NIH Public Access
Author Manuscript
Trends Neurosci. Author manuscript; available in PMC 2012 March 1.

Published in final edited form as:
Trends Neurosci. 2011 March ; 34(3): 114–123. doi:10.1016/j.tins.2010.11.002.

Temporal coherence and attention in auditory scene analysis

Shihab A. Shamma[1], **Mounya Elhilali**[2], and **Christophe Micheyl**[3]

[1]Department of Electrical and Computer Engineering and Institute for Systems Research, University of Maryland, MD 20742, USA

[2]Electrical Engineering Department, John Hopkins University, Baltimore, MD, USA

[3]Department of Psychology, University of Minnesota, MN 55105, USA

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 argue 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 towards a particular feature (e.g., pitch) 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.

The auditory “scene analysis” problem

Humans and other animals routinely detect, identify, and track sounds coming from a particular source (e.g., someone’s voice, a conspecific call) amid sounds emanating from other sources (e.g., other voices, heterospecific calls, ambient music, or street traffic) (Figure 1). The apparent ease with which they determine which components and attributes in a sound mixture arise from the same source belies the complexity of the underlying biological processes. By analogy with the “scene segmentation” problem in vision, this is referred to as the “auditory scene analysis” problem [1](Glossary) or, more colloquially, the “cocktail party” problem [2-4]. Understanding how the brain solves this problem is a fundamental challenge facing auditory scientists as it will shed light on the difficulties afflicting the hearing-impaired in multi-source environments [9], and give rise to more effective front-ends for auditory prostheses and automatic speech recognition [10].

Recent studies have inspired numerous hypotheses and models concerning the neural underpinnings of perceptual organization in the central auditory system, and especially the auditory cortex (see [3,7-8,11-20] for reviews). One prominent hypothesis that underlies most investigations is that sound elements segregate into separate “streams” whenever they activate well separated populations of auditory neurons that are selective to frequency or any other sound attributes that have been shown to support stream segregation (21-30). We shall

Corresponding author: Shihab Shamma, Electrical and Computer Engineering and Institute for Systems Research, University of Maryland, College Park, MD 20742, Tel: 301-405-6842, sas@umd.edu.

refer to this hypothesis as the “population-separation hypothesis”. Another influential hypothesis is that streams are formed automatically or *pre-attentively*, in or below the primary auditory cortex (30,60-61).

In this Opinion piece, we point out shortcomings of these two hypotheses and propose an alternative to each within an overall framework for understanding auditory scene analysis and its neural basis. Based on a combination of neurophysiological data, psychophysical observations, and computational studies, we argue that the formation of auditory streams depends fundamentally on the *temporal coherence* of responses of neural populations selective to various sound attributes (e.g. frequency, pitch, timbre, spatial location) in the auditory cortex. In addition, we suggest that attention plays a key role in *stream formation*, as it biases the auditory system toward a particular grouping or *binding* of sound-source attributes depending on the listener’s current behavioral or perceptual goals.

Temporal coherence in auditory scene analysis

Problems inherent to auditory scene analysis are largely common to those found in visual scene analysis. However, there are a few notable unique aspects. In particular, whereas natural and artificial visual scenes often contain a large proportion of static or slow-moving elements, auditory scenes are essentially dynamic, containing many fast-changing, relatively brief acoustic events (referred to as “tokens” in Box 1) [5,6]. Therefore, an essential aspect of auditory scene analysis entails the linking over time, or “streaming” of tokens produced by the same sound source, while simultaneously separating them from others produced by other sources. We shall explain here why we are of the opinion that the key first step to this process of streaming is the *temporal coherence* of the tokens within a stream (or equivalently, their incoherence across streams) and not the widely assumed population-separation hypothesis. Please see **Glossary** for precise definition of *Coherence*.

The population-separation theory of auditory streaming

Over the last decade, numerous psychophysical and neurophysiological studies of auditory streaming have concluded that the perceptual organization of sounds into streams is determined by the spatial overlap between responsive neural populations in the peripheral and/or central auditory system. Simply stated, under this hypothesis, sounds that activate distinct (or weakly overlapping) neural populations are heard as separate streams. Since the tonotopic axis is a major organizational principle throughout the auditory system, most models based on this “population-separation” theory of auditory streaming have focused on the frequency dimension [21-23] and have successfully accounted for many important aspects of the perceptual organization of simple tone sequences [24-30] (Box 2). Within a broader framework, however, this hypothesis has been extended to account for stream formation based on other features, such as spectral shape (timbre), periodicity (pitch), or spatial sound localization [7,30-31]. The necessary “multi-feature” analysis presumably arises from neural responses in the central auditory system that are selective to attributes other than frequency, e.g., to various spectral and temporal characteristics of sounds [32-40], sound-source location [41,42,102], and pitch [43].

However, the separation of neural responses cannot account for the observed influence of the relative timing of sounds on the streaming percepts. For example, the “population-separation” hypothesis predicts that both alternating and synchronous tones (see Figure IIa,b in Box 2) that differ widely in frequency should be heard as separate streams. This prediction is contradicted by psychophysical and neurophysiological data [44], which demonstrate that sequences of tones that are separated by an octave or more are still heard as a *single* stream if the tones are synchronous or, more precisely, fully coherent in time (Box 2 and Glossary). Numerous other psychoacoustical findings indicate that coherence strongly promotes perceptual grouping [45]. To account for these findings, it is necessary to consider the relative timing of the neural responses, or more specifically their temporal coherence.

Temporal coherence and auditory streaming

By combining multi-feature representations and temporal-coherence analysis, one is led to a general and flexible scheme for explaining the formation of auditory streams. This framework is illustrated in Figure 2. It begins with frequency analysis in the cochlea, followed by extraction of a wide variety of spectral and temporal features, including a multi-resolution representation of spectral shapes, harmonicity, temporal periodicity, and inter-aural time and level differences. Some of these features (e.g. harmonicity and inter-aural differences) are related directly to perceptual attributes (e.g. pitch and location).

We next postulate the existence of a temporal coherence-analysis stage that computes correlations among the outputs of the different feature-selective neurons. The correlations are computed over relatively long time windows, ranging in duration between 50 and 500 ms. This range is consistent with the slow dynamics of stimulus-induced fluctuations in spike rate in the auditory cortex (<20 Hz) [35,103]. It is also consistent with the sound-presentation rates over which the formation of streams usually occurs, as well as with the rates of temporal-envelope fluctuations typically encountered in speech (syllabic rate) and music (tempo). While these coherence computations may take place automatically (pre-attentively), we postulate that active listening or attention is necessary to exploit the results and *bind* coherent channels into a perceptual stream, while segregating them from the remaining incoherent channels. Clearly, more complex patterns of streaming would arise if channels were partially coherent, and hence may belong to none or to more than one stream simultaneously [101].

Temporal coherence solves the auditory binding problem

The principle of "grouping by temporal coherence" provides an elegant solution to the auditory "binding problem", i.e., the problem of associating different sound features (loudness, pitch, timbre, and spatial location) with the "correct" (i.e. corresponding) sound source, and of linking these features together to produce a unified percept, while keeping them separate from the features of other sources. This is because features of a particular source will, in general, be present whenever the source is active, and absent when it is silent. Furthermore, different sound sources (with all of their associated features) will rarely fluctuate in strength at *exactly* the same times. However, it is important to point out that the plausibility of coherence-dependent computations in mediating the perceptions of streaming, and their biological underpinnings, still need to be investigated. Experimental support for this hypothesis might include the identification of neurons whose responses depend strongly on the temporal coherence of their input spike trains, or of "combination sensitive" neurons that respond selectively to particular combinations of inputs; for instance, neurons that respond strongly to two simultaneously presented tones, even though they respond weakly to either tone alone [46-50,105]. The identification of such neurons would provide a substrate for the integration of temporally coherent responses across spatially distributed neural populations.

