Chapter 46
Rate Versus Temporal Code? A Spatio-Temporal Coherence Model of the Cortical Basis of Streaming

Mounya Elhilali, Ling Ma, Christophe Micheyl, Andrew Oxenham, and Shihab Shamma

Abstract A better understanding of auditory scene analysis requires uncovering the brain processes that govern the segregation of sound patterns into perceptual streams. Existing models of auditory streaming emphasize tonotopic or "spatial" separation of neural responses as the primary determinant of stream segregation. While partially true, this theory is far from complete. It overlooks the involvement of and interaction between both "sequential" and "simultaneous" grouping mechanisms in the process of scene analysis.

Here, we describe a new neuro-computational model of auditory streaming. Inspired by recent psychophysical (cf. abstract by Micheyl et al.) and physiological findings, this model is based on the premise that perceived segregation results from spatio-temporal incoherence, rather than just tonotopic separation. While tonotopic separation still plays an important role in this model, it is an indirect one: tonotopic overlap tends to reduce temporal incoherence, which in turn impedes segregation. The model simulates responses at the level of the primary auditory cortex and performs a correlative analysis of cortical responses in order to assess how different sound elements evolve in time in relation to each other. An eigenvector decomposition of this coherence analysis is used to predict how the input stimulus is organized into streams. The model is evaluated by comparing its neural and perceptual predictions under various stimulus conditions to physiological and psychophysical results.

Keywords Auditory streaming • Temporal coherence • Auditory cortex • Multirate model

M. Elhilali (✉)
Department of Electrical and Computer Engineering, Johns Hopkins University, Baltimore, MD, USA
e-mail: mounya@jhu.edu

E.A. Lopez-Poveda et al. (eds.), *The Neurophysiological Bases of Auditory Perception*, DOI 10.1007/978-1-4419-5686-6_46, © Springer Science+Business Media, LLC 2010

46.1 Introduction

A well established Gestalt principle that has been often evoked in visual perception is that of *common fate*; i.e., the tendency to group together objects that move together with the same motion pattern and speed (Blake and Lee 2005). In the auditory domain, this principle simply translates into the observation that features which “move” together in time will likely group together perceptually (Bregman 1990). While simple enough in its postulate, this idea has not been explored in studies of neural correlates of streaming. Until now, the prevalent view, based on data recorded mostly in the primary auditory cortex, has focused on a “spatial” (i.e., tonotopic) explanation of how the brain solves the segregation problem (Fishman et al. 2004, 2001; Micheyl et al. 2005). This view postulates that neuronal populations with spatially segregated average responses will likely give rise to perceptually segregated streams. While this principle holds for simple sequential organization conditions such as alternating tone sequences, it does not generalize to other stimuli. In particular, this principle does not address the interaction between synchrony and sequential grouping cues. Mounting perceptual evidence, most recently from the accompanying paper by Micheyl et al. (this volume; see also: Micheyl et al. 2010), indicates that these principles do indeed interact in guiding how our brain segregates sound.

Based on these results, we explore the idea of *temporal coherence* as a new framework for understanding the neural correlates of streaming. We present physiological experiments from recordings in single units, which support a spatio-temporal basis of stream segregation. In addition, we propose a model that successfully validates the perceptual data using tone sequences, based on response properties in cortical neurons.

46.2 Neurophysiological Basis of Stream Organization in AI

We set out to explore the neurophysiological basis of the organization of streams as guided by perceptual grouping principles. In this study, we focused on the organization of synchronous and sequential tones at the level of primary auditory cortex (AI), and explored the nature of the neural code to both stimulus types in order to account for their very different percepts.

In this experiment, we examined the distribution of responses to two pure tones, the frequencies of which were adjusted relative to the best frequency (BF) of an isolated single unit in AI of awake ferrets in five steps (labeled 1–5), where positions 1 and 5 correspond to one of the tones being at BF. The frequency separation (ΔF) between the tones was fixed at 1, 0.5, or 0.25 octaves, corresponding to 12, 6, and 3 semitones, respectively. Tones A and B were shifted coherently relative to BF, with tone B starting at the BF and tone A ending at the BF. ΔF between the tones was 0.25, or 0.5, or 1 octave, which was fixed within a trial and varied among different trials. The total number of conditions was: 5 positions × 3 ΔF × 2 modes.

