Neuron
Article

Temporal Coherence in the Perceptual Organization and Cortical Representation of Auditory Scenes

Mounya Elhilali,[1,5,*] Ling Ma,[2,5] Christophe Micheyl,[3,5] Andrew J. Oxenham,[3] and Shihab A. Shamma[2,4]
[1]Department of Electrical and Computer Engineering, Johns Hopkins University, Barton Hall 105, 3400 North Charles Street, Baltimore, MD 21218, USA
[2]Department of Bioengineering, University of Maryland, College Park, MD 20742, USA
[3]Department of Psychology, University of Minnesota, Minneapolis, MN 55455, USA
[4]Department of Electrical and Computer Engineering, University of Maryland, College Park, MD 20742, USA
[5]These authors contributed equally to this work
*Correspondence: mounya@jhu.edu
DOI 10.1016/j.neuron.2008.12.005

SUMMARY

Just as the visual system parses complex scenes into identifiable objects, the auditory system must organize sound elements scattered in frequency and time into coherent "streams." Current neurocomputational theories of auditory streaming rely on tonotopic organization of the auditory system to explain the observation that sequential spectrally distant sound elements tend to form separate perceptual streams. Here, we show that spectral components that are well separated in frequency are no longer heard as separate streams if presented synchronously rather than consecutively. In contrast, responses from neurons in primary auditory cortex of ferrets show that both synchronous and asynchronous tone sequences produce comparably segregated responses along the tonotopic axis. The results argue against tonotopic separation per se as a neural correlate of stream segregation. Instead we propose a computational model of stream segregation that can account for the data by using temporal coherence as the primary criterion for predicting stream formation.

INTRODUCTION

When listening to someone at a crowded cocktail party, or trying to follow the second violin line in a symphonic orchestra, we rely on our ears' and brain's extraordinary ability to parse complex acoustic scenes into individual auditory "objects" or "streams" (Griffiths and Warren, 2004). Just as the decomposition of a visual scene into objects is a challenging and mathematically ill-posed problem, requiring top-down and bottom-up information to solve (Marr, 1983; Zeki, 1993), the auditory system uses a combination of acoustic cues and prior experience to analyze the auditory scene. A simple example of auditory streaming (Bregman, 1990; Carlyon, 2004) can be demonstrated and explored in the laboratory using sound sequences like those illustrated in Figure 1. These sequences are produced by presenting two tones of different frequencies, A and B, repeatedly (Figure 1A). Many psychophysical studies have shown that this simple stimulus can evoke two very different percepts, depending on the frequency separation, ΔF, between the A and B tones, and the time interval, ΔT, between successive tones (for a review, see Bregman, 1990). In particular, when ΔF is relatively small (<10%), most listeners perceive and describe the stimulus as a single stream of tones alternating in frequency, like a musical trill. However, when ΔF is large, the percept is that of two parallel but separate streams, each containing only tones of the same frequency (A-A- and B-B-; see Supplemental Data available online for an auditory demonstration). The perceptual separation of sound components into distinct streams is usually referred to as stream segregation; the converse process is variously known as stream integration, grouping, or fusion. Manifestations of auditory streaming have been observed in various nonhuman species, including birds, fish, and monkeys, suggesting that streaming is a fundamental aspect of auditory perception, which plays a role in adaptation to diverse ecological environments (Bee and Micheyl, 2008; Fay, 1998, 2000; Hulse et al., 1997; Izumi, 2002; MacDougall-Shackleton et al., 1998).

Inspired by the observation that frequency-to-place mapping, or tonotopy, is a guiding anatomical and functional principle throughout the auditory system (Eggermont, 2001; Pickles, 1988), current models of auditory streaming rely primarily on frequency separation for sound segregation (Beauvois and Meddis, 1991, 1996; Hartmann and Johnson, 1991; McCabe and Denham, 1997). These models predict that consecutive sounds will be grouped perceptually into a single auditory stream if they activate strongly overlapping tonotopic channels in the auditory system. In contrast, sounds that have widely different spectra will activate weakly overlapping (or nonoverlapping) channels, and be perceptually segregated (i.e., heard as separate streams). In this way, models based on tonotopic separation can account for behavioral findings that show an increase in perceived segregation with increasing frequency separation (Hartmann and Johnson, 1991). By additionally taking into account neural adaptation and forward suppression of responses to consecutive tones, these models can also account for the influence of temporal stimulus parameters, such as the intertone interval or the time since sequence onset, on auditory streaming (Beauvois and Meddis, 1991, 1996; Bee and Klump,

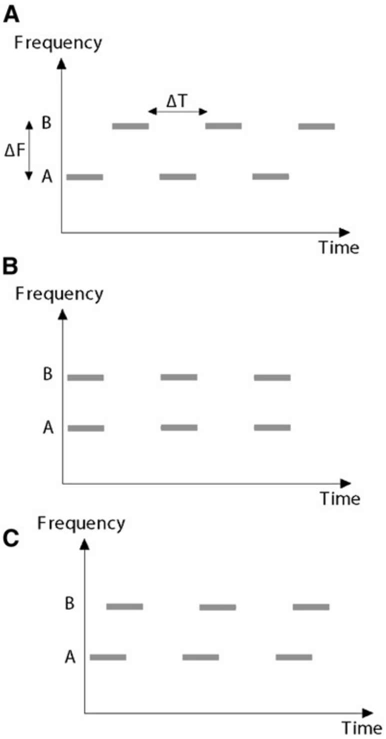

Figure 1. Schematic Spectrograms of Stimuli Used to Study the Perceptual Formation of Auditory Streams

(A) The typical stimulus used in many psychophysical and physiological studies of auditory streaming; a sequence of tones alternating between two frequencies, A and B. The percept evoked by such sequences depends primarily on the frequency separation between the A and B tones, ΔF, and on the intertone interval, ΔT. For small ΔFs and relatively long ΔTs, the percept is that of a single stream of tones alternating in pitch (ABAB); for large ΔFs and relatively short ΔTs, the percept is that of two separate streams of tones of constant pitch (A-A and B-B).

(B) A variation on the traditional stimulus, used in this study. Here, the A and B tones are synchronous, rather than alternating. Such sequences usually evoke the percept of a single stream, regardless of ΔF and ΔT.

(C) An alternating sequence of tones that is partially overlapped (40 ms onset asynchrony or about 50% overlap). This sequence is usually heard like the nonoverlapping tone sequence (see panel [A]).

2004, 2005; Fishman et al., 2001, 2004; Kanwal et al., 2003; McCabe and Denham, 1997; Micheyl et al., 2005, 2007; Pressnitzer et al., 2008).

Although tonotopic separation is important, it is clearly not the only determinant of auditory perceptual organization. Another factor is the relative timing of sounds. Sounds that start and end at the same time are more likely to be perceived as a single event than sounds whose onsets and offsets are staggered by several tens or hundreds of milliseconds (Darwin and Carlyon, 1995). Accordingly, if the AB tone pairs were presented synchronously (as in Figure 1B) instead of sequentially (as in Figure 1A), they might form a single perceptual stream, even at large frequency separations. This prediction poses a serious problem for purely tonotopic models of auditory streaming. Unfortunately, nearly all perceptual studies of auditory streaming so far have used strictly sequential, temporally nonoverlapping, stimuli (Figure 1A), although one informal description of an experiment involving partially overlapping stimuli exists (Bregman, 1990, page 213). On the physiological side, it is unclear how synchrony affects neural responses in the primary auditory cortex (AI), where previous studies have identified potential neural correlates of auditory streaming using purely nonoverlapping stimuli (Fishman et al., 2001, 2004; Gutschalk et al., 2005; Kanwal et al., 2003; Micheyl et al., 2005, 2007; Snyder et al., 2006; Wilson et al., 2007). The complexity of auditory cortical responses makes it difficult to predict how responses of single AI units will be influenced by stimulus synchrony: depending on the position of the tones relative to the unit's best frequency, responses might be facilitated (i.e., enhanced), inhibited (i.e., reduced), or left unchanged by the synchronous presentation of a second tone within the unit's excitatory receptive field.

Here we use a combination of psychophysics in humans, cortical physiology in ferret, and computational modeling to address these questions. We first report psychoacoustic findings, which reveal that synchronous and nonsynchronous sound sequences are perceived very differently, with synchronous tone sequences heard as a single stream, even at very large frequency separations. We then present physiological findings that show synchronous and nonsynchronous tone sequences evoke very similar tonotopic activation patterns in AI. Together, these findings challenge the current view that tonotopic separation in AI is necessary and sufficient for perceptual stream segregation. Finally, we describe a computational model of stream segregation that uses the temporal coherence of responses across tonotopic (or other) neural channels to predict perception, and demonstrate that this model can account for the present and other psychophysical findings. By combining simultaneous and sequential perceptual organization principles that have traditionally been studied separately, the model proposed here provides a new and more general account of auditory perceptual organization of any arbitrary sound combinations. More generally, the present findings suggest that the principle of grouping information across sensory channels based on temporal coherence may play a key role in auditory perceptual organization, just as has been proposed for visual scene analysis (Blake and Lee, 2005).

RESULTS

Psychophysics

Experimental Results

Informal listening to sound sequences like those illustrated in Figure 1A reveals that, for relatively large frequency separations between the A and B tones (e.g., six semitones or more), the alternating-tone sequence (Figure 1A) usually evokes a percept of two separate streams, each with a constant pitch (A-A- and B-B-). In contrast, the sequence of synchronous tones (Figure 1B) evokes the percept of a single stream, even at large Δ*F*s. To confirm and quantify this subjective impression, we asked listeners to discriminate between two sequences similar to those shown in Figure 1B, except that in one of those two sequences, the last B tone was slightly shifted temporally (forward or backward) so that it was no longer exactly synchronous with the corresponding A tone. We measured the smallest temporal shift that listeners could correctly detect 79.4% of the time. Based on earlier results indicating that listeners can detect onset shifts of as little as a few milliseconds between spectral components of complex tones (Zera and Green, 1993a, 1993b, 1995), but cannot accurately judge the relative timing of tones that fall into separate auditory streams (Bregman and Campbell, 1971; Broadbent and Ladefoged, 1959; Formby et al., 1998; Neff et al., 1982; Roberts et al., 2002; Warren et al., 1969), we reasoned that if listeners heard the A and B tones as a single fused stream, their thresholds in the asynchrony detection task would be relatively small (i.e., a few milliseconds), whereas if the listeners perceived the A and B tones as two separate streams, their thresholds should be substantially larger.

