues through the noise.

Figure 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 from Bregman and Rudnicky (1975).

Figure 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 spectra 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 from Steiger (1980).

Figure 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 from Halpern (1977).

Figure 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 from Tougas and Bregman (1990).

Figure 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 from Deutsch (1975).

Figure 8: Block diagram of the ARTSTREAM auditory streaming model. See text for further details.

Figure 9: 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. In addition, this feed-forward activation is contrast-enhanced. There is also competition 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 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, there is asymmetric competition for each pitch so that one stream is biased to win and the same pitch can not be represented in another stream. Finally, 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.

Figure 10: 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.

Figure 11: (a) spectrogram and (b) result of energy measure for the two tone stimulus.

Figure 12: 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.

Figure 13: Schematic of how the model segregates the two inharmonic tones into two different streams. See text for explanation.

Figure 14: (a) spectrogram and (b) result of energy measure for the tone-silence-tone stimulus.

Figure 15: 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.

Figure 16: (a) spectrogram and (b) result of energy measure for the tone-noise-tone stimulus.

Figure 17: 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.

Figure 18: The (a) spectral and (b) pitch stream layers for stream 3 for the tone-noise-tone stimulus.

Figure 19: (a) spectrogram and (b) result of energy measure for the ramp stimulus.

Figure 20: 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.

Figure 21: (a) spectrogram and (b) result of energy measure for the crossing glide stimulus.

Figure 22: 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.

Figure 23: (a) spectrogram and (b) result of energy measure for the crossing glide stimulus with silence replacing the intersection point.

Figure 24: 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.

Figure 25: (a) spectrogram and (b) result of energy measure for the crossing glide stimulus with noise replacing the intersection point.

Figure 26: 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.

Figure 27: (a) spectrogram and (b) result of energy measure for the Steiger (1980) diamond stimulus.

Figure 28: 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.

Figure 29: (a) spectrogram and (b) result of energy measure for the Steiger (1980) diamond stimulus with noise bursts replacing the bifurcation points.

Figure 30: 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.

Figure 31: Geometric representation of spatial lateralization using interaural timing differences (ITD).

Figure 32: Block diagram of an ARTSTREAM model that incorporates both pitch and spatial location cues.

Figure 33: 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 $\tau = 0$.

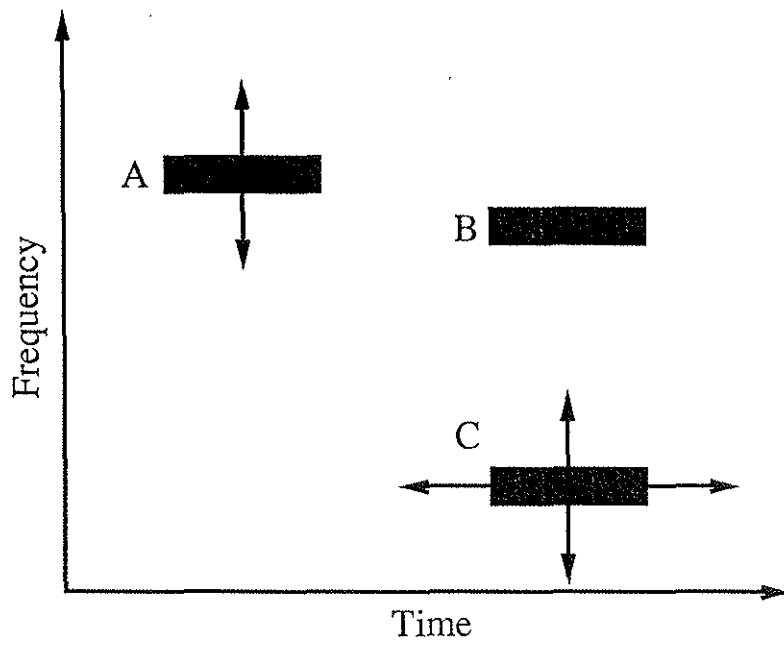


Figure 1

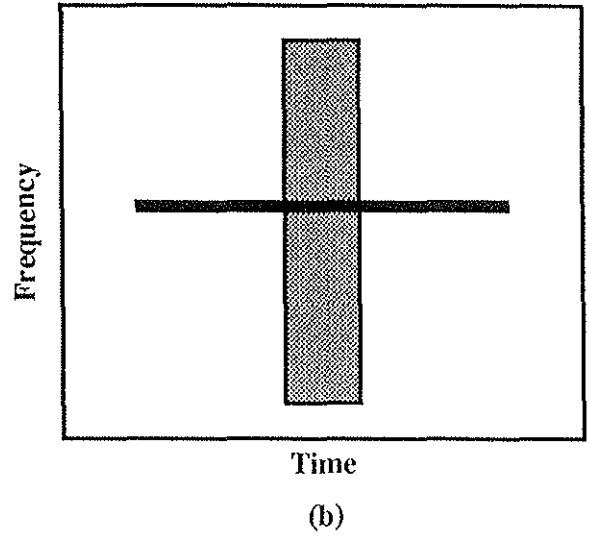
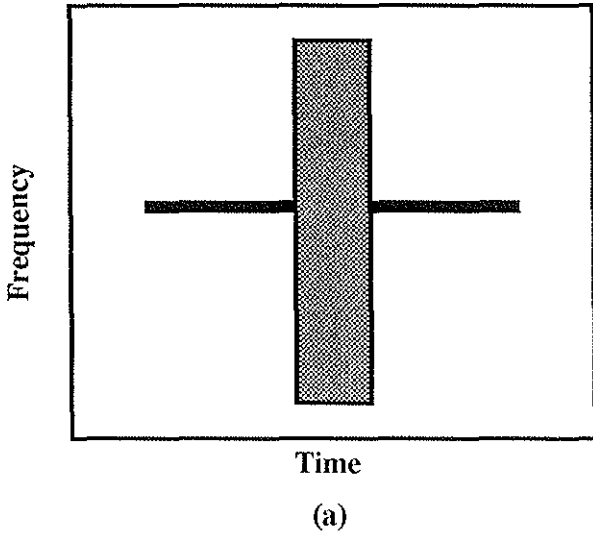


Figure 2

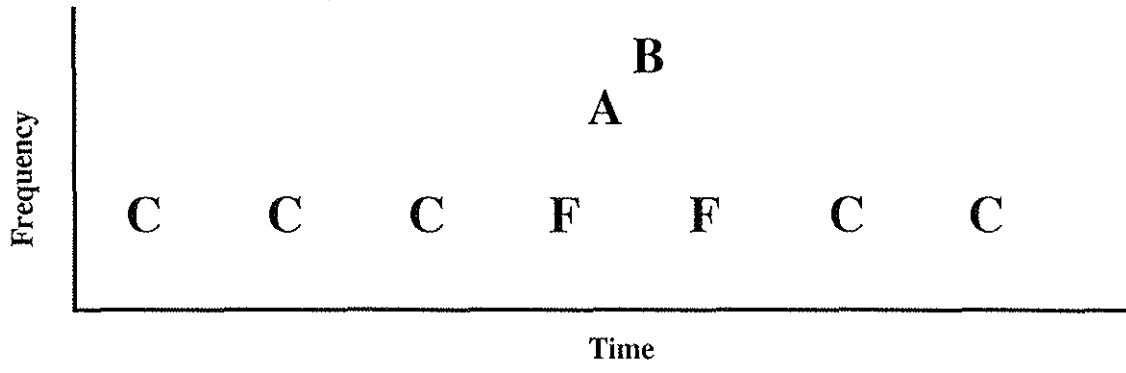
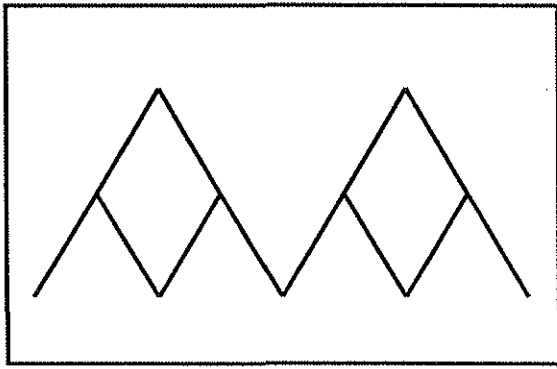
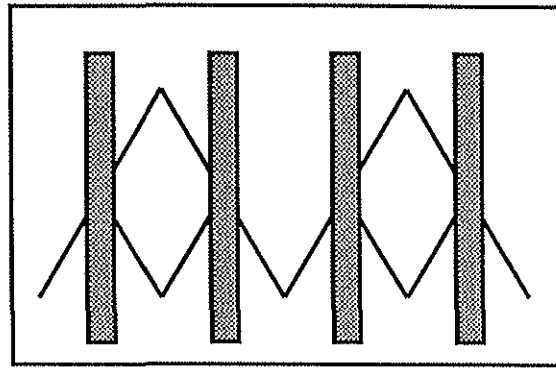


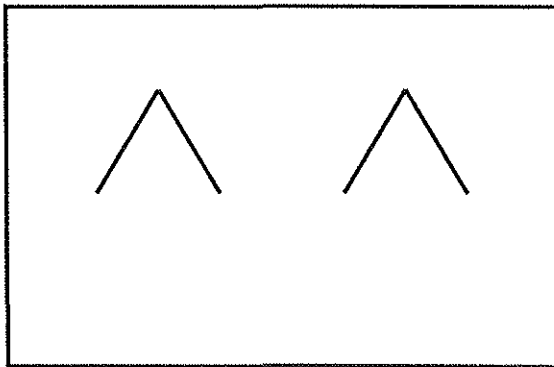
Figure 3



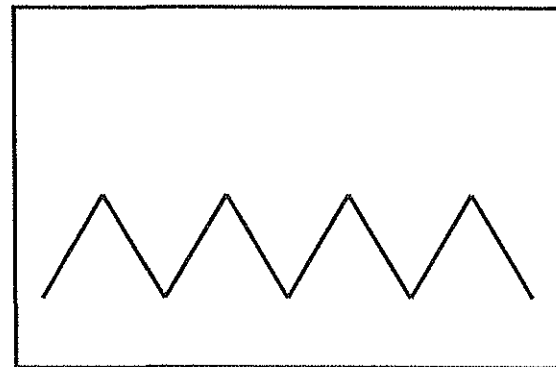
(a)



(b)



(c)



(d)