The results from a population of 122 units in the AI of four ferrets are shown in Fig. 46.1. We analyzed the average rate profiles of each unit to each tone sequence under all frequency separations and frequency position. When the tones are far apart ($\Delta F = 1$ octave; right panel of Fig. 46.1), and when either tone is near BF, responses are strong (positions 1 and 5); they diminish considerably when the BF is midway between the tones (position 3), suggesting relatively good spatial separation between the representations of each tone. When the tones are closely spaced ($\Delta F = 0.25$ octave; left panel of Fig. 46.1), the responses remain relatively strong at all positions, suggesting that the representations of the two tones are not well separated. More importantly, the average rate profiles are similar for both presentation modes; in all cases, the responses are well-segregated with significant dips when the tones are far apart ($\Delta F = 1$ octave), and poorly separated (*no* dips) when the tones are closely-spaced ($\Delta F = 0.25$ octaves). Thus, based on average rate responses, the neural data mimic the perception of the asynchronous but not the synchronous tone sequences. Therefore, the distribution of average rate responses does not appear to represent a general neural correlate of auditory streaming.

Overall, the results from the physiological experiments in awake ferrets reveal that a simple rate profile of responses in primary auditory cortical neurons is not sufficient to explain the perceptual difference between synchronous and alternating tone sequences. Clearly, a model that is successfully able to predict perception from these neural data will need to incorporate the time dimension.

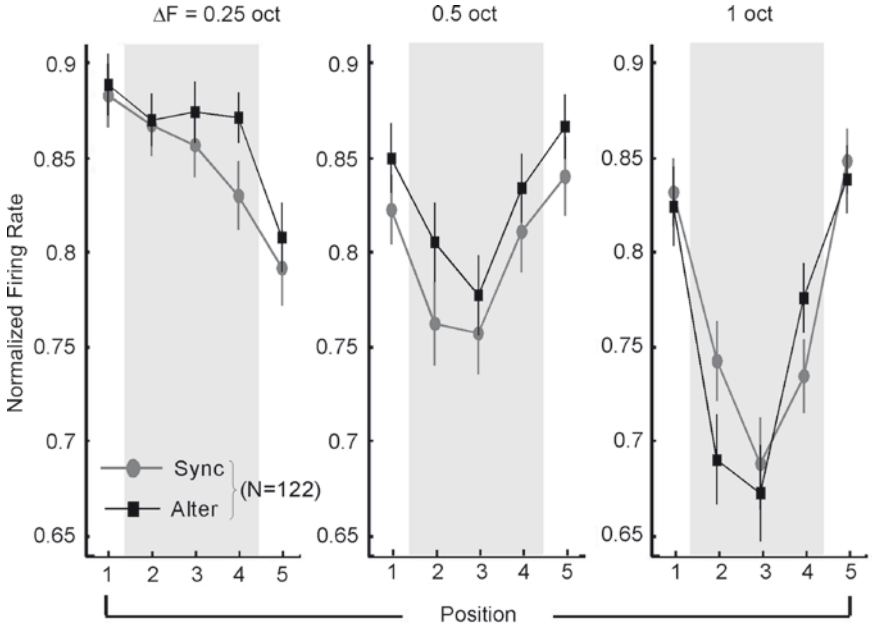

Fig. 46.1 Single unit recordings in AI using synchronous and alternating tone sequences

46.3 Spatio-Temporal Coherence Model

In this work, we emphasize the need for incorporating the temporal dimension in any model of stream organization. The basic premise of the model is that temporal coherence of sound features is an important principle for organizing sound mixtures into streams.

46.3.1 Auditory Processing from Periphery to Cortex

It is well established that acoustic signals undergo a series of transformations as they journey up the auditory system starting at the periphery all the way to cortex, mapping signals into a higher-dimensional representation. In order to capture this mapping in a mathematical formulation, we employ a model of auditory processing which abstracts from existing physiological data in animals and psychoacoustical data in human subjects as explained in details by Chi et al. (1999, 2005) and Elhilali et al. (2003).

