Open access • Journal Article • DOI:10.1051/JP4:1994576

Computational modelling of auditory streaming phenomena — Source link

M. Beauvois, M. Beauvois, R. Meddis

Institutions: IRCAM, Paris Descartes University, Loughborough University

Published on: 01 May 1994 - Journal De Physique Iv (EDP Sciences)

Topics: Auditory scene analysis

Computational modelling of auditory streaming phenomena

M. Beauvois, R. Meddis

Computational modelling of auditory streaming phenomena

M. BEAUVOIS et R. MEDDIS*

Laboratoire de Psychologie Expérimentale, Université René Descartes, Paris V, URA 316 du CNRS, 28 rue Serpente, 75006 Paris, France
and
IRCAM, 31 rue St-Merri, 75004 Paris, France
* *Department of Human Sciences, University of Technology, Loughborough LE11 3TU, U.K.*

Abstract: A computer model is described that uses simple physiological principles that operate mainly at a peripheral level to account for perceptual coherence among successive pure tones of changing frequency. Using a single set of parameter values, the model is able to reproduce a number of fundamental auditory streaming phenomena. These include the build-up of auditory stream segregation over time, and the temporal coherence and fission boundaries of human listeners. Whereas these streaming phenomena are generally accounted for in terms of a high-level auditory scene-analysis process, the success of the model in reproducing experimental data obtained from humans justifies the potential value of a low-level analysis for explaining auditory grouping phenomena, and suggests that some auditory grouping may be the product of low-level auditory processing.

1. INTRODUCTION

If listeners are presented with isochronous alternating-tone sequences composed of two pure tones of different frequencies (ABAB...), two percepts are possible depending on the frequency separation (Δf) and the time interval (TRT) between the A and B tones. These are 1) alternating A and B tones at a rate R (a percept of temporal coherence, similar to a musical trill), or 2) two separate streams of tones with one stream containing only A tones and another with only B tones. Here, both streams have a rate of 1/2 R, with each stream apparently coming from a separate sound source. This latter phenomenon has become known as auditory stream segregation. In ABAB tone sequences, the attended stream is subjectively louder than the unattended stream, producing an auditory figure-ground percept.

Van Noorden [1] investigated stream segregation in ABAB pure-tone sequences, and described the existence of two perceptual boundaries related to Δf and TRT (Fig. 1). Above the temporal coherence boundary (TCB) it is impossible to integrate the A and B tones into a single perceptual stream. The listener typically hears two separate pulsing tones; one clearly louder than the other. Below the fission boundary (FB) 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 depending upon the observer's attentional set.

2. THE MODEL

A computer model has been constructed that accounts for perceptual coherence in alternating-tone sequences by using simple physiological principles that operate mainly at a peripheral level. The current model is similar to the model described by Beauvois & Meddis [2], except that further components have been added to make the model more physiologically plausible. A summary of the model features follows.
1. The acoustic signal is first subjected to a peripheral frequency analysis which establishes 'channels' characterised by a bandpass frequency response to stimuli.
2. The output of each bandpass filter is fed into a simulation of an auditory hair cell whose output is characterised by a cumulative random element that is proportional to the firing rate of the hair cell.

Article published online by EDP Sciences and available at http://dx.doi.org/10.1051/jp4:1994576

3. Each model channel subdivides into three pathways:

3.1. A temporal fine-structure path preserves all aspects of the hair-cell output to enable the signal to be processed by higher levels, and to preserve all information carried by the auditory nerve for pitch extraction purposes.

3.2. An amplitude-information path temporally integrates the hair-cell output (3-ms time constant).

3.3. An excitation-level path subjects the output of the amplitude-information path to a slower temporal integration process (500-ms time constant).

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 system output is the sum of the outputs of the attenuated and non-attenuated amplitude-information paths.

7. The model assesses stream segregation on the basis of the relative amplitude levels of the two tones. If the levels are comparable, the percept is judged to be coherent. Otherwise, a 'stream segregation' percept is reported.

In the model itself, stream segregation is assumed to occur in ABAB sequences when there is an amplitude imbalance between the two tones in the system output. This amplitude difference between the A and B tones is created by a suppression mechanism where the action of a preceding tone suppresses the output of a following tone. The attentional mechanism that attenuates all frequency-selective channels except the dominant channel gives rise to an auditory figure-ground effect, where the information coming through the attenuated channels is heard 'in the background'. The channel activity is not wholly deterministic, but is conditioned by the random nature of auditory-nerve spike generation.

3. BUILDUP OF SEGREGATION OVER TIME

A study by Anstis & Saida [3] 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. Their methodology was suitable for computer simulation, and the response of the model to the stimuli used in their experiment was recorded and compared with their data.