The results shown in Figure 2 (filled squares) support these predictions. In the condition where all the A and B tones before the last were synchronous (with intertone intervals of 50 ms), thresholds were small, in the 2–4 ms range. This is true at all of the three A-B frequency separations tested, including the very large one (15 semitones, which is larger than an octave). This outcome is consistent with the hypothesis that synchrony between the A and B tones promotes the perceptual integration of these tones into a single stream, even for relatively large frequency separations.

To check that thresholds in the asynchrony-detection task provided a valid marker of the listener's auditory percept of streaming, three control conditions were run. The first two control conditions involved stimulus sequences in which the silent gap between consecutive B tones was either shorter (30 ms) or longer (70 ms) than that between consecutive A tones (50 ms), but with the stimulus parameters chosen such that the last A and B tones in one of the two stimulus sequences presented on a given trial would be synchronous (see Figure 2 inset). If thresholds in the asynchrony detection task are a faithful indicator of the listener's auditory percept, these thresholds should be larger in these control conditions than in the main condition because (1) the asynchrony between the A and B tones would promote the perceptual segregation of the stimulus sequence into two separate streams, and (2) stream segregation should hamper listeners' ability to accurately compare the timing of the A and B tones in that pair. The aim of the third control condition was to measure listeners' sensitivity to changes in the timing of the last B tone, even when the A tones were not present. We turned off the A tones and asked listeners to decide in which of the two presented sequences of (B only) tones the last tone was temporally shifted (either backward or forward, as in the main experiment). Therefore, in this control condition, listeners had to detect which of two presented sequences of B tones contained a temporal irregularity near the end.

Considerably larger thresholds (10–20 ms) were observed in the two control experiments, where the nominal duration of the intertone interval in the A tone stream was different from that in the B tone stream, being either shorter (30 ms) or longer (70 ms) [F(1,10) = 7.394, $p < 0.001$]. This outcome is consistent with the idea that asynchrony between the A and B tones promotes the perceptual segregation into two streams and makes it difficult, if not impossible, to accurately judge the timing of events in separate streams. In fact, the thresholds measured in those conditions were not significantly different from those measured in the third control condition, in which the A tones were turned off, so that the only cue listeners could use to perform the task was to listen for an irregularity in the timing of the B tone stream. This indicates that listeners were able to use the presence of the A tones to improve performance only when the A tones were all gated synchronously with the B tones. Overall, the psychophysical results confirm that synchronous and asynchronous tone sequences produce very different percepts, with the synchronous tones being perceived as a single stream and the asynchronous tones being perceived as two streams at large frequency separations.

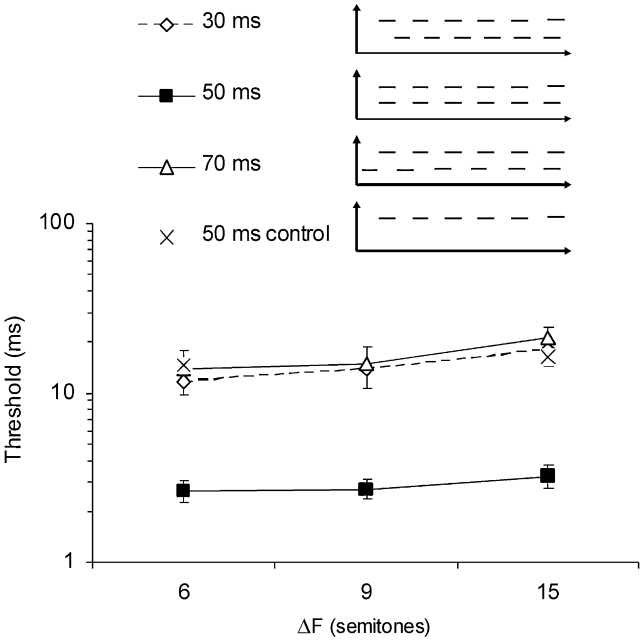

Figure 2. Thresholds for the Detection of a Temporal Shift Imposed on the Last B Tone in Various Types of Stimulus Sequences

The different symbols indicate different sequence types, which are represented schematically in the inset. Polygonal symbols correspond to sequences of A and B tones, with the duration of the silent gap between consecutive A tones set to 30, 50, or 70 ms, as indicated in the legend. Note that because the duration of the silent gap between consecutive B tones (excluding the last two) was kept constant at 50 ms, the use of a 50 ms gap for the A tones yielded synchronous A and B tones with identical tempi; in contrast, when the gap between consecutive A tones was equal to 30 or 70 ms, these tones were not synchronous with the B tones, and had a different (slower or faster) tempo. Crosses are used to indicate the results of a control condition, in which the A tones were turned off, and the listener's task was to indicate in which of the two presented sequences of B tones the last tone was shifted in time, creating an heterochrony. The numbers on the abscissa indicate the frequency separation between the A and B tones, in semitones. For the control condition in which only the B tones were present, this parameter was used to determine the frequency of the B tones so that it was equal to that used in corresponding conditions where A tones were also present. The error bars are geometric standard errors.

Neurophysiology

The psychophysical results raise the question of whether neural responses to sequences of synchronous and sequential tones in the central auditory system differ in a way that can account for their very different percepts. To answer this question, we

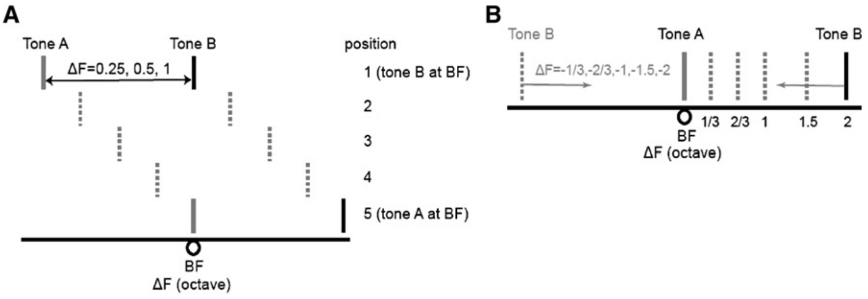

Figure 3. Schematic of the Tone Frequencies and Conditions Used in the Physiological Experiments

Both alternating and synchronous tone sequences were tested in all conditions.

(A) Experiment I: The two tone frequencies were held fixed at one of three intervals apart (ΔF = 0.25, 0.5, 1 octaves), and then shifted through five equally spaced positions relative to the BF of the isolated cell.

(B) Experiment II: Tone A is fixed at the BF of the isolated unit, and tone B is shifted closer to BF in several steps.

performed two experiments in which we recorded the single-unit responses in AI to sequences such as those illustrated in Figure 1 in the awake (nonbehaving) ferret. In the first experiment, we explored directly the extent of segregation between the responses to the A and B tones. In the second experiment, we assessed the range of frequencies over which the tones interacted (or mutually influenced their responses).

Experiment I: Segregation between Two-Tone Responses

This experiment examined the distribution of responses to the two tones by translating them together, relative to the best frequency (BF) of an isolated single unit in AI of awake ferrets in five steps (labeled 1–5 in Figure 3A), where positions 1 and 5 correspond to one of the two tones being at BF of the unit. The frequency separation (ΔF) between the tones in each test was fixed at 1, 0.5, or 0.25 octaves, corresponding to 12, 6, and 3 semitones, respectively. As described previously, alternating tone sequences are usually perceived as two streams at separations of 12 and 6 semitones (1 or 0.5 octaves), but are only marginally segregated at a separation of 3 semitones (0.25 octaves). In contrast, synchronous tone sequences are always heard as one stream (Figure 2). Therefore, if the spatial segregation hypothesis were valid, alternating sequences should evoke well-segregated neural responses to the far-apart tones (1 and 0.5 octaves), whereas synchronous sequences should evoke spatially overlapping responses in all cases.

The results from a population of 122 units in the AI of four ferrets are shown in Figure 4. In Figure 4A, the average rate profiles for the synchronous, overlapping, and alternating presentation modes are constructed from the responses as described in Methods. All 122 units were tested with the synchronous and alternating modes; 75/122 units were also tested with the overlapping sequences. When the tones are far apart (ΔF = 1 octave; right panel of Figure 4A), responses are strongest when either tone is near BF (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 (ΔF = 0.25 octave; left panel of Figure 4A), 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 all presentation modes: in all cases, the responses are well-segregated with significant dips when the tones are far apart (ΔF = 1 octave), and poorly separated (no dips) when the tones are closely spaced (ΔF = 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.