The *early auditory stages* process the incoming acoustic signal through a sequence of stages representing cochlear filtering, hair cell transduction, and lateral inhibition to yield a final auditory spectrogram output (Fig. 46.2). This sequence of operations effectively computes a spectrogram of the signal using a bank of constant-Q filters, with a bandwidth tuning Q of about 12 (or just under 10% of the center frequency of each filter).

The *central cortical stages* further analyze the auditory spectrum into more elaborate representations and separate the different cues and features associated with different sound percepts. Electrophysiological evidence shows that cortical neurons are tuned to a variety of sound features, including BF, spectral bandwidth, and temporal dynamics. In the present study, we focus on the temporal integration tuning of cortical neurons. The time-scales of cortical dynamics are commensurate with dynamics of stimuli used in streaming experiments, as well as the dynamics of speech (Chi et al. 1999; Elhilali et al. 2003), musical melodies, and many other sensory percepts (Carlyon and Shamma 2003; Viemeister 1979). Mathematically, this analysis is achieved via an affine wavelet analysis of the auditory spectrogram.

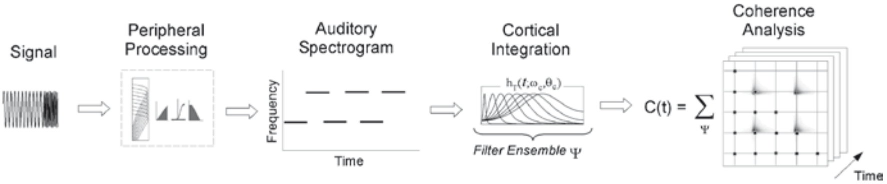

Fig. 46.2 Schematic of the spatio-temporal coherence model

The cortical temporal model estimates the temporal modulation content of the auditory spectrogram via a bank of modulation-selective filters (the wavelets) centered at each frequency along the tonotopic axis. Each filter is tuned ($Q=1$) to a range of temporal modulations (also referred to as rates or velocities (in Hz)), and is constructed by a temporal gamma function. This mother wavelet is scaled and shifted at different rates (Wang and Shamma 1995). Effectively, the model analyzes the time-sequence from each frequency-scale channel by convolving it with a temporal receptive field, effectively integrating the signal energy over a multiple scales of time ranging from 4 Hz to 64 Hz in logarithmic steps. It is worth noting that 64 Hz is a relatively high upper limit to known cortical dynamics. It is however used in the present study to ensure that short sounds (of the order of tens of milliseconds, such as one short tone) do indeed induce a response through the cortical integration stage.

46.3.2 Coherence Analysis

The focal proposition of this model is that features are grouped based on their coherence over time. The premise is extensible to a range of dimensions, and should be valid when applied to acoustic features, including frequency, spectral shape (or timbre), pitch, spatial location, etc. For details of correlation analysis based on spectral shape, (Elhilali and Shamma 2007) describes an analysis of informational masking stimuli.

In the present paper, we focus on applying the model along the frequency dimension. Specifically, a signal is processed through the peripheral processing stage (described in sect. 46.3.1) yielding a time-frequency representation. The outcome is then passed through the cortical temporal analysis described in sect. 46.3.1, where each spectral channel is integrated through multirate analysis windows. A correlation analysis is then performed on the channel against each other, as described in the steps below:

1. Map the signal $x(t)$ into a time-frequency spectrogram $y(t,x)$
2. Perform the cortical multirate analysis: $r\left(t,x;\omega_c,\theta_c\right)=y\left(t,x\right)*_t h_T(t;\omega_c,\theta_c)$
3. For each

$$R(t_0,\omega_c,\theta_c)=\left[r(x_1;t_0,\omega_c,\theta_c),y(x_2;t_0,\omega_c,\theta_c),\ldots,y(x_N;t_0,\omega_c,\theta_c)\right]^T,$$

   perform a correlation analysis:RR* (where * denotes complex-conjugate).
4. Integrate all matrices over the range of rate filters Ψ:

