Ann. N.Y. Acad. Sci. ISSN 0077-8923

ANNALS OF THE NEW YORK ACADEMY OF SCIENCES
Issue: *The Year in Cognitive Neuroscience*
REVIEW

Recent advances in exploring the neural underpinnings of auditory scene perception

Joel S. Snyder[1] and Mounya Elhilali[2]

[1]Department of Psychology, University of Nevada, Las Vegas, Las Vegas, Nevada. [2]Department of Electrical and Computer Engineering, The Johns Hopkins University, Baltimore, Maryland

Address for correspondence: Joel S. Snyder, Department of Psychology, University of Nevada Las Vegas, 4505 South Maryland Parkway Box 455030, Las Vegas, NV 89154-5030. joel.snyder@unlv.edu; Mounya Elhilali, Department of Electrical and Computer Engineering, The Johns Hopkins University, 3400 North Charles Street, Barton Hall Room 307, Baltimore, MD 21218. mounya@jhu.edu.

Studies of auditory scene analysis have traditionally relied on paradigms using artificial sounds—and conventional behavioral techniques—to elucidate how we perceptually segregate auditory objects or streams from each other. In the past few decades, however, there has been growing interest in uncovering the neural underpinnings of auditory segregation using human and animal neuroscience techniques, as well as computational modeling. This largely reflects the growth in the fields of cognitive neuroscience and computational neuroscience and has led to new theories of how the auditory system segregates sounds in complex arrays. The current review focuses on neural and computational studies of auditory scene perception published in the last few years. Following the progress that has been made in these studies, we describe (1) theoretical advances in our understanding of the most well-studied aspects of auditory scene perception, namely segregation of sequential patterns of sounds and concurrently presented sounds; (2) the diversification of topics and paradigms that have been investigated; and (3) how new neuroscience techniques (including invasive neurophysiology in awake humans, genotyping, and brain stimulation) have been used in this field.

Keywords: auditory scene analysis; concurrent sound segregation; auditory stream segregation; informational masking; change deafness

The study of auditory scene analysis, pioneered by Bregman and others,[1–4] has traditionally sought to reveal the principles underlying how listeners segregate patterns coming from different physical sound sources and perceive distinct sound patterns. We distinguish two types of psychological representation: (1) *auditory objects*, which are typically relatively brief and temporally continuous (e.g., perception of a single word, musical note, or frog croak); and (2) *auditory streams*, which are series of events that are perceived as connected to each other across time (e.g., perception of a sentence, melody, or series of croaks from a single frog). Decades of work have revealed notable parallels between the psychological principles underlying visual and auditory scene analysis; both rely to varying degrees on Gestalt principles.[5–8] In particular, Gestalt psychology highlights our perceptual ability to segregate entire scenes (both auditory and visual) into figure and ground elements based on specific patterns in various stimulus features.[5,9,10]

Complementing these findings are parallel efforts to unravel the neural and computational mechanisms of auditory scene analysis in the auditory system. Much is known about how the brain processes individual events and features. Specifically, studies using traditional neurophysiological methods, such as single-unit recordings, shed light on the tuning properties of neurons at various stages of the auditory pathway. Similarly, theoretical models inspired by these findings (e.g., theories based on tonotopy, forward masking, and adaptation in single neurons[11–13]) have been developed to explain perception. However, by placing too much emphasis on feature analysis, a danger is that too much of the forest is being missed for the trees. Therefore, we believe

doi: 10.1111/nyas.13317

that also studying how patterns are computed in the brain will lead to important insights into the processes of segregation and object formation.

Recently, progress has been made in the study of auditory scene analysis by integrating psychophysical, neural, and computational approaches, resulting in a more complete understanding of complex scene parsing. To some extent, these studies take a fresh look at auditory scene analysis (e.g., by using more complex time-varying stimuli). Moreover, the use of a wide array of neuroscience techniques (including genetics and both noninvasive and invasive electrophysiology in humans and other species) allows us to ask new questions or answer old ones that have not been solved using older techniques. Recent findings are complementing our understanding of the complex processes underlying auditory scene analysis and are becoming test beds for true neuromechanistic theories. Beyond understanding the parsing of auditory scenes, recent advances in the field are also connecting with theories of consciousness typically developed by vision scientists, thus contributing to the debate about common principles of consciousness in the brain that might apply across modalities.

Naturally, much work remains to further our understanding of the neural mechanisms of auditory scene analysis. Thus, in the remainder of this paper, we will provide further detail about new developments in the field of auditory scene analysis. We review individual studies that have been published relatively recently in some detail and attempt to integrate their findings with older findings and theories, while also pointing out how future studies could make further progress to move beyond our current limited understanding. It should be noted that, despite the increased prevalence of studies using neural and computational approaches, there are many more empirical studies than quantitative theoretical studies. Even rarer are computational models that explain what particular neurophysiological mechanisms implement particular computations. We describe one such model in this paper—on bistable perception.

