Chapter 59
Temporal Coherence and the Streaming of Complex Sounds

Shihab Shamma, Mounya Elhilali, Ling Ma, Christophe Micheyl, Andrew J. Oxenham, Daniel Pressnitzer, Pingbo Yin, and Yanbo Xu

Abstract Humans and other animals can attend to one of multiple sounds, and follow it selectively over time. The neural underpinnings of this perceptual feat remain mysterious. Some studies have concluded that sounds are heard as separate streams when they activate well-separated populations of central auditory neurons, and that this process is largely pre-attentive. Here, we propose instead that stream formation depends primarily on temporal coherence between responses that encode various features of a sound source. Furthermore, we postulate that only when attention is directed toward a particular feature (e.g., pitch or location) do all other temporally coherent features of that source (e.g., timbre and location) become bound together as a stream that is segregated from the incoherent features of other sources. Experimental

S. Shamma (✉) • P. Yin • Y. Xu
Department of Electrical and Computer Engineering,
Institute for Systems Research, University of Maryland,
College Park, MD 20742, USA
e-mail: sas@umd.edu

L. Ma
Bioengineering Program, University of Maryland, College Park, MD 20742, USA

Department of Electrical and Computer Engineering,
Institute for Systems Research, University of Maryland,
College Park, MD 20742, USA

M. Elhilali
Department of Electrical and Computer Engineering,
Johns Hopkins University, Baltimore, MD, USA

C. Micheyl • A.J. Oxenham
Department of Psychology, University of Minnesota, Minneapolis, MN, USA

D. Pressnitzer
Département d'études Cognitives, Equipe Audition,
Ecole Normale Superieure, Paris, France

Laboratoire de Psychologie de la Perception (UMR CNRS 8158),
Université Paris Descartes, Paris, France

B.C.J. Moore et al. (eds.), *Basic Aspects of Hearing*,
Advances in Experimental Medicine and Biology,
DOI 10.1007/978-1-4614-1590-9_59,

neurophysiological evidence in support of this hypothesis will be presented. The focus, however, will be on a computational realization of this idea and a discussion of the insights learned from simulations to disentangle complex sound sources such as speech and music. The model consists of a representational stage of early and cortical auditory processing that creates a multidimensional depiction of various sound attributes such as pitch, location, and spectral resolution. The following stage computes a coherence matrix that summarizes the pair-wise correlations between all channels making up the cortical representation. Finally, the perceived segregated streams are extracted by decomposing the coherence matrix into its uncorrelated components. Questions raised by the model are discussed, especially on the role of attention in streaming and the search for further neural correlates of streaming percepts.

1 Introduction

Listening in a complex acoustic environment fundamentally involves the ability to parse out and attend to one sound stream as the foreground source against the remaining background. In this view, streaming is an active listening process that engages attention and induces adaptive neural mechanisms that reshape the perceptual scene, presumably by enhancing responses to the target while suppressing responses to the background.

It is often conceptually useful to think of auditory streams as sequences of events or "tokens" that constitute the primitives of hearing, analogous to an alphabet. A token, such as a tone, a vowel, or a syllable, may have many concurrent perceptual attributes that arise very quickly through mechanical and hardwired neural mechanisms. Examples include a vowel's pitch, harmonic fusion, location, loudness, and the timbre of its spectral envelope. To segregate a sequence of tokens (be they phonemes or tones), it is necessary to satisfy a key condition – that the tokens be perceptually distinct from those associated with competing sequences, e.g., the pitches of two talkers or of two alternating tone sequences must be sufficiently different. This well-known principle of steaming has often been referred to as the "channeling hypothesis" implying that streams form when they activate distinct neuronal populations or processing channels (Bregman 1990; Hartmann and Johnson 1991). This requirement, however, is insufficient to explain stream formation, as we discuss next.

