Computer simulation of auditory stream segregation in alternating-tone sequences

Michael W. Beauvois[a)]
Laboratoire de Psychologie Expérimentale (CNRS URA 316), Université René Descartes, 28, rue Serpente, F-75006 Paris, France and IRCAM, 31, rue St-Merri, F-75004 Paris, France

Ray Meddis
Department of Human Sciences, University of Technology, Loughborough LE11 3TU, United Kingdom

(Received 15 June 1993; revised 6 July 1995; accepted 21 November 1995)

A computer model is described that takes a novel approach to the problem of accounting for perceptual coherence in alternating pure-tone sequences by using simple physiological principles that operate at a low level. Using the same set of parameter values, the model is able to reproduce a number of phenomena associated with auditory stream segregation. These are (1) the buildup of stream segregation over time, (2) the temporal coherence and fission boundaries obtained from human listeners, and (3) the trill threshold. Whereas these phenomena are generally accounted for in terms of an auditory scene-analysis process that works on the basis of Gestalt perceptual principles, the operation of the model suggests that some Gestalt auditory grouping may be the product of low-level processes.

PACS numbers: 43.66.Ba, 43.66.Mk

INTRODUCTION

An important issue confronting auditory science today is the question of how listeners isolate one sound source from a background of competing noise. That is, how does the auditory system decompose an acoustic mixture into its constituent parts, and then assign these parts to the individual, original sound sources that created the mixture? For example, how is a person listening to an orchestra through a single loudspeaker able to hear an individual instrument? The listener must group across time a subset of the sonic events emanating from the loudspeaker into a coherent stream of sound that corresponds to the attended instrument. The segregation of a single sound source from an acoustic mixture is analogous to the visual separation of scenes into ''figure'' and ''ground'' (Koffka, 1935). We present, below, a model which aims to show that a number of auditory figure/ground phenomena can be accounted for by a small number of low-level processing principles compatible with the known physiology of the peripheral auditory system.

In auditory terms, figure–ground effects can be produced using an isochronous alternating-tone sequence composed of two pure tones of different frequencies (A and B) which are repeated continuously (i.e., ABAB...). At long tone-repetition times (TRT, or the time interval between the onset of consecutive tones in the sequence), or if there is a small frequency separation (Δf) between the tones, an observer will perceive the sequence as a connected series of tones, or a musical trill [Fig. 1(a)]. This property of continuity is known as *temporal coherence* (van Noorden, 1975). However, at short TRTs, or when there is a large Δf, the trill seems to split into two parallel sequences or ''streams,'' one high and one low in pitch, as if there were two different, but interwoven, sound sources. Here, the observer's attention is focused on only one tone stream (A or B), and the stimulus appears to have a longer periodicity equal to twice the TRT [Fig. 1(b)]. The attended stream is subjectively louder than the unattended stream, producing an auditory figure–ground percept in terms of loudness. This phenomenon is known as *auditory stream segregation* (Bregman and Campbell, 1971).

A. Factors influencing auditory stream segregation

Van Noorden (1975) plotted the occurrence of segregated and coherent percepts in ABAB sequences as a function of Δf and TRT, and found three separate perceptual areas defined by two perceptual boundaries dependent on frequency proximity and temporal proximity (Fig. 2). Above the *temporal coherence boundary* it is impossible to integrate A and B into a single perceptual stream. Below the *fission boundary* it is impossible to hear more than one stream, and the tone sequence forms a coherent whole. The region between the two boundaries is an ambiguous region perceptually, since either a segregated or an integrated percept may be heard. The choice of alternative percepts can be influenced by the observer's attentional set. However, a spontaneous alternation between the two percepts often occurs when the listener makes no conscious effort to influence the process. All ABAB stimuli begin by sounding coherent, but the probability of segregation occurring increases steadily over time as a function of total sequence duration (Anstis and Saida, 1985).

B. Auditory scene analysis and stream formation

Auditory scene analysis theory (Bregman, 1990) states that the auditory system decomposes a sound mixture by assigning the frequency components of the mixture to the separate perceptual structures that are used to represent individual sound sources. The grouping mechanisms used by the

a)Please address correspondence to M. Beauvois at IRCAM; E-mail: beauvois@ircam.fr

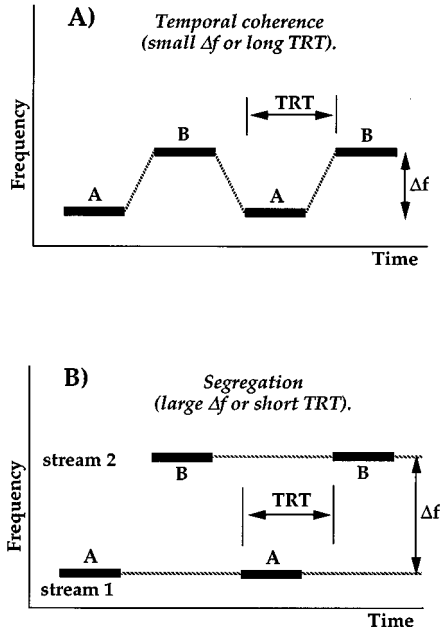

FIG. 1. (a) The percept of a temporally coherent ABAB tone sequence, and (b) the percept of a segregated ABAB tone sequence.

auditory system are broadly claimed to operate according to Gestalt perceptual principles. For example, the Gestalt principle of *proximity* states that nearer elements are grouped together in preference to those that are spaced further apart. This is demonstrated by the tendency of tones to form streams when they are in the same frequency region and there is a small Δf (e.g., below the fission boundary).

Bregman (1990) suggests that auditory stream segregation occurs as a consequence of the scene analysis process, because the auditory system is attempting to group auditory components into streams, each of which represents a separate sound source. For example, the increase of segregation reports later in the stimulus presentation is explained by Bregman (1990) in terms of a primitive segregation mechanism which gradually accumulates evidence that a tone sequence contains different subsets of sounds with distinct properties, and that these subsets should be sorted into separate streams. Similarly, the focusing on one particular stream by an observer is caused by a schema-based segregation mechanism that groups elements on the basis of the Gestalt principle of (frequency) *proximity*—i.e., the high tones in a sequence tend to group with other high tones if brought close to them in time by a decrease in TRT.

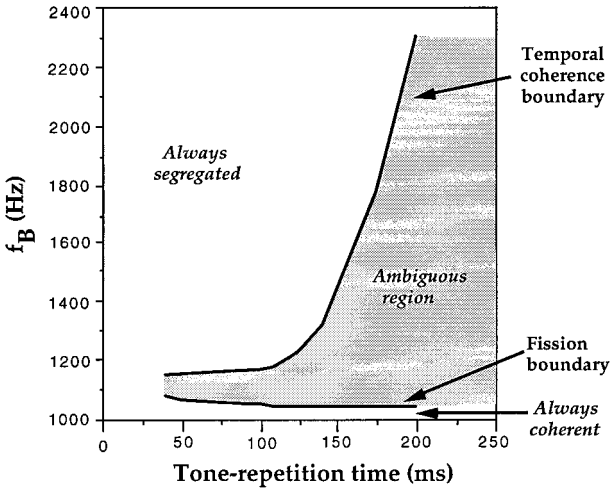

FIG. 2. The temporal coherence boundary (upper curve) and the fission boundary (lower curve) for an ABAB sequence composed of alternating 40-ms pure tones (redrawn from McAdams and Bregman, 1979). The percept of the listener is determined by the frequency separation between the two tones ($f_B - f_A$) and the tone-repetition time (see text). Here, $f_A = 1000$ Hz.