Stimulus-driven segregation mechanisms in the ascending auditory system

Sequential segregation

Sequential auditory scene analysis refers to the ability to hear two separate sequences (or streams) of sounds when two or more different sounds are presented in a repeating fashion.[14,15] A well-studied stimulus for this consists of low (A) and high (B) tones that are alternated in time repeatedly, often with every other high tone omitted (ABA–ABA–. . .)[3] or in a repeating fashion (ABABAB . . .) (Fig. 1A, left). Such patterns are usually first heard as one stream of sounds, but after several repetitions listeners often report hearing segregated patterns of low and high tones (A–A–A–. . . and B–B–. . .).[16] While most of our knowledge about auditory stream segregation comes from studies of such simple tone patterns, more recent studies have used a wider variety of stimuli, which has led to fundamentally new ideas about the computational and neural bases of sequential segregation.

A tacit assumption on the part of many in the field is that segregating sound sources requires cues that facilitate the differentiation of target sounds from the background.[17] A longstanding theory about the neural basis of segregation invokes the tonotopic organization (neural maps based on frequency) of the auditory pathway in a major role.[12,18,19] Specifically, this theory posits that segregation of auditory streams depends on the activation of sufficiently distinct neural populations somewhere along the auditory hierarchy, including as early as the cochlea, auditory nerve, and cochlear nucleus.[12,20] As acoustic cues are revealed over time, the separation of neural populations driven by each source would allow processes downstream to integrate events from these sources into distinct perceptual streams. In contrast to frequency-based segregation, neural populations that encode features such as bandwidth, amplitude modulation, and spatially informative cues are likely to facilitate segregation in the midbrain, thalamus, and auditory cortex. Evidence from single-unit electrophysiological recordings from the cochlear nucleus all the way to the primary auditory cortex has found activation of distinct neural populations in a manner that would support perceptual segregation of auditory objects.[20–22] Studies in humans using functional magnetic resonance imaging (fMRI), electroencephalography (EEG), and magnetoencephalography (MEG) are also consistent with the role of feature selectivity and tonotopic organization along the auditory pathway in facilitating stream segregation.[23,24]

One of the main limitations of this population separation theory is that it does not take into account

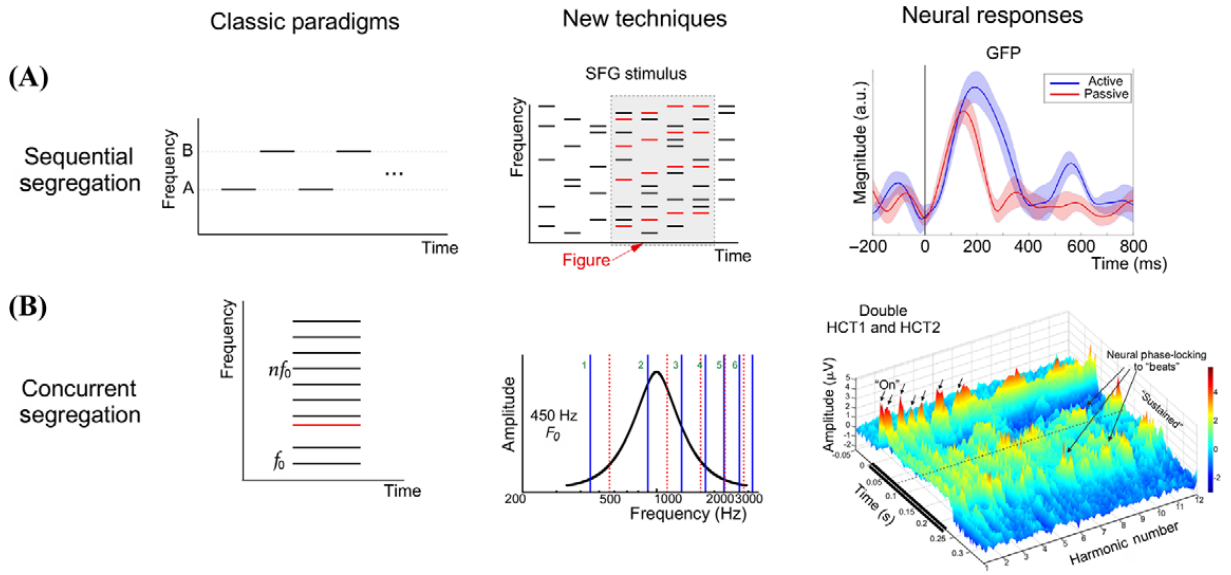

Figure 1. Examples of classic paradigms and newer techniques used in studies of auditory scene perception. (A) Left: schematic shows high and low notes repeating over time. Middle: a newer stimulus (stochastic figure ground (SFG)) employs a sequence of random inharmonic chords. If a subset of these tones repeats or changes slowly over time (shown in red), they pop out as a "figure." Right: example neural response for active (blue) and passive (red) conditions listening to SFG stimuli reproduced from Ref. 77. (B) Left: schematic of harmonic complex with a mistuned component. Middle: schematic of double harmonic complex tones (HCT) reproduced from Ref. 65 (Fig. 1A). Two HCT stimuli (solid blue and dashed red lines) are presented simultaneously. The neural frequency response function (i.e., tuning curve) is shown in black. Right: An example rate-place neural response to concurrent harmonic stimuli reproduced from Ref. 65 (Fig. 5a) from a recording site exhibiting phase-locked activity.