2 Feature Binding and Temporal Coherence

Forming a stream also requires *binding* of the parallel perceptual attributes of its tokens, to the exclusion of those belonging to competing streams. The simplest principle that explains how this phenomenon comes about is *temporal coherence* (Shamma et al. 2011). It asserts that any sequences of attributes that are temporally correlated will bind and form a stream segregated from uncorrelated tokens of perceptually different attributes. A simple example is the alternating two-tone

sequences that stream apart when their pitches are sufficiently different (Bregman 1990). When the tones are made fully correlated (synchronous sequences), the streaming fails because the two pitch percepts bind together forming a repeating complex perceived as one stream (Elhilali et al. 2009).

We postulate that temporal coherence is the organizing principle necessary to make the correct perceptual assignments as to which tokens form a stream. More specifically, correlated tokens form a single stream regardless of the diversity of their associated percepts, e.g., whether they are simple synchronized tones of different pitches, or the far more complex voices of a choir of soprano and bass pitches all singing in unison. The importance of temporal coherence in streams is a natural consequence of the fact that environmental sources normally produce sounds with temporally coherent attributes. For instance, a speech signal typically fluctuates in amplitude at temporal rates of a few Hertz. Consequently, the salience of all instantaneous estimates of its attributes would fluctuate similarly, be it the salience of its pitch, its location, or its spectral envelope. This temporal pattern is unlikely to be correlated with that of another signal emanating from an independent source, and hence the lack of temporal coherence is the simplest direct cue to the segregation of the two signals. When multiple "physical sources" become correlated as in the example of the choir, or when an orchestra plays the same melody, the entire group is treated perceptually as one source (Shamma et al. 2011).

In this chapter, we briefly review a mathematical model of this idea (Elhilali et al. 2009; Ma 2011) and discuss its biological realization and results of physiological experiments to test its predictions. We also discuss some of the psychoacoustic implications of this model and relate it to earlier formulations of the streaming process based on the Kalman prediction (Elhilali and Shamma 2008).

3 The Temporal Coherence Model

The proposed computational scheme emphasizes two distinct stages in stream formation (Fig. 59.1): (1) extracting auditory features and representing them in a multidimensional space mimicking early cortical processing and (2) organizing the features into streams according to their temporal coherence. Many feature axes are potentially relevant including the tonotopic frequency axis, pitch, spectral scales (or bandwidths), location, and loudness. All these features are usually computed very rapidly (<50 ms). Tokens that evoke sufficiently distinct (nonoverlapping) features in a model of cortical responses are deemed perceptually distinguishable and hence potentially form distinct streams *if* they are temporally anti-correlated or uncorrelated over relatively long time periods (>100 ms), consistent with known dynamics of the cortex and stream buildup.

Figure 59.1 illustrates these processing stages. Inputs are first transformed into auditory spectrograms (Lyon and Shamma 1997) followed by a multiresolution analysis analogous to that thought to occur in the primary auditory cortex (Chi et al. 2006). For the purposes of this model, this transformation is implemented in two steps: (1) a multiscale (spectral) analysis that maps incoming spectrograms into multiscale (bandwidth) representations, followed by (2) temporal rate analysis in