The hypothesis that temporal coherence across neural populations solves the binding problem is not unique to the auditory modality [51,52]. Temporal coherence across different sensory modalities might support cross-modal binding (as in lip-reading where both visual and auditory inputs are used). However, relatively little is known concerning interactions between auditory and visual or somatosensory inputs in auditory streaming (see [53] for an exception). Finally, variants of the principle of grouping by temporal coherence have been applied earlier to sensory perception problems [54,55], including models of auditory scene analysis [56,57]. However, our current approach outlined here differs in that all temporal properties of the responses stem entirely from the relatively slow-varying stimulus features (<20 Hz) that induce phase-locked cortical responses, and not, for example, from any intrinsic (i.e., not stimulus driven) oscillatory activity in the nervous system (e.g. local field-potential oscillations in the gamma frequency range [58]). Two recent computational studies have implemented some of these ideas to successfully simulate the formation of auditory streams for a wide variety of stimuli [44,59], including simple sequences of regularly repeating tones, stochastic tone sequences, and concurrent speech sounds.

The role of attention in auditory stream formation

Is streaming a pre-attentive process?

A widely held view, which has emerged from electrophysiological studies in humans [60-65], is that auditory streams are formed “pre-attentively” in the auditory system, much like the extraction of low-level features in early pre-cortical stages. Dependent on the listener’s intentions, and guided by representations of previously encountered auditory “objects” (or streams) that are now stored in memory, attention would simply serve to enhance the perception of a particular stream in the auditory scene, while suppressing others [66-70]. Thus, in this view, attention is involved in “stream selection”, rather than in “stream formation” [71] which remains essentially a pre-attentive process. This is reminiscent of similar views proposed earlier in the visual modality [72,73].

We refer to this as the “object-based attention” theory of auditory scene analysis. One challenge to this hypothesis is that complex auditory scenes can often be organized perceptually in many different ways. For instance, when listening to an orchestra, one can listen to the ensemble, to a particular instrument (e.g. the trumpets or flutes), or to a group of instruments (e.g. the strings or the woodwinds). In the first case, the orchestra will be heard as a single stream; in the other cases, different streams will be heard, corresponding to individual instruments, or to groups of instruments. It seems unlikely that the brain would waste resources representing large numbers of potential decompositions of auditory scenes into streams *prior* to (and independently of) attentional selection.

Attention influences stream formation

The hypothesis that attention can only operate on neural representations of already formed auditory “objects” is contradicted by psychophysical findings. Firstly, when listening to sound sequences such as those illustrated in Figure IIb in Box 2, the frequency separation required to induce a percept of two separate streams is usually much smaller if the listener is actively trying to “hear out” the high-pitch tones than if he/she is listening less selectively [74]. This finding indicates that active engagement in the task, and the implicit attention brought to bear during it, does not merely serve to select one among several already formed streams; instead, attention can influence the stream-formation process itself [75, 76].

At the neural level, attention may influence auditory stream formation in at least two important ways. First, it can modulate responses to different features, thus modifying the neural representation–and ultimately, the perceptual saliency of these features. During the last decade, several studies (reviewed in [77]) have demonstrated such rapid task- and attention-dependent changes in the spectro-temporal receptive fields of the auditory cortex. Preliminary results of a study that sought to test this hypothesis in awake-behaving animals performing streaming tasks indicate that during behavior, responses to the attended stream become better segregated compared to those in response to the background sounds [78].

In addition, attention can influence streaming by modulating the temporal coherence of neural populations [79]. Recent findings that indicate that temporal coherence between distinct populations of neurons tuned to a target is augmented during attention (Figure 3) are consistent with this hypothesis. Enhanced phase coherence between distributed neuronal clusters helps to resolve the competition between different acoustic features in a sound mixture by facilitating the temporal coherence analysis, thereby heightening the perceptual boundary between the currently attended stream and the background. Evidence for such a general mechanism by which attention influences the timing of neural responses has been found in the auditory system (e.g. [70]), and also in the visual [80] and somatosensory modalities [81].

Temporal coherence reconciles feature- and object-based attention

Temporal coherence can help bind the diverse features of a stream in a manner that highlights an elegant synergy between object-based and feature-based attention in stream formation. To elaborate when a feature is selectively attended to, it effectively serves as the *anchor* that points to, and binds the remaining features that are coherent with it. For example, when attempting to "hear out" a female talker in the presence of a concurrent male talker, a listener may choose to attend to the high-pitch, and then through that particular aspect perceptually access all other voice attributes that are coherent with it (e.g. location and timbre). If, instead, the listener has access to the approximate location of the female talker (e.g. based on visual information), s/he could attend selectively to the corresponding region of auditory space, and subsequently access other coherent attributes (e.g. pitch and timbre) of the female voice. Thus, as long as one distinctive feature of a target stream is sufficiently salient to be attended to by the listener, s/he could have access to and the ability to distinguish all other features of the target stream. This process is, in some ways, similar to that invoked during perceptual learning studies in which observers attend selectively to task-relevant visual features, and learn not just these features, but also all other task-irrelevant features that occur concomitantly - even when the irrelevant features are too weak to be consciously perceived [82].

Memory in auditory scene analysis

As outlined here, our focus has been on the postulate that sequential processes utilize *dynamic* cues to stream sounds and render them perceptually as auditory objects. One might ask why has the emphasis been placed on *dynamic* cues, given that *static* scenes such as images have been the primary vehicle for the study of segmentation and identification of visual objects. We propose that in the absence of dynamic cues, recognition of objects in static scenes must combine memory (i.e., priors or heuristics) with low-level perceptual primitives such as edges, edge-continuity, texture analysis and color. Such perceptual primitives are analogous to the percepts of harmonicity and binaural disparities in audition (referred to as "instantaneous percepts" in Box 1). For example, identifying a complex assemblage of oval shapes, straight and curved edges, and multiple colors and textures on a canvas as a face or a tree must invoke pre-existing (either learned or hardwired) templates of these objects. The same logic applies to *static* auditory scenes: determining whether a sustained (or steady) sound from a throat-singer or from two simultaneous choir singers is either one source or two is essentially arbitrary and depends on the listener's expectations and contextual cues (memory) and not sensory evidence alone. However, once dynamic cues are introduced, as when the two voices become dynamically modulated (coherently or incoherently) in pitch, loudness, or timbre, the sensory evidence becomes the key to the perceptual streaming of the sound either into one complex (composed of two elements) or into two separate sources (Box 3).

To summarize, listening for sources in natural environments often engages hardwired preferences of conspecific vocalizations and memories of familiar sounds that are important to the animal for survival or reproduction [83-84,98-99]. But in many common situations when sources are novel (such as speech produced by an unfamiliar speaker or musical notes of a novel melody), or when the acoustic environment is complex and cluttered, dynamic cues (temporal coherence) play the primary role in enabling attention to bind coherent attributes and organize them into streams.

Summary

Here, we proposed two ideas within an overall framework to explain the perception of auditory scenes. The first is that auditory stream formation is critically dependent on the temporal coherence between neural responses to sounds in the auditory cortex. Specifically, when stimulus-induced cortical responses are temporally coherent, the features they represent can potentially become perceptually unified (or bound) as one stream, distinct from other temporally incoherent responses. This principle explains stream formation and perception of a wide range of stimuli including spectrally and temporally complex natural sounds such as voices and music. The second hypothesis is that attention influences stream-formation by initiating the binding process and modulating the neural representations of the acoustic features and/or of temporal coherence patterns among these features. Both of these hypotheses remain under intensive scrutiny and experimentation. Nevertheless, they are already proving useful as a theoretical framework to broaden and guide future investigations (Box 4) of the neural basis of auditory scene analysis.