the relative timing of the activation of these neural populations as the scene is processed over time. A more recent theory emphasizing the importance of temporal coherence complements the population separation theory by incorporating both the selectivity of neuronal populations in the auditory system and information about the relative timing across neural responses[25] (for older examples of the importance of timing for segregation, see Ref. 5). It has been proposed that the temporal coherence mechanism tracks the evolution of acoustic features over the course of hundreds of milliseconds and that sounds that covary in time should be grouped together. In tracking temporal trajectories of sound features, temporal coherence extends the concept of common onset (i.e., frequency components that start together group together[26]). The theory posits that sound patterns that unfold in a temporally correlated fashion over hundreds of milliseconds are likely to be perceived as a group.[27] The idea of temporal coherence has been tested in computational models that have indeed shown its potential role in using cues emanating from a target source and segregating it from other sound streams that are incoherent (uncorrelated) with it.[28,29]

Recent neurophysiological evidence has provided support to the claim that the population separation theory is indeed insufficient to explain perceptual separation of auditory objects.[28] This evidence based on single-unit recordings in awake nonbehaving ferrets suggests that temporal coherence may be computed downstream from the primary auditory cortex. Along the same lines, recent work in humans using fMRI found no evidence of coherence-related blood oxygen–level dependent (BOLD) activity in the primary auditory cortex but reported significant activation of the intraparietal sulcus (IPS) and the superior temporal sulcus.[30] The experimental paradigm in this study employed a novel stimulus, known as stochastic figure ground (SFG), which consists of randomly selected inharmonic chords comprising several pure tones. When a number of these tones are changed coherently over time (by keeping them fixed or changing them slowly over several consecutive chords), a spontaneous figure pops out against the random background (Fig. 1A, middle). An EEG study using a slightly modified version of the SFG stimulus reported evidence of early and automatic computations of temporal coherence that peaked between 115 and 185 ms[31] (Fig. 1A,

right). Linear regression revealed a clear neural signature of temporal coherence in the passive listening condition that localized bilaterally to temporal regions. This evoked response was corroborated in an MEG study, in which the response pattern was stable even in the presence of noise, although its amplitude and latency varied systematically with the coherence of the figure.[32]

While temporal coherence computations appeared to evoke neural activity in the temporal cortex during passive listening, its basic profile was maintained even under attentional control (Fig. 1A, right). O'Sullivan *et al.*[31] reported a similar response pattern, but with a longer persistence, a later peak, and greater amplitude in an active condition during which listeners were engaged in detecting the figure patterns, compared with a passive condition in which participants ignored the stimuli. The topographies of both active and passive responses were similarly localized to bilateral temporal regions, suggesting a common locus for coherence computation that is modulated by attention. Together, these studies indicate that temporal coherence can be computed to some extent without focused attention, but that paying attention enhances processing. Furthermore, numerous studies point to a role of the planum temporale and the IPS in the computation of temporal coherence. However, recent work suggests a possible contribution of the auditory cortex during performance of a task involving temporal coherence as a mechanism for auditory object segregation. In an EEG study, human listeners engaged in target detection amid a competing background showed covariation of neural signatures, likely arising from both within and outside the auditory cortex.[33] Along the same lines, preliminary work in awake behaving ferrets trained to attend to a two-tone ABAB sequence revealed cortical responses with notable changes in the bandwidth of receptive fields of single neurons consistent with the postulates of the temporal coherence theory.[34] It remains unknown to what extent these neural responses reflect coherence computation taking place at the level of the auditory cortex versus projections from other brain areas. The engagement of participants in a task suggests the engagement of a broader neural network, potentially spanning the planum temporale and the IPS, two loci linked to temporal coherence processing.[26,31,32] However, the exact neural circuitry underlying such computations remains unknown.

Concurrent segregation

Complementing sequential segregation processes are mechanisms that facilitate the grouping of acoustic components that are simultaneously present into auditory objects.[35,36] A popular laboratory paradigm used to investigate concurrent segregation presents a harmonic complex consisting of simultaneous pure tones (e.g., 100, 200, 300, 400 Hz) with a common fundamental frequency (f_0, e.g., 100 Hz). Such a complex tone is almost always heard as a single auditory object and bears important similarities to naturalistic sounds, such as vowels and many musical sounds.[37] Both the f_0 and the harmonics are strong determinants of the pitch of a complex tone, with the perceived pitch typically matching the pitch of a pure tone that has the same frequency as f_0.[38]