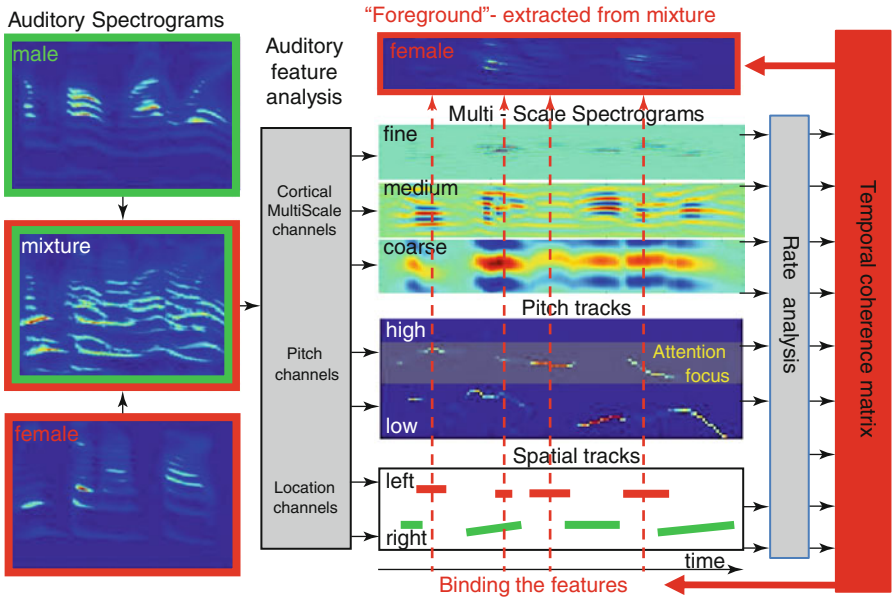

Fig. 59.1 Temporal coherence model. The mixture (sum of one male and one female sentences) is transformed into an auditory spectrogram. Various features are extracted from the spectrogram including a multiscale analysis that results in a repeated representation of the spectrogram at various resolutions; pitch values and salience are represented as a pitch-gram; location signals are extracted from the interaural differences. All responses are then analyzed by temporal modulation band-pass filters tuned in the range from 2 to 16 Hz. A pair-wise correlation matrix of all channels is then computed. When attention is applied to a particular feature (e.g., female pitch channels), all features correlated with this pitch track become bound with other correlated feature channels (indicated by the *dashed straight lines* running through the various representations) to segregate a foreground stream (female in this example) from the remaining background streams

which the temporal modulations of the (fine to coarse) multiscale spectrograms are analyzed by a filter bank tuned to rates from 2 to 16 Hz. In addition, other features such as pitch and location are estimated from the input spectrograms and the resulting tracks are later analyzed through the same rate analysis as for other channels, as illustrated in Fig. 59.1.

Subsequent to the feature and rate analysis, a pair-wise correlation matrix is computed among all scale-frequency-pitch-location channels, which is then used to group the channels into two sets representing the foreground and background streams. The responses are maximally correlated within each stream and least correlated across the two streams. One such factorization procedure is illustrated for the simple two-tone alternating (ALT) and synchronized (SYNC) sequences shown in Fig. 59.2. The correlation matrix cross-channel entries induced by these two sequences are quite different, being strongly positive (negative) for the SYNC (ALT) tones. A principal component analysis would then yield an eigenvector that can function as a "mask" to segregate the anti-correlated channels of the ALT stimulus, while grouping them together for the SYNC sequence, in agreement with their usual percept.