Box 1: Principles of stream formation and perception

Percepts and processes underlying auditory perceptual-organization can be conceptually divided into two categories: *instantaneous* (sometimes referred to as “simultaneous”) percepts and *sequential* processes (or “stream formation”) [1].

An *instantaneous* percept refers to that of a sound epoch or *token* that arises rapidly after its onset and continues throughout its duration. Natural sounds are dynamic and can be conceptualized as *sequences of tokens*, each token having associated perceptual attributes (pitch, loudness, timbre, and location) that reflect its frequency components and their relationships, e.g. whether harmonically related or what their relative amplitudes are. Sound tokens encountered in our environment are endowed with richly varied and complex percepts (some are illustrated in Figure Ia). For instance, a sound token may consist of one or two tones, a perceptually fused harmonic complex, or an inharmonic complex with a “fractured” multi-tone percept. Tokens may also have attributes other than frequency, such as the pitch of musical notes or a whole chord (Figure Ib), and the perceived location of a point source (Figure Ic). Finally, tokens may have complex attributes such as the timbre of one or more simultaneous vowels (Figure Id), a highly diffuse sound in a large reverberant hall, or that of a large choir singing in unison. All these percepts are extracted relatively early and rapidly in the auditory system by basic neural structures (within a few tens of milliseconds - hence the term “instantaneous percepts”), and there is a large body of psychoacoustic and neurophysiological results that relates the acoustic parameters of a complex sound to these attributes (e.g. see [85] for a review).

Sequential organization specifically refers to the sorting of interleaved sound tokens arriving from a mixture of sources into streams that can be selectively attended to, and tracked over time. Examples of auditory streams are: (i) two independent interleaved melodies played by a violin and a piano, (ii) the melody of a piano within an orchestra, or (iii) someone’s voice in a crowd. Each stream can be thought of as a sequence of tokens that the listener can attend to and perceive as the target “stream” or melody. To do so, the listener must distinguish the attributes of the different tokens (instantaneous percepts), and organize them into separate streams (sequential process).

This process has a few basic properties that are addressed in this article. One is that tokens in different streams must be sufficiently incoherent in time, and must also be perceptually distinct enough to reflect the different acoustic characteristics of their sources. These percepts should remain relatively stable over time within a stream. For instance, the timbre and pitch of sounds within a stream should not change drastically and quickly over time, or these sounds will fail to form a coherent auditory stream. This is essentially identical to the "continuity" principle, which is often invoked as a key ingredient for the learning of object invariance along various dimensions [86,100]. Since sequences of tokens unfold relatively slowly over time (> 50 ms), sequential organization (or formation of a stream) is a slow process that may take several seconds to complete, especially when the tokens to be segregated are perceptually close. Finally, it is argued that unlike the instantaneous processes that have been demonstrated even in anesthetized animals [104], stream formation engages cognitive processes, such as attention and expectations [19].

Box 2: Coherence and attention in streaming: Examples with tone-sequences

The simplest stimuli to illustrate the role of temporal coherence and attention in stream formation are the much-studied sequences of pure tones [1]. To start, tones that *alternate* repeatedly between two far-apart frequencies are usually heard as two streams (Figure II*a*). This, we claim, is *not* because the responses are widely spaced on the tonotopic axis, but rather because they induce incoherent responses (i.e. as illustrated in the separate auditory channels of A and B). The evidence for this statement is that when channel responses in A and B are made temporally coherent, e.g. when the tones are synchronized (Figure II*b*), the tones are heard as *one* stream despite their large separation in frequency [41].

While temporal coherence computations could occur without significant cognitive control (e.g. similar to cochlear frequency analysis), we propose that attentive listening is necessary for subsequent exploitation of the results to bind coherent attributes or group channels into different streams. An experimental finding that is consistent with this claim is that when one attends to the incoherent responses of the alternating tones illustrated in Figure II*a*, one initially hears a unified percept that only gradually gives way to two streams (known as the build-up) [76], suggesting that the incoherence is ignored prior to the onset of attention.

To explain further the relationship between coherence, binding, and streaming within the context of the model, consider the percepts evoked by the alternating and synchronous tones (Figure II*a,b)* when presented in *separate* ears (e.g. A-Right; B-Left ear). Each tone now has two coherent attributes, pitch and location, and so by attending to one (e.g. pitch) it binds perceptually with the other (location) to form one stream. The alternating tones (e.g. as illustrated in Figure II*a*) are incoherent, and hence their attributes are also incoherent and will stream apart, making it easy to distinguish and associate each tone with its pitch and ear-of-entry. By contrast, the synchronous tones (e.g. as illustrated in Fig*ure IIb*) and *all* their attributes are coherent, and hence all will bind together into one stream. In this case, we predict that listeners would find it difficult to determine which tone is in which ear even if the frequencies are well separated.

We should emphasize that *synchronicity* and *coherence* are different notions. The first is an instantaneous property, whereas the latter is an average measure (a windowed cross-correlation). We propose that only *coherence is key to streaming*. To illustrate this distinction, consider the closely-spaced alternating tones of Figure IIc. These tones are asynchronous, but their channels (A and B) are sufficiently overlapped in their frequency ranges that they carry similar (coherent) responses and hence the tones are heard as *one* stream [25,93,96]. By contrast, we predict that the *synchronous* tone sequences illustrated in Figure II*d* stream apart because no pairs of channels (i.e. A and B) have coherent responses. This latter example in fact immediately generalizes to the so-called *informational masking* (IM) stimulus (see Figure II*e*) in which the *target* tone (illustrated by the green responses in channel A) streams apart from the surrounding (synchronous) *masker* tones when the responses in the target and masker channels are sufficiently incoherent. A final well-known example is that which is illustrated in Figure II*f* where the two *synchronous* tones are perceived in *separate* streams [1] because the two different frequency channels have incoherent responses.

Box 3: Hearing out sounds within a stream

The perceptual segregation of streams - a process of sequential organization - should not be confused with the hearing out of a component out of many simultaneous components in a sound complex such as a musical chord. When listening to a complex sound token, one can listen “analytically”, “hear out” individual sound components, and even attend selectively to one of these components. For example, normal listeners can readily hear out a mistuned component in a harmonic complex [e.g. Figure IIIa(iii)], perceptually attend to the different notes in a chord or to one of a pair of synchronous pure tones that are far apart in frequency (see Figure IIb in Box 2). These percepts, however, are *not* examples of streams since they do not arise through any sequential processes or organization, and moreover, they fail *objective* psychoacoustic criteria of streaming percepts (discussed below). A simple example is the case of the simultaneous tone sequences discussed earlier (see Box 2), which, despite being readily heard as distinct tones, are nevertheless perceived as a *single* stream [44]. We claim that the same arguments apply to the sequences of mistuned harmonics illustrated in Fig*ure IIIa*(iii), the double-vowels illustrated in Fig*ure IIIb*, and the two directional sounds shown in Figure III*c*. In each of these cases, the distinct sound heard out of the complex mixture of sounds is nevertheless part of the same one stream because it produces coherent responses, the *fundamental* criterion for streaming.

A more complex example is the musical fragment illustrated in Figure III*d*, where we predict that the two opening bars are heard as a single, rich stream with all instruments playing in a temporally coherent fashion just like an orchestra playing in unison. In the subsequent bars, two streams diverge as the oboe and the violins play incoherently. Another example involves multiple talkers (see Figure IIIe), as might occur during a “cocktail party”. It is generally agreed that the segregation of simultaneous voices in this case is largely facilitated by the temporal “incoherence” of their syllabic segments which enables the listener to “glimpse” (or gather “snapshots”) of the target voice during “dips” in the other voice [88]. Viewed abstractly, the alternating bursts of the perceptually distinct green and pink speech patterns illustrated in Figure III*e* are analogous to the alternating tones illustrated in Figure IIa of Box 2.