Instead of averaging the responses from all cells, we tabulated the number of cells indicating a significant segregation in the responses (implying a percept of two streams) or no segregation (a percept of one stream) by examining whether a significant dip occurred in each cell's profile during the two extreme presentation modes (synchronous versus alternating tones). The determination of a dip was derived for each condition by finding a significant difference (one-tailed t test; $p < 0.025$) between the distributions of the maximum response at either of the BF sites (1 or 5) compared with the minimum response at any of the non-BF sites (2, 3, or 4). For the purposes of this analysis, we used a population of 66 units for which positions 1 or 5 were BF sites, and measurements were completed at all positions (1–5). In most experiments, several units with diverse BFs were recorded simultaneously with multiple electrodes, and hence it was only possible to match the tone frequencies to the BF of one or two of the cells. The percentage of cells with a significant dip in their profiles is shown in the histograms of Figure 4B. We also calculated the magnitude of the dip (see Experimental Procedures) for each unit and established that there was no significant difference in neural responses between synchronous and alternating modes (two-tailed t test, $p = 0.54$ at 0.25 octave, $p = 0.37$ at 0.5 octave, and $p = 0.42$ at 1 octave), and that spatial segregation increases significantly with increasing ΔF (one-tailed t test, shown in Figure 4B). The results show that (1) segregation is strongest at 1 octave separation and weakest at 0.25 octaves, and that (2) there is little difference between the patterns of responses to the synchronous and alternating sequences. Thus, this alternative individual-cell response measure also fails to predict the different streaming percepts of the alternating and synchronous tones.

Experiment II: Frequency Range of Interactions

The key question of interest in this experiment was whether the range of interactions between the two tones was significantly different in the three presentation modes (alternating, overlapping, or synchronous). We measured the frequency range of interactions between the two tones by fixing tone A at the BF of the isolated unit, while placing tone B at ±1/3, ±2/3, ±1, ±1.5, and ±2 octaves around the BF (Figure 3B). We also estimated the unit's frequency tuning by measuring the isointensity response curve with a single tone sequence (curve with open diamonds in

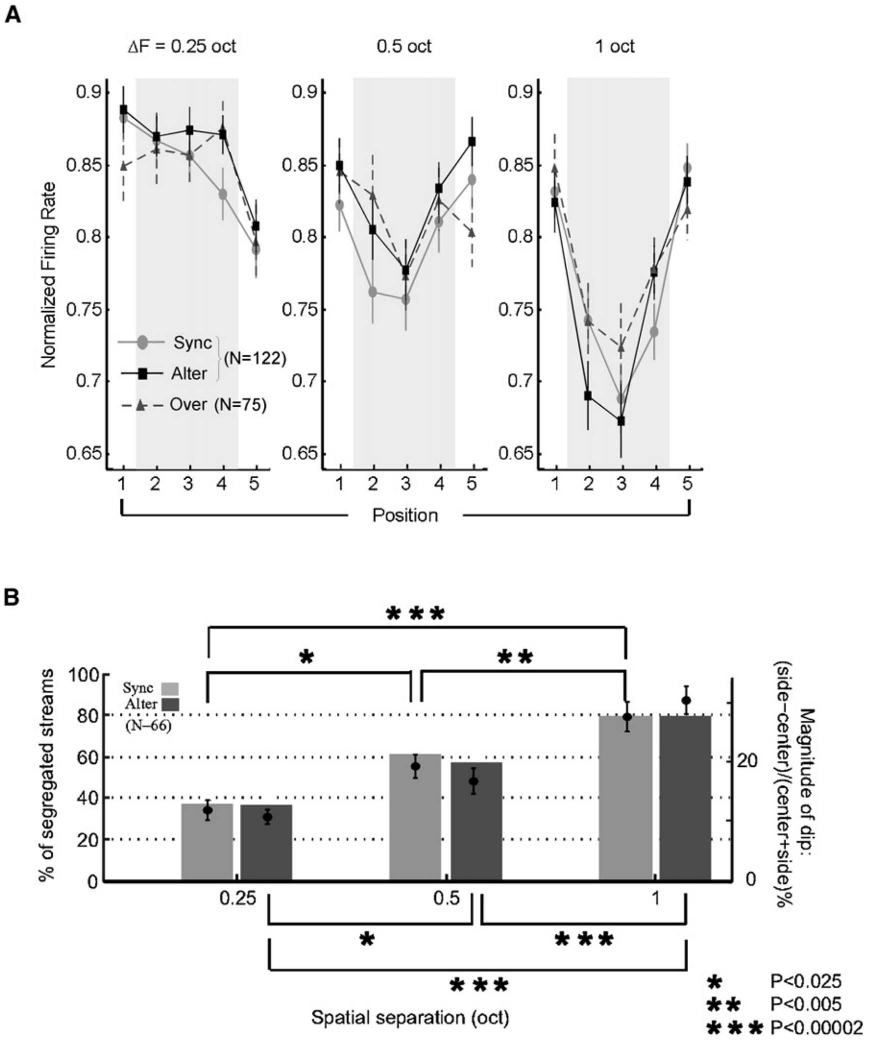

Figure 4. Responses of Single Units to Alternating (Nonoverlapping and Partially Overlapping) and Synchronous Two-Tone Sequences at Three Different Intervals (ΔF = 0.25, 0.5, 1 Octaves)

The two tones were shifted relative to the BF of the cell in five equal steps, from tone B being at BF (position 1) to tone A at BF (position 5), as described in experiment I paradigm.

(A) Average firing rates from a total of 122 single units in the five frequency positions in the synchronous and nonoverlapping modes. Overlapping tones were tested in only 75/122 units. Error bars are standard errors. Responses in all presentation modes exhibited a significant dip in response when tones were further apart (0.5 and 1 octaves), and neither was at BF (shaded positions 2–4).

(B) The percentage of cells that exhibited a significant dip in their responses were similar in the two extreme presentation modes (synchronous and nonoverlapping alternating). Only the 66 single units that were tested at all five positions were included in this analysis (as responses from all positions are necessary to compile such histograms). The magnitude of dip showed significant difference across ΔF but nonsignificant difference across presentation mode. Error bars represent standard errors.

Figure 5A). Other methodological details can be found in Experimental Procedures.

The average spike counts are shown in Figure 5A from a population of 64 single units (in the synchronous and alternating modes) and 41 units (overlapping mode) that were recorded separately from experiment I. All data were combined by computing the iso-intensity response curve of each unit, centering it around the BF of the unit, and normalizing it by the response of the unit to the single BF tone. We then kept only the half of the tuning curve above or below the BF from which the full two-octave range was tested. Such (half-tuning) curves from all units were then averaged for each condition. The results highlight the interactions observed as the tones approached each other in frequency. For instance, when tone B was far from tone A at BF (e.g., at ±2 octaves), the effects of the B tone on the cell are relatively small, and the firing rate in all modes was similar to that of the single tone at BF (the normalized rate of 1, indicated by the dotted line). As tone B approached BF, the responses become modulated, first decreasing and then increasing steeply beyond about one octave on either side of the BF. Apart from differences in absolute firing rates, the pattern of interactions was similar in all three presentation modes. For example, the frequency separations at which significant interactions ensue are similar, implying that the units' receptive fields (or their tuning curves) are similar whether they are driven by synchronous, alternating, or partially overlapping sequences.

To further quantify the population responses, we computed the effective bandwidth of interactions for each unit, defined as the furthest frequency on either side of the BF at which response interactions between the two tones were significant (see Experimental Procedures). The data from all units in the synchronous and alternating (nonoverlapping) modes are displayed in the histogram of the differences between the two measured ranges in Figure 5B. The scatter is mostly symmetric, with a mean not significantly different from zero (two-tailed t test, p = 1). Hence, the bandwidth differences for individual units fail once more to account for the different streaming percepts evoked by the alternating and synchronous presentation modes. Similar comparisons were also performed for the overlapping versus synchronous and overlapping versus alternating modes. The bandwidth differences in both cases were also mostly symmetric, with a mean not significantly different from zero.

Conclusions

The results from the two physiological experiments in awake ferrets contradict the hypothesis that segregation of AI responses to two-tone sequences is sufficient to predict their perceptual streaming. Instead, our findings reveal that synchronous and nonsynchronous sequences do not differ appreciably

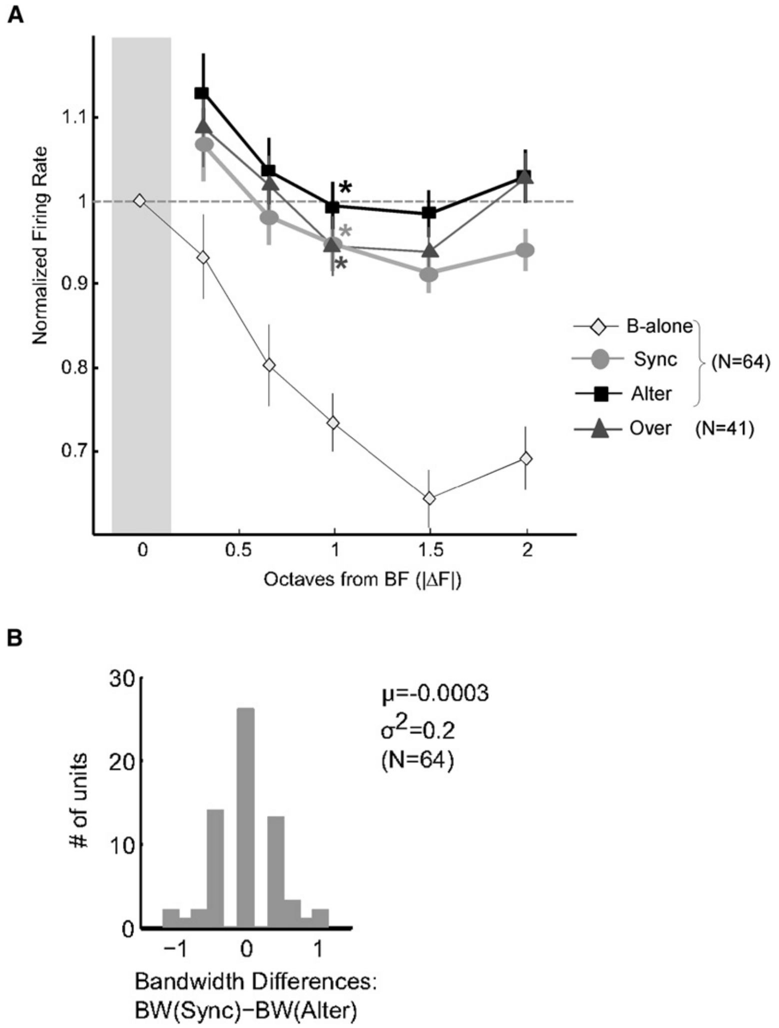

Figure 5. Averaged Responses from a Total of 64 Units Tested for Alternating, Synchronous, and Overlapping (Tested in Only 41/64 Units) Sequences Using the Paradigm of Experiment II

(A) The tuning near the BF averaged from all units. The average isointensity response curve is with open diamonds for comparison. To increase the number of cells included in the average, we folded the responses from above and below BF, but included only units that were tested with the entire two-octave range from BF. Error bars are standard errors. All presentation modes show some suppression of responses as tone A approaches the BF (1–1.5 octaves), and a significant increase closer to BF (about 1 octave, marked by asterisks).