Concurrent segregation paradigms present various stimuli that can result in segregation into two objects. One variant is the *mistuned harmonic paradigm*, which presents a single complex harmonic tone in which one of the pure tones is changed in frequency by some percentage such that it is no longer an integer multiple of the f_0 (Fig. 1B, left). This can result in the perception of two auditory objects, especially when the mistuning is relatively large: one object corresponding to the tuned portions of the complex and a second object corresponding to the mistuned tone.[39,40] Another variant is the *concurrent harmonic sounds paradigm*, which simultaneously presents two separate complex harmonic tones, each with a different f_0 (Fig. 1B, middle). The larger the difference in f_0, the more likely that the two sounds can be heard separately.[41] Similarly, the *double vowel paradigm* presents two concurrent harmonic sounds, but the amplitudes of the individual tone frequencies are shaped with a multipeaked function in order to approximate the formants in natural spoken vowels.[42–44]

A number of different computational mechanisms have been proposed to play important roles in concurrent segregation (for a review, see Ref. 41). These include place models that estimate the f_0 using peaks in the output of a cochlear frequency–based filter, autocorrelation models that use temporal fluctuations in the peripheral neural activity, and

models that suppress activity corresponding to one of the sounds in order to better perceive the nonsuppressed sound. Importantly, these different mechanisms are not mutually exclusive and have been combined in some models of segregation (e.g., see Refs. 42, 45, and 46). Additionally, models of mistuned harmonic perception have proposed the idea of harmonic templates with slots corresponding to expected values of integer multiple harmonics[47] or the importance of regular spacing of harmonics[48,49] to explain why, if a harmonic is mistuned enough, it will pop out as a separate auditory object.

Much like sequential segregation, it is clear that concurrent segregation is likely to be achieved through a number of transformations of sensory input in subcortical and cortical regions of the auditory system. Invasive neurophysiological studies have provided evidence for temporal fluctuations in neural activity, starting in the auditory nerve and continuing up to the primary auditory cortex, that could be used to estimate the presence of multiple harmonic sounds.[50–55] Meanwhile, noninvasive neurophysiological studies of the auditory cortex have identified a so-called *object related negativity* (ORN). The ORN increases in amplitude when cues for segregating concurrent sounds are more potent and occurs regardless of attention.[56–64]

A more recent study of concurrent harmonic tones in the monkey primary auditory cortex provides evidence that neurons in this region show action potential firing in response to lower harmonics of the tones, as well as beat frequencies that result from interactions of harmonics that are close in frequency.[65] This demonstrated the importance of low-frequency harmonics in segregation (Fig. 1B, right). Furthermore, the f_0s of both tones could be estimated by temporal fluctuations in firing rate that matched the frequencies of the two f_0s. This suggests that the auditory cortex can help identify the presence of two concurrent tones on the basis of pitch. Interestingly, this study also varied whether the two tones had synchronous onsets and found that the neural representations of the two concurrent tones were enhanced when they were asynchronous, as would be expected owing to the importance of this cue, as discussed above. A model of neural processing of the concurrent tones was able to closely reproduce the patterns of firing rate to different harmonics, and these patterns were used in a template-matching procedure that was able to accurately estimate the f_0s of the concurrent tones. Finally, this study found lower-frequency activity (i.e., local-field potentials) that was potentially related to the ORN discussed above. In particular, responses were isolated by comparing presentation of concurrent complex tones with an f_0 difference of four semitones to presentation of a single complex tone. As expected from human ORN studies, the concurrent tones elicited a more negative response at time points after the initial onset response to the tones, compared with the presentation of a single tone. Important issues for future research include the nature of the relationship between the action potential firing and the ORN and the role each plays in the behavioral ability to segregate sounds.

Another recent study of concurrent sound segregation examined subcortical and cortical neural activity measured using scalp electrodes in humans during a mistuned harmonic task.[66] Subcortical frequency-following responses (FFRs), likely arising from the inferior colliculus,[67] showed less phase locking with larger mistunings, and less phase locking was negatively correlated with perception of two objects. As in previous studies,[56,57] ORN amplitude increased with larger mistuning. An additional later negative response around 500 ms (N5) after the tone onset also occurred, which was largest for clearly tuned or clearly mistuned tones but smaller for more ambiguous tones. As with behavior, brain stem phase locking was negatively correlated with ORN amplitude.[66] Regression analysis showed that ORN and N5 amplitude and latency were better predictors of behavioral judgments of one versus two objects and reaction times, compared with brain stem phase locking. While this could be due to cortical processing being more directly related to conscious perception and behavioral responding, it is also possible that this is due to greater signal-to-noise ratios associated with cortical activity in comparison with brain stem activity, when measured at the scalp. The authors also used a model of auditory nerve activity to estimate neural representations of harmonicity (i.e., the extent to which a set of concurrent pure tones reinforced the same f_0 for different levels of mistuning of the second harmonic of complex tones). This simulated quantity was able to predict behavioral judgments of one versus two objects as well as ORN amplitude, consistent with the importance of some sort of place-based or time-based harmonic

template-matching process in detecting pop-out of mistuned harmonics.