The proposed distinction between hearing-out components in a complex *versus* streams raises the important question of how to objectively measure listeners’ perception of streams. A commonly used approach involves measuring listeners’ ability to detect (or discriminate) differences in the relative timing of sounds. It has been found that when listeners hear different sounds as belonging to separate streams (subjectively), they lose the ability to detect (or discriminate) small differences in the relative timing of those sounds [89-93,102]. For instance, they can no longer tell if one sound, which is perceived as part of one stream, starts before or after another sound, which is perceived as part of another stream. Other approaches involve the detection or discrimination of changes in some attribute (e.g. pitch) of sounds in one stream in the presence of irrelevant (e.g. random) changes in the same, or a different attribute, in another stream [94,95].

Box 4: Future directions

A number of questions regarding the perceptual organization of complex auditory scenes remain unresolved, ranging from neuronal mechanisms to behavior. Here, we highlight several of the key topics that are the subject of current and future investigations.

Neural circuitry of auditory scene analysis

- What are the neural underpinnings of streaming in non-primary auditory and non-sensory cortical areas?
- Is there explicit evidence for temporal-coherence computations carried out at some level of auditory cortical processing?
- What is the neural signature of emergence of auditory streams?

Role of attention and behavior

- What are the neural correlates of streaming in behaving animals?
- Do attention-induced neuronal changes at the level of the auditory cortex show a causal effect with improved behavioral performance during streaming tasks?
- How does attention modulate the binding of acoustic features into perceptual streams?

Scene analysis across modalities

- If confirmed by empirical evidence, does the principle of temporal coherence reveal a fundamental principle underlying scene analysis across sensory modalities?

Acknowledgments

This work was supported by the following grants to the authors: NIH R0107657, MURI N000141010278, AFOSR FA9550-09-1-0234 and NSF CAREER award IIS-0846112.

Glossary

Auditory scene analysis: The processes by which sequential and concurrent acoustic events are analyzed and organized into auditory streams.

Auditory stream: A series of sounds that is perceived by the listener as a coherent entity and, as such, can be selectively attended to amid other sounds. The word “stream” emphasizes the fact that sounds usually unfold over time. While sounds coming from different physical sound sources typically form separate streams, this is not always the case. For example, a choir singing in unison consists of multiple sources heard as a single stream, while an audio speaker is a single physical source that usually creates multiple streams. Several *objective* criteria exist by which one can determine if the series of sounds is perceived as a stream.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Coherence:	Temporal coherence between two channels is defined here in the following specific sense: It denotes the average similarity or coincidence of their responses measured over a given time-window. It is computed as the running cross-correlation coefficient *at zero-lag* between the channel responses integrated over relatively long time-windows (50-500 ms). Therefore channels with similar activity over this time interval are highly coherent (a correlation coefficient near 1) as with the synchronous tone-pairs in Figure II.b in Box 2. Anti-coherence therefore refers to the relationship between opposite or inverted responses (cross-correlation coefficient near −1) as with the alternating tones in Figure II.a in Box 2.
Complex tone:	A periodic sound that contains multiple frequencies.
Frequency:	The number of cycles per unit of time. It is usually expressed in cycles/s, or Hertz (Hz)
Fundamental frequency (F_0):	The inverse of the period of a harmonic complex tone. It is the highest frequency of which all other frequency components in a harmonic complex tone are integer multiples.
Harmonic:	A spectral component in a harmonic complex tone.
Noise:	Strictly speaking, an*aperiodic* sound. More broadly, it is any undesirable sound.
Pure-tone:	A tone that consists of a single frequency.
Spectrum:	A representation of the frequency content of a signal. It is usually obtained using the Fourier transform, and shows the amplitude and/or phase of the different frequency components in the signal.
Spectrogram:	A visual representation of the spectrum of a sound as a function of time. Time is usually shown along the abscissa, frequency along the ordinate, and the sound energy (or amplitude) at each time-frequency point is indicated using color, or shades of gray.
Sound Token:	Defined in this article as a burst of sound that rapidly evokes a percept. A token can be as simple as a pure tone, a harmonic complex, or a transient acoustic event like a click, or as complex as a vowel, a syllable, or a musical chord. It usually has one or more of the common attributes of sound such as a pitch, loudness, location, or timbre.
Streaming:	The process of forming segregated percepts of auditory sources. In the hearing-research literature, the words “streaming” and “stream formations” are most often reserved to describe sequential grouping or organization of sound segments or tokens over time. In this article, we exclusively employ “streaming” in this sense. There are both subjective and objective criteria to determine whether a stream is perceived or not, although these are not universally agreed upon.
Synchronous stimuli:	Stimuli that always have a common onset in time when they co-occur.
Tone:	A periodic sound.

Tone or token sequence:	Refers to a sequence of sound elements that occur at relatively slow rates (< 20 Hz). Examples are experimental sequences of pure tones, notes of a musical melody, or syllables in running speech.
Tuning curve:	Usually refers to the auditory neuron's selectivity to acoustic frequencies, often measured using pure tones. It is analogous to the receptive field of a visual neuron.

References

1. Bregman, AS. Auditory scene analysis: the perceptual organization of sound. MIT Press; 1990.
2. Cherry EC. Some experiments on the recognition of speech, with one and two ears. J. Acoust. Soc. Am. 1953; 25:975–979.
3. Bee MA, Micheyl C. The cocktail party problem: what is it? How can it be solved? And why should animal behaviorists study it? J. Comp. Psychol. 2008; 122:235–251. [PubMed: 18729652]
4. McDermott JH. The cocktail party problem. Curr. Biol. 2009; 19:R1024–1027. [PubMed: 19948136]
5. Attias, H.; Schreiner, CE. Advances in Neural Information Processing Systems. MIT Press; 1997. Temporal low-order statistics of natural sounds.
6. Singh NC, Theunissen FE. Modulation spectra of natural sounds and ethological theories of auditory processing. J. Acoust. Soc. Am. 2003; 114:3394–3411. [PubMed: 14714819]
7. Moore BCJ, Gockel H. Factors influencing sequential stream segregation. Acta Acustica. 2002; 88:320–333.
8. Fay, RR. Auditory Perception of Sound Sources. Springer; 2008. Sound source perception and stream segregation in nonhuman verterbrate animals.
9. Marrone N, et al. Evaluating the benefit of hearing aids in solving the cocktail party problem. Trends Amplif. 2008; 12:300–315. [PubMed: 19010794]
10. Wang, D.; Brown, GJ. Computational Auditory Scene Analysis: Principles, Algorithms, and Applications. Wiley-IEEE Press; 2006.
11. Nelken I. Processing of complex stimuli and natural scenes in the auditory cortex. Curr. Opin. Neurobiol. 2004; 14:474–480. [PubMed: 15321068]
12. Sinex DG. Spectral processing and sound source determination. Int. Rev. Neurobiol. 2005; 70:371–398. [PubMed: 16472640]
13. Alain C. Breaking the wave: effects of attention and learning on concurrent sound perception. Hear. Res. 2007; 229:225–236. [PubMed: 17303355]
14. Micheyl C, Oxenham AJ. Pitch, harmonicity and concurrent sound segregation: Psychoacoustical and neurophysiological findings. Hearing Res. 2010; 266:36–51.
15. Carlyon RP. How the brain separates sounds. Trends Cogn. Sci. 2004; 8:465–471. [PubMed: 15450511]
16. Micheyl C, et al. The role of auditory cortex in the formation of auditory streams. Hear. Res. 2007; 229:116–131. [PubMed: 17307315]
17. Snyder JS, Alain C. Toward a neurophysiological theory of auditory stream segregation. Psychol. Bull. 2009; 133:780–799. [PubMed: 17723030]
18. Bidet-Caulet A, Bertrand O. Neurophysiological mechanisms involved in auditory perceptual organization. Front Neurosci. 2009; 3:182–191. [PubMed: 20011140]
19. Winkler I, et al. Modeling the auditory scene: predictive regularity representations and perceptual objects. Trends Cogn, Sci. 2009; 13:532–540. [PubMed: 19828357]
20. Shamma S, Micheyl C. Behind the scenes of auditory perception. Curr. Opin. Neurobiol. 2010; 20:361–366. [PubMed: 20456940]
21. Hartmann W, Johnson D. Stream segregation and peripheral channeling. Mus. Percep. 1991; 9:155–184.