(B) Histogram of the difference in bandwidth of interactions between the tones during the two extreme presentation modes (synchronous and alternating) is roughly symmetric, indicating no systematic bias in the scatter.

in the spatial representations of their temporally averaged responses in AI despite the substantial differences in their streaming percepts. Clearly a model that is successfully able to predict perception from these neural data will need to incorporate the time dimension.

Computational Modeling

Model Structure

Based on our physiological and psychophysical results, we propose a new model of the relationship between cortical responses and auditory streaming, which can account for the finding that synchronous and nonsynchronous tone sequences evoke very different percepts.

The model is based on the premise that temporal coherence between different sound features (e.g., frequency) is a fundamental organizing principle underlying the formation of perceptual streams. Specifically, auditory channels whose activity is positively correlated over time are assigned to the same perceptual stream, whereas channels with uncorrelated or anticorrelated activation patterns are assigned to different streams. In this way, the model combines, in a general framework, aspects of sequential and simultaneous auditory grouping, which in the past have often been treated as separate areas of research (e.g., Darwin and Carlyon, 1995).

The model consists of two stages, which are schematically depicted in Figure 6A. The first stage (temporal integration) takes as input an auditory spectrogram of a physical stimulus. The signal in each frequency band or ‘‘channel’’ of this spectrogram is passed through an array of bandpass filters tuned to frequencies between 2 and 32 Hz (see Experimental Procedures for details); these ‘‘rate filters’’ perform temporal integration with time constants ranging from 50 to 500 ms, consistent with the multiscale dynamics of cortical responses (Chi et al., 1999). In the second stage (coherence analysis), a windowed correlation between each pair of channels is computed by multiplying the outputs from filters corresponding to different channels with each other. The result is represented as a dynamic coherence matrix (denoted C), i.e., a correlation matrix that evolves over time. Effectively, the model computes the coincidence between all pairs of channels viewed over a range of timescales of the order of tens to hundreds of milliseconds, consistent with experimentally observed cortical temporal responses (Kowalski et al., 1996a, 1996b; Miller et al., 2002). Cortical responses typically phase lock only to relatively slow temporal modulations of less than 30 Hz (Miller et al., 2002). Consequently, measuring correlations between cortical responses must be commensurate with these time scales, allowing for a window long enough to include multiple periods of such responses.

Model Predictions

Figure 6B shows simulated coincidence matrices corresponding to alternating (upper panel) and synchronous (lower panel) tone sequences (depicted in Figure 6B, left). The right panels of Figure 6B represent dynamic coherence matrices averaged over time, and capture both the average spatial distribution of activity as well as temporal coherence within and across channels. The diagonal entries merely reflect the average power in the input channels and are not predictive of the perceptual

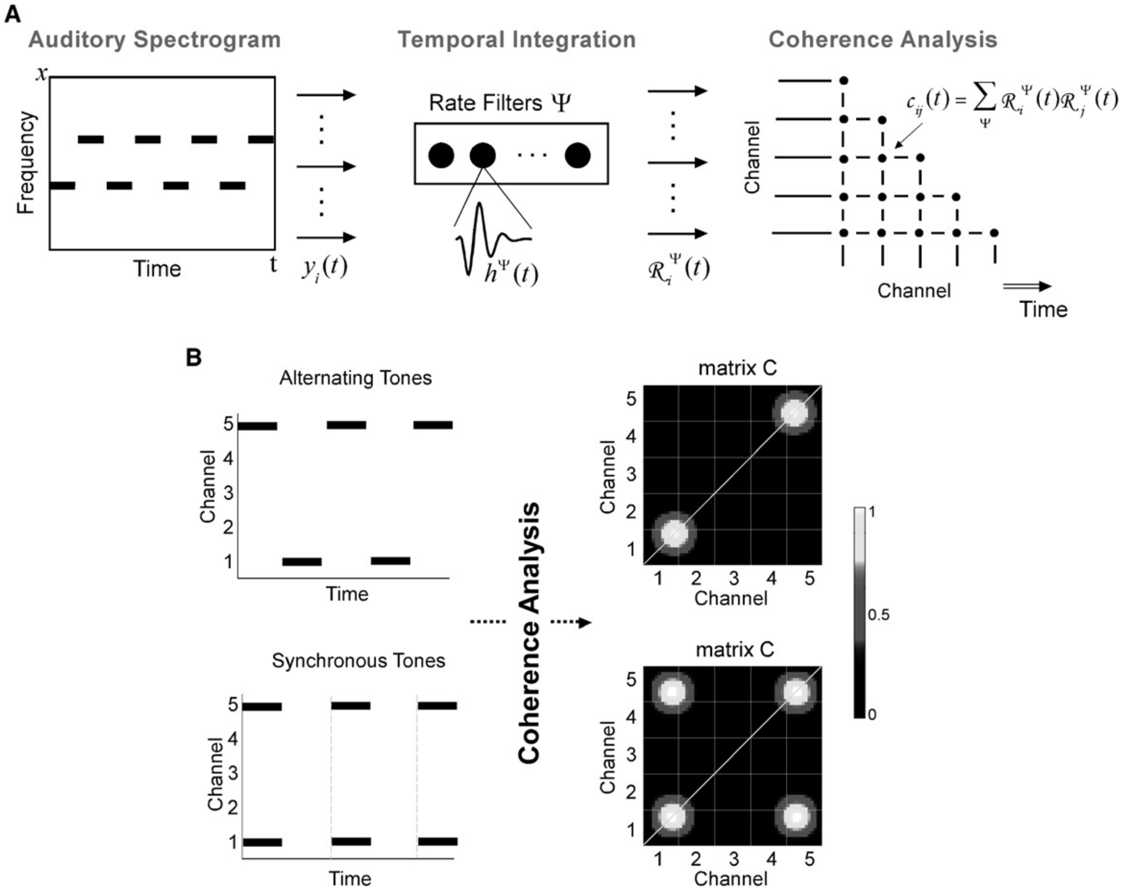

Figure 6. Schematic of the Coherence Analysis Model

(A) The model takes as input a time-frequency spectrographic representation of sound. The signal in each channel $y_i(t)$ is then processed through a temporal integration stage, implemented via a bank of filters (Ψ) operating at different time constants. Finally, the output of each rate analysis is correlated across channels, yielding a coherence matrix that evolves over time.

(B) A stimulus consisting of an alternating (upper) and synchronous (lower) tone sequence is generated with the two tones located at channels 1 and 5 of a five-channel spectrogram. The correlation matrices corresponding to these two sequences are generated and averaged over time (rightmost panels).

organization of the sequences. The off-diagonal entries are indicative of the correlation (or lack thereof) across different channels, and are predictive of how the stimulus sequences are perceptually organized. These entries are at zero for the alternating sequence because in this case the activation patterns in the channels corresponding to the A and B tones are out of phase (i.e., anticorrelated). In contrast, for the synchronous sequence, the off-diagonal entries corresponding to channels 1 and 5 are nonzero, reflecting the fact that these channels are activated in a coherent way (i.e., in phase).

To quantify the difference between these matrices, we used an eigenvalue analysis to decompose each matrix into its maximally coherent components (Golub and Van Loan, 1996). Intuitively, this spectral decomposition of the coherence matrix allows us to determine which channels are positively correlated with each other (hence possibly forming one stream), and anticorrelated with different channels (which would form a separate stream). By performing an eigen decomposition of the coherence matrix, we are effectively determining the number of axes (or independent dimensions) that best capture the data and, by analogy, the number of streams present in the stimulus. Hence, the rank of the matrix (or number of independent dimensions) can be interpreted as the number of auditory streams into which the stimulus sequence is likely to be perceptually organized by human listeners. Thus, a matrix of rank 1 (i.e., a matrix that can be fully decomposed using a single eigenvalue) is interpreted as reflecting a single perceived stream, and a matrix of rank 2 is associated with a percept of two streams.