Higher-level aspects of auditory scene analysis

Attention

One construct intimately related to studies of auditory scene analysis is attention. Attention can favor detection or tracking of a particular sound target, presumably by enhancing neural activity related to processing events associated with the target.[68,69] Numerous studies have indeed shown that attention plays a crucial role in scene analysis, with some evidence even suggesting that attention may be a prerequisite for segregation, although this is still a topic of debate.[70–73]

At the neurophysiological level, attention has been shown to induce rapid changes to neural responses at the level of the primary auditory cortex, altering the brain's responses to sensory cues in a direction that boosts the representation of task-relevant sounds.[74,75] In the context of auditory scene analysis, we have only recently begun understanding the impact of these rapid changes on neuronal properties for parsing complex scenes. One recent study on speech segregation recorded cortical activity using high-density intracranial electrode arrays in human participants undergoing clinical treatment for epilepsy.[76] This provided the rare opportunity to shed light on both the spatial and temporal characteristics of attention-related neural processing in listeners. Listeners were presented with two simultaneous utterances from different speakers with distinct pitches (male versus female), spectral profiles (vocal tract shape), and speaking rates. By maintaining the acoustic stimulus and manipulating which speaker was the target of attention, the experimental paradigm elucidated how much of the neural response was driven by acoustic properties of the signal versus perception-driven or attention-driven factors. The study employed a powerful new decoding technique[77–80] that estimates the input stimulus from the neural data using a reverse analysis method (Fig. 2A). This technique offers a theoretical approach to solve a decoding problem: estimating the stimulus on the basis of neural responses. By combining neural recordings obtained in response to the same input stimulus, one can compose the

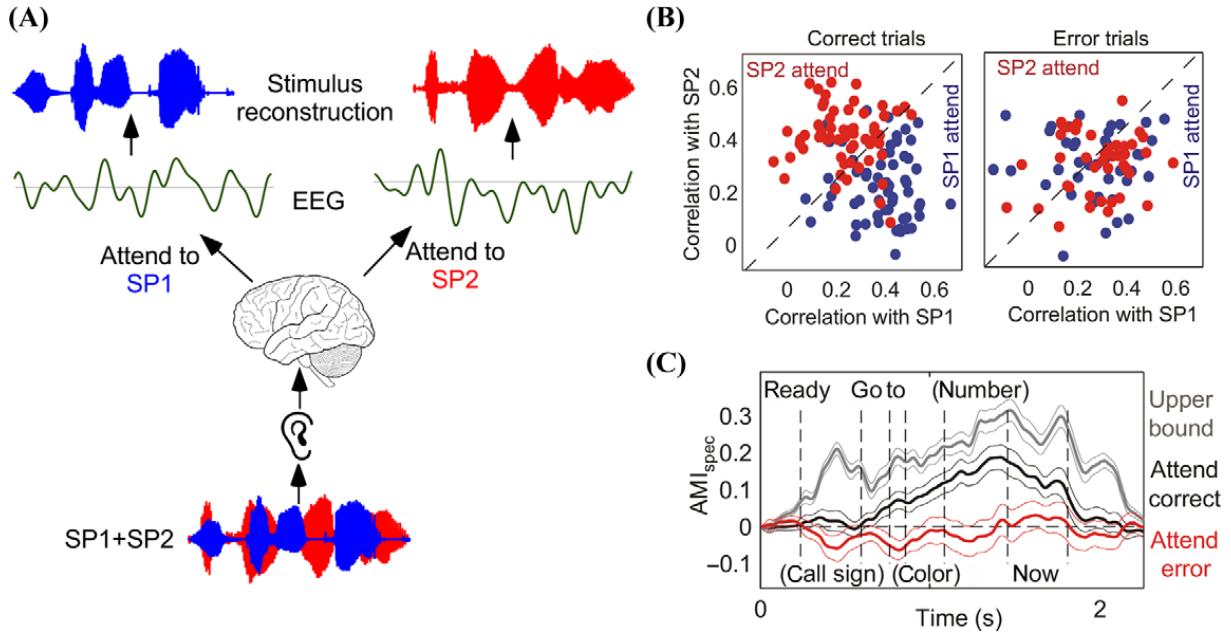

Figure 2. (A) Diagram of stimulus reconstruction technique, adapted from Ref. 148. The envelope of the speech from an attended speaker is decoded from the neural response. (B,C) Figures reproduced from Ref. 76. The plot shows correlation coefficients between the spectrogram of single speaker and reconstructed spectrograms of speaker mixture under different attentional conditions in correct and error trials. The time course of an attentional modulation index (AMI) calculated on the basis of the correlation between reconstructed spectrograms from mixtures and the original attended speaker spectrogram. Positive values of the AMI indicate shifts toward the target, while negative values indicate shifts toward the masker.

total effective stimulus that excites the neural population. In the context of attention paradigms, reverse analysis helps discern which of two concurrent speakers the subject was attending. The decoding problem is typically optimized to maximize the correlation between the reconstructed decoded signal and the speech envelope of the attended speaker. Using this technique, the findings revealed that neural responses in the nonprimary auditory cortex (posterior superior and middle temporal gyrus) were driven almost solely by the attended speaker. Specifically, stimulus reconstruction using neural responses to speech mixtures produced audible patterns that mostly reflected the attended speaker while suppressing irrelevant competing speech. However, on trials in which participants failed to track the correct target, stimulus reconstruction contained a greater presence of the nontarget speaker (Fig. 2B). An analysis of the time courses of attention-induced neural modulations revealed that error trials showed early and inappropriate attentional focus on the wrong speaker (Fig. 2C).