Figure 4

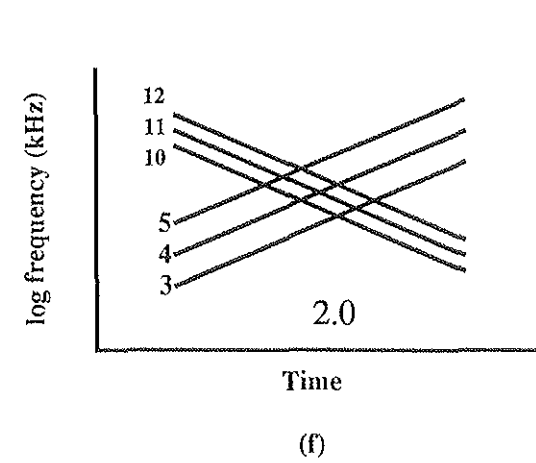
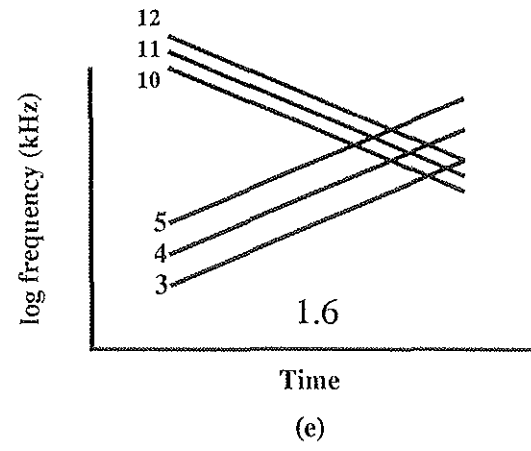
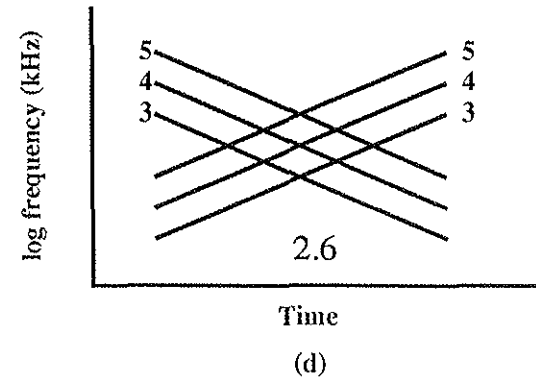
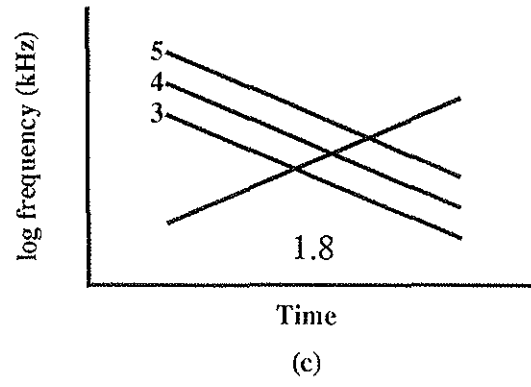
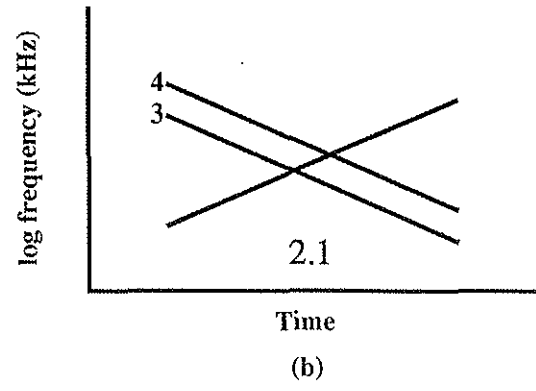
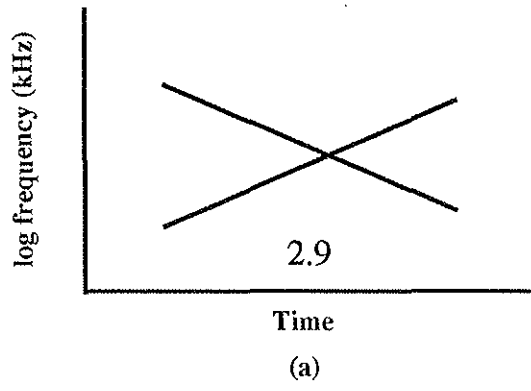
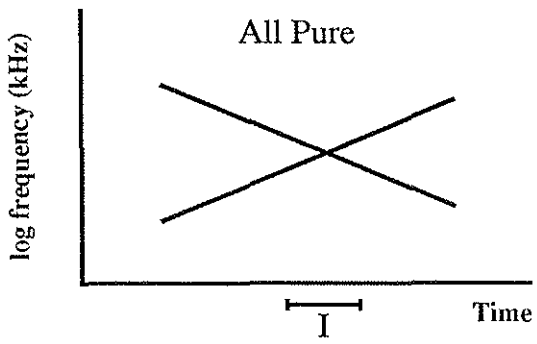
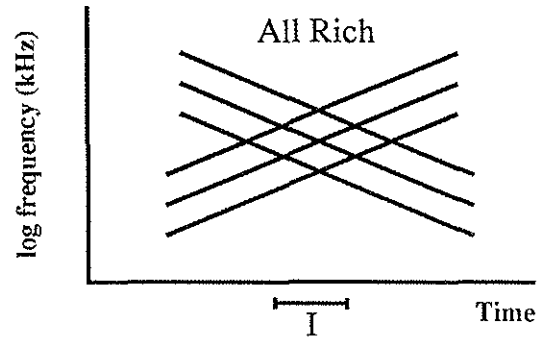


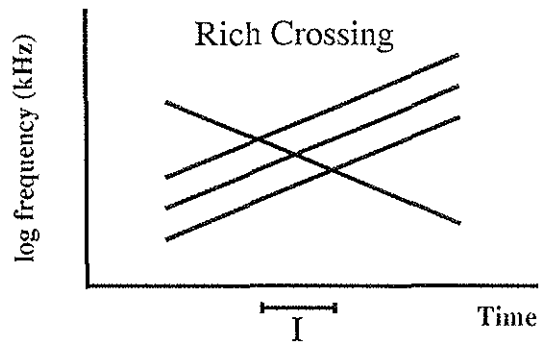
Figure 5



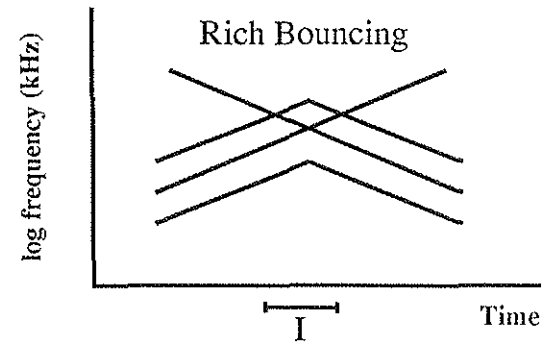
(a)



(b)



(c)



(d)

Figure 6

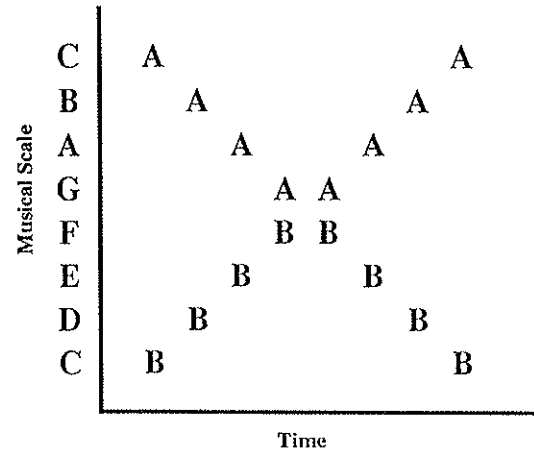
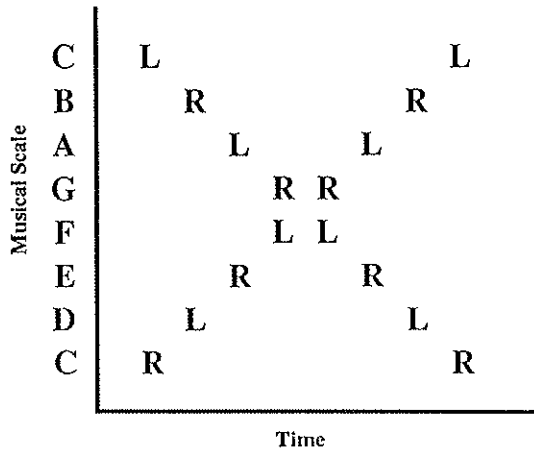