$$C(t_0)=\sum_{\Psi}R(t_0,\omega_0,\theta_0)$$

46.3.3 Decomposing the Coherence Matrix

The matrix C captures the degree of coherence in the neural responses at different frequency locations along the tonotopic axis. A high correlation value between two channels indicates a strong coherent activity at these two locations, while a low correlation value indicates lack of coherent activity. In order to determine the optimal factorization of the matrix in terms of maximally correlated channels, we perform an Eigen Value Decomposition (EVD). The structure of the significant eigenvectors is informative about which channels should be grouped together as one stream, and which should belong to a different stream. If a sound mixture contains only mutually coherent activity, its EVD will yield one strong eigenvalue corresponding to the channels that are maximally coherent. If, instead, a sound mixture contains two streams, their uncorrelated activity over time will emerge in the coherence matrix as two main directions, which will give rise to two strong eigenvalues. In the simulations presented in this work, we will use the ratio of second to the first eigenvalue as a correlate of this segregation: The smaller the ratio, the more likely the original sound contains *one* stream. In order to explore the actual structure of the streams associated with these eigenvector, we use the corresponding eigenvectors as weights on the different frequency channels.

46.3.4 Model Validation

In order to test the model's performance, we simulate a range of stimuli consisting of tone sequence with various spectro-temporal organizations. In the first simulation, we vary the degree of synchrony between 2 tone sequences, spanning the continuum from fully synchronous to fully asynchronous. In the second and third simulations, we explore the interaction between synchrony and sequential grouping cues, following the perceptual studies presented in the accompanying study by Micheyl et al. (2009).

46.3.4.1 Varying Degrees of Synchrony

In the first simulation, we use a sequence of a low A tone fixed at 300 Hz, and a high B tone at 952 Hz. Both tones were 75 ms long, with 10 ms onset and offset raised cosine ramp. We vary the onset to onset delay between the A and B tones from $\Delta T = 0\%$ (for fully synchronous) to $\Delta T = 100\%$ (fully asynchronous) with graduate steps in between. Figure 46.3a shows the ratio of eigenvalues as a function of ΔT. At the lowest end ($\Delta T = 0\%$), the coherence matrix maps to almost one main eigenvalue, hence the eigen-ratio is very small correlating with a percept of one stream (Elhilali et al. 2009). In this case, the coherence matrix can be mapped onto one main dimension, which yields an almost zero second eigenvalue. At the other