The stimuli used were 30-s ABAB sequences, where fA=800 Hz and fB=1200 Hz. The TRT was either 62.5 or 125 ms, and there were no silences between tones. Each sequence was played 500 times and, during the presentation of the stimulus, decisions by the model in favour of temporal coherence were totalled for each second along the length of the sequence and expressed as a percentage. It was found that the model was successful in replicating the build-up of stream segregation over time found by Anstis & Saida. However, one feature of their results not replicated by the model was the apparently regular fluctuation in the percentage of coherent responses over time shown by their subjects (period ≈ 8 s): this is probably due to the difference in the number of trials used by Anstis & Saida (n=25) and the model simulation (n=500).

4. SIMULATION OF TEMPORAL COHERENCE AND FISSION BOUNDARIES

ABAB sequences of 40-ms tones were presented to the model. The TRT varied between 50 and 270 ms, fB varied between 1060 and 1780 Hz, and fA was kept constant at 1000 Hz. The sequences were 15 s long and the number of coherent responses for the last second of each sequence were totalled over 1000 trials for each combination of TRT and fB and then converted to a percentage. Fig. 2 shows the % coherent responses for each TRT value as a function of fB.

To map the model output onto the responses of listeners, we take 35% coherent responses as being the model segregation threshold (MST) - below which a segregated percept is heard, and above which a temporally coherent percept is heard - and assume that the MST is equivalent to the criterion employed by listeners to define the TCB. 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 FB, which defines a region where the percept is always one of permanent temporal coherence.

The TCB asymptotes when TRT is ≈ 200 ms. If we relate this to the model output, it implies that above a certain TRT value in this range, the model output will never drop below the MST, whatever the Δf between the two tones. This is shown in Fig. 2, where all the TRTs investigated, except for the 190 and 270-ms conditions, drop below the MST. These results support the hypothesis that, for large TRT values, the % coherent responses will never drop below the MST and justify our assumption that the MST can be taken as being equivalent to the criterion employed by listeners to define the TCB. This suggests that the ambiguous region can be defined as the region over which the percentage of coherent responses varies between 100% (the MFB) and 35% (the MST).

The ambiguous region can be illustrated by plotting the fB values where each 2-dimensional TRT response surface in Fig. 2 intersects the MST and the MFB. Where the model's output fluctuated around the MST, the average of the crossover points was used as the value of fB. This procedure gives the model equivalents to the TCB and the FB, with the area between them corresponding to the ambiguous region of Van Noorden [1] - see Fig. 3. Note that both TCBs show the same gradual increase with TRT, and that the FBs are relatively constant with respect to TRT. However, no model value for the TCB can be obtained for a TRT value >170 ms, as the model's % coherence output is always greater than the MST (35%), resulting in the characteristic TCB asymptote. The close correspondence between the experimental and model data further justifies taking fixed percentage coherence levels as being equivalent to the criteria employed by listeners to define the TCB and the FB.

5. GENERAL DISCUSSION

Bregman's [4] theory of auditory scene analysis postulates that the grouping mechanisms used by the auditory system accumulate evidence for a stream over time, and operate according to Gestalt perceptual principles. For example, the principle of proximity is demonstrated by the tendency of tones to form streams when they are in the same frequency region and there is a small Δf between the A and B tones [1, 4].

However, the low-level processing of the model reproduces the buildup of stream segregation over time, and is also able to simulate auditory grouping phenomena normally thought to occur due to the Gestalt auditory grouping principle of frequency proximity. This is shown by the model's ability to simulate the FB. In addition, the same model parameter values were used to reproduce the build-up of stream segregation over time found by Anstis & Saida [3], and the TCB and FB of human listeners found by Van Noorden [1]. This suggests that, for sequential pure-tone stimuli, the model is imitating the responses of human listeners fairly accurately, and may be processing the stimuli in a similar manner to the human auditory system. If this is the case, then it may be that Gestalt auditory grouping is, in part, the product of processes that are located at a relatively peripheral level of the auditory system.

6. REFERENCES

[1] Van Noorden, L.P.A.S., "Temporal coherence in the perception of tone sequences", unpublished doctoral dissertation, Institute for Perception Research, Eindhoven, The Netherlands (1975).

[2] Beauvois, M.W., & Meddis, R., "A computer model of auditory stream segregation", Quarterly Journal of Experimental Psychology, 43A (3) (1991) 517-541.

[3] Anstis, S., & Saida, S., "Adaptation to auditory streaming of frequency-modulated tones", Journal of Experimental Psychology: Human Perception & Performance. 11 (3) (1985) 257-271.