Figure 7

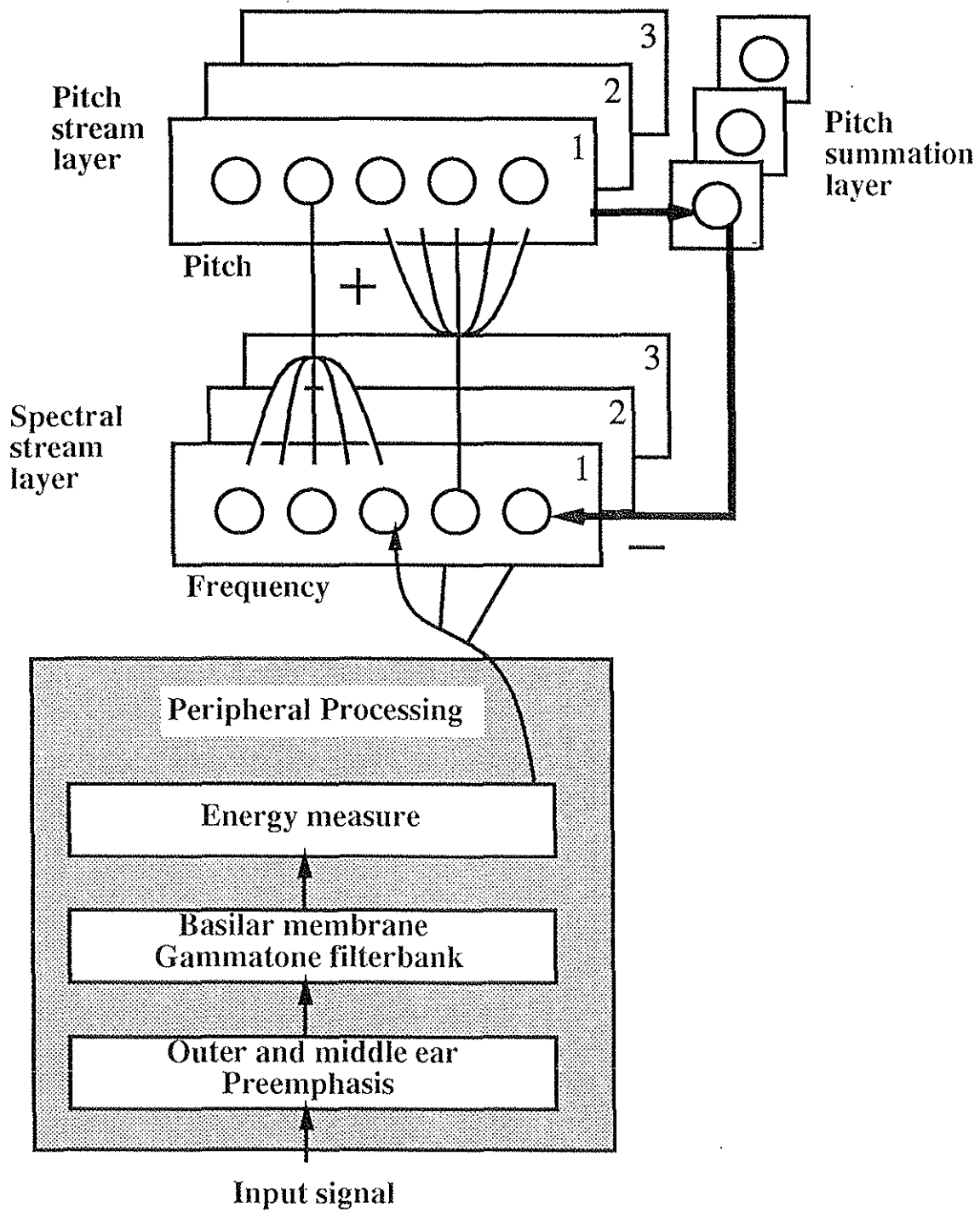


Figure 8

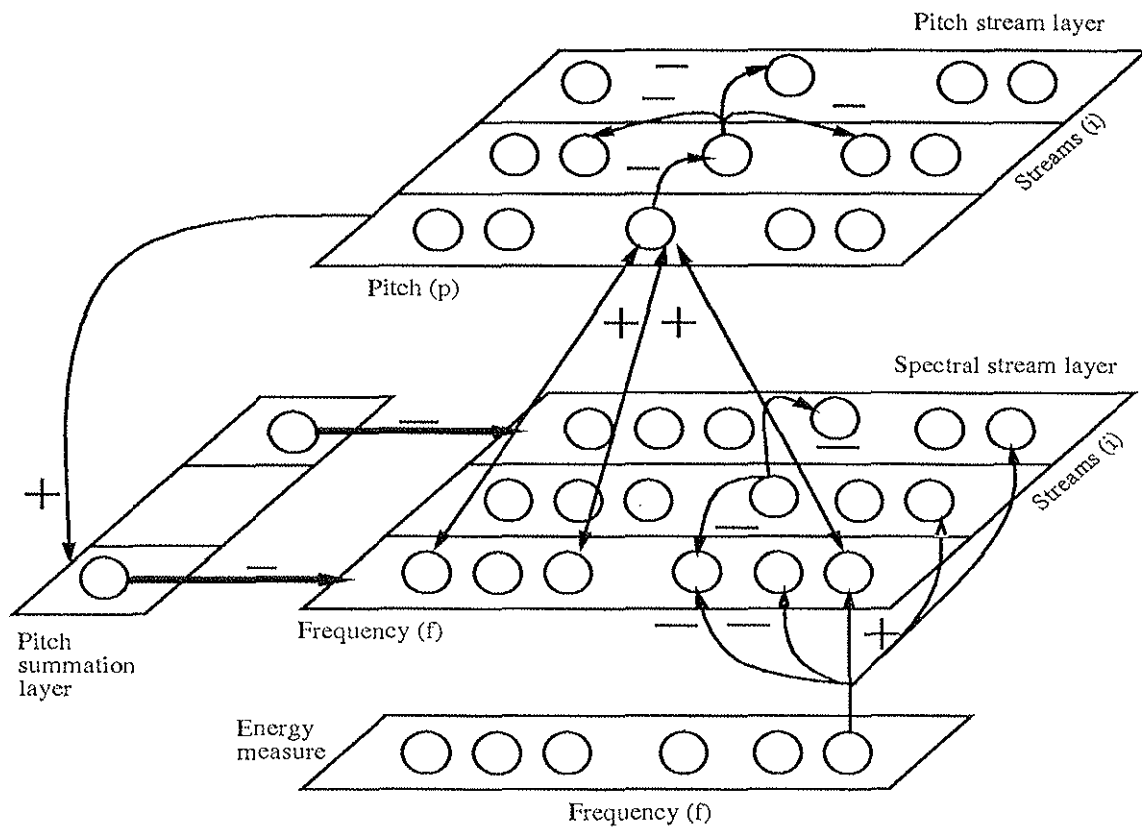


Figure 9

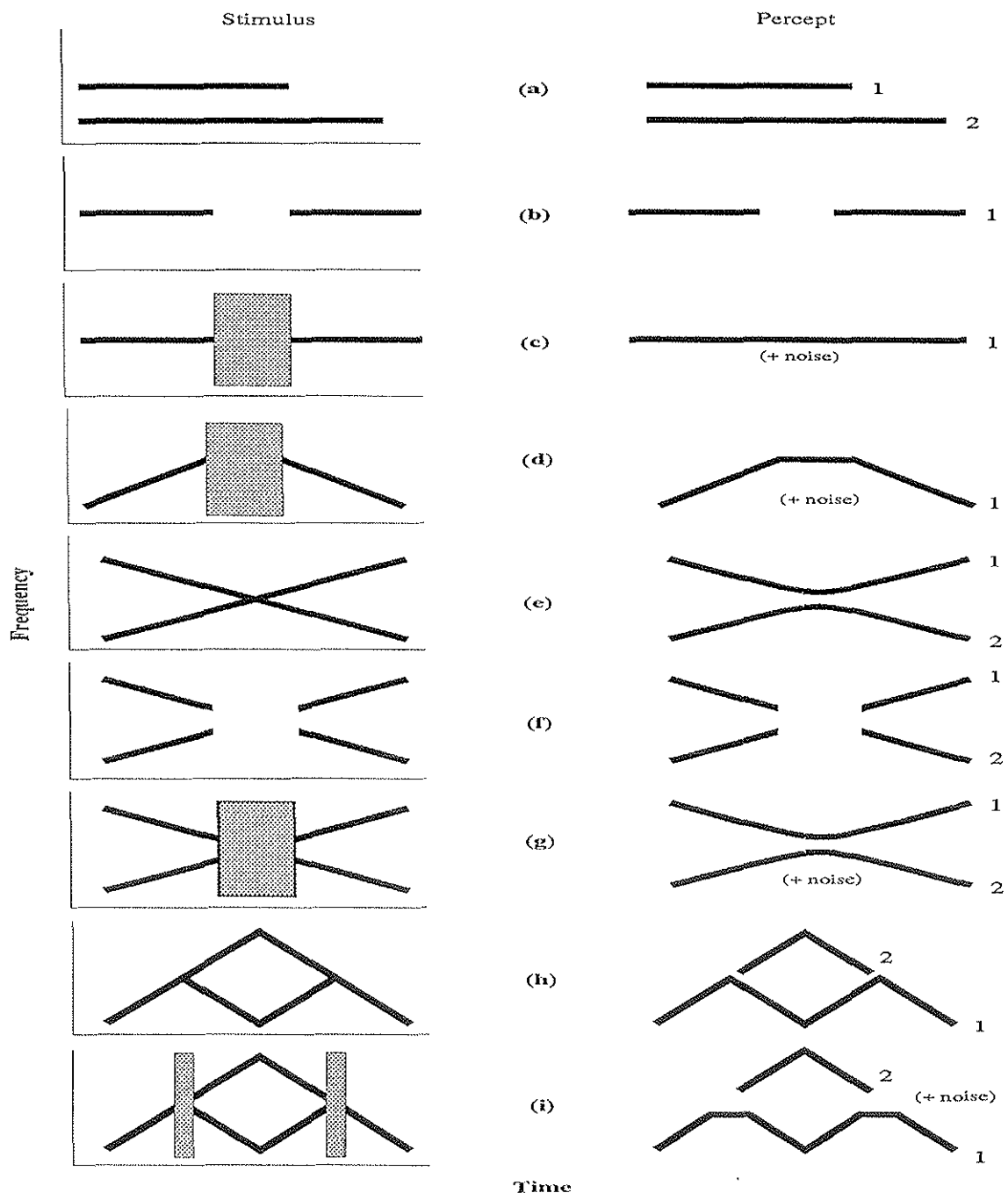
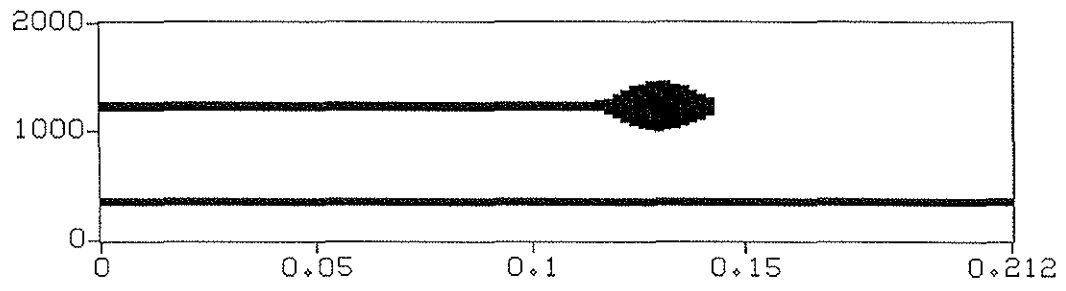
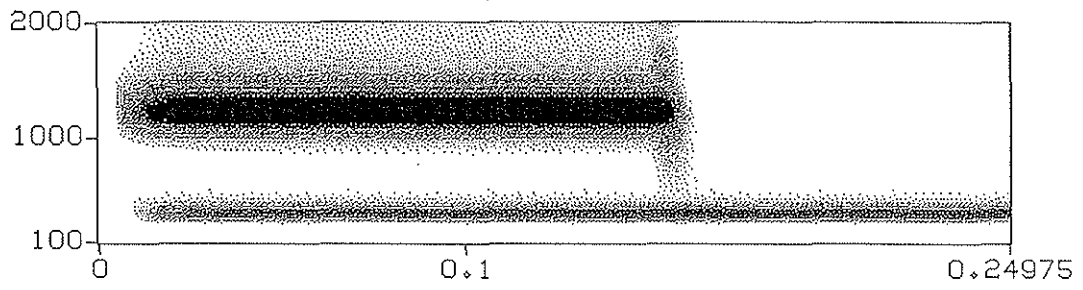


Figure 10

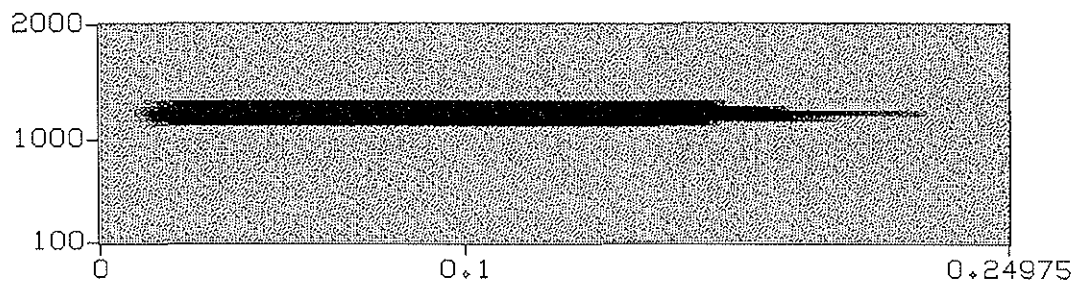


(a)

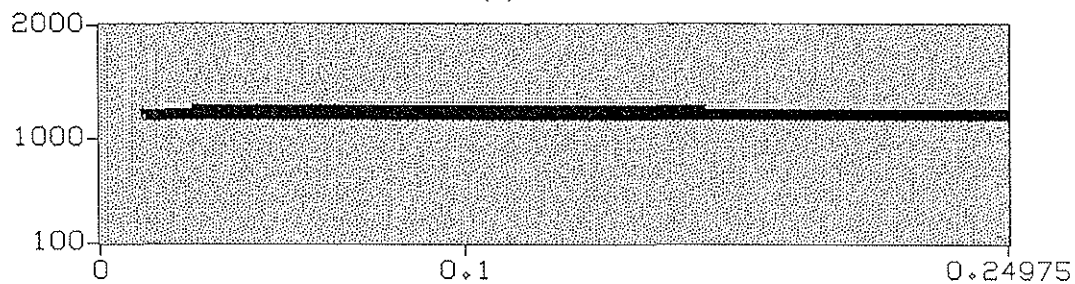


(b)

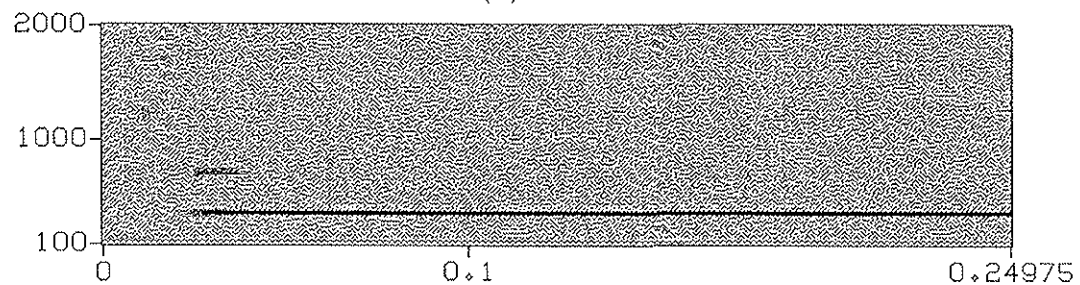
Figure 11



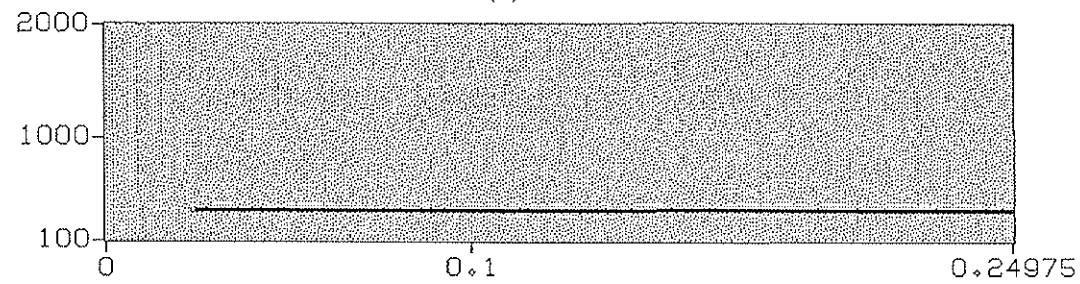
(a)



(b)



(c)



(d)