Similar observations have also been reported in other work using noninvasive techniques, such as MEG and EEG. The powerful stimulus reconstruction approach has enabled the tracking of attentional selection during sound mixtures. Recent work[77] has even shown that one can decode single trials of EEG activity to determine the attentional state of listeners in complex multispeaker environments. The analysis revealed a strong correlation between stimulus-decoding accuracy and participant behavior, hence establishing a link—albeit indirect—between behavior and brain responses in cases of sustained attentional deployment.

The distributed nature of neural circuitry underlying attentional selection during concurrent speech processing has been investigated in another recent study using intracranial electrode recordings.[80] The study revealed a strong modulation of neural responses in lower-level auditory cortex in the area of the superior temporal gyrus (STG); in particular, the neural response reflected a mixture of two concurrent speech narratives with a bias toward the attended target. The analysis of neural responses, both in low-frequency phase profiles (corresponding to time scales of fluctuations of the speech envelope) and high gamma power, appeared to maintain representations of the sound mixture, again with a bias for the target. The authors of the study argued that this modulation at the level of STG can be taken as evidence for an early selection process. This process is possibly then complemented by further selection in high-order regions, such as the inferior frontal cortex, anterior and inferior temporal cortex, and inferior parietal lobule. This selection process entrains low-frequency responses to the attended speech stream, allowing segregation from the interfering stream. This entrainment appeared to improve over the time course of the stimulus, suggesting a dynamic modification of the selection process as the auditory system refined its representation of the attended target.

Formation of auditory objects

Most neuroscience studies of auditory scene analysis have focused on how various stages of the auditory system process stimuli that physically vary in terms of how likely they are to be perceived as segregated. As a result, we know far less about the neural processes that most directly give rise to conscious percepts of auditory objects and streams. A major challenge to addressing this question is difficulty in directly relating perception and brain activity. In nonhuman animals, it is difficult and time consuming to train participants to report their perception, although there have been some successful efforts in purely behavioral studies to convincingly measure animals' perception[81–83] (for a review, see Ref. 84). Intrancranial recordings in animals during behavioral performance are important because, in humans, the low spatial resolution of noninvasive recordings makes it difficult to identify the potentially small populations of neurons and subtle changes in activity that correlate with perception.[85–90]

A further limitation in the literature is that relatively little is known about the role of subcortical brain areas in conscious perception of auditory objects and streams. A recent study recorded the brain stem FFR and the middle-latency response (MLR) that come from the primary auditory cortex, using scalp electrodes in humans while they performed an auditory stream–segregation task with low and high tones.[91] Unlike the responses typically measured in human studies of streaming,[72,86,92] both the FFR and MLR are brief enough to be elicited by each tone with little to no response overlap in time. Consequently, the authors were able to show that the FFR and MLR of particular tones

in the ABA– pattern were larger when participants reported hearing two segregated streams compared with when they heard one integrated stream. The authors also examined the time course of amplitude change of the FFR and MLR, showing that both the FFR and MLR changed in amplitude around the time of perceptual switches. A cross-correlation analysis indicated that MLR changes preceded FFR changes, suggesting a possible role for top-down projections from the auditory cortex to subcortical regions around the time of changes in conscious perception. The precise role of such projections remains unknown, although a number of theories of conscious visual perception point out important roles for top-down activations and other forms of connectivity.[93–96] There is also some behavioral evidence for top-down processing during auditory perception.[97]

At the cortical level, Ding and Simon[98] evaluated criteria for whether the auditory cortex processed auditory objects while participants were listening to competing speech streams (i.e., two different people talking at the same time). Using a reverse decoding method (Fig. 2A) that reconstructs the temporal envelope of the signal from the neural response,[31] the study revealed that cortical activity selectively tracks the spectrotemporal properties of the attended stream even in the presence of concurrent background speech that is spectrally overlapping with the target. Furthermore, the study showed that the neural representation of the target speech is robust against changes in the intensity of the background speaker. This reinforces the distinction between acoustic-driven neural activity and object-based or perception-based representations. Specifically, the invariant encoding of the target speech regardless of manipulations of the background is consistent with the ideas of Griffiths and Warren,[99] who claimed that auditory objects result from encoding individual sound sources as segregated from background sounds. These findings using two concurrent speakers are consistent with target-focused attentional responses reported using other complex scene paradigms, including speech with interfering background noise[100] and a regular tone stream in the presence of background tone clouds[28] or in the presence of competing tone streams with different presentation rates.[101] Moreover, the auditory object representation appears to evolve at successive stages of auditory processing with greater correspondence to perception (as opposed to stimulus encoding) at later stages of processing.