Using this model, we can relate the perceptual organization of the synchronous and alternating sequences to the neural responses of the cortical units obtained in neurophysiological experiment I. The responses for each stimulus position (numbered 1–5) are equivalent to responses from different cortical sites, corresponding to five different ‘‘channels’’ along the spectral axis. We accumulated the peristimulus time histograms (PSTHs) for each position and each stimulus condition by averaging over the ten presentation times at a resolution of 1 ms. These five-channel histograms for each stimulus condition

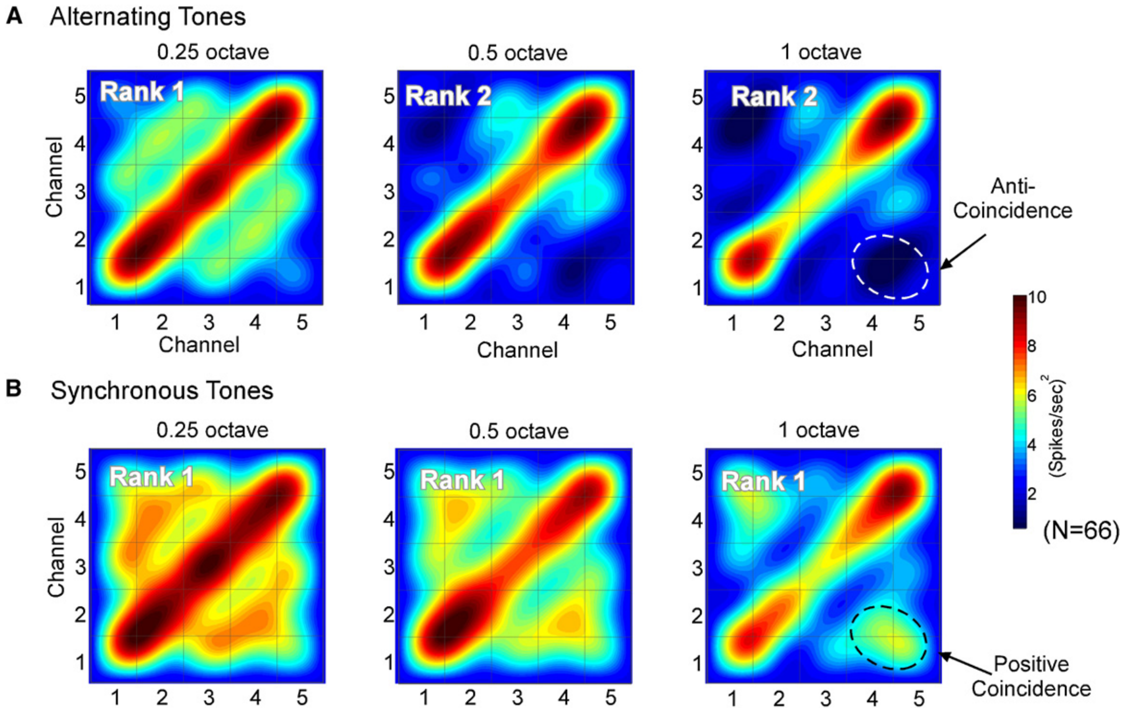

Figure 7. Coherence Analysis from Neural Population

The neural responses of n = 66 neurons are averaged for each tone configuration (alternating, A, and synchronous, B, tones), and each frequency separation (ΔF = 0.25, 0.5, and 1 octave). For each condition, a coherence matrix is derived for each neuron and averaged across the population. The final population coherence matrix has a resolution of 5 × 5 (five stimulus positions along the spectral axis). For display purposes, we interpolate each matrix into 500 × 500 points using MATLAB (MathWorks Inc., MA). The (5 × 5) matrices have been interpolated for display purposes only. The singular value decomposition for each matrix (from left to right) yields the values (0.97, 0.14, 0.11, 0.10, 0.10), (0.97, 0.15, 0.12, 0.12, 0.10), and (0.93, 0.25, 0.21, 0.15, 0.13) for synchronous sequences, and (0.92, 0.30, 0.17, 0.15, 0.13), (0.78, 0.55, 0.19, 0.16, 0.15), and (0.78, 0.52, 0.23, 0.21, 0.17) for alternating sequences. The noise floor is estimated at about 0.45.

(presentation mode and ΔF) are then given as input to the temporal integration and coherence analysis model, to derive coherence matrices similar to those described in Figure 7.

Figure 7A shows the mean coherence matrices (across all 66 neurons in the recorded BF sites sample from experiment I) for alternating (top row) and synchronous (bottom row) tone sequences with frequency separations (ΔFs) of 0.25 (left), 0.5 (middle), and 1 (right) octave. As explained previously, the positive diagonal entries reflect overall activity in each frequency channel. These positive diagonal entries show decreasing activity in the intermediate channels (2–3) with increasing frequency separation between the tones, reflecting increasing tonotopic separation in the measured cortical activation patterns. The fact that this pattern is observed for both synchronous and alternating tones confirms that tonotopic separation per se is not a valid indicator of the perceptual organization of these stimulus sequences. In contrast, the activation of the off-diagonal entries in these coherence matrices follows a pattern that is more closely aligned to perception, with more activation found in conditions that are perceived as a single stream.

The predicted number of streams, as determined by the number of "significant" eigenvalues (see Experimental Procedures) is shown in the upper left corner of each panel in Figure 7. For the synchronous conditions, the coherence matrices always yielded a single significant singular value, even at the largest two ΔFs (0.5 and 1 octave; singular values for each matrix are included in the figure caption), in line with the perception of a single stream. In contrast, in the alternating conditions, a second significant eigenvalue was observed at frequency separations of 0.5 and 1 octave, in line with the perception of two streams.

The model presented here can be adapted so as not to rely only on the rank of the coherence matrix to predict the perceptual organization from the input. The size of the eigenvalues, as well as the shape of the eigenvectors, is also a strong indicator of the different dimensions (or streams) in the scene. To illustrate this claim, we performed a simulation using sequences of two tones (A and B). The low tone was fixed at 300 Hz, and the high tone was fixed at 952 Hz. Both tones were 75 ms long. The onset delay between the A and B tones was varied from 0 ms (ΔT = 0%) to fully alternating (ΔT = 100%). Figure 8A shows the coherence analysis for the latter case, and reveals that the coherence matrix has rank 2 (indicating a two-stream percept). The ratio of the second-to-first singular values (λ_2/λ_1) equals 0.93, indicating that both λ values are almost equal. In contrast, Figure 8C shows the case of complete synchrony and reveals that the coherence matrix can be mapped on one main dimension, hence correlating with the percept of one stream. In this case, the ratio λ_2/λ_1 is equal to 0.01, revealing that the second singular value is close to zero. Using the relative sizes of the first and second singular values (not just the rank of the matrix), we

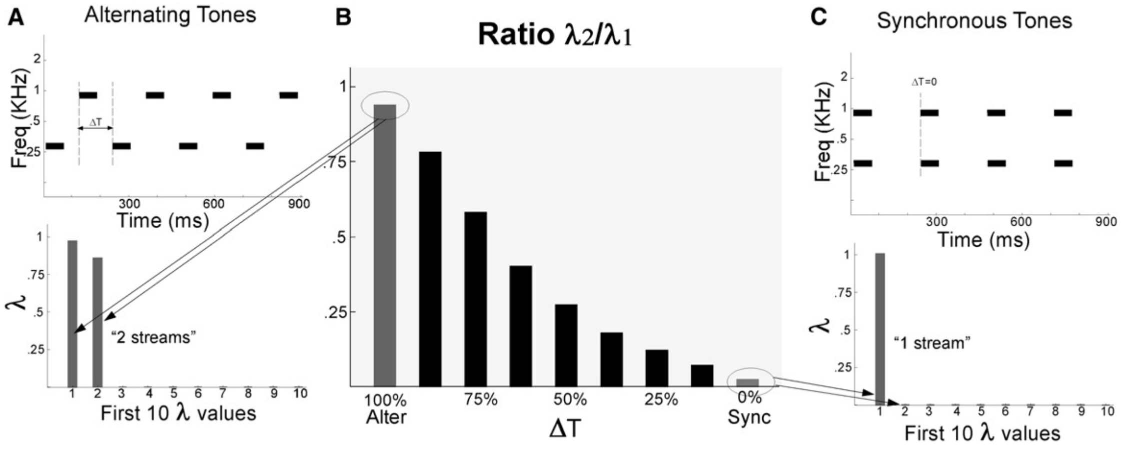

Figure 8. Simulation of Two-Tone Sequences with Varying Asynchrony

(A) A sequence of two alternating tones is presented as input to the model. The coherence analysis and singular value decomposition of the matrix C reveals a rank 2 matrix, as indicated by the two singular values (lower panel).

(B) Ratio of second-to-first (λ2/λ1) singular values as the value of ΔT is changed from 100% (alternating) to 0% (synchronous).

(C) A sequence of two synchronous tones is presented as input to the model. The coherence analysis and singular value decomposition of the matrix C reveals a rank 1 matrix, as indicated by one nonzero singular value (lower panel).

can explore the "strength" or "confidence" of the percept of one or two streams as we vary the degree of asynchrony. Figure 8B shows the decrease in this ratio as ΔT is gradually varied from 100% to 0%, 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.

DISCUSSION

Evidence Against a Purely Tonotopic or Spatial Model of Auditory Streaming

We examined the hypothesis that acoustic stimuli exciting spatially segregated neural response patterns are necessarily perceived as belonging to different perceptual streams. This "spatial" hypothesis underlies (explicitly or implicitly) previous interpretations of the neural correlates of streaming in the physiological investigations and the computational models of streaming (Beauvois and Meddis, 1991, 1996; Bee and Klump, 2004, 2005; Fishman et al., 2001, 2004; Kanwal et al., 2003; McCabe and Denham, 1997; Micheyl et al., 2005, 2007; Pressnitzer et al., 2008). One of the elegant aspects of the spatial hypothesis is that it can be generalized to predict that separate streams will be perceived whenever sounds evoke segregated responses along any of the representational dimensions in the auditory cortex, including not just the tonotopic axis but also a fundamental frequency (F0) or virtual pitch axis (Bendor and Wang, 2005, 2006; Gutschalk et al., 2007) as well as, perhaps, temporal and spectral modulation rate axes (Bendor and Wang, 2007; Kowalski et al., 1996a, 1996b; Schreiner, 1998; Schreiner and Sutter, 1992, 2005; Versnel et al., 1995), thereby accounting for psychophysical findings of stream segregation induced by differences in F0 or modulation rate in the absence of tonotopic cues (Grimault et al., 2002; Roberts et al., 2002; Vliegen and Oxenham, 1999).