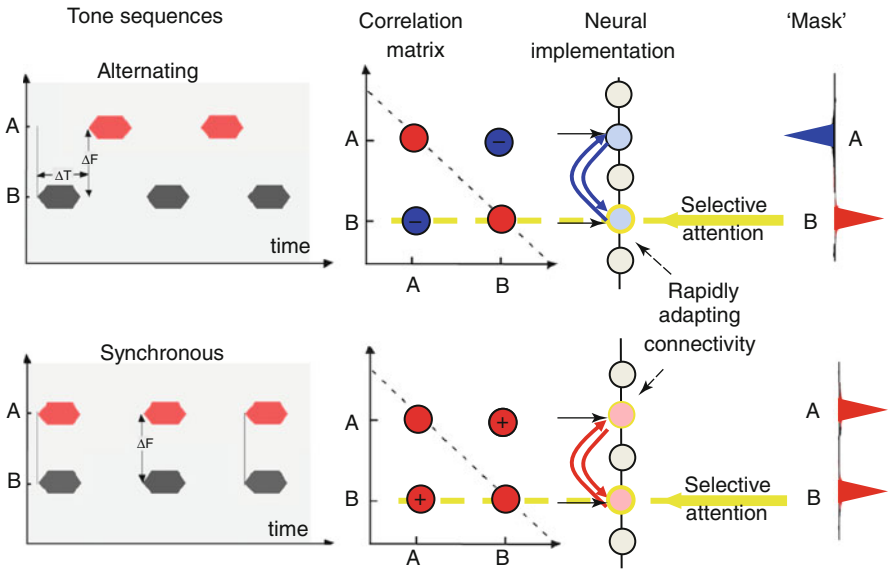

Fig. 59.2 Streaming of two-tone sequences. Alternating tone sequences are perceived as two streams when tones are far apart (large ΔF) and rates are relatively fast (small ΔT). Synchronous sequences are perceived as a single stream regardless of their frequency separation. The correlation matrices induced by these two sequences are different: pair-wise correlations between the two tones (A, B) are negative for the alternating sequence and positive for the synchronous tones. Neural implementation of this correlation computation can be accomplished by a layer of neurons that adapts rapidly to become mutually inhibited when responses are anti-correlated (alternating tones) and mutually excitatory when they are coherent (synchronous tones). When selective attention (*yellow arrow*) is directed to one tone (B in this example), the "row" of pair-wise correlations at B (along the *yellow dashed line*) can be used as a mask that indicates the channels that are correlated with the B stream. For the alternating sequence, tone A is *negatively* correlated with B, and hence, the mask is negative at A and eliminates this tone from the attended stream. In the synchronous case, the two tones are correlated, and hence, the mask groups both tones into the attended stream

4 Attention and Binding

It remains uncertain if the representation of streams in the brain requires attention or is simply modulated by it (Carlyon et al. 2001; Sussman et al. 2007). But it is intuitively clear that attending *selectively* to a specific feature such as the pitch of a voice (symbolized by the yellow-shaded pitch region in Fig. 59.1) results in binding the pitch with all other voice attributes in the foreground stream while relegating the rest of the concurrent sounds to the background. To explain how this process may occur, we consider the simpler two-tone stimulus in Fig. 59.2. When attention is directed to a particular channel (e.g., yellow arrow to tone B), the entries in the correlation matrix along the row of the selected channel can readily point to all the other channels that are highly correlated and hence may bind with it. Basically, this row is an approximation of the eigenvector of the correlation

matrix and can be used as "mask" to assign the channels to the different streams (rightmost panel). Note that in such a model, the attentional focus is essential to bring out the stream, and without it the correlation matrix remains unused. This idea is implemented to segregate the two-talker mixture in Fig. 59.1. Specifically, the female speech could be readily extracted by simply focusing on the rows of the correlation matrix corresponding to the female pitch (shaded yellow in Fig. 59.1) and then using the correlation values as a mask to weight all correlated channels from the mixture.

5 Biological Realizations and Evidence for Temporal Coherence

The temporal coherence model suggests that streaming is a dynamic process in which responses of the attended stream become enhanced relative to the background. This requires computing a correlation matrix whose entries change rapidly according to the ongoing correlational structure of the stimulus. A simple biologically plausible neural implementation of these computations is depicted in Fig. 59.2, where an ordered array of feature channels (e.g., the tonotopic axis) project to a layer of neurons. Each pair of neurons is reciprocally connected with a sign and strength which is continuously updated to reflect the ongoing correlation between their inputs ("Hebb's rule"). If the inputs are anti-correlated, the connectivity is mutually inhibitory (top panels, Fig. 59.2); if highly correlated, it is mutually excitatory (bottom panels, Fig. 59.2).

When neuronal connections change, they effectively alter the response selectivity of the neurons or their receptive field properties. It has been shown that engagement in an auditory task with attention to the stimulus is essential for such rapid changes to occur (Fritz et al. 2007). Therefore, in the context of the coherence model, we postulate that the mutual connectivity would not adapt to reflect the correlation matrix in a passively listening animal. Once the animal attends to the stimuli, connectivity begins to form, partly influenced by the focus of the attention. Thus, if attention is *global*, then connectivity adapts to reflect the mutual correlations among all units. If attention, however, is directed to a particular neuron, then only the mutual connections to this neuron are adapted, thus gating the input of the neuronal layer by allowing through only those that are positively correlated to it while suppressing others.