Figure 12

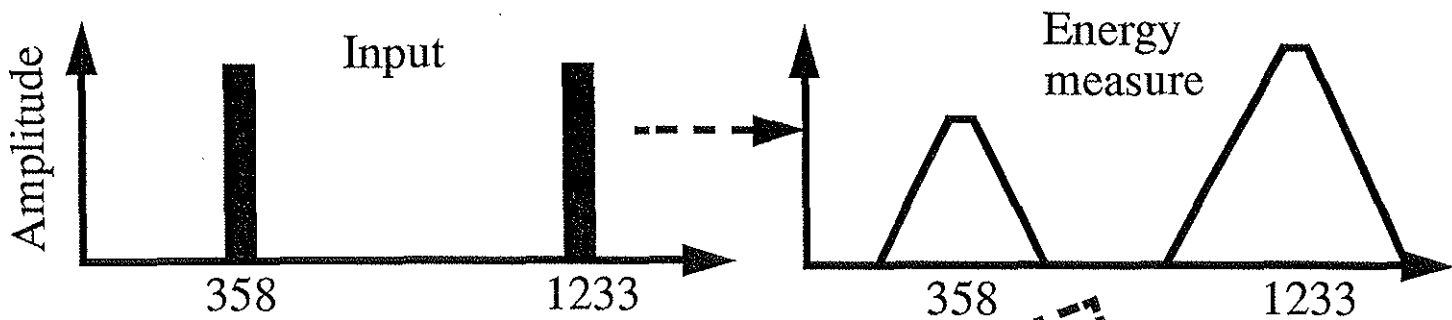
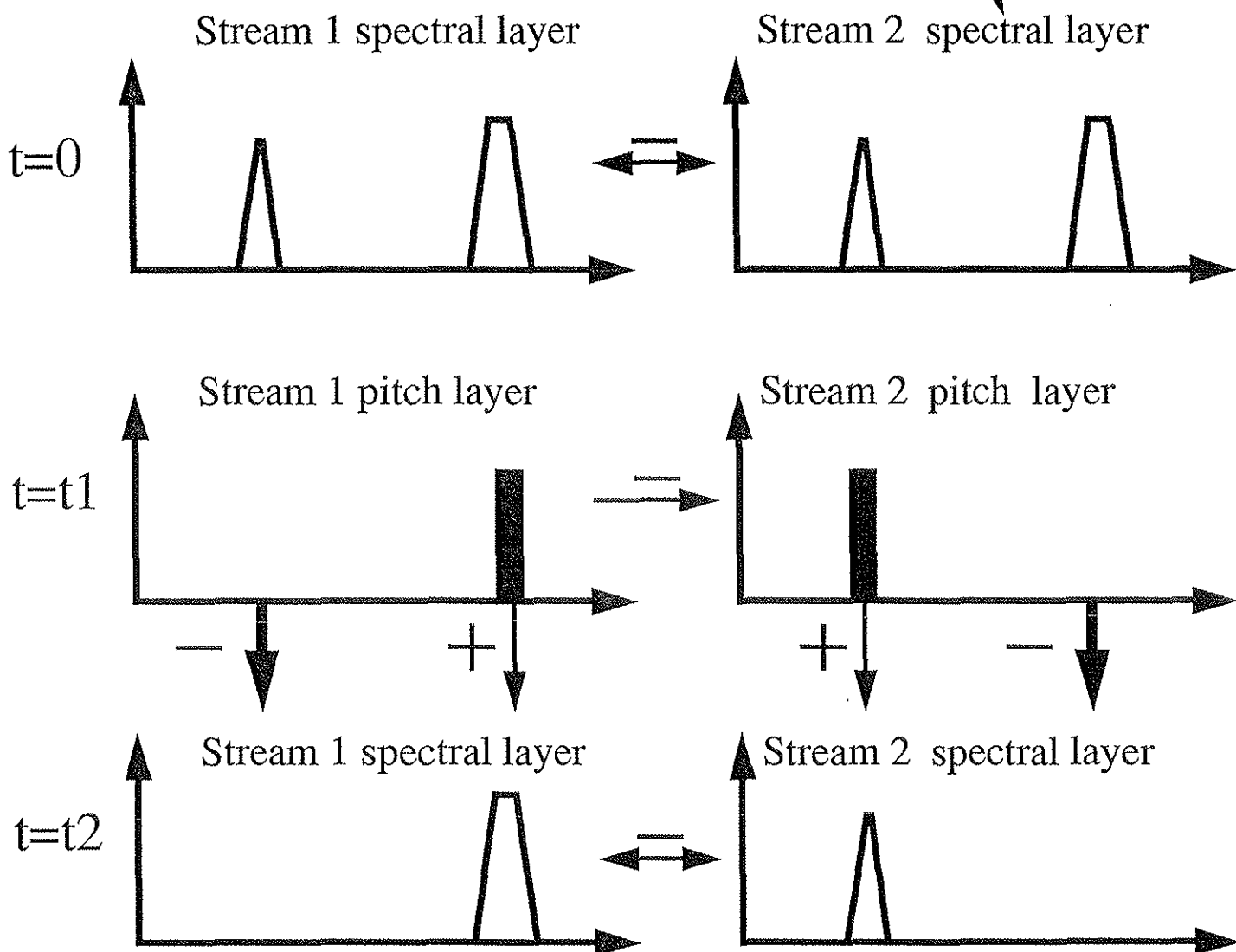
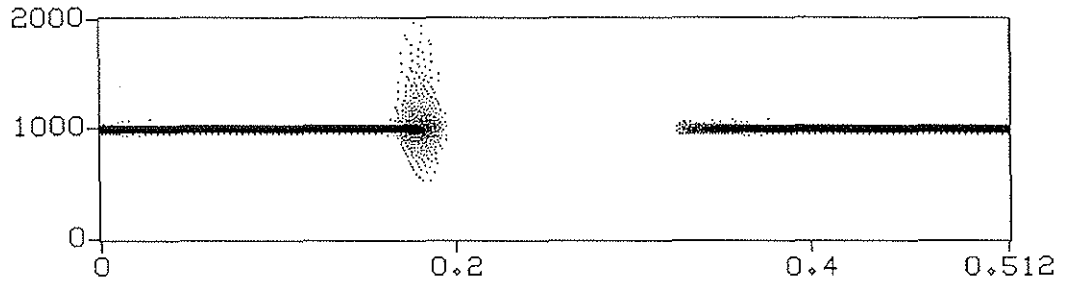
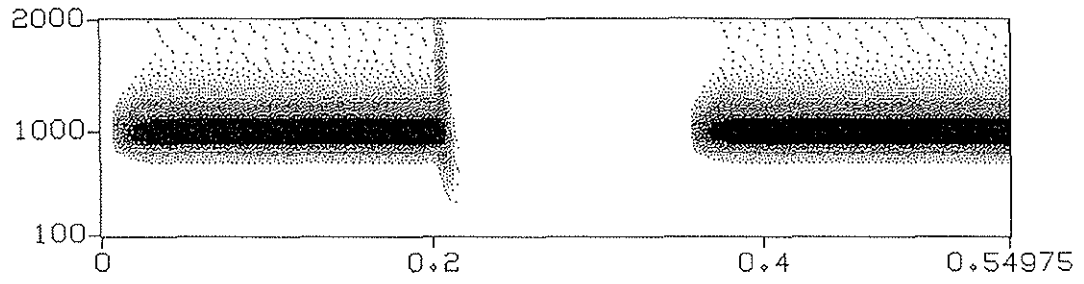


Figure 13



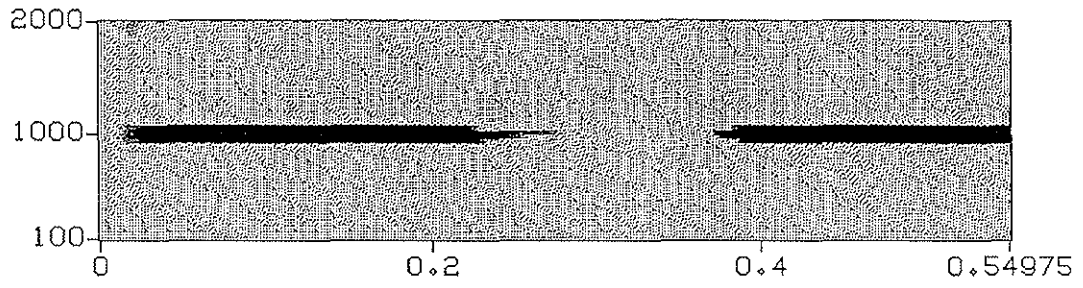


(a)

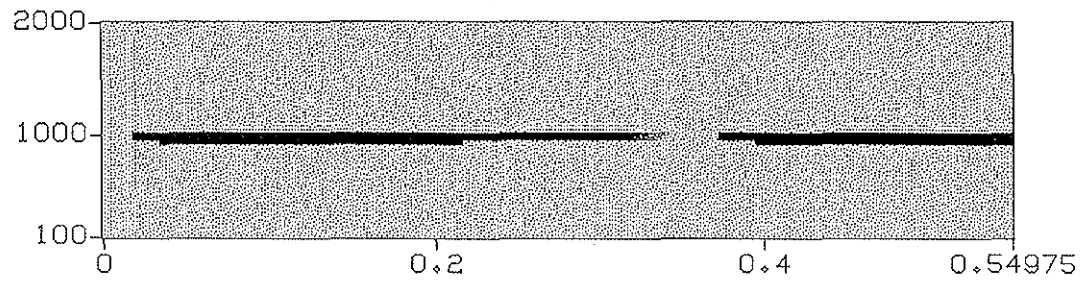


(b)

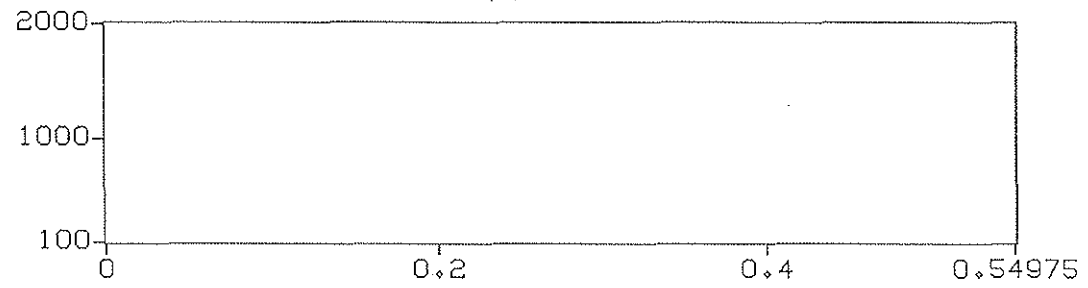
Figure 14



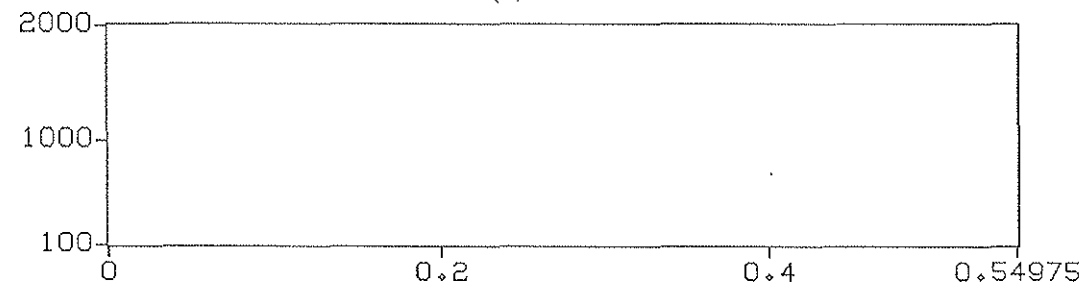
(a)



(b)

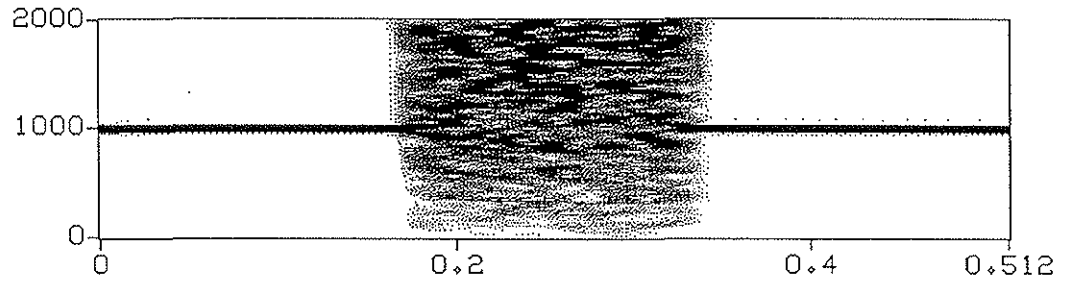


(c)

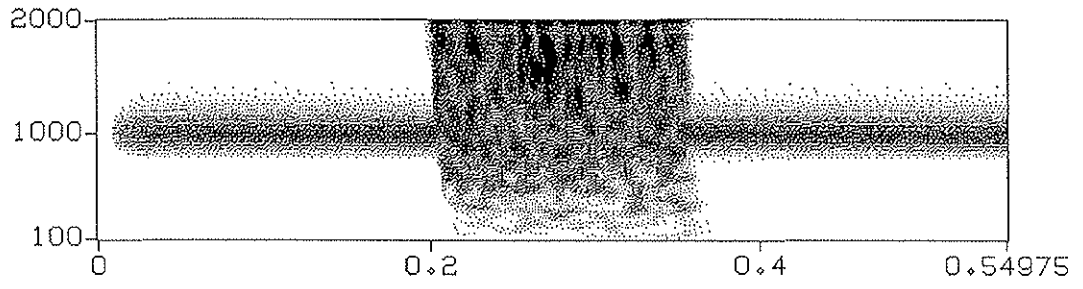


(d)

Figure 15

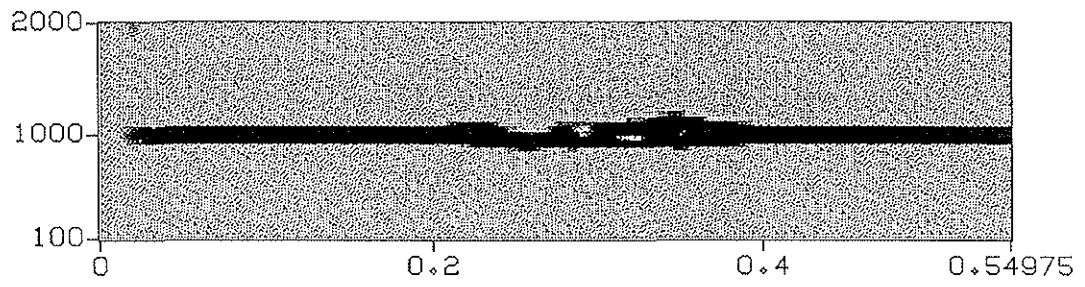


(a)

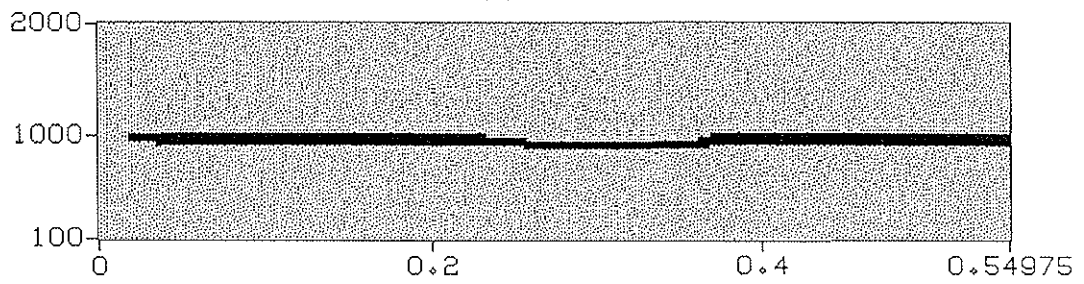


(b)