22. Beauvois MW, Meddis R. Computer simulation of auditory stream segregation in alternating-tone sequences. J. Acoust. Soc. Am. 1996; 99:2270–2080. [PubMed: 8730073]
23. McCabe S, Denham MJ. A model of auditory streaming. J. Acoust. Soc. Am. 1997; 101:1611–1621.
24. Fishman YI, et al. Neural correlates of auditory stream segregation in primary auditory cortex of the awake monkey. Hear. Res. 2001; 151:167–187. [PubMed: 11124464]
25. Fishman YI, et al. Auditory stream segregation in monkey auditory cortex: effects of frequency separation, presentation rate, and tone duration. J. Acoust. Soc. Am. 2004; 116:1656–1670. [PubMed: 15478432]
26. Kanwal JS, et al. Neurodynamics for auditory stream segregation: tracking sounds in the mustached bat's natural environment. Network. 2003; 14:413–435. [PubMed: 12938765]
27. Bee MA, Klump GM. Auditory stream segregation in the songbird forebrain: effects of time intervals on responses to interleaved tone sequences. Brain Behav. Evol. 2005; 66:197–214. [PubMed: 16127270]
28. Bee MA, Klump GM. Primitive auditory stream segregation: a neurophysiological study in the songbird forebrain. J. Neurophysiol. 2004; 92:1088–1104. [PubMed: 15044521]
29. Micheyl C, et al. Perceptual organization of tone sequences in the auditory cortex of awake macaques. Neuron. 2005; 48:139–48. [PubMed: 16202714]
30. Pressnitzer D, et al. Perceptual organization of sound begins in the auditory periphery. Curr. Biol. 2008; 18:1124–1128. [PubMed: 18656355]
31. Grimault N, et al. Auditory stream segregation on the basis of amplitude-modulation rate. J. Acoust. Soc. Am. 2002; 111:1340–1348. [PubMed: 11931311]
32. Schreiner CE, Sutter ML. Topography of excitatory bandwidth in cat primary auditory cortex: single-neuron versus multiple-neuron recordings. J. Neurophysiol. 1992; 68:1487–1502. [PubMed: 1479426]
33. Schreiner CE. Order and disorder in auditory cortical maps. Curr. Opin. Neurobiol. 1994; 5:489–496. [PubMed: 7488851]
34. Versnel H, et al. Ripple Analysis in the Ferret Primary Auditory Cortex. III. Topographic and Columnar Distribution of Ripple Response Parameters. Aud. Neurosci. 1995; 1:271–286.
35. Kowalski N, et al. Analysis of dynamic spectra in ferret primary auditory cortex. I. Characteristics of single-unit responses to moving ripple spectra. J. Neurophysiol. 1996; 76:3503–3523. [PubMed: 8930289]
36. Kowalski N, et al. Analysis of dynamic spectra in ferret primary auditory cortex. II. Prediction of unit responses to arbitrary dynamic spectra. J. Neurophysiol. 1996; 76:3524–3534. [PubMed: 8930290]
37. Schreiner CE, et al. Temporal processing in cat primary auditory cortex. Acta Otolaryngol. Suppl. 1997; 532:54–60. [PubMed: 9442845]
38. Schreiner CE. Spatial distribution of responses to simple and complex sounds in the primary auditory cortex. Audiol. Neurootol. 1998; 3:104–122. [PubMed: 9575380]
39. Miller LM, et al. Spectrotemporal receptive fields in the lemniscal auditory thalamus and cortex. J. Neurophysiol. 2002; 87:516–527. [PubMed: 11784767]
40. Sutter ML. Spectral processing in the auditory cortex. Int. Rev. Neurobiol. 2005; 70:253–298. [PubMed: 16472637]
41. Middlebrooks JC, et al. Binaural response-specific bands in primary auditory cortex (AI) of the cat: topographical organization orthogonal to isofrequency contours. Brain Res. 1980; 181:31–48. [PubMed: 7350963]
42. Mrsic-Flogel TD, et al. Encoding of virtual acoustic space stimuli by neurons in ferret primary auditory cortex. J. Neurophysiol. 2005; 93:3489–3503. [PubMed: 15659534]
43. Bendor D, Wang X. The neuronal representation of pitch in primate auditory cortex. Nature. 2005; 436:1161–1165. [PubMed: 16121182]
44. Elhilali M, et al. Temporal coherence in the perceptual organization and cortical representation of auditory scenes. Neuron. 2009; 61:317–329. [PubMed: 19186172]

45. Sheft, S. Envelope processing and sound-source perception. In: Yost, WA.; Fay, RR.; Popper, AN., editors. Auditory perception of sound sources. Springer; 2007. p. 233-280.
46. Pienkowski M, Eggermont JJ. Nonlinear cross-frequency interactions in primary auditory cortex spectrotemporal receptive fields: a Wiener-Volterra analysis. J. Comput. Neurosci. 2010; 28:285–303. [PubMed: 20072806]
47. Atencio CA, et al. Hierarchical computation in the canonical auditory cortical circuit. Proc. Natl. Acad. Sci. U.S.A. 2009; 106:21894–21899. [PubMed: 19918079]
48. Bizley JK, et al. Interdependent encoding of pitch, timbre, and spatial location in auditory cortex. J. Neurosci. 2009; 29:2064–2075. [PubMed: 19228960]
49. Luczak A, et al. Multivariate receptive field mapping in marmoset auditory cortex. J. Neurosci. Methods. 2004; 136:77–85. [PubMed: 15126048]
50. Sadagopan S, Wang X. Nonlinear spectrotemporal interactions underlying selectivity for complex sounds in auditory cortex. J. Neurosci. 2009; 29:11192–11202. [PubMed: 19741126]
51. Blake R, Lee SH. The role of temporal structure in human vision. Behav. Cogn. Neurosci. Rev. 2005; 4:21–42. [PubMed: 15886401]
52. Alais D, et al. Visual features that vary together over time group together over space. Nat. Neurosci. 1998; 1:160–164. [PubMed: 10195133]
53. Rahne T, et al. A multilevel and cross-modal approach towards neuronal mechanisms of auditory streaming. Brain Res. 2008; 1220:118–131. [PubMed: 17765207]
54. Eggermont, JJ. The correlative brain: Theory and experiment in neural interaction. Springer; 1990.
55. von der Malsburg, C. Models of Neural Networks. Springer; 1994. The correlation theory of brain function.
56. von der Malsburg C, Schneider W. A neural cocktail-party processor. Biol. Cybern. 1986; 54:29–40. [PubMed: 3719028]
57. Wang DL, Brown GJ. Separation of speech from interfering sounds based on oscillatory correlation. IEEE Trans. Neural Netw. 1999; 10:684–697. [PubMed: 18252568]
58. Lakatos P, et al. An oscillatory hierarchy controlling neuronal excitability and stimulus processing in the auditory cortex. J. Neurophysiol. 2005; 94:1904–1911. [PubMed: 15901760]
59. Elhilali M, Shamma SA. A cocktail party with a cortical twist: how cortical mechanisms contribute to sound segregation. J. Acoust. Soc. Am. 2008; 124:3751–3771. [PubMed: 19206802]
60. Sussman E, et al. An investigation of the auditory streaming effect using event-related brain potentials. Psychophysiol. 1999; 36:22–34.
61. Sussman ES, et al. The role of attention in the formation of auditory streams. Percept. & Psychophys. 2007; 69:136–152.
62. Alain C, Arnott SR. Selectively attending to auditory objects. Front. Biosci. 2000; 5:D202–212. [PubMed: 10702369]
63. Alain C, et al. Neural activity associated with distinguishing concurrent auditory objects. J. Acoust. Soc. Am. 2002; 111:990–995. [PubMed: 11863201]
64. Dyson BJ, Alain C. Representation of concurrent acoustic objects in primary auditory cortex. J. Acoust. Soc. Am. 2004; 115:280–288. [PubMed: 14759021]
65. Snyder JS, Alain C. Age-related changes in neural activity associated with concurrent vowel segregation. Brain Res. Cogn. Brain Res. 2005; 24:492–429. [PubMed: 16099361]
66. Hillyard SA, et al. Electrical signs of selective attention in the human brain. Science. 1973; 182:177–180. [PubMed: 4730062]
67. Tiitinen H, et al. Selective Attention Enhances the Auditory 40-Hz Transient-Response in Humans. Nature. 1993; 364:59–60. [PubMed: 8316297]
68. Woldorff MG, et al. Modulation of early sensory processing in human auditory cortex during auditory selective attention. Proc. Natl. Acad. Sci. U. S. A. 1993; 90:8722–8726. [PubMed: 8378354]
69. Bidet-Caulet A, et al. Effects of selective attention on the electrophysiological representation of concurrent sounds in the human auditory cortex. J. Neurosci. 2007; 27:9252–9261. [PubMed: 17728439]