C. A physiologically based approach to auditory stream formation

A computer model of the auditory system is demonstrated below, which is designed to work specifically on alternating pure-tone sequences (ABAB...), and which uses low-level auditory analysis to account for some stream segregation phenomena associated with these stimuli. The computer model is based upon simple circuits analogous to physiological systems present in the lower auditory system and auditory periphery, and exhibits similar behavior to that of human listeners for certain simple stimuli. The approach taken here can be seen as being complementary to the Gestalt approach, but at a different level of explanation, inasmuch as it explores the possibility that the physiology of the lower auditory system is responsible for some segregation phenomena.

In this account, we build on earlier work (Beauvois and Meddis, 1991, 1995) by refining the peripheral model to include adaptation of the auditory-nerve (AN) response to tones, and we extend the range of stimulus paradigms evaluated. At the heart of the model lie two general ideas that we seek to evaluate. The first principle is the idea that ''streaming'' can be construed usefully as a selective accentuation of one auditory object. That is, one of two simultaneous sounds is heard as louder than the other, even though both sounds have the same physical intensity. This is not the same as saying that the two objects are allocated to two separate streams and only one is heard. Both are clearly heard, but one has greater perceptual salience than the other. The question then arises as to how this salience is internally managed and why foreground/background shifts appear to occur spontaneously.

The second principle suggests that these shifts are an inevitable consequence of the stochastic nature of the neural activity in the auditory periphery. The probabilistic nature of spike generation in the AN and elsewhere in the brain stem means that there is random variation in the response of the system to stimuli of otherwise similar intensity. When this variation passes through neural systems characterized by low-pass filtering, we obtain random-walk phenomena with properties which, we claim, can explain some of the effects associated with auditory foreground/background perceptual experiences. Other principles are undoubtedly at work, such as descending influences that, presumably, are involved in conscious control of the percept and lateral inhibition. We

acknowledge these as possible contributors, but they are not studied in the current model.

I. MODEL DESCRIPTION

A. Summary of model features

(1) The acoustic signal is first subjected to a peripheral frequency analysis which establishes ''channels'' characterized by a bandpass frequency response to stimuli.

(2) The output of each bandpass filter is fed into a simulation of a group of inner-hair-cell/AN-synapse units.

(3) Each model channel subdivides into three pathways:

(3.1) A temporal fine-structure path preserves all aspects of the AN output to enable the signal to be processed by higher levels, and to preserve all AN information for pitch extraction purposes.

(3.2) A temporal-integration path temporally integrates the AN output.

(3.3) An excitation-level path adds a cumulative random element to the output of the temporal-integration path, and then subjects it to a slower temporal integration process.

(4) The output of the excitation-level paths are examined to see which one has the highest excitation level compared to the other channels. The channel with the highest excitation level is then defined as the ''dominant channel.''

(5) The activity in all amplitude-information pathways is then attenuated by a factor of 0.5, except for the dominant channel.

(6) The model amplitude output is the sum of the outputs of the attenuated and nonattenuated amplitude-information paths.

(7) Stream segregation is assessed on the basis of the relative amplitude levels of tones A and B in the model output. If the levels are comparable, the percept is judged to be coherent. Otherwise, a ''segregated'' percept is reported.

A diagram showing the construction of one model channel is shown in Fig. 3. All the stimuli presented to the model are composed of pure tones of a fixed input level. Most stages of the model (apart from the stimulus, bandpass frequency filter, and IHC/AN simulation) are evaluated using a sampling rate of 1 kHz. The stimulus, bandpass frequency filter, and IHC/AN simulation use a sampling rate of 20 kHz.

B. Stage 1: Stimulus

The stimuli presented to the model consist of number sequences composed of two alternating pure tones (with frequencies f_A and f_B) and intervening silences. The tones are computed as

$$s_t = \alpha \cos(2\pi f_j t), \tag{1}$$

where s_t is the instantaneous amplitude level of the tone, t is the time, f_j is the frequency of the tone, and α is the input signal amplification factor scaled to give an arbitrary rms of 1.0 at 30 dB. In this implementation, one stimulus level (75 dB) was used for all tones, and was taken to be equivalent to the level of a pure tone at 35 dB above threshold. However, various tone durations, sequence durations, and onset/offset ramps were used for each model simulation (described below).

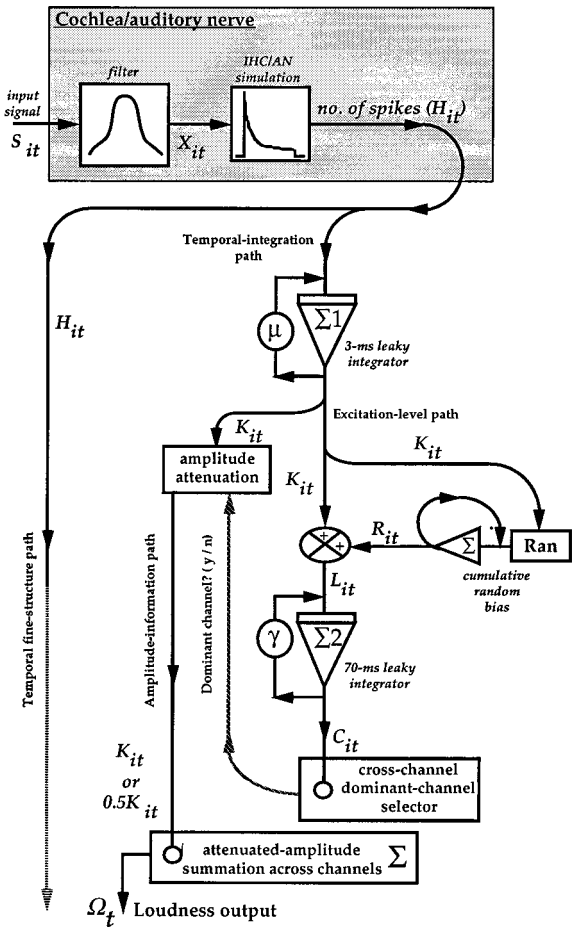

FIG. 3. Diagram showing the construction of one model channel.

C. Stage 2: Filter bank

The second stage of the model is a bandpass auditory filter bank, which calculates the output from the ith filter in response to a tone with frequency f_j using formulas suggested by Glasberg and Moore (1983)—see also Patterson and Moore (1986) for a detailed discussion of this topic. The formula used was

$$\sqrt{W_{ij}} = (1 + (p_i g_i))\exp(-p_i g_i), \tag{2}$$

where $\sqrt{W_{ij}}$ is the amplitude attenuation in the ith channel of the frequency f_j (f_A or f_B),

$$p_i = 4 f_j / \mathrm{ERB}(\mathrm{CF}_i) \tag{3}$$

and

$$g_i = (\mathrm{CF}_i - f_j)/\mathrm{CF}_i, \tag{4}$$

where CF_i is the center frequency of the ith filter in Hz. The ERB (equivalent rectangular bandwidth) of the filter was calculated using the formula given by Glasberg and Moore (1990):

$$\mathrm{ERB}(\mathrm{CF}_i) = 107.939(\mathrm{CF}_i/1000) + 24.7. \tag{5}$$

The output from each filter (x_{it}) is the amplitude of the input signal attenuated by the filter function. Hence, when a tone is active,

$$x_{it}=s_{t\surd}W_{ij} \tag{6}$$

and when no tone is active

$$x_{it}=0. \tag{7}$$

Note that j identifies the tone active at time t.

