BNAM2012
Joint Baltic-Nordic Acoustics Meeting
June 18th - 20th 2012 Odense, Denmark

A computational model of auditory stream segregation based on a temporal coherence analysis

Simon Krogholt Christiansen, Morten Løve Jepsen and Torsten Dau
Center for Applied Hearing Research, Technical University of Denmark, DK-2800, Denmark, skch@elektro.dtu.dk

The ability to perceptually separate acoustic sources and focus one's attention on a single source at a time is essential for our ability to use acoustic information. In this study, a physiologically inspired model of human auditory processing was used as a front end of a model for auditory stream segregation. A temporal coherence analysis was applied at the output of the preprocessing, using the coherence across tonotopic channels to group activity across frequency. Using this approach, the described model is able to quantitatively account for classical streaming phenomena relying on frequency separation and tone presentation rate, such as the temporal coherence boundary and the fission boundary. The same model also accounts for the perceptual grouping of distant spectral components in the case of synchronous presentation. The most essential components of the front-end and back-end processing in the framework of the presented model are analysed and future perspectives discussed.

1 Introduction

In a natural acoustic environment, the sound that reaches our ears is a complex mixture of different sound sources. In order to parse this complex stimulus in useful information, our central auditory system segregates the signal into separate auditory objects or "streams" [1]. This allows us to selectively attend to a single auditory stream, and thus to focus on a conversation or piece of music while ignoring competing acoustic information. Models of auditory stream segregation relying primarily on frequency separation for stream segregation have been proposed, and physiological studies in animals have also shown a correlation between tonotopic separation and psychophysical stream segregation (e.g. [2]). While tonotopic separation may be necessary for stream segregation, it cannot explain the grouping or fusion of distant spectral components due to e.g. common pitch or onset/offset synchrony. These cues facilitate the fusion of spectral components into the same perceptual stream, despite tonotopic separation [3]. Elhilali et al. [3] suggested a conceptual model that could account for the grouping of distant spectral components due to synchrony. In the present study, the model by [3] is combined with the computational auditory signal-processing and perception (CASP) model [4], to create a physiologically inspired model of auditory stream segregation. The proposed model extends the functionality of the conceptual model of [2] by enabling it to account for the classical streaming phenomena of van Noorden [5] relying on frequency separation and tone repetition rate [5]. In the present study, the model is presented and evaluated in the experiments from [5].

2 Model Description

The model consists of two parts: A decomposition stage (peripheral processing and modulation filtering) based on CASP [4], and a grouping stage (temporal coherence analysis) based on the conceptual model by [3].

The peripheral processing stage consists of a basilar-membrane filterbank, a hair-cell transduction stage, and an adaptation stage. The basilar membrane filterbank is implemented as a 4^{th} order gamma-tone filterbank [6] with one ERB [7] spacing. The hair-cell transduction stage is realized by half-wave rectification followed by low-pass filtering at 1 kHz. Neural adaptation is modelled by five feedback loops connected in series, with time-constants ranging from 5 to 500 ms [8]. The output of the peripheral stage is processed by a first-order low-pass filter with a cut-off frequency of 150 Hz, simulating the decreasing sensitivity to sinusoidal modulation as a function of modulation frequency. The low-

pass filter is followed by a modulation filterbank. This is functionally similar to the temporal integration stage used in [3]. The modulation filterbank consists of band-pass filters with center frequencies ranging from 0 (low-pass filter) to 1000 Hz [4].

The output from the decomposition stage is processed by the grouping stage. Stream segregation is determined based on correlation between auditory channels. Channels with positively correlated activity over time are assigned to the same perceptual stream. As in Elhilali et al. [3], a windowed correlation between each pair of peripheral channels is computed by multiplying each pair of modulation filtered peripheral channels. The result is presented as a dynamic coherence matrix that shows the correlation between the peripheral channels over time. To quantify the coherence matrix, an eigenvalue decomposition is performed. The decomposition shows channels that are positively correlated with each other (and form a stream). The eigenvalue decomposition determines the number of independent dimensions of the coherence matrix, and by analogy, the number of streams present in the stimulus [3].

In the current study it is of interest whether a stimulus is perceived as one or two streams, and thus, whether there are one or two significant eigenvalues. The ratio of the second largest eigenvalue (λ_2) to the largest eigenvalue (λ_1) is therefore used as a measure of the "strength" of the two-stream percept. If the coherence matrix can be decomposed into one main component, the ratio λ_2/λ_1 will be very low (close to zero), corresponding to a one-stream percept. If the ratio λ_2/λ_1 is high, this indicates that there are (at least) two significant dimensions, and thus, at least two streams.

3 Method

The model is applied to the stimuli used by [5], and a schematic representation of the stimuli is shown in Figure 1. The stimuli consisted of two pure tones, A and B, presented in an ABA-ABA pattern. Each tone was 40 ms long and gated on and off using 5 ms raised cosine ramps. The onset-to-onset time between alternating tones was controlled by the tone repetition time (TRT). The frequencies of tones A and B were set to 1 kHz and *N* semitones higher. Combinations of ten TRT values (60 - 150 ms in steps of 10 ms) and 31 levels of *N* (0 – 15 semitones) were tested with the model (310 different conditions in total). The eigenvalue ratio λ_2/λ_1 was calculated for each condition.