While the studies reviewed above suggest the importance of rhythmic brain activity that phase locks with the rhythm of auditory patterns, there is very little causal (as opposed to correlational) evidence tying such brain responses to processes of scene segregation and perception. A recent study, however, provided evidence that rhythmic brain activity is indeed important for auditory segregation of tone patterns[102] (for additional commentary, see Ref. 127). They used rhythmic patterns of transcranial electrical current stimulation directed through both auditory cortices that was either in phase or out of phase with an isochronous pattern of target tones embedded in background noise. When the electrical current stimulated the auditory cortex in phase with the tones, participants were better able to detect the target tones, compared with when the current was out of phase with the tones.

Another novel approach to studying auditory perception that was recently used is the analysis of how variation in genotypes predicts perception.[103] In this study, the authors qualified the dopamine-related catechol-*O*-methyltransferase (COMT) gene and the serotonin 2A receptor (HTR2A) gene in healthy volunteers, who also performed several auditory and visual bistable perception tasks, including auditory stream segregation. However, instead of quantifying the likelihood of perceiving one or two objects or streams, this study measured the number of switches between percepts during prolonged exposure to stimuli (cf. Refs. 104 and 105). The number of perceptual switches in different tasks was significantly correlated both within and across modalities, suggesting common or similar brain mechanisms underlying the tendency to switch. For auditory bistable tasks, the number of perceptual switches was greater for those with the *COMT* genotype that had two copies of the Met allele, compared with those with one or no Met allele. However, for the visual tasks, there was no difference between people with different *COMT* genotypes and there was only a marginal effect of *HTR2A* genotype for one of the visual tasks. This suggests that the two genes investigated in this study may not play substantial roles in the common mechanisms underlying bistable perceptual switching across the senses that were suggested by the correlations in behavior.

However, it is currently unclear which brain areas exhibit differences in dopamine and serotonin function that might be associated with altered perceptual switching.

The dynamics of bistable perception in the context of auditory stream segregation were further explored in recent work by Rankin *et al.*,[106] which described one of the only recent neuromechanistic models of stream segregation and bistable percepts. This study simulated activity underlying behavioral responses of listeners using alternating ABA– tone sequences, particularly the alternation between percepts of one stream versus two streams as the stimulus unfolds over time. Unlike previous models of bistable auditory perception, in which stimulus elements were first mapped into discrete perceptual units before some form of competition between these units takes place (e.g., Ref. 107), the Rankin *et al.* model operated directly on the stimulus features and incorporated a number of processes possibly related to neuronal competition and perceptual encoding. The model's architecture incorporated a tonotopic organization with recurrent excitation (using NMDA-like synaptic dynamics) to embody neuronal memory and stability of percepts as the stimulus evolves over time. It also included a form of global inhibition, empirically found to best predict the relationship between the frequency difference between the high and low tones and the relative durations of the integrated and segregated percepts. The model sheds light on the dynamic nature of neuronal responses in the auditory cortex and the role of multiple mechanisms with different time constants in giving rise to bistable percepts with two-tone ABA triplet sequences. The interplay of adaptation, especially at intermediate and long time constants with the presence of intrinsic noise, explains the ambiguous interpretation of tone sequences presented over long times as switching between a grouped single-stream percept and a segregated two-stream percept. This back-and-forth toggling between a segregated and a grouped percept while listening to tone sequences can be conceptually viewed as the brain's way of weighing evidence about both interpretations of the scene. A recent study by Barniv and Nelken[108] presented a theoretical formulation in support of this evidence-accumulation view.

Informational masking

Informational masking (IM) is a perceptual phenomenon describing how the brain fails to detect suprathreshold target tones relative to other masker tones, even though they are not processed by the same channels in the periphery as maskers.[109–111] In one IM paradigm, a sound target is embedded in a cloud of maskers that does not overlap with the target in time or frequency, yet can mask its presence depending on the choice of parameters of the stimulus (e.g., density of masker tone cloud, spectral separation between target and neighboring maskers) (Fig. 3A).

Though extensively used in studies of auditory perception and masking, IM has become a popular tool to probe perceptual awareness of auditory objects. Its appeal in studies of auditory scene