The filter bank has, in principle, a large number of filters. However, for ABAB stimuli, it seems reasonable for evaluation purposes to consider only those channels with CFs equal to f_A and f_B, and one channel midway between the two (arithmetic mean frequency). This arrangement has the virtue of saving a considerable amount of computing time for signals of long duration.

D. Stage 3: IHC/AN simulation

When the inner hair cells (IHCs) associated with an auditory filter are stimulated by a tone, they initiate electrical impulses (spikes) in the AN. The greatest rate of impulses occurs at the onset of the tone. After a few ms, however, the firing rate of the IHCs very rapidly declines along a steep slope, which then levels off into a gentle slope which continues until the end of the tone. During a period of silence there is a low continuous firing-rate level—the spontaneous firing rate—until another tone is presented, whereupon the process starts all over again (Westerman and Smith, 1984). The response of a population of AN fibers can be simulated by a computer model of a coupled IHC/AN-fiber unit. A study by Hewitt and Meddis (1991) which evaluated a number of such models showed that the Meddis IHC model (Meddis, 1986, 1988; Meddis *et al.*, 1990) was the most effective, and this model was incorporated into the present model by passing the output of each filter into it. The IHC model converts the filter output into an estimate of the probability of spike occurrence in the postsynaptic AN. The parameters of the IHC model were set to simulate a fiber with a spontaneous rate of about 35 spikes/s, a saturated rate of about 150 spikes/s, and a limited (30 dB) dynamic range (Hewitt *et al.*, 1992).

Individual AN spikes are generated from the IHC firing probability using pseudorandom number techniques (Hastings and Peacock, 1975, p. 41). Due to the probabilistic nature of the technique, a different pattern of AN spikes is generated with each stimulus presentation. In this implementation, 60 AN-fiber units, all with the same CF, were simulated for each model channel. The output of this stage (H_{it}) represents the number of active (spiking) AN fibers in each model channel.

Although the IHC model uses a 20-kHz sampling rate, later stages of the model use a 1-kHz sampling rate, which considerably reduces computing time. The IHC/AN simulation output (H_{it}) is, therefore, the number of spikes over the previous 1-ms period.

E. Stage 4: Division into separate pathways

Each model channel divides into two separate pathways. One pathway (the temporal-integration path) temporally integrates the AN output. A second pathway (the temporal fine-structure path) preserves the AN output for possible later use in determining the various qualities of the model output. In principle, this allows higher-level processes to extract pitch information from the signal on the basis of the temporal fine structure of the signal, or the place of origin on the basilar membrane. This process is not explicitly addressed in this implementation, however, because, as will be seen, the temporal fine-structure path plays no active role in the generation of the stream segregation phenomena reported here.

F. Stage 5: Temporal integration of signal

Temporal integration of the signal is achieved by passing the AN activity in each channel (the number of spikes, H_{it}) through an exponential accumulation and decay function (a leaky integrator). A leaky integrator sums the energy occurring within a given time period or temporal ''window,'' and functions as a low-pass filter. It shows a gradual accumulation in excitation while a stimulus is on, and a gradual decline in excitation once the stimulus has ended. The time constant of the integrator ($\mu=3$ ms) controls the accumulation and decline of excitation. The output of the temporal-integration pathway (K_{it}) is computed using the formula

$$K_{it}=K_{i(t-1)}e^{-\Delta t/\mu}+H_{it}. \tag{8}$$

G. Stage 6: Random bias and further leaky integration of the signal

At the heart of the model are the two critical principles of stochastic operation and low-pass filtering (or leaky integration). We propose that both are characteristic of neural processing and critical to an understanding of the spontaneous stream segregation effects which occur in response to alternating pure tones of different frequency. These are implemented schematically in the model by sending the output from each leaky integrator (K_{it}) through two separate pathways. One pathway (the amplitude-information path) preserves the temporally integrated AN output for determining the loudness of the signal. In the other pathway (the excitation-level path), the stochastic nature of the input to the auditory brain stem is represented by varying the temporally integrated AN output (K_{it}) by adding a smoothed random value R_{it}. This is computed by choosing a new random number (r_{it}) at each step of the evaluation of the model. This random number has a mean of zero and a range of $+/-MK_{it}$. The random bias is thus made proportional to the activity in each channel (K_{it}), in order to approximate the properties of a Poisson process where the variance of the random variable is directly proportional to its expected value. M is an important parameter of the model whose value (0.006) was found by fitting the model output to the data of Anstis and Saida (1985) as described below. The random sequence is smoothed by passing it through the following low-pass filter:

$$R_{it}=R_{i(t-1)}+r_{it} \tag{9}$$

and the result is added to temporally integrated AN output

$$L_{it}=K_{it}+R_{it}. \tag{10}$$

It is critical to the functioning of the model that the random sequence (r_{it}) is unique to each channel. It gives rise to a random walk which will differ from one channel to another. The effect of this becomes obvious at the next stage of the model where the activity in each channel is integrated in a manner thought to be involved in normal loudness-summation processes. We represent this here as another low-pass filter

$$C_{it}=C_{i(t-1)}e^{-\Delta t/\gamma}+L_{it}, \tag{11}$$

where γ is the time constant of the filter and is relatively long (γ=70 ms).

H. Stage 7: Dominant-channel selection and channel attenuation

The channel selector, at 1-ms intervals, chooses the channel with the highest excitation level (C_{it}). This is called the ''dominant channel.'' This channel experiences no attenuation in its amplitude-information-path output, whereas the activity level in the amplitude-information paths of all other channels (K_{it}) is attenuated by half. This attenuation value seems reasonable for the purposes of developing the model. The attenuated output of the amplitude-information path (q_{it}) of the *dominant* channel is therefore

$$q_{it}=K_{it} \tag{12}$$

and the attenuated output of the amplitude-information path (q_{it}) of all the other channels is

$$q_{it}=0.5K_{it}. \tag{13}$$

The loudness output (Ω_t) of the model is found by adding together the resulting levels of all the amplitude-information paths over the three channels,

$$\Omega_t=\sum_{i=1}^{k} q_{it}/k, \tag{14}$$

where the number of channels (k) equals three. Equation (14) yields a high-level output if the dominant channel is also the channel carrying the active signal. However, complex sequences of tones can give rise to situations where the signal-carrying channel is not dominant. In this case, the loudness output for that tone is weakened, and the reduced contribution of the tone to the model output will create a ''background''-type signal. The channel-attenuation mechanism will therefore simulate the perceived loudness difference between attended and unattended streams (van Noorden, 1975, 1977).