6 Physiological Correlates of Streams in Behaving Ferrets

To explore these hypotheses, recordings were made in the auditory cortex of ferrets trained to attend globally to ALT or SYNC two-tone sequences and to detect a transition to a random cloud of tones by licking a waterspout for reward, as illustrated

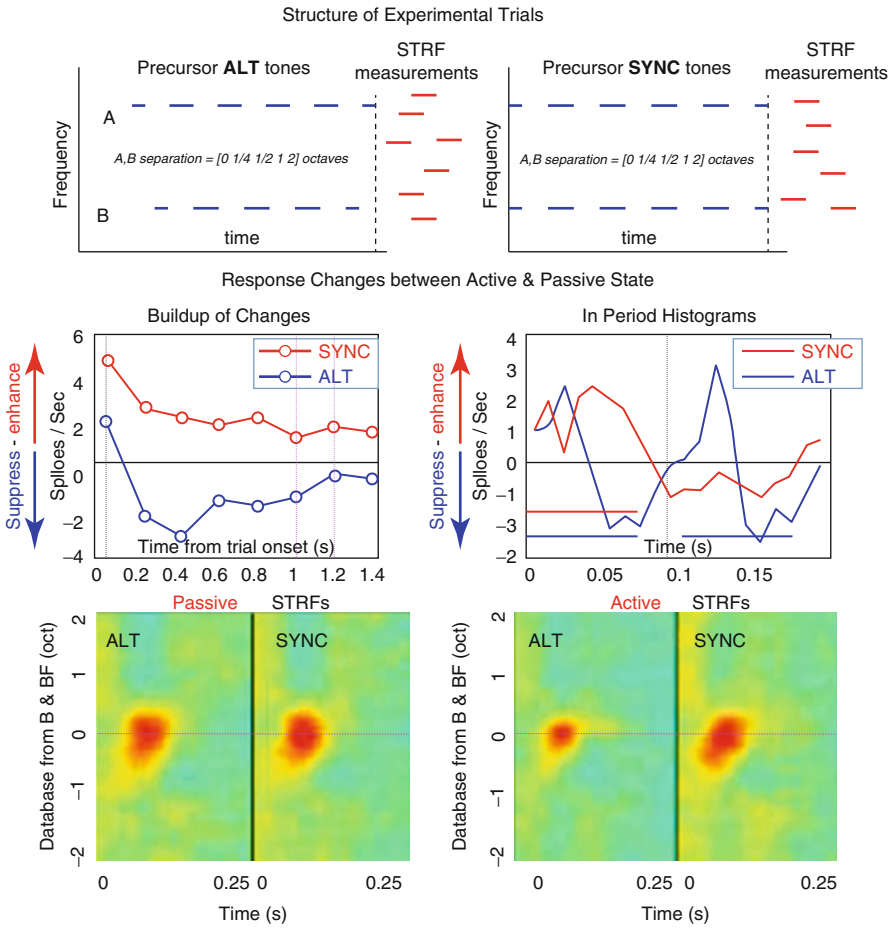

Fig. 59.3 Behavioral neurophysiology. (*Top Panels*) Structure of experimental trials. Ferrets listened to ALT or SYNC tone sequences presented for 1–3 s followed by a cloud of random tones (*red*) used to measure the STRF of the recorded neuron. (*Middle Panels*) Responses change when animals begin to listen attentively and globally to all tone sequences, i.e., not selectively to one tone. The responses become enhanced for the SYNC sequences (*red*) and attenuated for the ALT sequences (*blue*). Response changes (*left panel*) start immediately after onset of the trial but reach a plateau after three to four tone bursts (~0.5 s). Period histograms of responses to the tones (*red* and *blue bars* in *right panel*) reveal that SYNC tone responses (*red*) become significantly enhanced, while those of ALT tones become suppressed (*blue*). (*Bottom Panels*) STRFs measured at the end of tone sequences during the passive state show very little differences (*left panel*). During active attentive listening, STRFs become depressed after ALT compared to SYNC tone sequences (*right panel*)

in Fig. 59.3. The structure of the experimental trials is depicted in the top panels of Fig. 59.3. Responses were measured throughout the tone sequences to examine changes after trial onset as well as in the period histograms. Responses to the final random tone cloud were used to estimate the spectrotemporal receptive fields (STRFs) (deCharms et al. 1998). The type of sequence (ALT or SYNC) and its

frequency combinations were randomly interleaved throughout a block of trials. Figure 59.3 (middle and bottom panels) displays results of recordings from 96 cells that were tuned at the frequency of the B tones, with A tone frequencies up to two octaves above and below that of the B tone.