However, the experimental data reported here cast doubt on the validity of an explanation of auditory streaming in terms of neural response separation that ignores temporal coherence as an important determinant of perceived segregation. Our human psychophysical results show very different perceptual organization of synchronous and asynchronous tone sequences, whereas the extent of segregation of the neural responses in ferret AI was essentially independent of the temporal relationships within the sequences. This finding emphasizes the fundamental importance of the temporal dimension in the perceptual organization of sound, and reveals that tonotopic neural response separation in auditory cortex alone cannot explain auditory streaming.

A Spatiotemporal Model of Auditory Streaming

Our alternative explanation augments the spatial (tonotopic) segregation hypothesis with a temporal dimension. It is a spatiotemporal view, wherein auditory stream segregation requires both separation into neural channels and temporal incoherence (or anticoherence) between the responses of these channels. This spatiotemporal hypothesis predicts that if the evoked neural responses are temporally coherent, a single stream is perceived, regardless of the spatial distribution of the responses. This prediction is consistent with our psychophysical findings using synchronous tone sequences. The prediction is also consistent with the introspective observation, confirmed in psychophysical studies, that synchronous spectral components generally fuse perceptually into a single coherent sound (e.g., a vowel or a musical chord), whereas the introduction of an asynchrony between one and the other components in a complex tone results in this component "popping out" perceptually (Ciocca and Darwin, 1993).

The present demonstration of a critical role of temporal coherence in the formation of auditory streams does not negate the role of spatial (tonotopic) separation as a factor in stream segregation. The extent to which neurons can signal temporal incoherence across frequency is determined in large part by their frequency selectivity. For example, the responses of two neurons tuned to the A and B tones in an alternating sequence (Figure 1A) can only show anticoherence if the frequency selectivity of the neurons is relatively high compared with the A-B frequency separation. If the neurons' frequency tuning is broader than the frequency separation, both neurons are excited by both tones (A and B) and respond in a temporally coherent fashion. In this sense, spatial separation of neural responses along the tonotopic axis may be necessary for stream segregation but, as this study shows, it is not sufficient.

The principle of channel coherence can be easily extended beyond the current stimuli (see Supplemental Data for further simulations) and the tonotopic frequency axis to include other auditory organizational dimensions such as spectral shape, temporal modulations, and binaural cues. Irrespective of the nature of the dimension explored, it is the temporal coherence between the responses along that dimension that determines the degree of their integration within one stream, or segregation into different streams.

Finally, there are interesting parallels between the present findings, which suggest an important role of temporal coherence across sensory channels in auditory scene analysis, and findings in other sensory modalities such as vision, where grouping based on coherence of temporal structure has been found to provide an elegant solution to the binding problem (e.g., Alais et al., 1998; Blake and Lee, 2005; Fahle, 1993; Treisman, 1999). Together, these findings suggest that although the perceptual analysis of visual and auditory ''scenes'' pose (at least, superficially) very different problems, they may in fact be governed by common overarching principles. In this regard, parallels can be drawn between prominent characteristics of auditory stream formation, such as the buildup of streaming and its dependence on frequency separation, and processes involved in the visual perception of complex scenes.

Do Percepts of Auditory Streams Emerge in or Beyond Primary Auditory Cortex?

For neural activity in AI to be consistent with the psychophysical observation that synchronous tones with remote frequencies are grouped perceptually while alternating tones are not, there should be cells in AI whose output is strongly influenced by temporal coherence across distant frequencies. Though such cells are likely to be present in AI (Barbour and Wang, 2002; Kowalski et al., 1996b; Nelken et al., 1999), we did not systematically find many that reliably exhibited the properties necessary to perform the coincidence operation. For example, all neurons sampled in this study followed the temporal course of the stimuli (with increased firing rates during epochs where at least one tone was present); the responses did not unambiguously increase in the presence of temporal coherence across tonotopic channels. Therefore, one possibility is that the percepts of stream segregation and stream integration are not determined in AI. Another possibility is that the coincidence and subsequent matrix decomposition described in the model are realized in a different, less explicit, form. For instance, it is theoretically possible to replace the spectral decomposition of the coherence matrix by a singular value decomposition directly upon the arrayed cortical responses. The spectral decomposition of the coherence matrix is equivalent to principal component analysis of the covariance matrix of the channel responses. Equivalent results can be computed by a singular value decomposition directly on the channel temporal responses (i.e., without computing the covariance matrix), obviating the need for the coincidence detectors. This leaves open the question of how and where, in or beyond AI, the detection of temporal coincidences across remote frequency channels is neurally implemented (Nelken, 2004).

The auditory streaming paradigm, with its relatively simple and well-controlled stimuli and extensively characterized percepts, may provide an excellent vehicle to explore a broader issue in brain function—that of the relationship between perception and neural oscillations, which reflects coherent responses across different regions in the brain. Coherence as an organizing principle of brain function has gained prominence in recent years with the demonstration that it could potentially play a role in mediating attention (Liang et al., 2003; Zeitler et al., 2006), in binding multimodal sensory features and responses (Lakatos et al., 2005; Schroeder et al., 2008), and in giving rise to conscious experiences (Fries et al., 1997; Gross et al., 2007; Meador et al., 2002; Melloni et al., 2007). Our results reinforce these ideas by emphasizing the importance of temporal coherence in explaining auditory perception. Specifically, the inclusion of the time dimension provides a general account of auditory perceptual organization that can in principle deal with any arbitrary combinations of sounds over time and frequency.

Attention and the Neural Correlates of Streaming

Interpretations of neural responses recorded in passive animals as ''correlates'' of auditory percepts are necessarily speculative, as behavioral measures of the animal's percepts during the recordings are not available. Under such conditions, the experimenter can, at best, assert that the neural responses differ across experimental conditions (e.g., different stimuli) in a way that is consistent with behavioral measurements obtained in the same (or a different) animal (or species) under similar stimulus conditions. In this respect, the present study suffers from the same limitation as previous investigations of the neural basis of auditory streaming in awake animals that were either passive (Bee and Klump, 2004, 2005; Fishman et al., 2001, 2004; Kanwal et al., 2003) or engaged in a task unrelated to streaming (Micheyl et al., 2005).

The possibility remains that AI responses to alternating and synchronous tone sequences in awake animals that are engaged in a task, which requires actively attending to the stimuli, might be substantially different from those recorded in passive animals. It is known that neural responses in AI are under attentional control, and can change rapidly as the task changes (Fritz et al., 2003, 2005a, 2005b). Such attentionally driven changes in receptive fields might differentially affect the neural responses to alternating tones and to synchronous tones, in a way that makes these responses more consistent with the percepts evoked by those sequences (Yin et al., 2007). However, the aspects of streaming investigated here—in particular the increased segregation with increasing frequency separation in asynchronous conditions—have been posited to be automatic or primitive and hence independent of attention (Macken et al., 2003; Sussman et al., 2007), although the matter is still debated (Carlyon et al., 2001).

The possible effects of attention could be investigated in future studies by controlling the attentional and behavioral state of the animal. Our model postulates the existence of units that should exhibit a dependence on temporal coherence. We have not found such units in AI, and therefore a future search may concentrate more fruitfully on other, supramodal, areas, such as the prefrontal cortex, where attentional modulation of AI responses may originate (Miller and Cohen, 2001).

EXPERIMENTAL PROCEDURES

Psychophysics

Listeners

Nine listeners took part in the study (none of authors participated in these tests). They all had normal hearing (defined as pure-tone hearing thresholds less than 20 dB HL at octave frequencies between 250 and 8000 Hz) and extensive experience with the test procedure and stimuli.

Stimuli

The stimuli were sequences of A and B tones, where A and B represent different frequencies. The frequency of the A tone was kept constant at 1000 Hz. The frequency of the B tone was set 0.5, 0.75, or 1.25 octaves above that of the A tone. Each tone was 100 ms in duration, including 10 ms raised cosine onset and offset ramps. Each sequence consisted of five precursor tones at each frequency (i.e., five A tones and five B tones), followed by two target tones (i.e., one A tone and one B tone). Depending on the condition being tested, the precursor A and B tones were either synchronous or asynchronous. In the synchronous case, all the tones were separated by silent gaps of 50 ms; in the asynchronous case, the gap between consecutive precursor B tones was still 50 ms, but the gap between consecutive A tones was either 30 ms or 70 ms, depending on the condition being tested. Depending on the observation interval, the target B tone was either separated from the preceding precursor B tone by the same 50 ms silent gap as consecutive precursor B tones (standard interval), or it was shifted forward or backward in time by an amount, ΔT, which was controlled by the adaptive threshold-tracking procedure (signal interval). The two parallel sequences of A and B tones in each interval were always positioned in time relative to each other in such a way that the target A and B tones were synchronous in the standard interval and shifted by ΔT in the target interval. In addition, a control condition was run in which the A tones were turned off and the B tones were generated in exactly the same way as described previously.

Procedure

Thresholds were measured using a two-interval, two-alternative forced-choice (2I-2AFC) procedure with an adaptive three-down one-up rule, which tracked the 79.4%-correct point on the psychometric function. On each trial, two sequences were presented, separated by a silent gap of 500 ms. In one of those sequences (the standard interval), the target A and B tones were synchronous; in the other (the target interval), they were asynchronous. The order of presentation of the two sequences was randomized, with each sequence being a priori as likely as the other to come first. The listener was informed of this fact and asked to indicate after each trial which of the two sequences (first or second) contained the asynchronous A and B tones at the end. Listeners gave responses by pressing keys (“1” or “2”) on a computer keyboard. At the beginning of each adaptive run, the tracking variable, ΔT, was set to 20 ms. It was divided by a factor *c* after two consecutive correct responses, and multiplied by that same factor after each incorrect response. The value of *c* was set to 4 at the beginning of the adaptive run; it was reduced to 2 after the first reversal in the direction of tracking (from decreasing to increasing), and to $\sqrt{2}$ after a further two reversals. The procedure stopped after the sixth reversal with the $\sqrt{2}$ step size. Threshold was computed as the geometric mean of ΔT at the last six reversal points. Each listener completed at least four threshold measurements in each condition. The psychophysical data shown in this article are geometric mean thresholds across listeners.