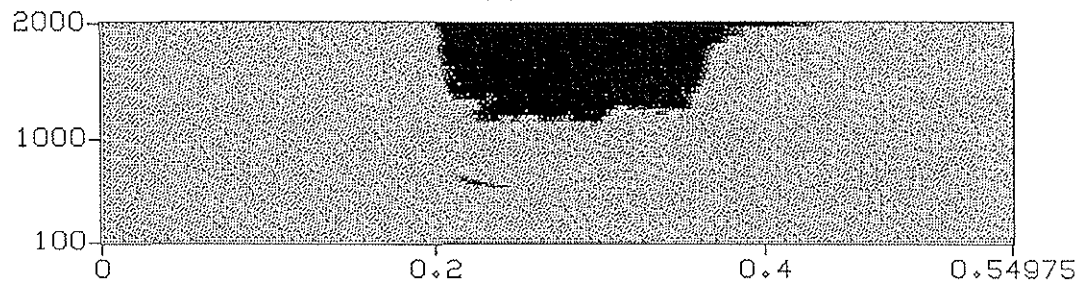
Figure 16



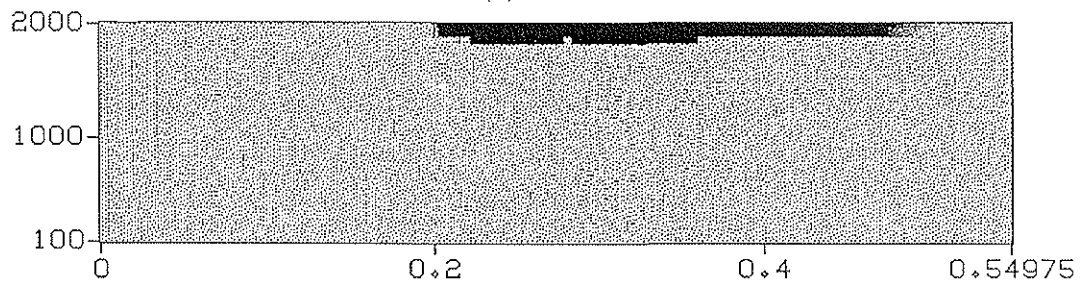
(a)



(b)

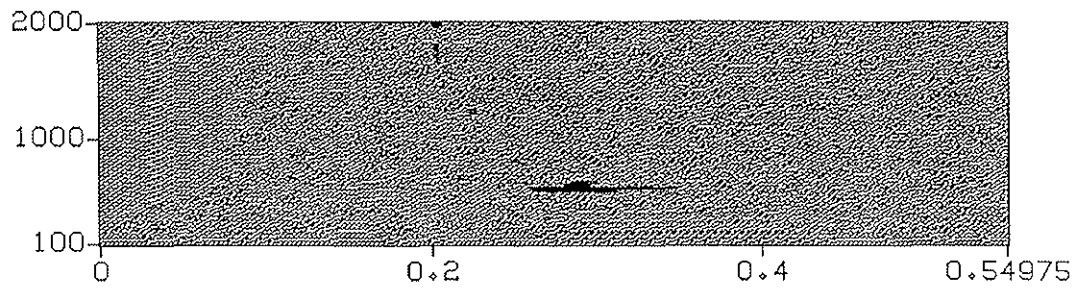


(c)



(d)

Figure 17

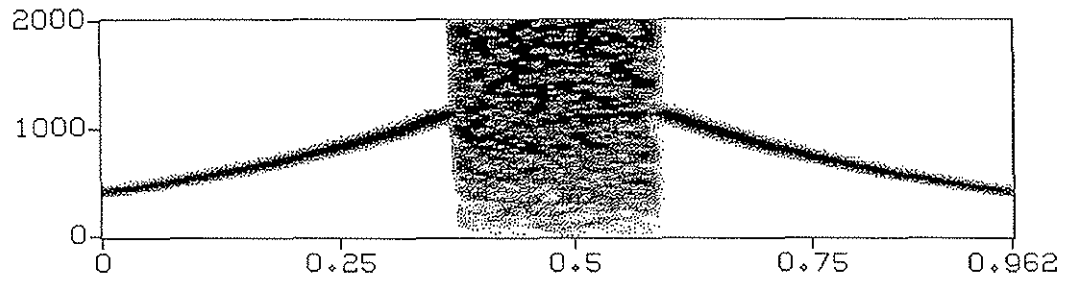


(a)

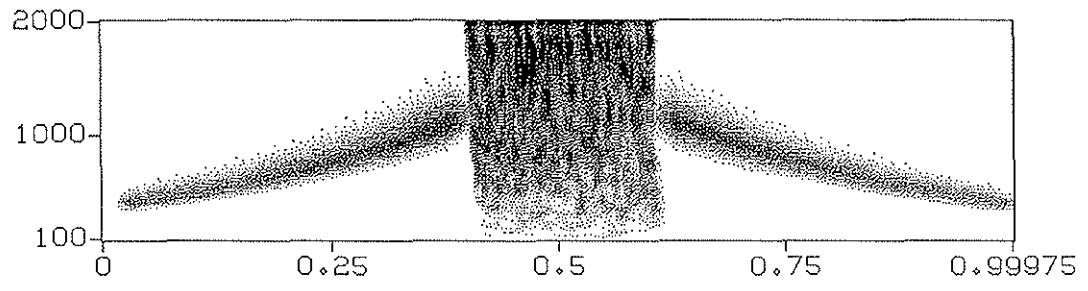


(b)

Figure 18

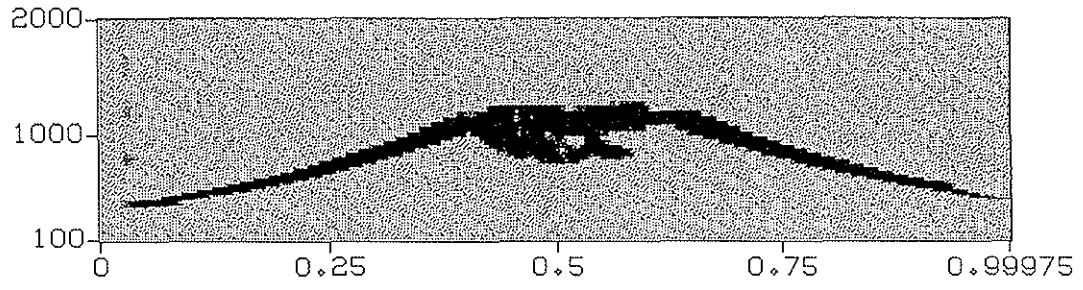


(a)

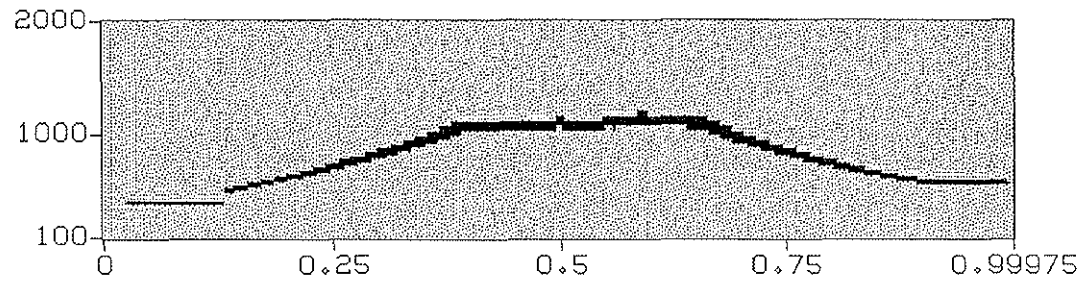


(b)

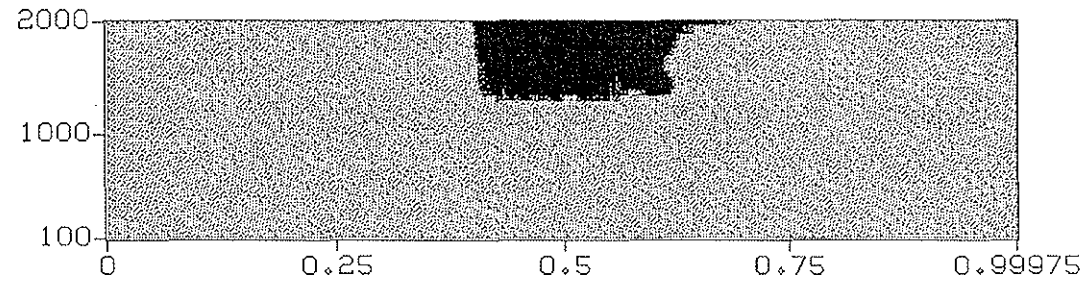
Figure 19



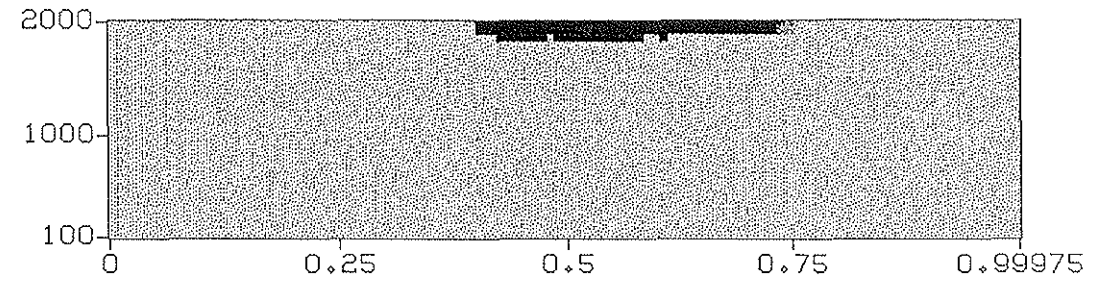
(a)



(b)

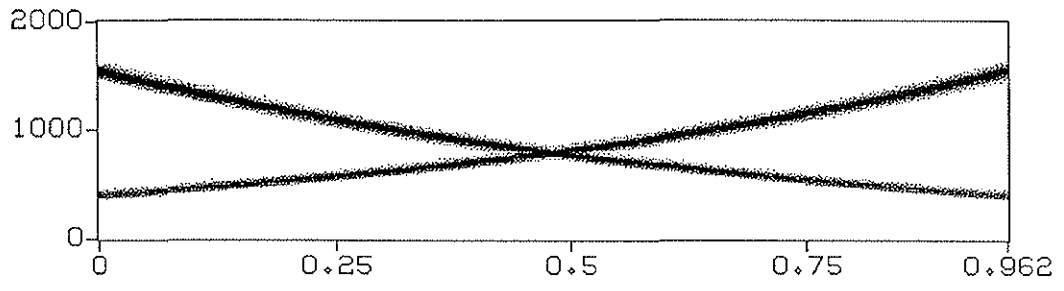


(c)

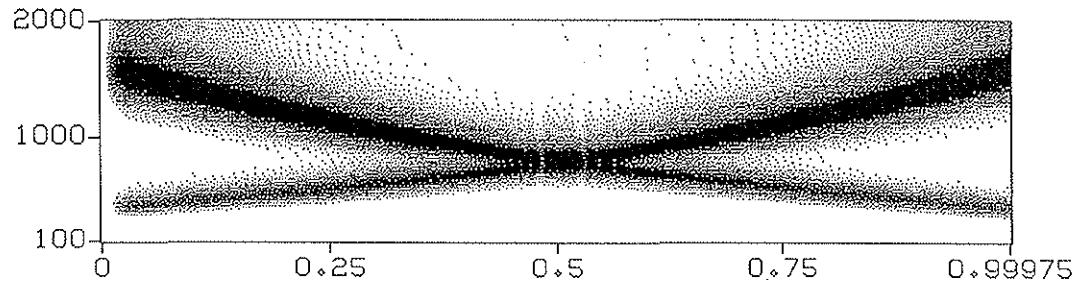


(d)

Figure 20

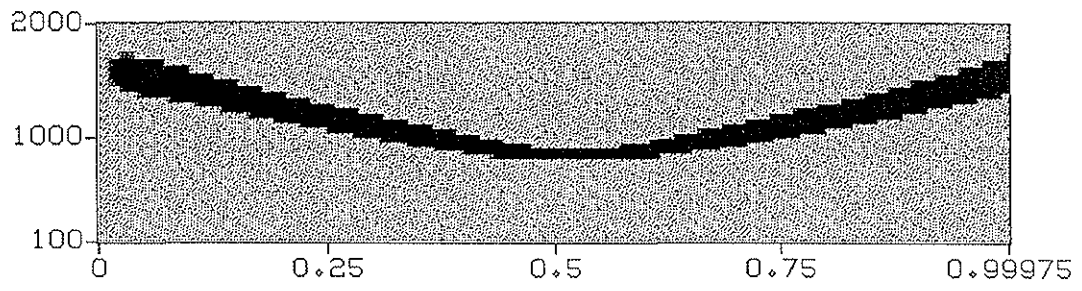


(a)