The average responses to the tone sequences changed dramatically when the passive animal began to attend globally to the stimuli. In both SYNC and ALT conditions, average responses adapted rapidly to a steady state by about the third burst period (*left-middle panel*; Fig. 59.3). SYNC responses were significantly enhanced compared to their passive level, whereas ALT responses were suppressed. The changes in period histograms between the active and passive states for the SYNC and ALT stimuli are compared in Fig. 59.3 (*right-middle panel*). The SYNC response increases significantly during behavior; by contrast, the ALT response displays a strong but slightly delayed suppression soon after each tone's onset response.

Finally, the *bottom panels* contrast the STRFs measured after the end of the SYNC and ALT sequences during the passive and active states. When the animal was passive (Fig. 59.3: *left-bottom panel*), the average STRFs were similar. During behavior, however, there was a strong suppression of the STRFs following the ALT sequences. The average STRF was slightly enhanced after the SYNC sequence. These STRF changes persist but gradually weaken over the next few seconds.

7 Discussion

The physiological results are consistent with the postulates of the temporal coherence model. During SYNC sequences, responses become enhanced possibly reflecting mutually positive interactions. The opposite occurs during ALT sequences, where neurons decrease their overall responsiveness and compete as expected from mutually inhibitory interactions. Furthermore, we postulate that if attention had been directed to one of the ALT competing tones, it would have enhanced (to the perceptual foreground) the attended responses at the expense of the competing tone, consistent with previously published experimental results (Yin et al. 2006).

Finally, the temporal coherence model bears a close relationship to the Kalman predictive clustering-based algorithm described in Elhilali and Shamma (2008). This is because the principal eigenvector of the correlation matrix acts as a reduced feature "template" (or "mask" in Fig. 59.2) which combines and extracts the input feature vectors that match it. In the Kalman prediction model, the same matching operation is performed, but the "template" is computed by a classic on-line gradual clustering of the input patterns. Under certain conditions (e.g., input pattern normalization), the two types of algorithms are equivalent and yield similar clusters (Duda and Hart 1973).

References

Bregman A (1990) Auditory scene analysis: the perceptual organization of sound. MIT Press, Cambridge

Carlyon R, Cusack R, Foxton J, Robertson I (2001) Effects of attention and unilateral neglect on auditory stream segregation. J Exp Psychol Hum Percept Perform 27:115–127

Chi T, Ru P, Shamma S (2006) Multiresolution spectrotemporal analysis of complex sounds. J Acoust Soc Am 118:887–906

deCharms R, Blake D, Merzenich M (1998) Optimizing sound features for cortical neurons. Science 280:1439–1443

Duda R, Hart P (1973) Pattern classification and scene analysis. John Wiley and Sons, New York

Elhilali M, Shamma S (2008) A cocktail party with a cortical twist: how cortical mechanisms contribute to sound segregation. J Acoust Soc Am 124:3751–3771

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

Fritz J, Shamma S, Elhiliali M (2007) Does attention play a role in dynamic receptive field adaptation to changing acoustic salience in A1? Hear Res 229:186–203

Hartmann W, Johnson D (1991) Stream segregation and peripheral channeling. Music Percept 9:155–184

Lyon R, Shamma S (1997) Computational strategies for pitch and timbre. In: Hawkins H, McMullen T, Popper A, Fay R (eds) Auditory computations. Springer, New York, pp 221–270

Ma L (2011) Auditory streaming: behavior, physiology, and modeling. PhD Thesis, Bioengineering Program, University of Maryland, College Park

Shamma S, Elhilali M, Micheyl C (2011) Temporal coherence and attention in auditory scene analysis. Trends Neurosci 34:114–123

Sussman E, Horvát J, Winkler I, Orr M (2007) The role of attention in the formation of auditory streams. Percept Psychophys 69:136–152

Yin P, Ma L, Elhilali M, Fritz J, Shamma S (2006) Primary auditory cortical responses while attending to different streams. In: Kollmeier B, Klump K, Hohmann V, Langemann U, Mauermann M, Uppenkamp S, Verhey J (eds) Hearing – from sensory processing to perception. Springer, New York, pp 257–266