I. Stage 8: Model output and the segregation/coherence decision

The final stage of the model compares the average loudness output during tone A (ΩA) with the average loudness output during tone B(ΩB) for the preceding 1-s period:

$$Z_t=\Omega\mathrm{A}/\Omega\mathrm{B}. \tag{15}$$

The model assumes that stream segregation occurs when a critical output difference between A and B is exceeded. That is, one tone dominates the model output in terms of loudness, similar to the figure/ground loudness percept of a segregated ABAB sequence. Therefore, when Z_t, or $1/Z_t$, exceeds a critical value, Z_{crit} (1.117), A and B are considered to have segregated into separate streams over the previous second; i.e., the model ''hears'' one tone as being consistently ''louder'' than the other. This condition occurs when one tone dominates the model output. This judgment is made at the end of each second from the beginning to the end of the stimulus. Z_t represents the output of the model, and is used in the model simulations of stream-segregation experiments described below.

II. MODEL OPERATION

We can illustrate the operation of the model by considering the main factors that affect the model output—the destabilization of the channel excitation levels created by the action of the random bias, and the effects of Δf and TRT.

A. Destabilization of dominant channel

For most stimuli, when a tone is played, the channel whose CF corresponds to the tone frequency will have the highest excitation level. In this situation, that channel's output will dominate the model output, because the output from all other channels will be attenuated. However, because the model responds sluggishly, the channel with the highest excitation level will not always be the one maximally stimulated by a tone. In such cases, the model output in response to that tone will be reduced, because that channel has an attenuated contribution. This situation may arise when the frequencies of the tone-A and tone-B channels are close together, and the middle channel accumulates more excitation over the whole stimulus presentation. It may also occur simply as the result of the action of the random bias, which may accumulate unusually large amounts of excitation in one channel so that it remains dominant, even when it is not being actively stimulated. In some instances, therefore, a channel that is not being directly stimulated will contain the information necessary for the system to determine what percept is heard. In others, the percept heard is dependent upon channel alternation between dominant channels.

When an ABAB sequence is presented to the model then, as each tone is presented, there is a gradual buildup of excitation in the 70-ms leaky integrators of the tone-A and tone-B channels. When the tone ends, there is a gradual decline in excitation level in the corresponding channels. Figure 4 shows such a situation. As A comes on, the excitation level in the corresponding tone-A channel rises to a peak, causing it to be the dominant channel in terms of excitation level. However, when A goes off and the excitation caused by it decays, the excitation caused by B rises to a peak and makes the tone-B channel the dominant channel. This situation will lead to the model's output being dominated alternately by the tone-A and tone-B channels, regardless of Δf and TRT. That is, $\Omega\mathrm{A}\approx\Omega\mathrm{B}$, and $Z<Z_{\mathrm{crit}}$. Here, a model output of temporal coherence results for the duration of the

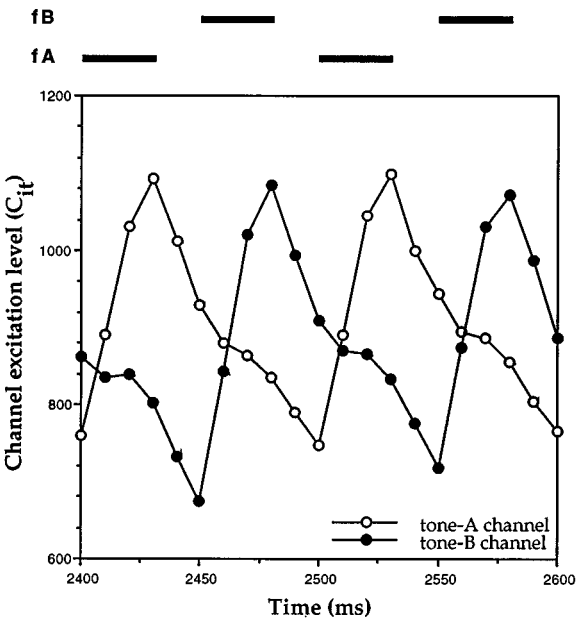

FIG. 4. Channel excitation levels (C_{it}) for the tone-A channel (CF=f_A) and the tone-B channel (CF=f_B) showing the overlapping vertical excursions of the channel excitation levels. C_{it} is measured in arbitrary units. Here f_A=1000 Hz, f_B=1200 Hz, and TRT=50 ms. The model output is sampled every 10 ms.

sequence. This situation will also occur if the middle channel dominates the system output, as $\Omega A \approx \Omega B$, and $Z \approx 1.0$.

However, the separate random biases applied to the excitation-level paths accumulate over time, so that sometimes a situation is reached where a large overall change results. As the excitation level is sometimes higher than usual, sometimes lower, the accumulation of excitation by the 70-ms leaky integrators will vary differently in each channel over time, due to the separate random walks in each channel. As a result, the channel with the greatest excitation level will switch randomly from the tone-A to the tone-B channel over the course of the sequence. As the channel with the greatest excitation level is taken to be the dominant channel, then the fluctuation between the tone-A and tone-B channels—in terms of highest excitation level—will imitate the apparently random attentional alternation between tone-A and tone-B streams over the duration of a tone sequence (Anstis and Saida, 1985). Should one channel have the highest excitation level for a long period of time, this channel will become the dominant channel for an extended period, with the result that ΩA and ΩB are unequal, and $Z > Z_{crit}$, giving a model output of segregation. It should be noted here that reducing the value of the random-bias control parameter (M) will cause less destabilization of the channel excitation levels, and consequently fewer segregation reports in the model output. However, this effect can be counterbalanced by altering one of the other model parameters to increase either the destabilization of the channel excitation levels, or the inequality of ΩA and ΩB in the system output, thereby creating more segregation reports. This can be achieved by either (1) increasing the time constant of the second leaky integrator (i.e., >70 ms), (2) decreasing the value of Z_{crit} (i.e., <1.117), or (3) decreasing the channel-attenuation value (i.e., <0.5).

A further consequence of random-bias accumulation is that as the random bias is proportional to channel activity, the onset of a tone which follows a silent period is preceded by a history of little accumulated bias. However, the effect of the bias and the random walk will accumulate over the first few presentations of the tone. As the random bias takes time to accumulate, the first few seconds of a stimulus will always be characterized by alternation between the tone-A and tone-B channels. That is, *irrespective of the stimulus parameters, temporal coherence will occur at the beginning of a sequence* (Bregman, 1990). After a period of time, however, the bias accumulation will begin to take effect, and the probability of one channel having a higher excitation level than the other, and consequently forming a stream, will increase (Bregman, 1990; Anstis and Saida, 1985). However, this effect will be counteracted by the occurrence of an extended silent period which, due to the absence of a stimulus, will create minimal channel-excitation and random-bias levels, and allow the channel excitation levels to decay to a baseline level.

The effect of random-bias accumulation is illustrated in Fig. 5, which shows the excitation levels (C_{it}) for each model channel in response to a 10-s ABAB sequence. Here f_A=1000 Hz, f_B=1250 Hz, TRT=100 ms, and tone duration=40 ms. When $t<5$ s, the highest excitation level alternates between the tone-A and tone-B channels (as in Fig. 4), so that $\Omega A \approx \Omega B$, and $Z < Z_{crit}$ (see Figs. 5 and 6). However, when $t>5$ s, the tone-A channel consistently has the highest excitation level, resulting in a large difference between ΩA and ΩB in the model output (see Fig. 6), so that $Z > Z_{crit}$ (see Figs. 5 and 6). The model will therefore give a segregation output in response to this stimulus only when $t>5$ s.