Apparatus

The stimuli were generated digitally and played out via a soundcard (LynxStudio L22; Costa Mesa, CA) with 24-bit resolution and a sampling frequency of 32 kHz, and presented to the listener via the left earpiece of Sennheiser HD 580 headphones (Sennheiser Electronic Corporation; Old Lyme, CT). Listeners were seated in a double-walled sound-attenuating chamber (Industrial Acoustics Company; Bronx, NY).

Neurophysiology

Experimental Design

The stimuli were sequences of A and B tones, where A and B represent different frequencies as illustrated in Figure 1. Both alternating (nonoverlapping and partially overlapping) and synchronous sequences were used (see details following). In experiment I, tones A and B were shifted equally in five steps relative to a unit’s BF, as shown in Figure 3A, with tone B starting at the BF and tone A ending at the BF. ΔF between the tones was 0.25, 0.5, or 1 octave, which was fixed within a trial and varied among different trials. The total number of conditions was 45 (five positions × 3 ΔF × 3 modes). In experiment II, tone A was set at the BF of the isolated unit, and tone B was placed to be ±1/3, ±2/3, ±1, ±1.5, and ±2 octaves away from tone A, as illustrated in Figure 3B. The stimuli also included a single tone sequence to measure the frequency tuning of the unit.

In both experiments I and II, each trial included 400 ms prestimulus silence, 3 s stimulus length, and 600 ms poststimulus silence. Tone duration was 75 ms, including 5 ms onset and offset ramps, and an intertone gap of 25 ms in the alternating sequence and 125 ms in the synchronous sequence. For the overlapped sequences, the tone onset asynchrony was 40 ms (i.e., about 50% overlap between the tones). All conditions were presented pseudorandomly 10 times at 70 dB SPL or at about 10 dB above threshold of the isolated units.

Data Analysis

For each unit and each condition, a period histogram was constructed from the PSTHs by folded (averaged) responses to the two tones over the duration of the trial from 0.6 to 3 s after the onset of the stimuli. Examples of such histograms from a single unit responding to stimuli of experiment I are shown in Figures S1 and S2. For each stimulus response, we excluded the first 0.6 s so as to avoid adaptation effects. The mean firing rate (spikes per second) was computed by taking the average value of the period histogram (averaged over 0.2 s). The overall firing rate patterns were obtained by averaging the normalized responses from all isolated units. To compensate for inherent differences in the relative strength of tone responses across units, firing rates were first normalized by dividing them by the maximum rate at each ΔF and at each stimulus mode in experiment I and by the mean firing rate at BF in experiment II.

The magnitude of dip was determined according to the following equation:

$$(\text{Side} - \text{Center})/(\text{Side} + \text{Center})\,\%$$

where “Side” is the maximum response at either of the BF sites (position 1 or 5); and “Center” is the minimum response at any of the non-BF sites (positions 2, 3, or 4).

To measure the effective bandwidth of interaction between tones, the mean firing rate at the frequency closest to BF (i.e., 1/3 or $-1/3$ octave) was compared with those at the other frequencies on the same direction (i.e., below BF or above BF). The frequency showing the significant difference (two-tailed t test, $p < 0.05$) in mean firing rate from the frequency closest to BF was the effective bandwidth of interaction.

Modeling

The neural responses to the shifting two tones (experiment I) from n = 66 (BF sites) neurons are pooled together and processed through a coherence analysis as follows.

A PSTH is constructed for each stimulus condition (i.e., a given tone synchrony configuration, a specific frequency separation, and a position relative to the spectral response of the neuron) by averaging the responses across ten stimulation trials using 1 ms bins. Each PSTH sequence is then convolved in time with an array of filters, with impulse response $h^{\Psi}(t)$, parameterized by $\Psi = (\omega_c, \theta_c)$, and defined as $h^{\Psi}(t;\omega_c,\theta_c) = \omega_c g(\omega_c t)\cos\theta_c + \omega_c \hat{g}(\omega_c t)\sin\theta_c$, where $g(t) = t^2 e^{-3.5t}\sin(2\pi t)$ and g(.) and $\hat{g}$(.) denote Hilbert transform pairs (Bracewell, 1999). Each filter $h^{\Psi}(t)$ is characterized by two tuning parameters:

ω_c, a characteristic rate, which varies over the range [2 4 8 16 32] Hz; and θ_c, a characteristic phase, which is set to span the entire range [0,2π] in steps of π/3. The characteristic rates are chosen to cover the range of temporal modulations observed in tuning properties of cortical neurons, whereas the phases vary to allow different phase configurations of the impulse response of the rate analysis. The seed function g(.) is chosen to be a gamma function (as shown in the inset of the model schematic in Figure 6A (Chi et al., 1999). Functionally, this temporal filtering stage integrates the temporal response from each channel over a range of analyses windows, yielding a three-dimensional representation covering time (t), channel (i), and rate (Ψ). The responses from all channels are pooled together in a vector representation, $R(t;\omega_c,\theta_c) = [R_1(t;\omega_c,\theta_c), R_2(t;\omega_c,\theta_c), \cdots, R_5(t;\omega_c,\theta_c)]'$, where $[.]'$ denotes the transpose operator. These responses are cross-correlated with each other (via an inner product operation) and averaged across the entire array of rate filters to yield a cross-correlation matrix C(t) whose entries are defined as

$$c_{ij}(t) = \sum_{\omega_c}\sum_{\theta_c} R_i(t;\omega_c,\theta_c)R_j(t;\omega_c,\theta_c)'$$

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, whereas a low correlation value indicates lack of coherent neural activity. To estimate the baseline level for average eigenvalue level in case of random coherence, we simulate activity of 66 different neurons with random PSTH activity over 3 s duration for five different positions. These random PSTHs are then processed through the coherence analysis, yielding a matrix C of random correlations among channels. The matrix from each "random" unit is first normalized to unit norm, before averaging across all units to yield one random coherence matrix. The singular value decomposition of this final matrix produces a full rank matrix, with five almost-equal eigenvalues.

SUPPLEMENTAL DATA

Supplemental Data include two figures and three audio files and can be found with this article online at http://www.neuron.org/supplemental/S0896-6273(08)01053-2.

ACKNOWLEDGMENTS

We thank Cynthia Hunter for assistance in collecting the psychophysical data, and Pingbo Yin and Stephen David for assistance with physiological recordings. We also thank three anonymous reviewers for their valuable comments. This work was 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).

Accepted: December 4, 2008
Published: January 28, 2009

REFERENCES

Alais, D., Blake, R., and Lee, S.H. (1998). Visual features that vary together over time group together over space. Nat. Neurosci. *1*, 160–164.

Barbour, D.L., and Wang, X. (2002). Temporal coherence sensitivity in auditory cortex. J. Neurophysiol. *88*, 2684–2699.

Beauvois, M.W., and Meddis, R. (1991). A computer model of auditory stream segregation. Q. J. Exp. Psychol. A *43*, 517–541.

Beauvois, M.W., and Meddis, R. (1996). Computer simulation of auditory stream segregation in alternating-tone sequences. J. Acoust. Soc. Am. *99*, 2270–2280.

Bee, M.A., and Klump, G.M. (2004). Primitive auditory stream segregation: a neurophysiological study in the songbird forebrain. J. Neurophysiol. *92*, 1088–1104.

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

Bee, M.A., and Micheyl, C. (2008). The cocktail party problem: what is it? How can it be solved? And why should animal behaviorists study it? J. Comp. Psychol. *122*, 235–251.

Bendor, D., and Wang, X. (2005). The neuronal representation of pitch in primate auditory cortex. Nature *436*, 1161–1165.

Bendor, D., and Wang, X. (2006). Cortical representations of pitch in monkeys and humans. Curr. Opin. Neurobiol. *16*, 391–399.

Bendor, D., and Wang, X. (2007). Differential neural coding of acoustic flutter within primate auditory cortex. Nat. Neurosci. *10*, 763–771.

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

Bracewell, R. (1999). The Fourier Transform and Its Applications, Third Edition (Boston: McGraw-Hill).

Bregman, A. (1990). Auditory Scene Analysis (Cambridge, MA: MIT Press).

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

Broadbent, D.E., and Ladefoged, P. (1959). Auditory perception of temporal order. J. Acoust. Soc. Am. *31*, 1539–1540.

Carlyon, R.P. (2004). How the brain separates sounds. Trends Cogn. Sci. *8*, 465–471.

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

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

Ciocca, V., and Darwin, C.J. (1993). Effects of onset asynchrony on pitch perception: adaptation or grouping? J. Acoust. Soc. Am. *93*, 2870–2878.

Darwin, C.J., and Carlyon, R.P. (1995). Auditory grouping. In Hearing, B.C.J. Moore, ed. (Orlando, FL: Academic Press), pp. 387–424.

Eggermont, J.J. (2001). Between sound and perception: reviewing the search for a neural code. Hear. Res. *157*, 1–42.

Fahle, M. (1993). Figure-ground discrimination from temporal information. Proc. Biol. Sci. *254*, 199–203.

Fay, R.R. (1998). Auditory stream segregation in goldfish (Carassius auratus). Hear. Res. *120*, 69–76.

Fay, R.R. (2000). Spectral contrasts underlying auditory stream segregation in goldfish (Carassius auratus). J. Assoc. Res. Otolaryngol. *1*, 120–128.

Fishman, Y.I., Reser, D.H., Arezzo, J.C., and Steinschneider, M. (2001). Neural correlates of auditory stream segregation in primary auditory cortex of the awake monkey. Hear. Res. *151*, 167–187.

Fishman, Y.I., Arezzo, J.C., and 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.