70. Elhilali M, et al. Interaction between attention and bottom-up saliency mediates the representation of foreground and background in an auditory scene. PLoS Biol. 2009; 7:e1000129. [PubMed: 19529760]
71. Shinn-Cunningham BG. Object-based auditory and visual attention. Trends Cogn. Sci. 2008; 12:182–186. [PubMed: 18396091]
72. Duncan J. Selective attention and the organization of visual information. J. Exp. Psychol. Gen. 1984; 113:501–517. [PubMed: 6240521]
73. Desimone R, Duncan J. Neural mechanisms of selective visual attention. Annu. Rev. Neurosci. 1995; 18:193–222. [PubMed: 7605061]
74. van Noorden, LP. Temporal coherence in the perception of tone sequences. Eindhoven University of Technology; 1975.
75. Cusack R, et al. Effects of location, frequency region, and time course of selective attention on auditory scene analysis. J. Exp. Psychol. Hum. Percept. Perform. 2004; 30:643–656. [PubMed: 15301615]
76. Carlyon RP, et al. Effects of attention and unilateral neglect on auditory stream segregation. J. Exp. Psychol. Hum. Percept. Perform. 2001; 27:115–127. [PubMed: 11248927]
77. Fritz JB, et al. Auditory attention--focusing the searchlight on sound. Curr. Opin. Neurobiol. 2007; 17:437–455. [PubMed: 17714933]
78. Yin, P., et al. Hearing: from sensory processing to perception. Springer; 2007. Primary auditory cortical responses while attending to different streams; p. 257-266.
79. Niebur E, et al. Synchrony: a neuronal mechanism for attentional selection? Curr. Opin. Neurobiol. 2002; 12:190–194. [PubMed: 12015236]
80. Kim YJ, et al. Attention induces synchronization-based response gain in steady-state visual evoked potentials. Nat. Neurosci. 2007; 10:117–125. [PubMed: 17173045]
81. Steinmetz PN, et al. Attention modulates synchronized neuronal firing in primate somatosensory cortex. Nature. 2000; 404:187–190. [PubMed: 10724171]
82. Watanabe T, et al. Perceptual learning without perception. Nature. 2001; 413:844–848. [PubMed: 11677607]
83. Suied C, et al. Why are natural sounds detected faster than pips? J. Acous. Soc. Am. Exp. Lett. 2010; 127:EL105–110.
84. VanRullen R, et al. Spike times make sense. Trends Neurosci. 2005; 28:1–4. [PubMed: 15626490]
85. Moore, BCJ. An Introduction to the Psychology of Hearing. Academic Press; 2003.
86. DiCarlo JJ, Cox DD. Untangling invariant object recognition. Trends Cogn. Sci. 2007; 11:333–341. [PubMed: 17631409]
87. Carlyon, RP.; Gockel, H. Effects of harmonicity and regularity on the perception of sound sources. In: Yost, WA.; Fay, RR.; Popper, AN., editors. Auditory Perception of Sound Sources. Springer; 2008. p. 191-214.
88. Qin MK, Oxenham AJ. Effects of simulated cochlear-implant processing on speech reception in fluctuating maskers. J. Acoust. Soc. Am. 2003; 114:446–454. [PubMed: 12880055]
89. Bregman AS, Rudnicky AI. Auditory segregation: stream or streams? J. Exp. Psychol. Hum. Percept. Perform. 1975; 1:263–267. [PubMed: 1202149]
90. Vliegen J, et al. The role of spectral and periodicity cues in auditory stream segregation, measured using a temporal discrimination task. J. Acoust. Soc. Am. 1999; 106:938–945. [PubMed: 10462799]
91. Roberts B, et al. Primitive stream segregation of tone sequences without differences in fundamental frequency or passband. J. Acoust. Soc. Am. 2002; 112:2074–2085. [PubMed: 12430819]
92. Roberts B, et al. Effects of the build-up and resetting of auditory stream segregation on temporal discrimination. J. Exp. Psychol. Hum. Percept. Perform. 2008; 34:992–1006. [PubMed: 18665740]
93. Micheyl C, et al. Auditory stream segregation and the perception of across-frequency synchrony. J. Exp. Psychol. Hum. Percep. Perf. 2010; 36:1029–1039.

94. Micheyl, C., et al. Performance measures of auditory organization. In: Pressnitzer, D.; se Cheveigne, A.; McAdams, S.; Collet, L., editors. Auditory Signal Processing: Physiology, Psychoacoustics, and Models. Springer; 2005. p. 203-211.

95. Micheyl, C., et al. Hearing out repeating elements in randomly varying multitone sequences: a case of streaming?. In: Kollmeier, B.; Klump, G.; Hohmann, V.; Langemann, U.; Mauermann, M.; Uppenkamp, S.; Verhey, J., editors. Hearing - From Basic Research to Applications. Springer; 2007.

96. Darwin CJ, et al. Grouping in pitch perception: evidence for sequential constraints. J. Acoust. Soc. Am. 1995; 98:880–885. [PubMed: 7642826]

97. Kidd G Jr. et al. Reducing informational masking by sound segregation. J. Acoust. Soc. Am. 1994; 95:3475–3480. [PubMed: 8046139]

98. Roye A, Schröger E, Jacobsen T, Gruber T. Is My Mobile Ringing? Evidence for Rapid Processing of a Personally Significant Sound in Humans. The Journal of Neuroscience. 2010; 30(21):7310–13. 26, 2010. [PubMed: 20505097]

99. Kriegstein K, Giraud A. Implicit Multisensory Associations Influence Voice Recognition. PLoS Biol. 2010; 4(10):e326.

100. Best V, Ozmeral EJ, Kopco N, Shinn-Cunningham BG. Object continuity enhances selective auditory attention. Proc Natl Acad Sci U S A. Sep 2; 2008 105(35):13174–8. 2008. [PubMed: 18719099]

101. Shinn-Cunningham BG, Lee AK, Oxenham AJ. A sound element gets lost in perceptual competition. Proceedings of the National Academy of Sciences of the United States of America. 2007; 104:12223–12227. [PubMed: 17615235]

102. Boehnke S, Phillips D. The relation between auditory temporal interval processing and sequential stream segregation examined with stimulus laterality differences. J Perception & Psychophysics. 2005; 67(6):1088–1101.