B. The effect of TRT on the model output

The action of the cumulative random bias has consequences for the effect of TRT on the model output. If we present subjects with an ABAB sequence, where tone duration=40 ms, then there will be a long period of silence (210 ms) between tones when TRT=250 ms. The random bias is proportional to channel activity, which is at its highest when a tone is played (see Fig. 4). This indicates that the cumulative random bias will be mainly ''fueled'' by the excitation-level peaks created by the tones. Therefore, the low tone density in this stimulus will give a minimal opportunity for cumulative differences between the energy in the channels. This will reduce the chances of one channel dominating the model output for an extended period, and give a low probability of a segregation response. In addition, the long silences between tones will also allow the channel excitation levels to decline to baseline levels between tone presentations, thereby producing the overlapping channel excitation levels found in Fig. 4. In this situation, the dominant channel will alternate between the tone-A channel (when A is active) and the tone-B channel (when B is active), with the result that when each tone is playing, it is being

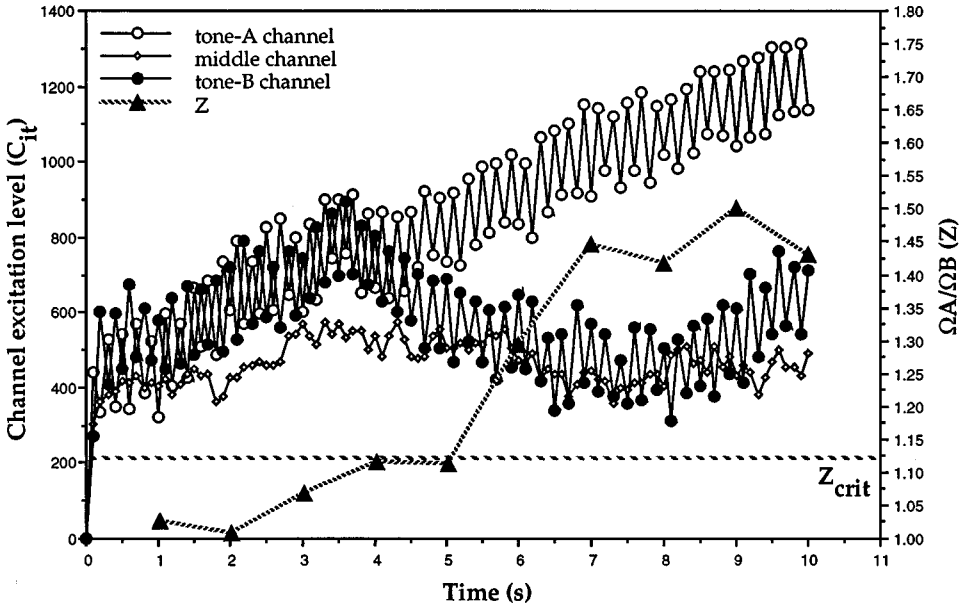

FIG. 5. Channel excitation levels (C_{it}) for all model channels in response to a 10-s sequence. C_{it} is measured in arbitrary units. Here f_A=1000 Hz, f_B=1250 Hz, TRT=100 ms, and tone duration=40 ms. Also shown are the Z values derived from the ΩA and ΩB values shown in Fig. 6 (see text). Z_{crit} is indicated by the dotted line. The model output is sampled every 100 ms.

relayed faithfully through its own channel without attenuation. Here the model output will be temporally coherent, as ΩA≈ΩB, and $Z<Z_{crit}$.

The opposite situation will apply at short TRTs (e.g., 50 ms). Here the high tone density in the stimulus sequence will give a maximal cumulative random bias over the duration of the stimulus (as in Fig. 5), consequently giving a high probability of a segregation output. Overall, then, we can say that the probability of one channel dominating the model output, and thereby producing a model output of segregation, is inversely related to TRT.

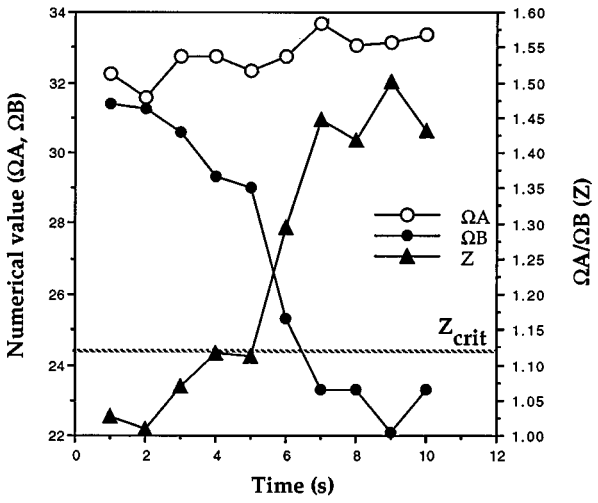

FIG. 6. The values of ΩA and ΩB (measured in arbitrary units) in the model output resulting from the channel excitation levels shown in Fig. 5 (see text). Also shown are the Z values derived from ΩA and ΩB (see text). Z_{crit} is indicated by the dotted line.

C. The effect of Δf on the model output

If a situation should arise where the tone-A channel is the dominant channel for an extended period, then the amplitude output for tone B (ΩB) will be reduced (as illustrated in Sec. II A). This creates an imbalance between ΩA and ΩB, and increases the value of Z. If Δf is small, then ΩB ≈ΩA, and $Z\approx 1.0$. This is because all channels respond roughly equally to the two tones at small Δf's. That is, the tone-B signal will pass through the tone-A channel to higher levels relatively unattenuated. However, if we increase Δf, we will attenuate the level of tone B in the tone-A channel, reduce the value of ΩB further, and increase the imbalance between ΩA and ΩB in the model output. This increases the likelihood of Z being greater than Z_{crit}, and results in a greater probability of the model giving a segregation output. This indicates that there is an increased probability of the model giving a segregation output as Δf increases, should one channel dominate the model output for an extended period.

D. Summary

The overall behavior of the model shows a two-factor interaction between the effects of TRT and Δf, with the lowest probability of the model giving a segregation output occurring at long TRTs and small Δf's, and the highest probability of the model giving a segregation output occurring at short TRTs and large Δf's. These latter conditions reflect the perceptual area circumscribed by the temporal coherence boundary, above which a percept of stream segregation is always heard (see Fig. 2).

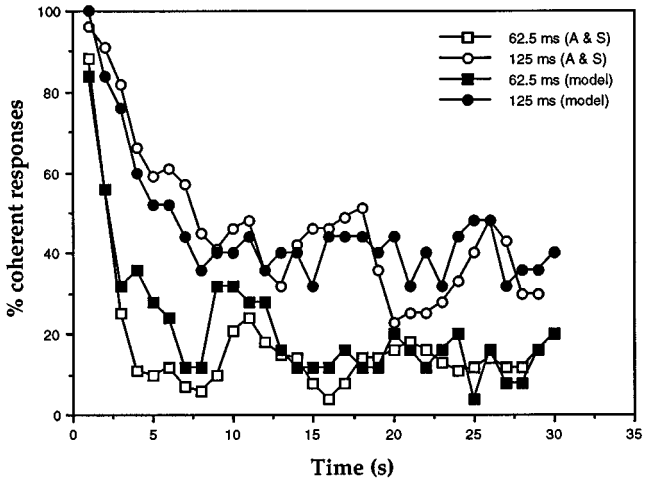

FIG. 7. Comparison of the results of Anstis and Saida (1985) with the output of the model (see text).