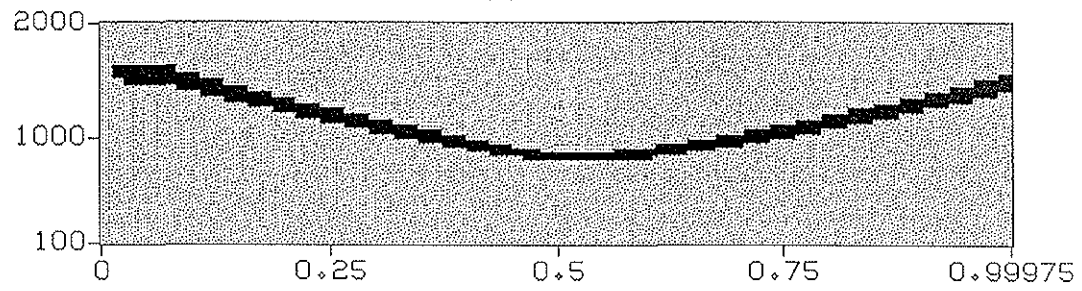


(b)

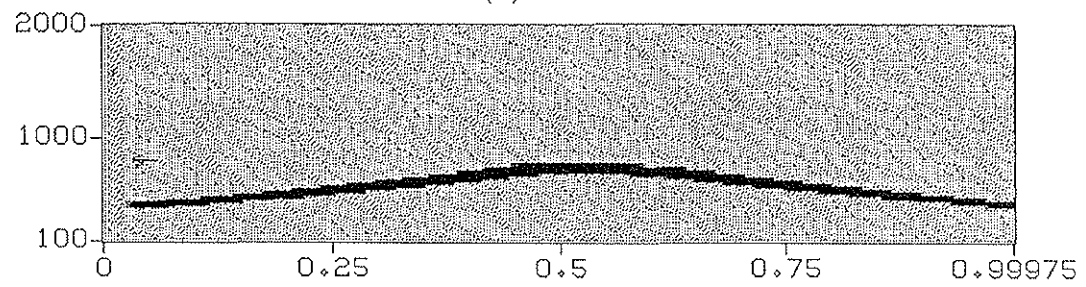
Figure 21



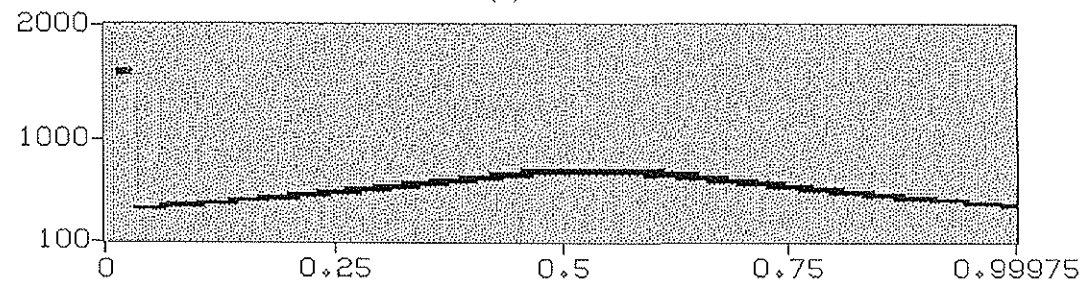
(a)



(b)

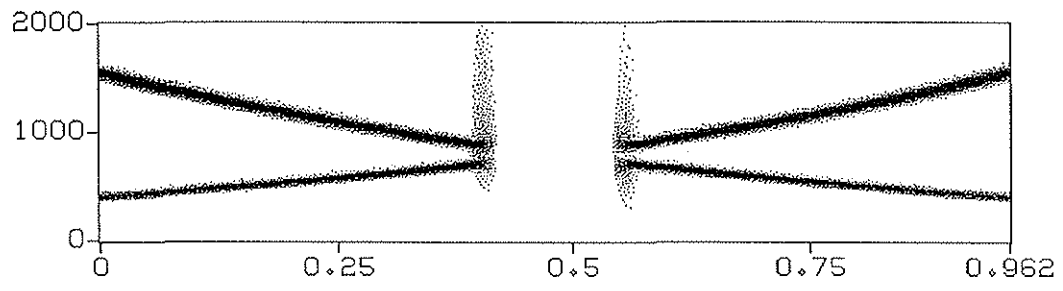


(c)

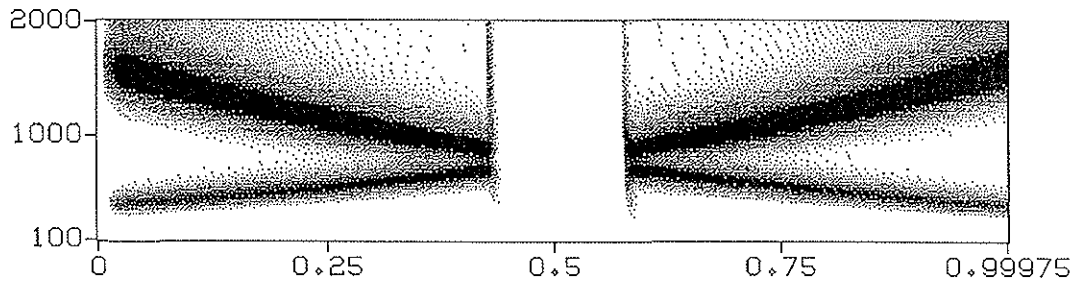


(d)

Figure 22

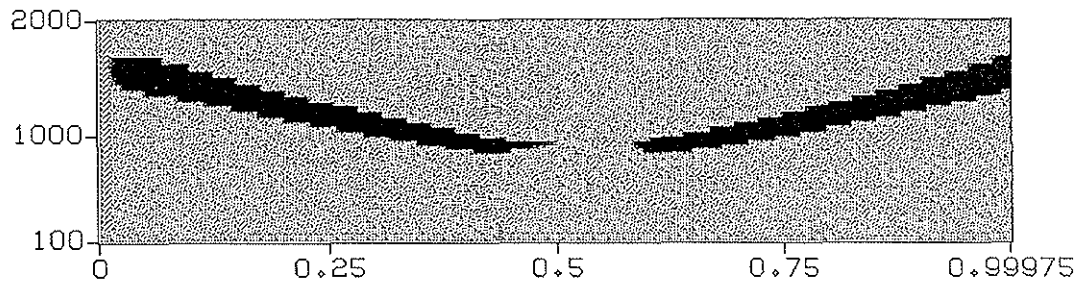


(a)

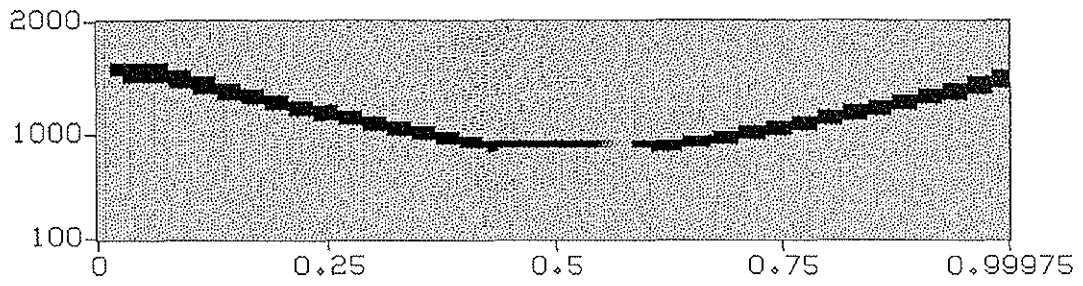


(b)

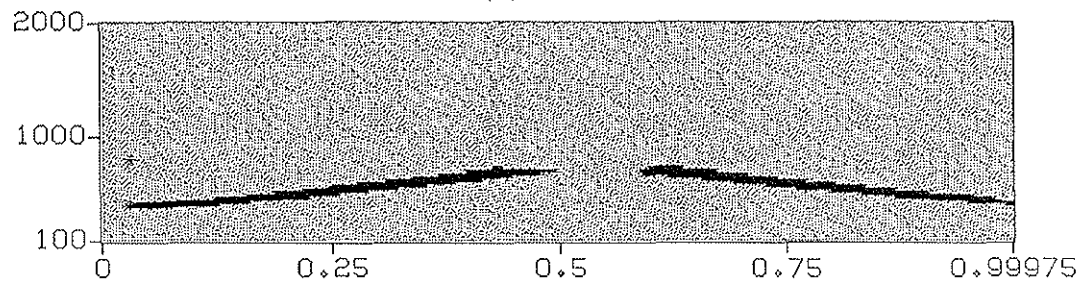
Figure 23



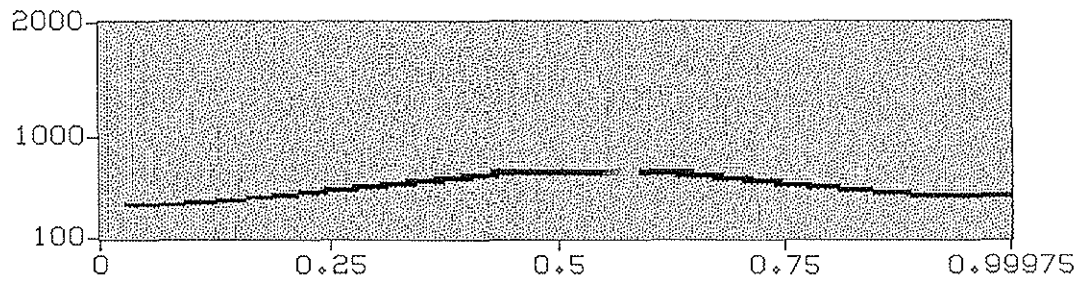
(a)



(b)

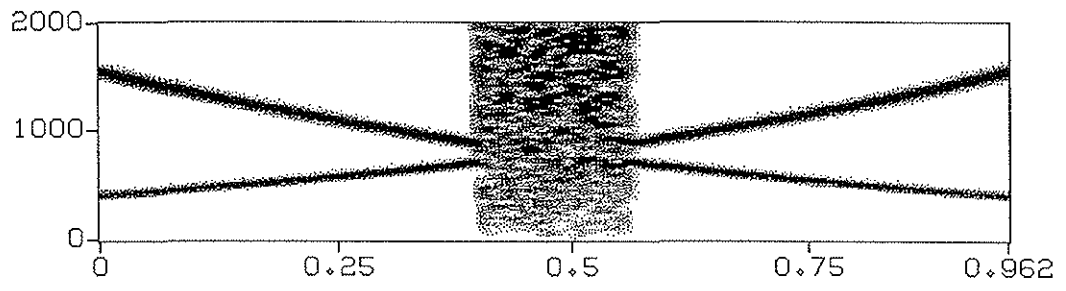


(c)

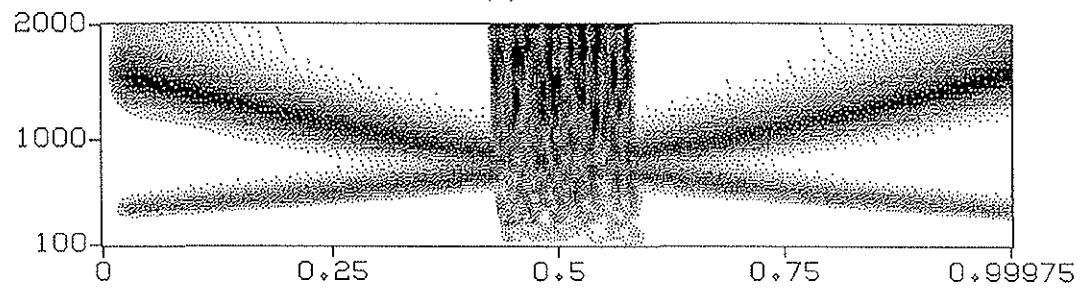


(d)

Figure 24

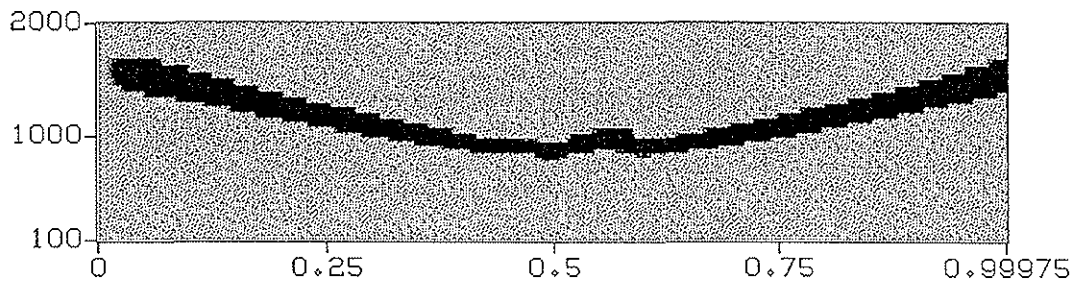


(a)