[4] Bregman, A.S., Auditory Scene Analysis (Cambridge, MA: MIT Press, 1990).

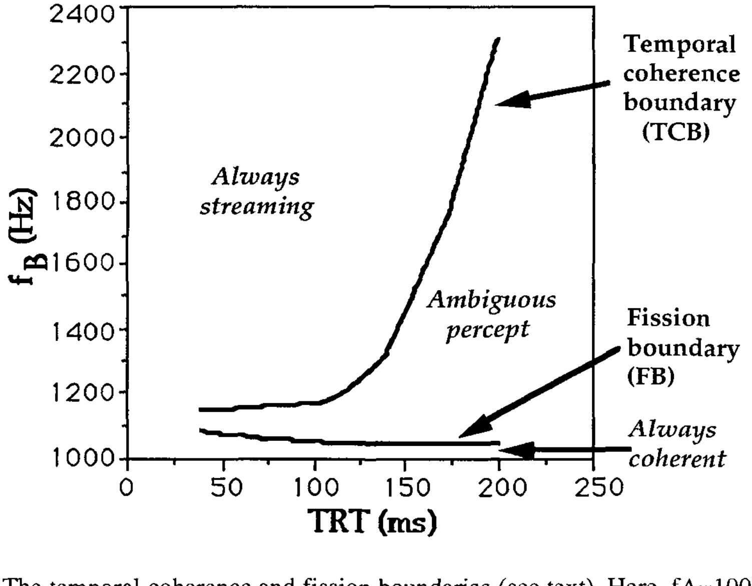

FIGURE 1. The temporal coherence and fission boundaries (see text). Here, fA=1000Hz.

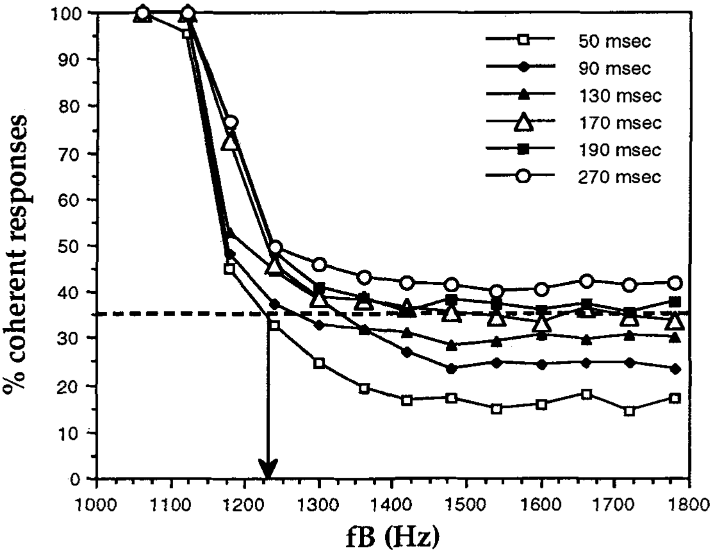

FIGURE 2. Average % coherent responses at the end of 15-s ABAB tone sequence (1000 trials). Here, fA = 1000 Hz, and tone duration = 40 ms.

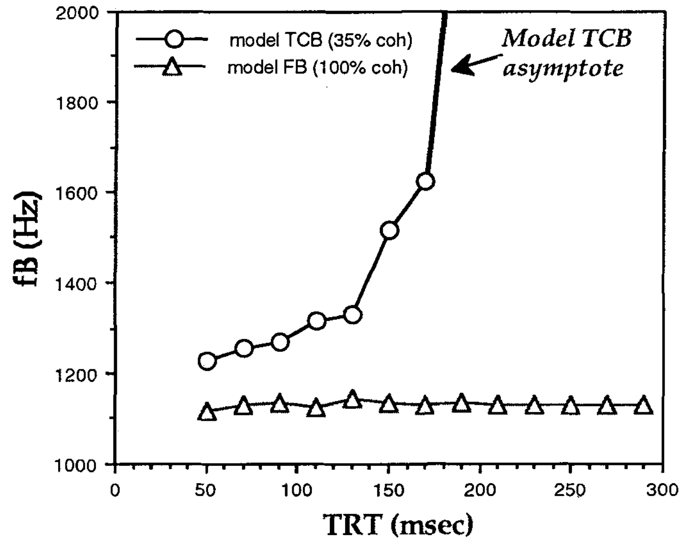

FIGURE 3. fB values where each 2-dimensional TRT response surface shown in Fig. 2 intersects the 35% & 100% coherent responses levels.