III. MODEL EVALUATION

To evaluate the effectiveness of the model, the results of a number of experimental studies carried out on human listeners were simulated using the model. In each model simulation, except where noted, the stimuli were 75-dB pure tones with 5-ms rise and fall times (implemented using raised cosine ramps). The model used the *same* parameter values in each simulation. The range value for the random bias module (M) was 0.006, the critical ratio value (Z_{crit}) was 1.117, the time constant of the first leaky integrator (μ) was 3 ms, and the time constant of the second leaky integrator (γ) was 70 ms. It should be noted here that the simulation of the results of Anstis and Saida (1985) described in Sec. III A was used to find rough values for the model parameters. These values were then optimised so that the model gave the best fit to the experimental data described below, which were gathered from 5 subjects (Anstis and Saida, 1985), 1 subject (van Noorden, 1975), and 14 subjects (Miller and Heise, 1950). Although a single model with one set of parameters could be considered as a ''single'' listener, the best-fit procedure described here suggests that the model simulations can be considered as the response of an ''average'' listener to the stimuli.

When the experimental studies used ABAB stimuli of indefinite duration, listeners' responses were simulated in the following way. The simulation of Anstis and Saida's (1985) study described below indicated that the model output in response to a stimulus, when averaged over a number of trials, is relatively invariant after 15 s (see Fig. 7). Therefore, the number of ''coherent'' responses given by the model for the last second of a 15-s sequence were recorded over a number of trials, and the ''coherent'' score was divided by the number of trials to give the percentage of coherent responses. The resultant percentage coherence value was then used to map the model output on to listeners' responses.

A. Buildup of stream segregation over time

Anstis and Saida (1985) showed that all ABAB sequences begin by sounding coherent, and that the probability of temporal coherence decreases steadily over time as a function of total sequence duration. The methodology of Anstis and Saida (1985) was suitable for computer simulation, and the response of the model to their stimuli was recorded and compared with their data.

The stimuli used were 30-s ABAB sequences, where f_A=800 Hz and f_B=1200 Hz. The TRT was either 62.5 or 125 ms, and the duration of each tone was equal to the TRT (i.e., no silences). Each sequence was played a total of 25 times, as in their study. During the presentation of the stimulus, decisions by the model in favor of temporal coherence were totalled for each second along the length of the sequence and expressed as a percentage. The results were then plotted and compared with those of Anstis and Saida (1985)—see Fig. 7. As can be seen, the model is clearly successful in replicating the buildup of stream segregation over time found by Anstis and Saida (1985).

Our claim that the model replicates the experimental data is limited to the broad ''buildup'' principle. We have no deterministic explanation for the undulations in the curves. These arise in the model output as a result of the stochastic nature of one stage of the process. However, when averaged over very large numbers of trials, the model gives a coherence function that shows a smooth decline with signal duration, and we assume that the same would be true of human listeners. Although, in principle, the variation in subject response should be predictable from the model, unfortunately no data on subject response is available, only mean response curves. If standard error values could be established for listeners, then these should be the same for the model. This provides one possible method for testing the model in future.

B. Temporal coherence and fission boundaries

To map the model output on to the responses of listeners and simulate van Noorden's (1975) temporal coherence and fission boundaries, we take fixed percentage coherence levels as being equivalent to the criteria employed by listeners to define these two perceptual boundaries. The model fission boundary (MFB) is taken to be the Δf value where the model output drops below 100% coherent responses. This criterion was suggested by the nature of the fission boundary, which defines a region where the percept is always one of permanent temporal coherence—i.e., a model output of 100% coherent responses. When the model output drops below 100% coherent responses, a segregated/coherent mix is indicated which corresponds to the percept heard in the ambiguous region. To find the model equivalent to the temporal coherence boundary, we take a value of 40% coherent responses (arrived at by experimentation) as being the model segregation threshold (MST), and assume that the MST is equivalent to the criterion employed by listeners to define the temporal coherence boundary. That is, for percentage coherence responses $<$40%, the percept ''heard'' by the model is equivalent to the percept of steady segregation heard by listeners above the temporal coherence boundary. The majority of van Noorden's (1975) experiments showed that the temporal coherence boundary asymptotes within a TRT range of 150–250 ms (Fig. 2). This suggests that above a certain TRT value in this range, the temporal coherence boundary will

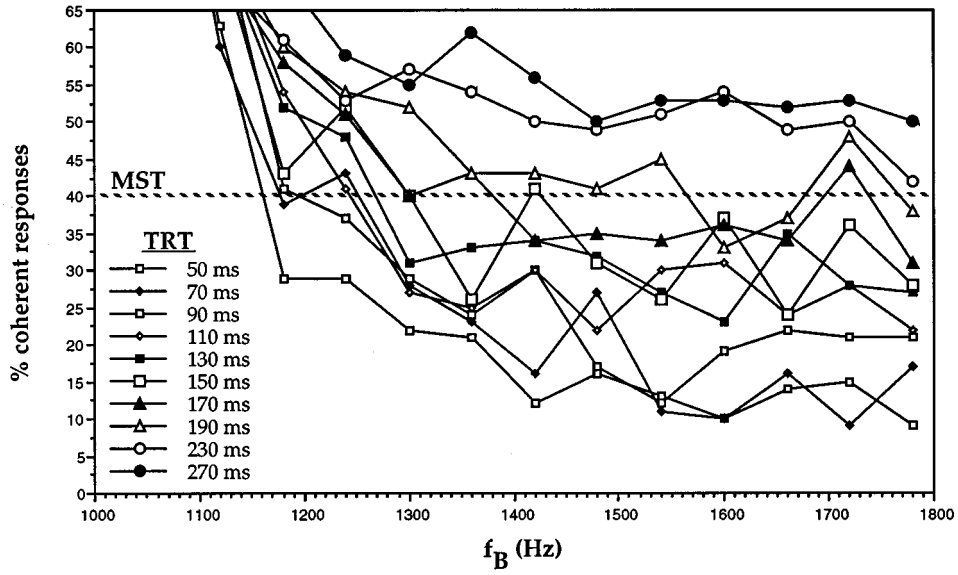

FIG. 8. Two-dimensional probability response surfaces illustrating the average % coherent responses given by the model at the end of a 15-s ABAB sequence using various combinations of TRT and f_B. Also shown is the model segregation threshold (MST) of 40% coherent responses. For all sequences, f_A=1000 Hz.

never be crossed, regardless of Δf, because the asymptote rises to infinity. This implies that above a certain TRT value in this range, the model output will never drop below the MST of 40% coherent responses, whatever the Δf between A and B.

To test this hypothesis, 15-s ABAB sequences of 40-ms tones were presented to the model. A number of TRT values and f_B values were used (TRT=50, 70, 90, 110, 130, 150, 170, 190, 230, and 270 ms, f_B=1060 to 1780 Hz in 60-Hz steps), and f_A was kept constant at 1000 Hz. There were 100 trials for each TRT/f_B combination, and the number of coherent responses for the last second of each sequence were totalled for each combination and converted to percentages. The MFB was found by repeating the above procedure using TRTs of 50, 100, 150, 200, and 250 ms for f_A=1000 Hz and finding the f_B value where the model output dropped below the 100% coherent responses level.