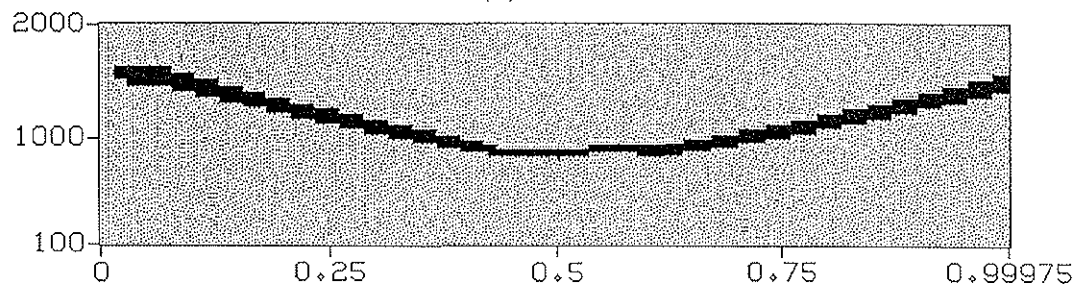


(b)

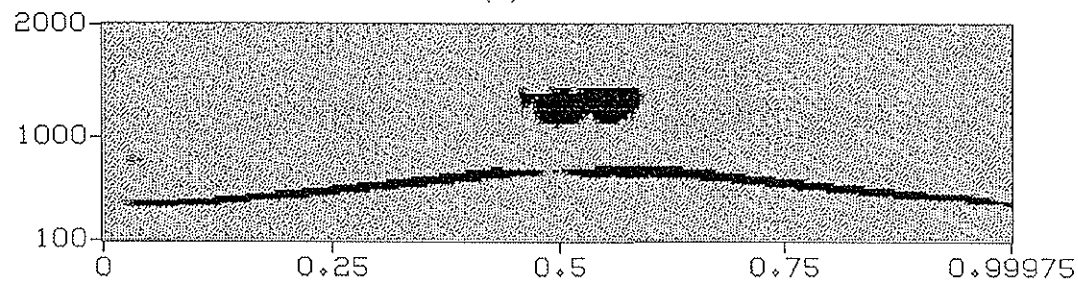
Figure 25



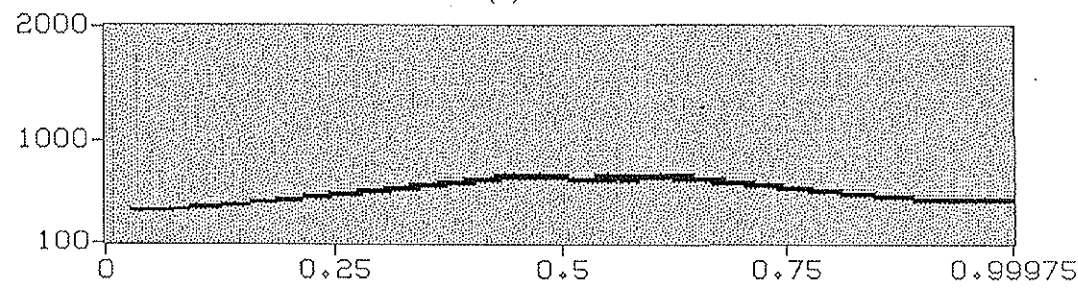
(a)



(b)

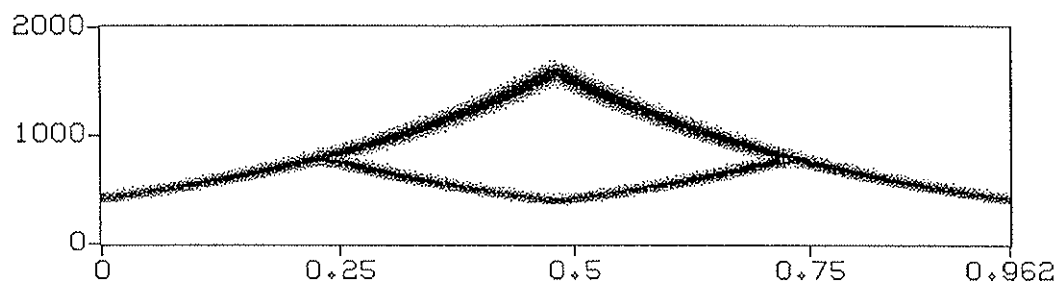


(c)

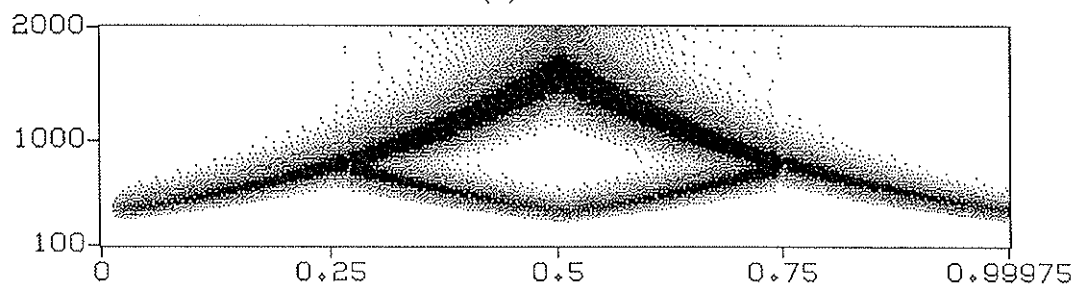


(d)

Figure 26

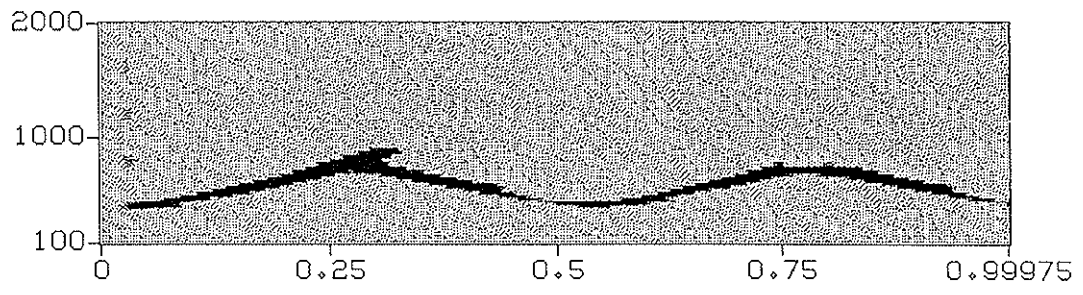


(a)

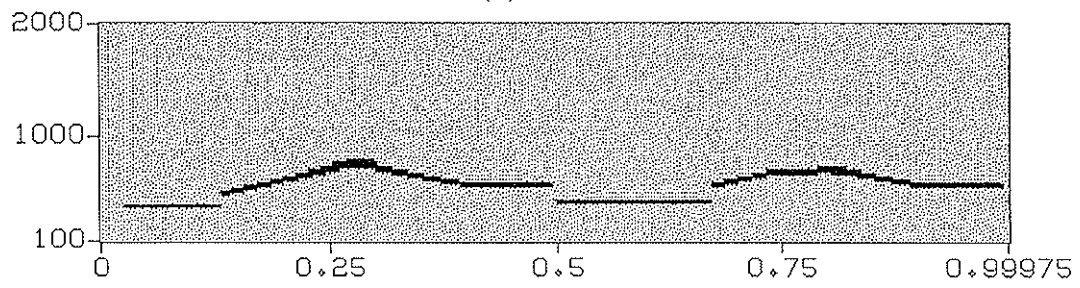


(b)

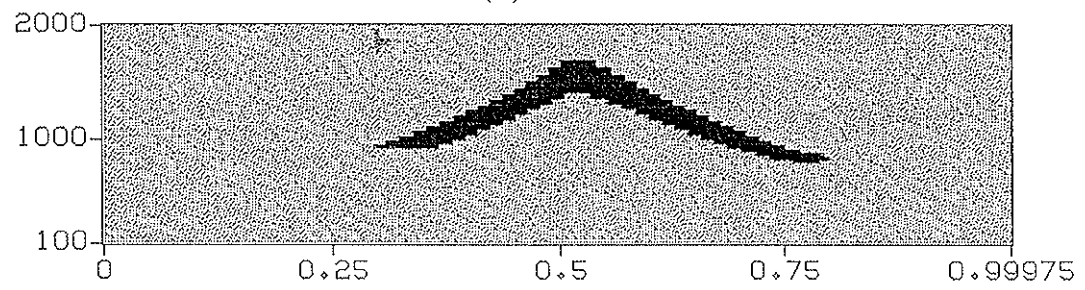
Figure 27



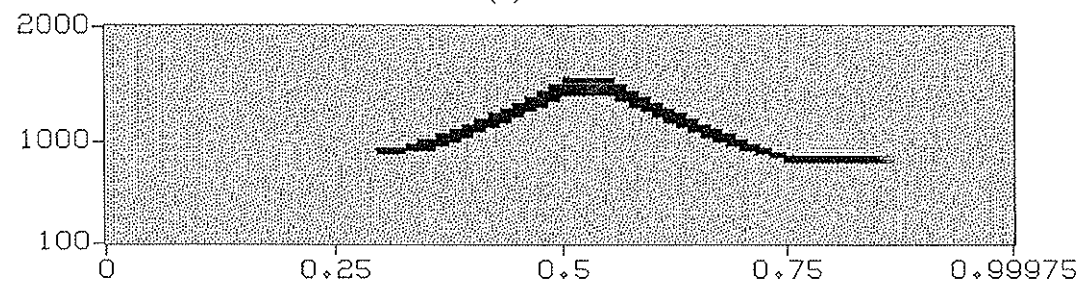
(a)



(b)

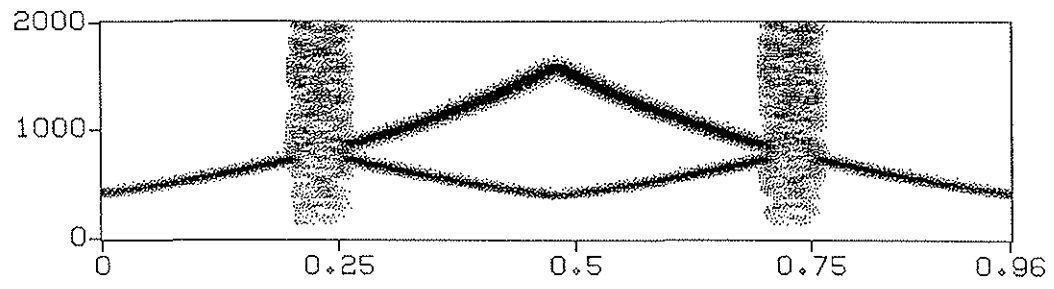


(c)

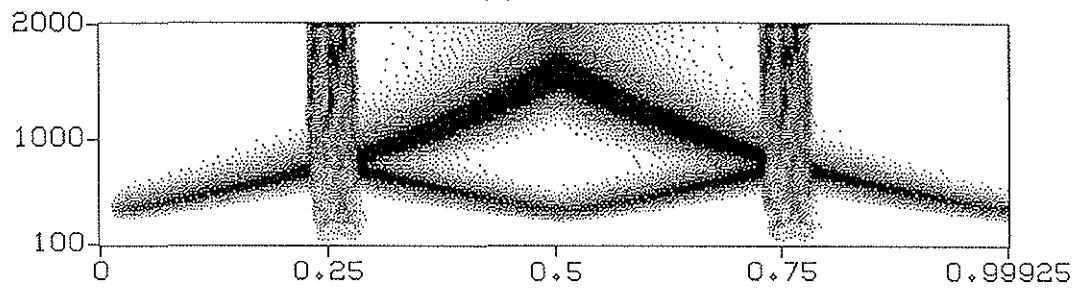


(d)

Figure 28

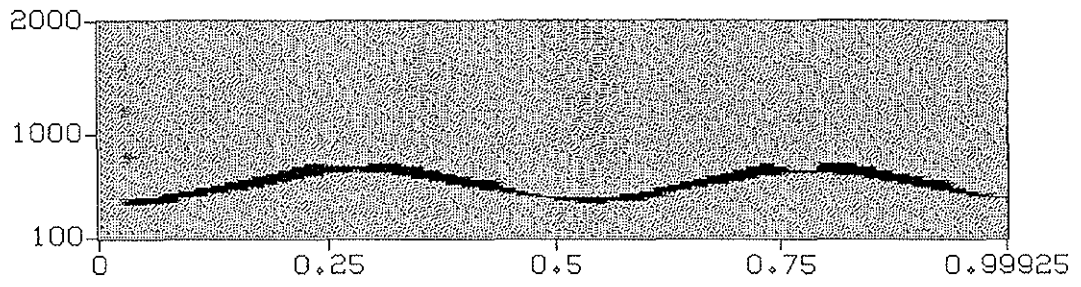


(a)