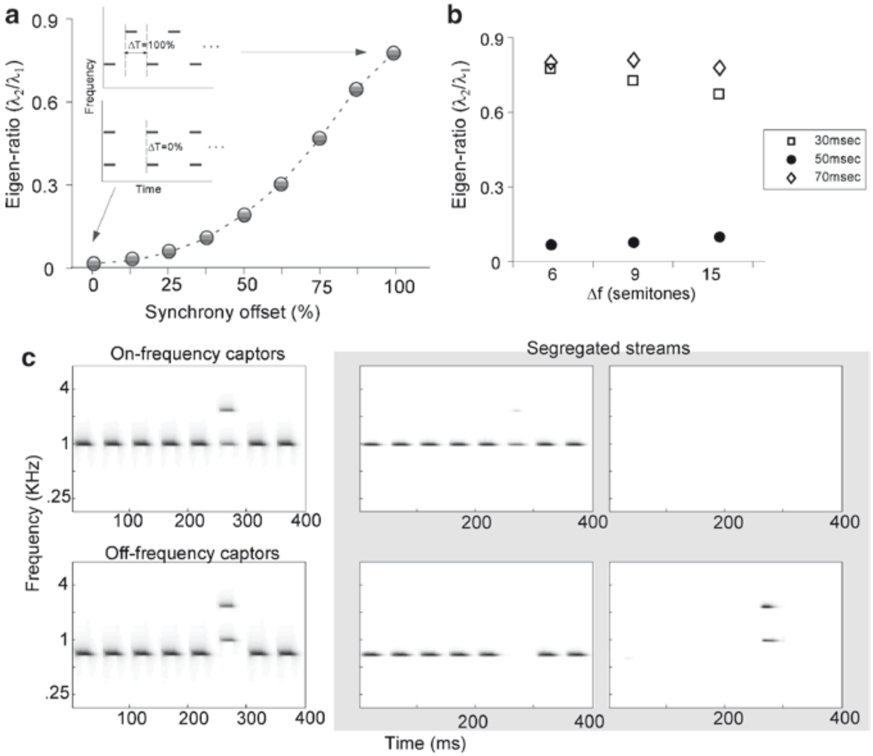

Fig. 46.3 Model simulation results

end of the continuum, the relative ratio of λ_2 to λ_1 reaches a high value indicating that both λ_1 and λ_2 are almost of equal value. In this case, the coherence matrix is in fact almost a rank 2 matrix, which can be mapped onto two main dimensions. In between these two extreme cases, we gradually vary the degree of synchrony between the two sequences. In this case, the relative ratio of λ_2 to λ_1 increases gradually; hence, allowing us to parametrically follow the influence of degree of asynchrony on grouping of two frequency streams, thereby allowing us to predict the transition between the percepts of one and two streams.

46.3.4.2 Experiment I: Synchrony Overrides Sequential Grouping

Next, we test the model using the same Experiment I paradigm used by Micheyl et al. in the accompanying paper (Micheyl et al. 2009). The stimuli consist of sequences of A and B, where A was fixed at 1,000 Hz and B was set at $\Delta F = 6$, 9, or 15 semitones above the A tone. Each tone was 100 ms in duration, including 10 ms raised-cosine onset and offset ramps. The silence interval between consecutive B tones was fixed at $\Delta T_B = 50$ ms. The silence between consecutive A tones

(ΔT_A) was varied across conditions. It was equal to 50 ms (in which case, the A and B precursors were synchronous), 30, or 70 ms (in which case, the A and B precursors were asynchronous).

In this simulation, we test how segregated are the two sequences under all variations of ΔF and ΔT_A. The psychoacoustic experiments show that the synchronous condition yields small thresholds even at the largest frequency separation (15 semitones). This low threshold is explained by the subjects' ability to make timing judgments within-stream, which is consistent with other results that synchronous tone sequences do indeed form a single perceptual stream even at large frequency separations exceeding 1 octave. In contrast, the asynchronous conditions where $\Delta T_A = 30$ or 70 ms yield larger thresholds, which in turn is consistent with an across-stream judgment (Bregman and Campbell 1971). The model simulations for these conditions are shown in Fig. 46.3b. The plot reveals that the synchronous condition (dark filled circles) does indeed yield a low eigen-ratio. This result is consistent with the perceptual finding that the 50 ms condition does indeed result from a percept of a single stream. It is worth noting that the eigen-ratio is low for all three frequency separation values of 6, 9, and 15 semitones. In contrast, the asynchronous conditions at $\Delta T_A = 30$ or 70 ms yield considerably higher eigen ratios, consistent with the perceptual findings.

46.3.4.3 Experiment II: Sequential Capture Overrides Synchrony Detection

In the next experiment, we explore the effect of sequential capture on synchrony judgments. This paradigm follows the design of Experiment II by Micheyl et al. (2009). The stimuli consisted of 3 tones, a single A tone at 1,000 Hz, a B tone at 6 or 15 semitones above A, and a C tone ("captor") at the same frequency as A ("On-frequency captor") or 6 semitones below A ("Off-frequency capture"). All tones were again 100 ms long with 10 ms raised-cosine onset and offset ramps. In the "On-frequency captor" condition, the A and B pair was surrounded by "captor" tones at the A frequency, with five captor tones before, and two captor tones after, the A–B pair. The captor tones were separated from each other, and from the target A tone, by a constant delay of 50 ms (Fig. 46.3c). In this condition, the target A tone formed part of a temporally regular sequence. In the "Off-frequency captor" condition, the frequency of the captor tones was set 6 semitones below the A tone, hence affecting the sequential grouping cues. The simulation results for this experiment are shown in Fig. 46.3c. Here, we show the results for frequency separation between A and B set at 15 semitones. In the first row, we show the simulation of the "On-frequency captors" condition. The leftmost panel shows the actual input analyzed by the model. After the coherence matrix is generated, we use the eigenvector structure to weigh the different frequency channels and group all coherent activity into one stream, and anticorrelated activity into a second stream (middle and rightmost panels). As shown in the figure, the sequential grouping cues override the presence of the synchronous A–B token, and groups the A tone with the preceding captor C tones. It is important to note that this result is only possible because of the cortical integration stage in the model, which gives the segregation of the streams inertia to look over relatively longer time scales; hence, allowing sequential cues to supersede the rules of synchrony. By this same inertia, a portion of the B tone is also "grabbed" along, though its energy is very weak because its presence was not long enough to drive responses from the cortical filters. The remaining energy of tone B is left in stream 2, though it does not show clearly in the spectrogram of the second stream. In contrast, the "off-frequency captors" condition results in a different organization (Fig. 46.3c, second row). In this case, the coherent activity from the C channel has no reason to group the A–B tones, hence segregating them into a separate stream.