103. Lu T, Liang L, Wang X. Temporal and rate representations of time-varying signals in the auditory cortex of awake primates. Nature neuroscience. 2001; 4(11):1131–38.

104. Winer, JA.; Schreiner, CE., editors. The Auditory Cortex. Springer; 2009.

105. Jiang D, Palmer AR, Winter IM. Frequency extent of two-tone facilitation in onset units in the ventral cochlear nucleus. J. Neurophysiol. 1996; 75(1):380–395. [PubMed: 8822565]

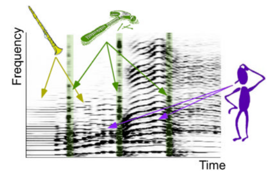


Figure 1.

A spectrogram of a complex scene with multiple objects.The figure shows a time-frequency analysis of an acoustic recording of a scene consisting of flute, a human voice and a hammer. The hammer hits are immediately visible as repetitive and transient broadband strips of energy spanning all frequencies. Both the flute and the human voice contain a rich harmonic structure that changes over time. The human voice reveals clear pitch variations and formant transitions, shown as time-course changes in both the pitch and formant locations. Note that the flute and speech give rise to clearly distinct acoustic events that are uncorrelated in time.

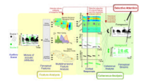


Figure 2.
Schematic of the proposed model of auditory stream formation. From left to right: Multiple sound sources constitute an auditory scene, which is initially analyzed through a feature-analysis stage. This stage consists of a cochlear frequency analysis followed by arrays of feature-selective neurons that create a multi-dimensional representation along different feature axes. The figure depicts timbre, pitch and spatial location channels. Note that for computational convenience and illustration purposes, these feature maps are shown with ordered axes when in fact such orderly representations are neither known nor are essential for the model. The outcome of this analysis is a rich set of cortical responses that explicitly represent the different sound features, as well as their timing relationships. The second stage of the model performs coherence analysis by correlating the temporal outputs of the different feature-selective neurons, and arranging them based on their degree of coherence; hence giving rise to distinct perceptual streams. Complementing this feed-forward bottom-up view are top-down processes of selective attention that operate by modulating the selectivity of cortical neurons. This feature-based selective attention translates onto object-based attentional mechanisms by virtue of the fact that selected features are coherent with other features that are part of the same stream.

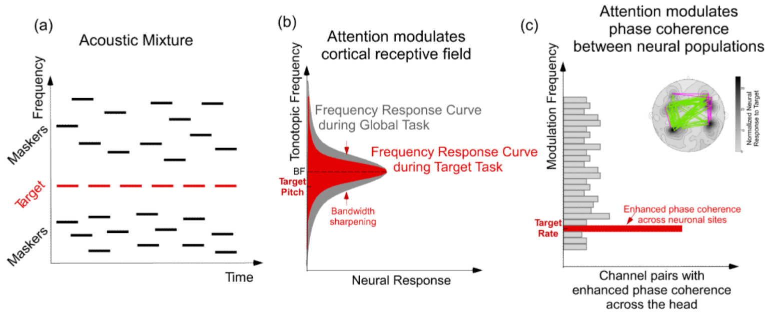
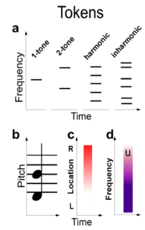


Figure 3.

Schematic of the influence of attention on the cortical selectivity of sound features and the representation of coherent features of an attended stream. **(a)** A schematic of the time-frequency distribution of an acoustic mixture with a regularly repeating tone sequence (target) amidst a background of random tones (maskers). The perception of the target depends critically on a number of parameters, including the frequency separation between the target and closest masker components, the repetition rate of the target, and the overall sequence duration. **(b)** An illustration of the frequency-response curve of a single-unit recorded in the primary auditory cortex of a behaving ferret and the changes that are observed under two different behavioral tasks. When the animal attends to the repeating target tone (“Target task” - red curve), the receptive field tuned to the target frequency sharpens in a direction that enhances the segregation of the target from the background of the maskers. When the animal performs a listening task that involves attending to the entire sound mixture (“Global task” - grey curve), the tuning curve shows a much broader tuning curve relative to the selective attention state (adapted from [78]). **(c)** The phase coherence between distinct neural populations as measured by distributed MEG (magnetoencephalography) channels recording neural activity in human subjects. The phase coherence contrasts a selective-attention task (where the subjects attended to the repeating target tone) versus a global-attention task (where the subjects paid attention to the background maskers). Such recordings reveal that an enhancement in phase coherence occurs exclusively at the attended target repetition rate (in this case 4Hz) (adapted from [70]). The inset represents an example of the MEG magnetic field distribution for a single listener, illustrating that the MEG channel pairs with robust phase coherence in response to the rate of the target tone sequence . Channel pairs with enhanced phase coherence are shown in green, while channel pairs with reduced coherence are shown in pink.

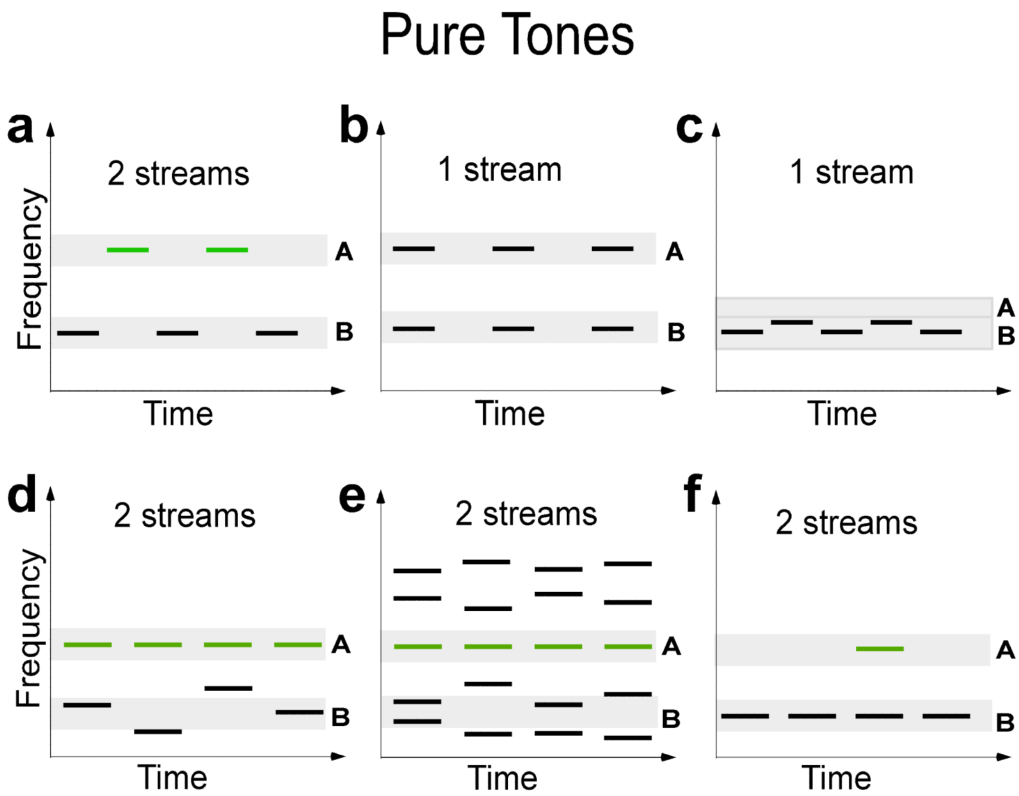


Box 1 Figure I. Principles and examples of auditory streaming: Instantaneous percepts (Tokens)