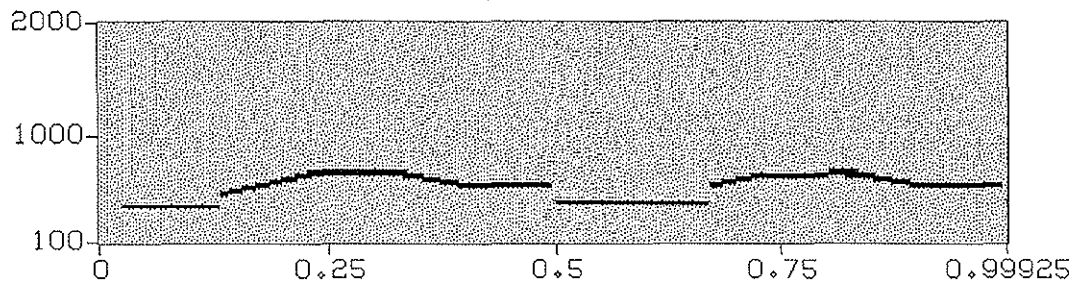


(b)

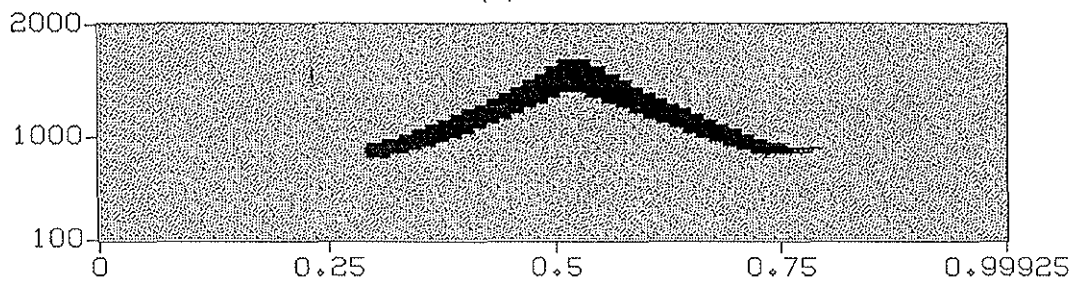
Figure 29



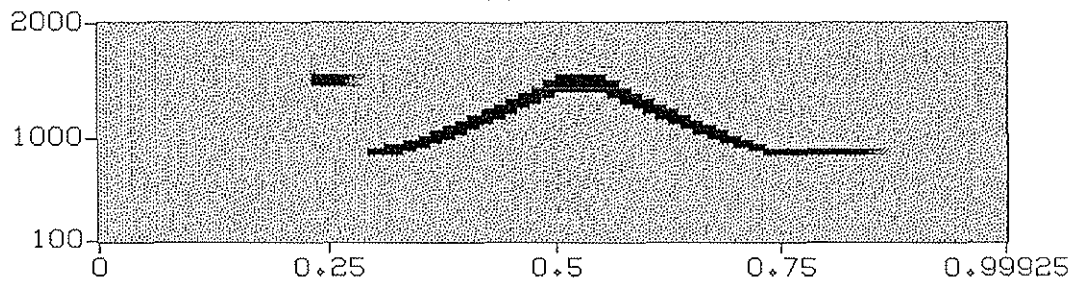
(a)



(b)



(c)



(d)

Figure 30

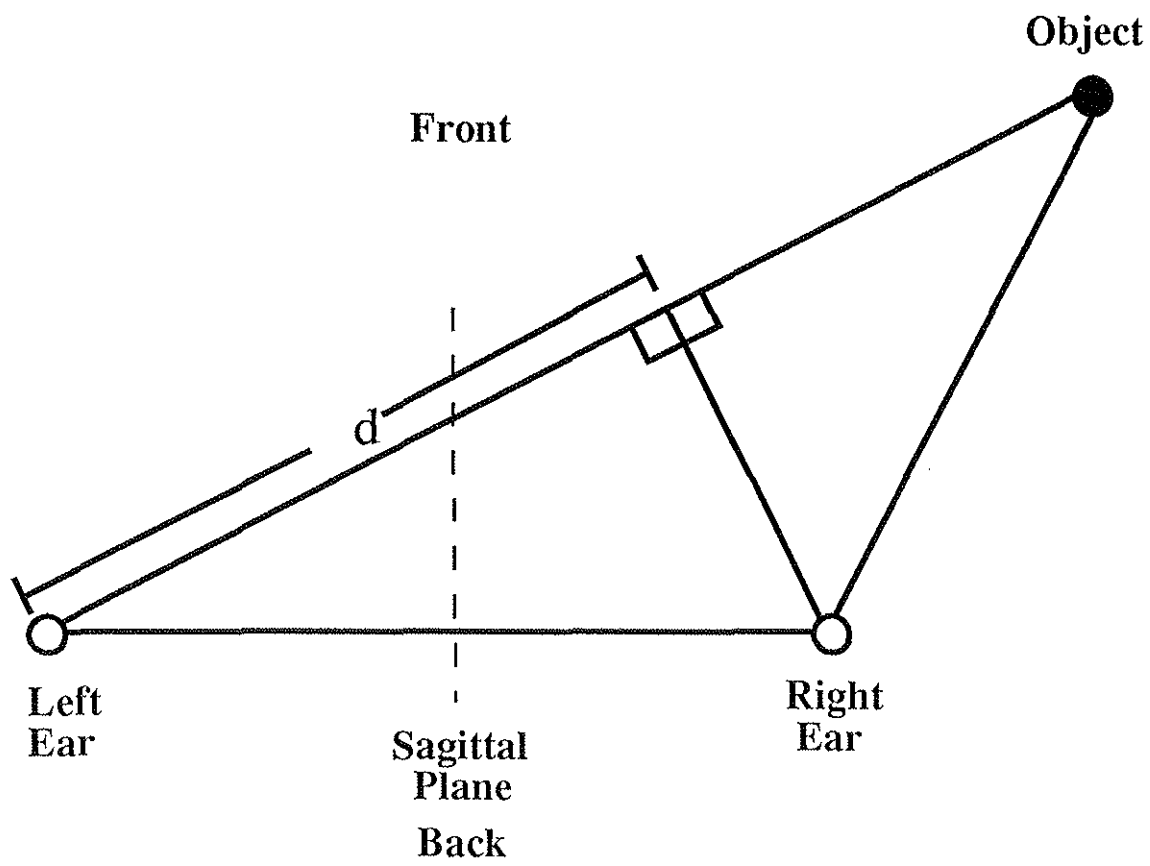


Figure 31

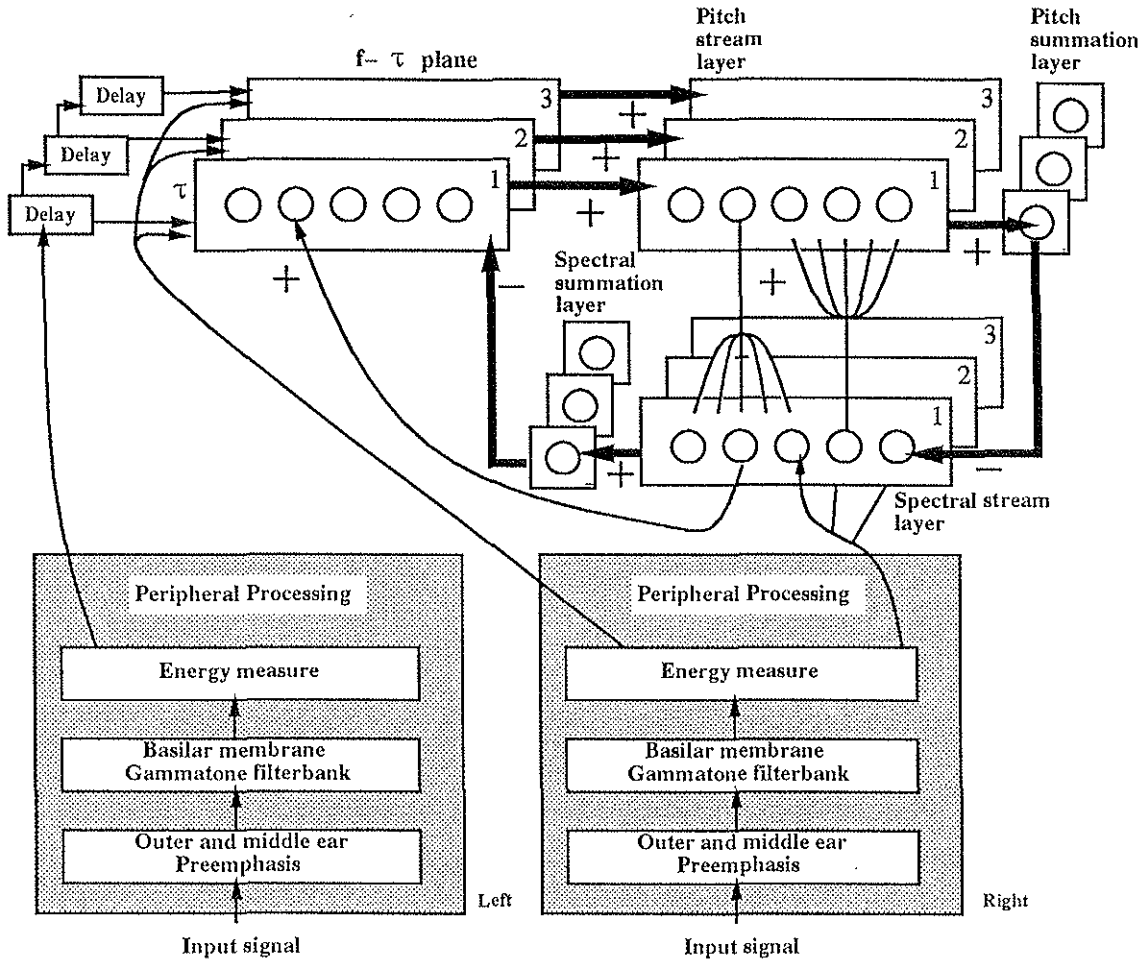


Figure 32

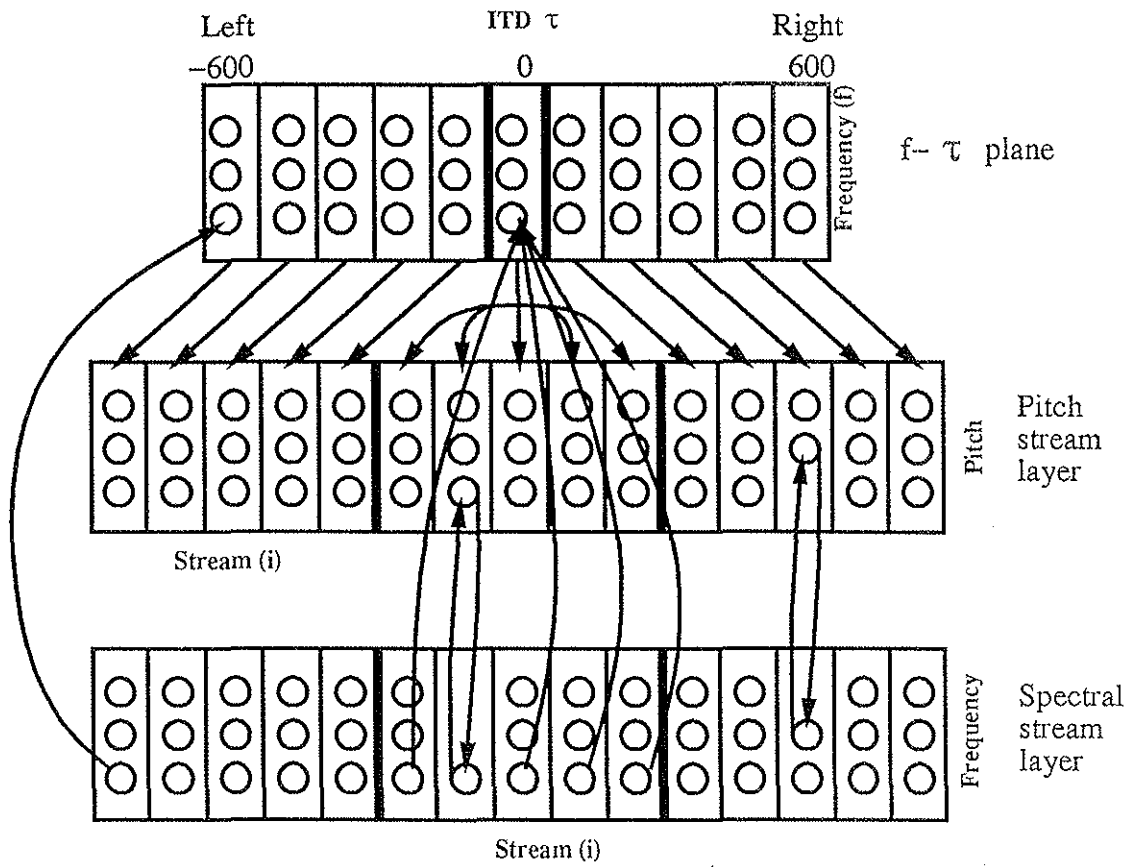


Figure 33