46.4 Conclusions

Overall, the physiological data supports the proposal that our current thinking of streaming in the auditory system needs to incorporate the temporal axis as a key principle in organizing acoustic scenes. It is however important to emphasize that this principle does negate the rule of spatial segregation. If two alternating tones are too close together in frequency, their activation pattern will not be distinct enough to be able to see their anti-correlated temporal coherence. Hence, the overall principle is truly a *spatio-temporal* model of stream segregation, as tested directly by our computational model. An outstanding question remains as to the exact biological mechanisms involved in the process of "matrix decomposition" (i.e., *detection* of temporal coherence), and whether it is indeed a process that occurs at the level of the primary auditory cortex or beyond.

Acknowledgments We thank Pingbo Yin and Stephen David for their assistance with physiological recordings. More details concerning the methods and results of the physiological experiment may be found in Elhilali et al. (2009). This work is supported by grants from the National Institute on Deafness and Other Communication Disorders (R01 DC 07657) and the National Institute on Aging, through the Collaborative Research in Computational Neuroscience program (R01 AG 02757301).

References

Blake R, Lee SH (2005) The role of temporal structure in human vision. Behav Cogn Neurosci Rev 4:21–42

Bregman AS (1990) Auditory scene analysis. (Cambridge MIT Press), MA

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

Carlyon RP, Shamma S (2003) An account of monaural phase sensitivity. J Acoust Soc Am 114:333–348

Chi T, Gao Y, Guyton MC, Ru P, Shamma S (1999) Spectro-temporal modulation transfer functions and speech intelligibility. J Acoust Soc Am 106:2719–2732

Chi T, Ru P, Shamma SA (2005) Multiresolution spectrotemporal analysis of complex sounds. J Acoust Soc Am 118:887–906

Elhilali M, Chi T, Shamma SA (2003) A spectro-temporal modulation index (STMI) for assessment of speech intelligibility. Speech Commun 41:331–348

Elhilali M, Ma L, Micheyl C, Oxenham AJ, Shamma S (2009) Temporal coherence in the perceptual organization and cortical representation of auditory scenes. Neuron 61:317–329

Elhilali M, Shamma SA (2007) The correlative brain: a stream segregation model. In: Kollmeier B, Klump G, Hohmann V, Langemann U, Mauermann M, Uppenkamp S, Verhey J (eds) Hearing: from Sensory processing to perception. Springer, New York

Fishman YI, Arezzo JC, Steinschneider M (2004) Auditory stream segregation in monkey auditory cortex: effects of frequency separation, presentation rate, and tone duration. J Acoust Soc Am 116:1656–1670

Fishman YI, Reser DH, Arezzo JC, Steinschneider M (2001) Neural correlates of auditory stream segregation in primary auditory cortex of the awake monkey. Hear Res 151:167–187

Micheyl C, Shamma S, Elhilali M, Oxenham A (2010) Sequential and simultaneous auditory grouping measured with synchromy detection. In: E.A. Lopez-Poveda, A-R. Palmer, R. Meddis (eds) The neurophysiological baser of auditory perfection. Springer, New York

Micheyl C, Tian B, Carlyon RP, Rauschecker JP (2005) Perceptual organization of tone sequences in the auditory cortex of awake macaques. Neuron 48:139–148

Viemeister NF (1979) Temporal modulation transfer functions based upon modulation thresholds. J Acoust Soc Am 66:1364–1380

Wang K, Shamma SA (1995) Spectral shape analysis in the central auditory system. IEEE Trans Speech Audio Process 3:382–395