Formby, C., Sherlock, L.P., and Li, S. (1998). Temporal gap detection measured with multiple sinusoidal markers: effects of marker number, frequency, and temporal position. J. Acoust. Soc. Am. *104*, 984–998.

Fries, P., Roelfsema, P.R., Engel, A.K., Konig, P., and Singer, W. (1997). Synchronization of oscillatory responses in visual cortex correlates with perception in interocular rivalry. Proc. Natl. Acad. Sci. U.S.A. *94*, 12699–12704.

Fritz, J., Shamma, S., Elhilali, M., and Klein, D. (2003). Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex. Nat. Neurosci. *6*, 1216–1223.

Fritz, J., Elhilali, M., and Shamma, S. (2005a). Active listening: task-dependent plasticity of spectrotemporal receptive fields in primary auditory cortex. Hear. Res. *206*, 159–176.

Fritz, J.B., Elhilali, M., and Shamma, S.A. (2005b). Differential dynamic plasticity of A1 receptive fields during multiple spectral tasks. J. Neurosci. *25*, 7623–7635.

Golub, G.H., and Van Loan, C.F. (1996). Matrix Computations, Third Edition (Baltimore: Johns Hopkins University Press).

Griffiths, T.D., and Warren, J.D. (2004). What is an auditory object? Nat. Rev. Neurosci. *5*, 887–892.

Grimault, N., Bacon, S.P., and Micheyl, C. (2002). Auditory stream segregation on the basis of amplitude-modulation rate. J. Acoust. Soc. Am. *111*, 1340–1348.

Gross, J., Schnitzler, A., Timmermann, L., and Ploner, M. (2007). Gamma oscillations in human primary somatosensory cortex reflect pain perception. PLoS Biol. *5*, e133.

Gutschalk, A., Micheyl, C., Melcher, J.R., Rupp, A., Scherg, M., and Oxenham, A.J. (2005). Neuromagnetic correlates of streaming in human auditory cortex. J. Neurosci. *25*, 5382–5388.

Gutschalk, A., Oxenham, A.J., Micheyl, C., Wilson, E.C., and Melcher, J.R. (2007). Human cortical activity during streaming without spectral cues suggests a general neural substrate for auditory stream segregation. J. Neurosci. *27*, 13074–13081.

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

Hulse, S.H., MacDougall-Shackleton, S.A., and Wisniewski, A.B. (1997). Auditory scene analysis by songbirds: stream segregation of birdsong by European starlings (Sturnus vulgaris). J. Comp. Psychol. *111*, 3–13.

Izumi, A. (2002). Auditory stream segregation in Japanese monkeys. Cognition *82*, B113–B122.

Kanwal, J.S., Medvedev, A.V., and Micheyl, C. (2003). Neurodynamics for auditory stream segregation: tracking sounds in the mustached bat's natural environment. Network *14*, 413–435.

Kowalski, N., Depireux, D.A., and Shamma, S.A. (1996a). Analysis of dynamic spectra in ferret primary auditory cortex. I. Characteristics of single-unit responses to moving ripple spectra. J. Neurophysiol. *76*, 3503–3523.

Kowalski, N., Depireux, D.A., and Shamma, S.A. (1996b). Analysis of dynamic spectra in ferret primary auditory cortex. II. Prediction of unit responses to arbitrary dynamic spectra. J. Neurophysiol. *76*, 3524–3534.

Lakatos, P., Shah, A.S., Knuth, K.H., Ulbert, I., Karmos, G., and Schroeder, C.E. (2005). An oscillatory hierarchy controlling neuronal excitability and stimulus processing in the auditory cortex. J. Neurophysiol. *94*, 1904–1911.

Liang, H., Bressler, S.L., Ding, M., Desimone, R., and Fries, P. (2003). Temporal dynamics of attention-modulated neuronal synchronization in macaque V4. Neurocomputing *52-54*, 481–487.

MacDougall-Shackleton, S.A., Hulse, S.H., Gentner, T.Q., and White, W. (1998). Auditory scene analysis by European starlings (Sturnus vulgaris): perceptual segregation of tone sequences. J. Acoust. Soc. Am. *103*, 3581–3587.

Macken, W.J., Tremblay, S., Houghton, R.J., Nicholls, A.P., and Jones, D.M. (2003). Does auditory streaming require attention? Evidence from attentional selectivity in short-term memory. J. Exp. Psychol. Hum. Percept. Perform. *29*, 43–51.

Marr, D. (1983). Vision (San Francisco: W. H. Freeman).

McCabe, S., and Denham, M.J. (1997). A model of auditory streaming. J. Acoust. Soc. Am. *101*, 1611–1621.

Meador, K.J., Ray, P.G., Echauz, J.R., Loring, D.W., and Vachtsevanos, G.J. (2002). Gamma coherence and conscious perception. Neurology *59*, 847–854.

Melloni, L., Molina, C., Pena, M., Torres, D., Singer, W., and Rodriguez, E. (2007). Synchronization of neural activity across cortical areas correlates with conscious perception. J. Neurosci. *27*, 2858–2865.

Micheyl, C., Tian, B., Carlyon, R.P., and Rauschecker, J.P. (2005). Perceptual organization of tone sequences in the auditory cortex of awake macaques. Neuron *48*, 139–148.

Micheyl, C., Carlyon, R.P., Gutschalk, A., Melcher, J.R., Oxenham, A.J., Rauschecker, J.P., Tian, B., and Courtenay Wilson, E. (2007). The role of auditory cortex in the formation of auditory streams. Hear. Res. *229*(1–2), 116–131.

Miller, E.K., and Cohen, J.D. (2001). An integrative theory of prefrontal cortex function. Annu. Rev. Neurosci. *24*, 167–202.

Miller, L.M., Escabi, M.A., Read, H.L., and Schreiner, C.E. (2002). Spectrotemporal receptive fields in the lemniscal auditory thalamus and cortex. J. Neurophysiol. *87*, 516–527.

Neff, D.L., Jesteadt, W., and Brown, E.L. (1982). The relation between gap discrimination and auditory stream segregation. Percept. Psychophys. *31*, 493–501.

Nelken, I. (2004). Processing of complex stimuli and natural scenes in the auditory cortex. Curr. Opin. Neurobiol. *14*, 474–480.

Nelken, I., Rotman, Y., and Bar Yosef, O. (1999). Responses of auditory-cortex neurons to structural features of natural sounds. Nature *397*, 154–157.

Pickles, J.O. (1988). An Introduction to the Physiology of Hearing, Second Edition (London: Academic Press).

Pressnitzer, D., Sayles, M., Micheyl, C., and Winter, I.M. (2008). Perceptual organization of sound begins in the auditory periphery. Curr. Biol. *18*, 1124–1128.

Roberts, B., Glasberg, B.R., and Moore, B.C. (2002). Primitive stream segregation of tone sequences without differences in fundamental frequency or passband. J. Acoust. Soc. Am. *112*, 2074–2085.

Schreiner, C.E. (1998). Spatial distribution of responses to simple and complex sounds in the primary auditory cortex. Audiol. Neurootol. *3*, 104–122.

Schreiner, C.E., and Sutter, M.L. (1992). Topography of excitatory bandwidth in cat primary auditory cortex: single-neuron versus multiple-neuron recordings. J. Neurophysiol. *68*, 1487–1502.

Schroeder, C.E., Lakatos, P., Kajikawa, Y., Partan, S., and Puce, A. (2008). Neuronal oscillations and visual amplification of speech. Trends Cogn. Sci. *12*, 106–113.

Snyder, J.S., Alain, C., and Picton, T.W. (2006). Effects of attention on neuroelectric correlates of auditory stream segregation. J. Cogn. Neurosci. *18*, 1–13.

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

Sutter, M.L. (2005). Spectral processing in the auditory cortex. Int. Rev. Neurobiol. *70*, 253–298.

Treisman, A. (1999). Solutions to the binding problem: progress through controversy and convergence. Neuron *24*, 105–110, 111–125.

Versnel, H., Shamma, S.A., and Kowalski, N. (1995). Ripple analysis in the ferret primary auditory cortex. III. Topographic and columnar distribution of ripple response. J. Aud. Neurosci. *1*, 271–285.

Vliegen, J., and Oxenham, A.J. (1999). Sequential stream segregation in the absence of spectral cues. J. Acoust. Soc. Am. *105*, 339–346.

Warren, R.M., Obusek, C.J., Farmer, R.M., and Warren, R.P. (1969). Auditory sequence: confusion of patterns other than speech or music. Science *164*, 586–587.

Wilson, E.C., Melcher, J., Micheyl, C., Gutschalk, A., and Oxenham, A.J. (2007). Cortical fMRI activation to sequences of tones alternating in frequency: relationship to perceived rate and streaming. J. Neurophysiol. *97*, 2230–2238.

Yin, P., Ma, L., Elhilali, M., Fritz, J., and Shamma, S.A. (2007). Primary auditory cortical responses while attending to different streams. In Hearing: From Sensory Processing to Perception, B. Kollmeier, G. Klump, V. Hohmann, U. Langemann, M. Mauermann, S. Uppenkamp, and J. Verhey, eds. (Berlin: Springer-Verlag).

Zeitler, M., Fries, P., and Gielen, S. (2006). Assessing neuronal coherence with single-unit, multi-unit, and local field potentials. Neural Comput. *18*, 2256–2281.

Zeki, S. (1993). Vision of the Brain (Oxford: Wiley-Blackwell).

Zera, J., and Green, D.M. (1993a). Detecting temporal asynchrony with asynchronous standards. J. Acoust. Soc. Am. *93*, 1571–1579.

Zera, J., and Green, D.M. (1993b). Detecting temporal onset and offset asynchrony in multicomponent complexes. J. Acoust. Soc. Am. *93*, 1038–1052.

Zera, J., and Green, D.M. (1995). Effect of signal component phase on asynchrony discrimination. J. Acoust. Soc. Am. *98*, 817–827.