Examples of acoustic tokens with different attributes are illustrated. **(a)** <u>Spectra of tonal tokens:</u> single, 2-tones**,** harmonic complex, and an inharmonic complex. Tokens are relatively brief and its constituents have a common onset. **(b-d)** <u>Complex tokens.</u> Sound tokens can have various attributes such as (b) the pitch of musical notes or chords, (c) location along the azimuth, or (d) the timbre of a vowel with a specific spectral shape (right panel). In each of these panels, the feature value is represented by the pattern of activation along the ordinate. For example, each note in (b) represents the place of activation along the low-to-high pitch values; the activation pattern in (c) has a peak on the *Right* along the Left-to-Right ordinate; the vowel in (d) is represented by its spectral shape along the frequency axis. All these features occur over a brief time interval.

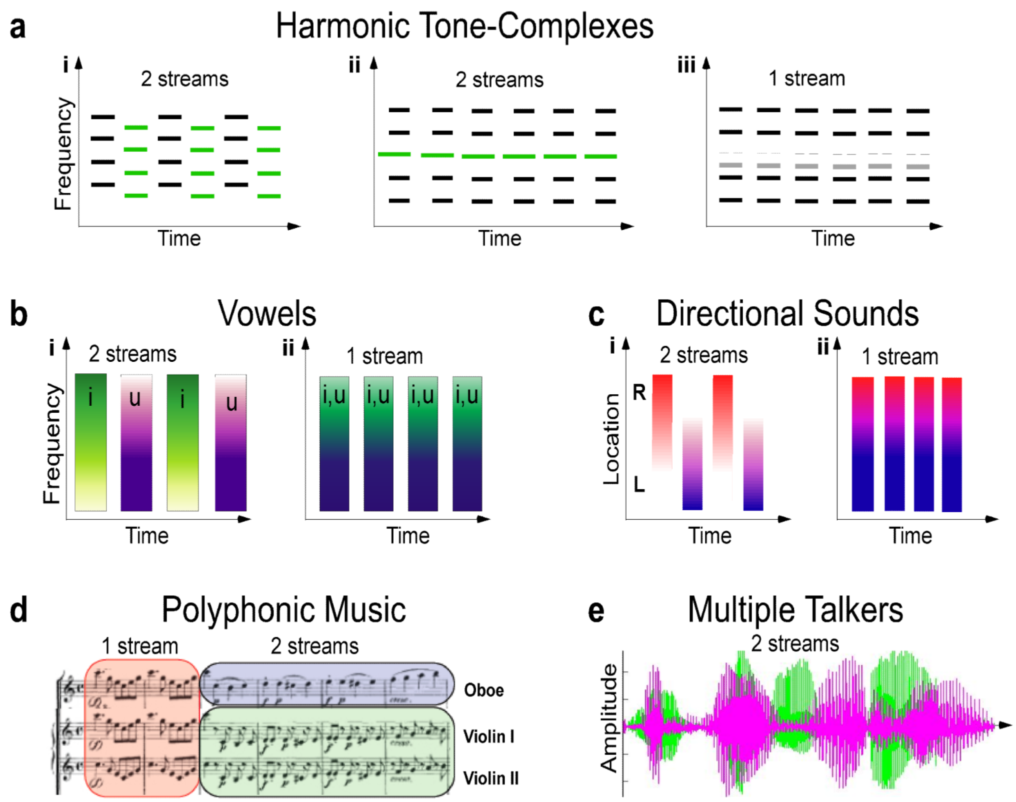
NIH-PA Author Manuscript

NIH-PA Author Manuscript



Box 2 Figure II. Streaming with pure tones

Examples of sequential organization of pure-tone sequences: **(*a*)** Two alternating tones of widely separated frequencies are usually perceived as two separate streams. The green color indicates a separate stream. The shaded regions denote two hypothetical neural auditory channels activated by the tones. The A,B channels are incoherent. **(b)** Two synchronous sequences are perceived as a single stream because the A,B channels are coherent. **(c)** Alternating (asynchronous) tones of nearby frequencies are usually heard as a single perceptual stream that oscillates in frequency regardless of tone presentation rates. The A,B channels here overlap and hence are driven by both tones and carry coherent responses. **(d)** Two synchronous tone sequences of fixed and variable frequencies. Two streams are predicted since the coherence between the A,B channels is weak. **(e) "**Release from informational masking" stimulus: when a target tone sequence is embedded in masker tones (surrounded by an empty or a protected zone), it evokes responses in channel A that are incoherent with channel B, and hence be heard streamed from the complex. **(f)** Capture and streaming of a simultaneous tone pair. A pair of simultaneous tones is normally heard as a single complex sound when presented in isolation. However, a preceding sequence of low tones (as illustrated in channel B) can perceptually "capture" the low tone, separating it from the high tone (illustrated in channel A), which is now heard clearly against the background of the low-tones.



Box 3 Figure III. Streaming with complex sounds

Principles of sequential organization apply equally well to complex stimuli that evoke responses in feature-selective channels (analogous to the frequency-tuned channels for tones). Examples illustrated are: **(a) Streaming with harmonic complexes**. Harmonic complexes are perceived usually as a fused sound with a pitch at the frequency of the fundamental (bottom) component of each complex. ***(i)*** Two alternating complexes (green and black) stream apart just like alternating pure tones [14]. ***(ii)*** A harmonic complex is perceptually fractured when one component begins earlier (e.g. the green harmonic). Because of its temporal incoherence, this component forms a separate stream from the rest of the complex (the black tones). ***(iii)*** A harmonic complex also becomes perceptually fractured when one component (the grey tone) is mistuned from a harmonic relationship and pops out from the complex. However, in this case, the two percepts within the token continue to belong to a single-stream as they remain temporally coherent. **(b) Streaming of vowels.** A sequence of vowel pairs is perceived either as two streams or one depending on the temporal coherence of the vowels; ***(i)*** The alternating pair of vowels, /i/ and /u/, are represented schematically by different spectra. These vowels (just like the alternating tones) segregate into two streams [3,15,17]; ***(ii)*** as with the synchronous tones, when the vowels are played simultaneously they may still be individually recognized but are nevertheless heard as a *single* stream. **(c) Streaming of sounds from different locations.** Two sounds from the left (L) and right (R) stream apart when ***(i)*** they are played alternately [15,102], but form a *single* stream when ***(ii)*** played coherently. In the latter case, we predict that the sound is heard as a single stream from (indeterminate) multiple locations. **d. Streaming of musical instruments.** The beginning of Mozart's Concerto K299 is illustrated here. The first two bars are heard as a *single* rich stream as all instruments are playing coherently despite the distinct timbres of the *oboe and* the *violin*, and the different notes (pitches) played by the two violins. In the subsequent bars, two streams diverge as the oboe and the violins play incoherently. **e. Streaming of two simultaneous talkers.** When the waveforms from two

NIH-PA Author Manuscript

NIH-PA Author Manuscript

different spoken sentences (represented by pink and green) are overlaid, they often appear as alternating sound tokens. This incoherence between the two waveforms (each with its own distinct timbre, pitch, or even location) facilitates their streaming apart. In a choir singing in unison, the waveforms from the all singers would completely overlap and hence are heard as one rich stream (analogous to a piano playing a sequence of chords).