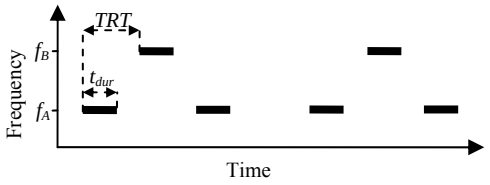
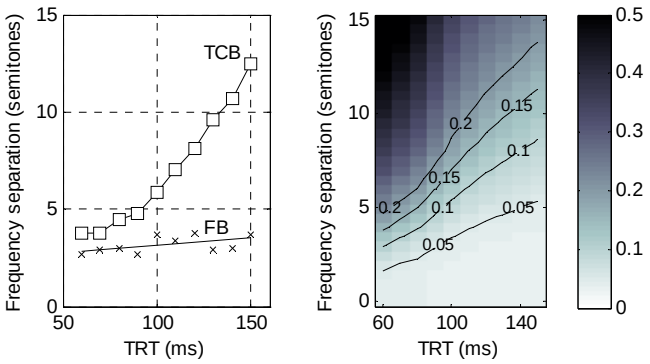


Figure 1: Schematic representation of the stimuli

4 Results

Data from van Noorden [5] are shown in Figure 2 (A). The curves indicate the temporal coherence boundary (TCB) and the fission boundary (FB). Above the TCB, the stimulus is always perceived as two streams, and below the FB the stimulus is always perceived as one stream. In the range between the curves, the percept can be controlled and either a fused or segregated percept can be achieved. Figure 2 (B) show the model results. The grey scale intensity indicates the eigenvalue ratio. A bright colour represents a low ratio, corresponding to a one-stream percept, and a dark colour represents a high ratio, corresponding to a two-stream percept. The lines indicate contours with fixed eigenvalue ratios.



(A) (B)

Figure 2: Results from the experiment. Part (A) shows experimental data from [5]. The upper curve shows the temporal coherence boundary (TCB), and the bottom line shows the fission boundary (FB). Part (B) shows the model simulation of the same experiment. The grey scale intensity indicates the eigenvalue ratio (λ_2/λ_1). The curves indicate contours with fixed eigenvalue ratios.

5 Discussion

The experimental data and the model simulation show that fast repeating tone sequences are more likely to split into two separate streams, whereas slowly repeating tone sequences can be perceived as a single stream for much larger frequency separations. In the modeling framework, a two-stream percept only occurs in the situation where (at least) two channels contain incoherent activity. Thus, in order to produce a two-stream percept, the stimuli must at least activate two separate peripheral filters. For the lowest non-zero frequency separation used in the simulation the A and B tones have frequencies of 1 kHz and 1.06 kHz. The bandwidth of the gammatone filter centered at 1 kHz is 133 Hz, and both tones will be processed by the same peripheral filters. Therefore, the model does not predict a two-stream percept. At larger frequency separations, e.g. 7 semitones, a substantial difference in the model results is observed between low and high TRTs. Since the frequency separation is the same, this cannot be explained by effects of spectral separation. Instead the different results are caused by the adaptation stage in the peripheral model which accounts for forward masking. The forward masking effect reduces the sensitivity of a peripheral channel after a tone has been presented, which effectively reduces the spread of excitation of the other tone. This in turn reduces the temporal coherence of the channels, causing the model to predict a two-stream percept. Physiological studies on animals [2] also suggest physiological forward masking as a possible cause of the stream segregation observed in the experimental paradigm utilized by van Noorden [5]. For tone sequences with small frequency separations, both tones in the stimuli were able to excite an auditory nerve tuned to one of the two frequencies. When the tone rate was increased (lower TRT), the excitation from the non characteristic-frequency tones was reduced, resulting in a reduced coherence of the auditory nerves tuned to the two frequencies. The observed behaviour corresponds to forward masking, with suppression of neural responses to a sound (the signal) following the presentation of a preceding sound (the masker). The results from our study support this hypothesis.

This study shows that by using the CASP model as a front-end of the model suggested by [2], the model can be extended to account for classical streaming phenomena relying on frequency separation and tone rate in addition to the grouping of spectral components due to synchrony. The functionality of the model is currently limited to providing an estimate of the "strength" of the two-stream percept. It cannot actually segregate sound sources. However, the suggested model may help to understand the underlying processes involved in primitive stream segregation, as was demonstrated with the influence of forward-masking on the perceptual organization of the tones.

References

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

[2] Bee, M. A., and Klump, G. M.: Auditory stream segregation in the songbird forebrain: effects of time intervals on responses to interleaved tone sequences. *Brain Behav. Evol*. 66 (2005). 197-214

[3] Elhilali, M., Ling, C., Micheyl, C., Oxenham, A. J., and Shamma, S. : Temporal coherence in the perceptual organization and cortical representation of auditory scenes. *Neuron*. 61 (2009). 317-329

[4] Jepsen, M. L., Ewert, S., and Dau, T.: A computational model of auditory signal processing and perception. *J. Acoust. Soc. Am*. 124 (2008). 422-438

[5] van Noorden, L. P. A. S.: Temporal coherence in the perception of tone sequences. Doctoral dissertation, Institute for Perception Research, Eindhoven, The Netherlands (1975).

[6] Patterson, R. D., Nimmo-Smith, I., Holdsworth, J., and Rice, P.: An efficient auditory filterbank based on the gammatone function. In paper presented at a meeting at the *IOC Speech Group on Auditory Modelling at RSRE*, December 14-15 (1997).

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

[8] Dau, T., Püschel, D., and Kohlrausch, A.: A quantitative model of the 'effective' signal processing in the auditory system. I. Model structure. *J. Acoust. Soc. Am*. (1996). 3615-3622