Figure 8 shows the percentage of coherent responses for each TRT as a function of f_B. For clarity, percentage of coherent responses >65% are omitted from the graph. All the TRTs investigated, except for the 230- and 270-ms conditions, drop below the MST. These results support the hypotheses that the percentage of coherent responses will never drop below the MST at long TRTs, and that an MST value of 40% coherent responses can be taken as being equivalent to the criterion employed by listeners to define the temporal coherence boundary. This suggests that the ambiguous region can be defined as the region over which the percentage of coherent responses varies between just below 100% (the MFB) to 40% (the MST).

The ambiguous region can be illustrated by plotting the f_B values where each two-dimensional TRT response surface in Fig. 8 intersects the MST. Where the model's output fluctuated around the MST, the average of the crossover points was used as the value of f_B. This procedure gives the model equivalent to the temporal coherence boundary. If the MFB is also plotted, then the area between them will correspond to the ambiguous region of van Noorden (1975)—see Fig. 9.

A comparison of van Noorden's (1975) results (Fig. 2) with the model data (Fig. 9) shows that both temporal coherence boundaries exhibit the same gradual increase with TRT, and that the fission boundaries are relatively constant with respect to TRT. However, no model value for the temporal coherence boundary can be obtained when TRT>190 ms, as the model's percentage coherence output is always greater than the MST (40%), resulting in the characteristic asymptote of the temporal coherence boundary. The resultant model asymptote is shown in Fig. 9. In addition, the variance around the MST for TRTs of 170 and 190 ms results in a variable value for the temporal coherence boundary in this range, thereby reproducing the observations of van Noorden

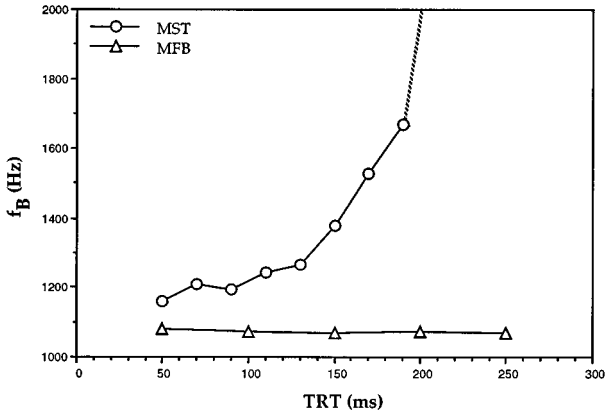

FIG. 9. The model equivalent to the temporal coherence boundary (MST) and its asymptote (dotted line) derived from Fig. 8. The graph also shows the model fission boundary (MFB). Here, f_A=1000 Hz.

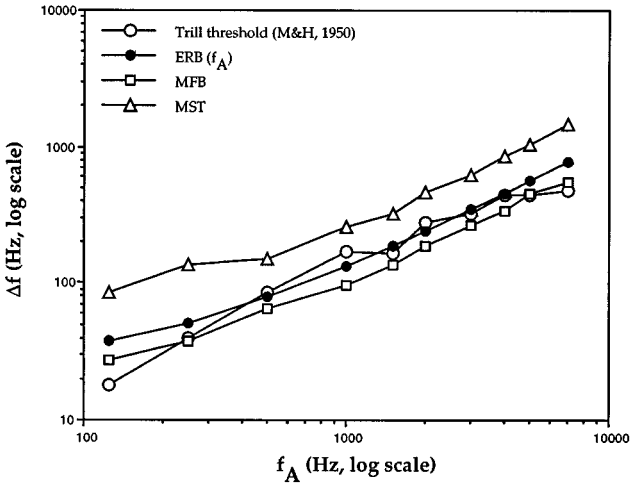

FIG. 10. Comparison of Miller and Heise's (1950) trill threshold, the equivalent rectangular bandwidth (ERB) of f_A, the model fission boundary (MFB), and the model temporal coherence boundary (MST). TRT=100 ms.

(1975) and Bregman (1990) that the temporal coherence boundary varies from subject to subject. The close correspondence between the experimental and model data justifies taking fixed percentage coherence levels as being equivalent to the criteria employed by listeners to define the temporal coherence and fission boundaries.

C. The trill threshold and the fission boundary

Miller and Heise (1950) defined the trill threshold as the Δf between two tones where the percept changes from that of a musical trill (i.e., a percept of temporal coherence) to a percept of "two unrelated and interrupted tones"—i.e., two separate streams of notes. Miller and Heise's (1950) experiment was simulated by studying both the fission boundary and the temporal coherence boundary, because their description of the trill threshold does not make it clear whether subjects increased Δf until the percept of permanent temporal coherence changed to the segregated/coherent mixture characteristic of the ambiguous region, or to the permanent-segregation percept found above the temporal coherence boundary. That is, whether the trill threshold is equivalent to the temporal coherence or fission boundary.

The model was presented with 15-s ABAB sequences where f_A was fixed at one of the f_A values used by Miller and Heise (1950), and f_B varied in frequency above it. The tone durations and TRTs of all sequences were 100 ms, corresponding to Miller and Heise's (1950) stimuli. There were 100 trials for each combination of f_A and f_B, and the number of coherent responses for the last second of each sequence were totalled for each combination and converted to percentages.

To find the fission boundary, Δf was gradually increased until the model output for the last second of the sequence dropped below 100% coherent responses (the MFB of Sec. III B). To find the temporal coherence boundary, Δf was gradually increased until the model output for the last second of the sequence dropped below 40% coherent responses (the MST of Sec. III B). Figure 10 shows the resultant Δf values for the MST and MFB for each f_A value compared with Miller and Heise's (1950) results and the value of ERB(f_A) calculated using Eq. (5). Figure 10 indicates that their trill threshold corresponds to the model's MFB (i.e., the fission boundary), with the MST (the temporal coherence boundary) showing higher Δf values over the range of frequencies investigated.

Van Noorden (1975) found a number of different values for the fission boundary, which varied with stimulus and experimental procedure, ranging from ≈1075 to ≈1155 Hz when f_A=1000 Hz. For f_A=1000 Hz, there is a close correspondence between the MFB of ≈1090 Hz, Miller and Heise's (1950) trill threshold of ≈1167 Hz, and van Noorden's (1975) fission boundary data. This indicates that the trill threshold corresponds to the fission boundary. Furthermore, the equivalence of the MFB and trill threshold with the value of ERB(f_A) strongly suggests that the fission boundary is connected with auditory-filter width (i.e., physiological factors).

IV. GENERAL DISCUSSION

Two very simple principles lie at the heart of the model: the stochastic nature of the system response to simple stimuli, and the low-pass filtering effect of neural signal processing. These have been grafted on to two generally accepted characteristics of the auditory periphery: bandpass mechanical filtering and the rapidly adapting response of AN fibers. At the output of the model, we have established a principle that the system attenuates activity in channels that are already less active—a process that could be seen as a method for enhancing signal-to-noise ratio. Despite the relative simplicity of the model, it has been shown above to behave in a manner qualitatively and quantitatively consistent with a range of stream segregation phenomena observed when listening to alternating pure-tone sequences. These include the effects of rate of presentation of stimuli and Gestalt auditory grouping on the basis of frequency *proximity* (Beauvois and Meddis, 1995). The behavior of the model is also consistent with two important observations: first, that the system begins in "coherent" mode, only gradually creating streams and, second, that shifts of attention can occur spontaneously without conscious effort (van Noorden, 1975; Anstis and Saida, 1985).

We accept that the model is a hybrid of physiology and electrical signal processing. We have done this because we know enough about the auditory periphery to be reasonably sure what the physiological response to tones is like at the level of the AN, while we are less sure about the details of the response at the level of the auditory brain stem. We can, however, be sure that the response is stochastic rather than deterministic. There are also many examples of temporal smearing as the signal is passed from one stage to another (Oertel *et al.*, 1988; Rees and Palmer, 1989). Our concern was to show that these two principles might have a role to play in explaining auditory stream segregation phenomena. We believe that a prima facie case has been made that this is the case. We do accept, however, that a detailed program of research lies ahead if we are to establish precisely how these principles are implemented in the auditory nervous system.

The work described above was greatly stimulated by experiments carried out in the Gestalt tradition (Bregman, 1990). Experiments in this tradition establish the principle that the mind has the ability to uncover structure or to impose structure on sensory experience. We see our modeling efforts as speculations concerning the kind of signal-processing hardware that might underpin the psychological principles that Gestalt psychology has established. The mechanisms that generate the behavior need not map directly on to the language of Gestalt psychology. For example, the concept of ''stream segregation'' is an apt description of the sensory experience, but need not imply that representations of auditory objects are physically separated by the mechanism, sent via different routes, and manipulated independently. In our model, there is no separation. On the contrary, the whole of the sensory response remains intact, but different parts of the representation are highlighted in such a way as to emphasize those parts of the representation arising from one of the ensemble of competing auditory objects. As a result, the organism should be able to organize a response to that auditory object without being confused by the presence of competing sounds. Later, the spotlight may shift and emphasize the features of a different object. The point is that the system does not contain streams, but it does behave in the same way as a human listener who experiences an alternation of foreground and background auditory objects.

The model further implies that the foreground/background segregation work takes place spontaneously at a very early stage. This proposal may appear to be contradicted by the fact that listeners can consciously switch attention between alternative auditory impressions. However, while conscious choice may require cortical involvement, we suggest that the formulation of alternatives, between which the choice must be made, is the work of low-level processes. While a computer model cannot, in itself, prove this hypothesis, it does show that it is a distinct possibility.

The model assumes the input to the system is noisy, and that the functioning of the system adds further noise. While it is widely recognized that the AN response to sound is a very noisy representation of the stimulus, there has been very little discussion of the implications of this fact for an understanding of auditory perceptual processes. It is natural to assume that statistical fluctuations are minimized by some averaging process at later stages of the system. However this need not always be the case. Our model proposes that the noise is exploited to facilitate the process of segregation. Our implementation of the model uses a very crude, but easily manipulated method to inject noise into the system. A desirable future development is a more detailed study of the stochasticity of the nerve fiber response, and a demonstration that this is consistent with the requirements of the model as described above.

ACKNOWLEDGMENTS

This research was supported by grants to the first author from SERC (UK), CNRS (France), and the Fyssen Foundation, Paris, France. Thanks go to Steve McAdams, Steve Colburn, and three anonymous reviewers for their comments on an earlier version of this paper.

Anstis, S., and Saida, S. (**1985**). ''Adaptation to auditory streaming of frequency-modulated tones,'' J. Exp. Psychol.: Hum. Percept. Perform. **11**(3), 257–271.

Beauvois, M. W., and Meddis, R. (**1991**). ''A computer model of auditory stream segregation,'' Q. J. Exp. Psychol. **43A**(3), 517–541.

Beauvois, M. W., and Meddis, R. (**1995**). ''Computer simulation of Gestalt auditory grouping by frequency proximity,'' in *Neural Computation and Psychology*, edited by L. S. Smith and P. J. B. Hancock (Springer-Verlag, Berlin), pp. 155–164.

Bregman, A. S. (**1990**). *Auditory Scene Analysis* (MIT, Cambridge, MA).

Bregman, A. S., and Campbell, J. (**1971**). ''Primary auditory stream segregation and perception of order in rapid sequences of tones,'' J. Exp. Psychol. **89**, 244–249.

Glasberg, B. R., and Moore, B. C. J. (**1983**). ''Suggested formulae for calculating auditory-filter bandwidths and excitation patterns,'' J. Acoust. Soc. Am. **74**, 750–753.

Glasberg, B. R., and Moore, B. C. J. (**1990**). ''Derivation of auditory filter shapes from notched-noise data,'' Hear. Res. **47**, 103–138.

Hastings, N. A. J., and Peacock, J. B. (**1975**). *Statistical Distributions* (Butterworths, London).

Hewitt, M. J., and Meddis, R. (**1991**). ''An evaluation of eight computer models of mammalian inner hair-cell function,'' J. Acoust. Soc. Am. **90**, 904–917.

Hewitt, M. J., Meddis, R., and Shackleton, T. M. (**1992**). ''A computer model of a cochlear-nucleus stellate cell: Responses to amplitude-modulated and pure-tone stimuli,'' J. Acoust. Soc. Am. **91**, 2096–2109.

Koffka, K. (**1935**). *Principles of Gestalt Psychology* (Routledge and Kegan Paul, London).

McAdams, S., and Bregman, A. S. (**1979**). ''Hearing musical streams,'' Comput. Music J. **3**, 26–43, 60.

Meddis, R. (**1986**). ''Comments on 'Very rapid adaptation in the guinea pig auditory nerve' by Yates, G. K. *et al.*, 1985,'' Hear. Res. **23**, 287–290.

Meddis, R. (**1988**). ''Simulation of auditory-neural transduction: Further studies,'' J. Acoust. Soc. Am. **83**, 1056–1063.

Meddis, R., Hewitt, M. J., and Shackleton, T. M. (**1990**). ''Implementation details of a computational model of the inner hair-cell/auditory-nerve synapse,'' J. Acoust. Soc. Am. **87**, 1813–1816.

Miller, G. A., and Heise, G. A. (**1950**). ''The trill threshold,'' J. Acoust. Soc. Am. **22**, 637–638.

Oertel, D., Wu, S. H., and Hirsch, J. A. (**1988**). ''Electrical characteristics of cells and neuronal circuitry in the cochlear nuclei studied with intracellular recording from brain slices,'' in *Auditory Function*, edited by G. M. Edelman, W. E. Gall, and W. M. Cowan (Wiley, New York), pp. 313–336.

Patterson, R. D., and Moore, B. C. J. (**1986**). ''Auditory filters and excitation patterns as representations of frequency resolution,'' in *Frequency Selectivity in Hearing*, edited by B. C. J. Moore (Academic, New York), pp. 123–177.

Rees, A., and Palmer, A. R. (**1989**). ''Neuronal responses to amplitude-modulated and pure-tone stimuli in the guinea pig inferior colliculus, and their modification by broadband noise,'' J. Acoust. Soc. Am. **85**, 1978–1994.

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

van Noorden, L. P. A. S. (**1977**). ''Minimum differences of level and frequency for perceptual fission of tone sequences,'' J. Acoust. Soc. Am. **61**, 1041–1045.

Westerman, L. A., and Smith, R. L. (**1984**). ''Rapid and short-term adaptation in auditory nerve responses,'' Hear. Res. **15**, 249–260.