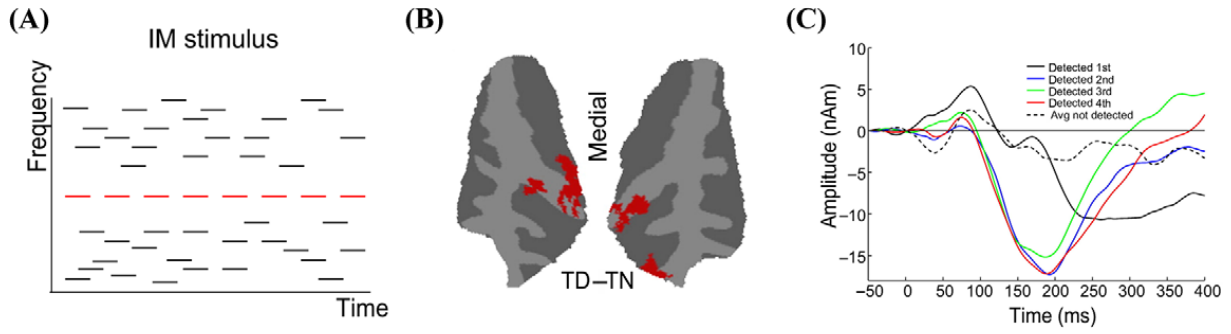

Figure 3. (A) Schematic of an informational masking stimulus. It consists of a tone cloud of masker notes with tone frequencies at randomly chosen time and frequency values. A repeating target note (shown in red) sometimes stands out from the background if no masker is present within a fixed frequency distance from the target (called a spectral protection region around the target). (B,C) Figures reproduced from Ref. 112. (B) A surface map of BOLD response for trials where the target was detected (TD) versus not detected (TN). (C) MEG source waves averaged across subjects and hemispheres for detected (solid lines) and undetected (dashed line) target tones. The figure highlights long-latency negativity (ARN) only for detected targets.

analysis is that the physical stimulus can remain unchanged while evaluating neural responses in cases where the target is detectable versus nondetectable. In doing so, we can begin to dissociate the neural responses driven by the physical cues in the stimulus versus perceptual abstractions of such a stimulus. In one study,[112] fMRI and MEG techniques were combined to provide a high-resolution temporal and spatial description of the emergence of auditory awareness in the auditory cortex. BOLD activity revealed a differential increase in neural activity in the auditory cortex (Fig. 3B): detected targets induced stronger activity than undetected targets in medial Heschl's gyrus (thought to be the location of the primary auditory cortex in humans). These undetected targets themselves induced greater activity in the posterior STG (containing portions of the secondary auditory cortex) than a random masker baseline without any target. Importantly, the contrast of neural activity between perceived and undetected targets in the auditory cortex supports claims of cortical involvement in conscious perception. The BOLD responses localized this effect to the primary auditory cortex and away from the secondary areas, although a region of interest–based analysis suggested activation of both the primary and secondary auditory cortices.[112]

To further reveal the involvement of the auditory cortex in conscious perception of targets, analysis of MEG recordings in the same study confirmed a specific neural signature for target detection consisting of a long-latency negativity called the awareness-related negativity (ARN) first reported by Gutschalk *et al.*[113] (Fig. 3C). This relatively long response to detected targets could reflect the auditory cortex receiving recurrent projections from higher-order cortical areas, in line with similar ideas about the visual system mentioned earlier.[93,95] This idea of recurrent feedback could help account for conflicting notions about the nature of sound representation in the primary auditory cortex, and particularly the extent to which neural responses in the core auditory cortex reflect sensory features or higher-level processing.[114] In particular, if the auditory cortex is principally encoding parameters of the stimulus, then possible recurrent feedback from higher-order cortical areas could modulate neural activity in the primary cortex in a manner that reflects higher-level perceptual representations and awareness of elements in the scene. This would agree with the delayed latency of the ARN typically observed around 150 ms. However, a major confounding factor is that of attention, which not only modulates responses in the auditory cortex but may also affect processing of a target sound in an IM paradigm.

Most studies of IM and some on auditory stream segregation have focused on auditory cortex activity without looking for activity in wider cortical or subcortical areas. However, a recent study found single neurons in monkey claustrum—an area previously theorized to be involved in consciousness[115]—that reflected whether a target tone was or was not presented in background noise.[116] Other studies have found activity in parietal and frontal areas,[117,118] which could reflect activation of attention networks. These, along with the studies on stream segregation discussed above, suggest the importance of considering multiple brain areas and their interactions in comprehensive theories of the neural basis of auditory scene analysis.

Understanding processing of realistic auditory scenes

While there is increasing interest within the research community in using more complex stimuli in studies of auditory scene perception, most stimuli remain relatively impoverished in that they rely on different configurations of tones or noises. In recent years, however, several lines of research have begun to illuminate perception of complex auditory scenes that bear more resemblance to real soundscapes (e.g., Refs. 119–121). A topic that has been studied by a few different research groups is change deafness, the failure to notice when a sound is added, removed, or replaced in a scene composed of multiple sounds (Fig. 4A; see reviews in Refs. 69, 122, 123). Typical studies of change deafness present several recognizable sounds at the same time, followed by the same set of sounds again or with one of the original sounds changing. An ERP study on change deafness found that the long-latency sensory N1 response (at around 120 ms) to the onset of the second scene was larger when listeners successfully detected changes, compared with when they did not detect changes.[124] In this study, a later P3b response was also larger for cases with detected compared with undetected changes. These N1 and P3b modulations may reflect perceptual awareness of the change and subsequent memory

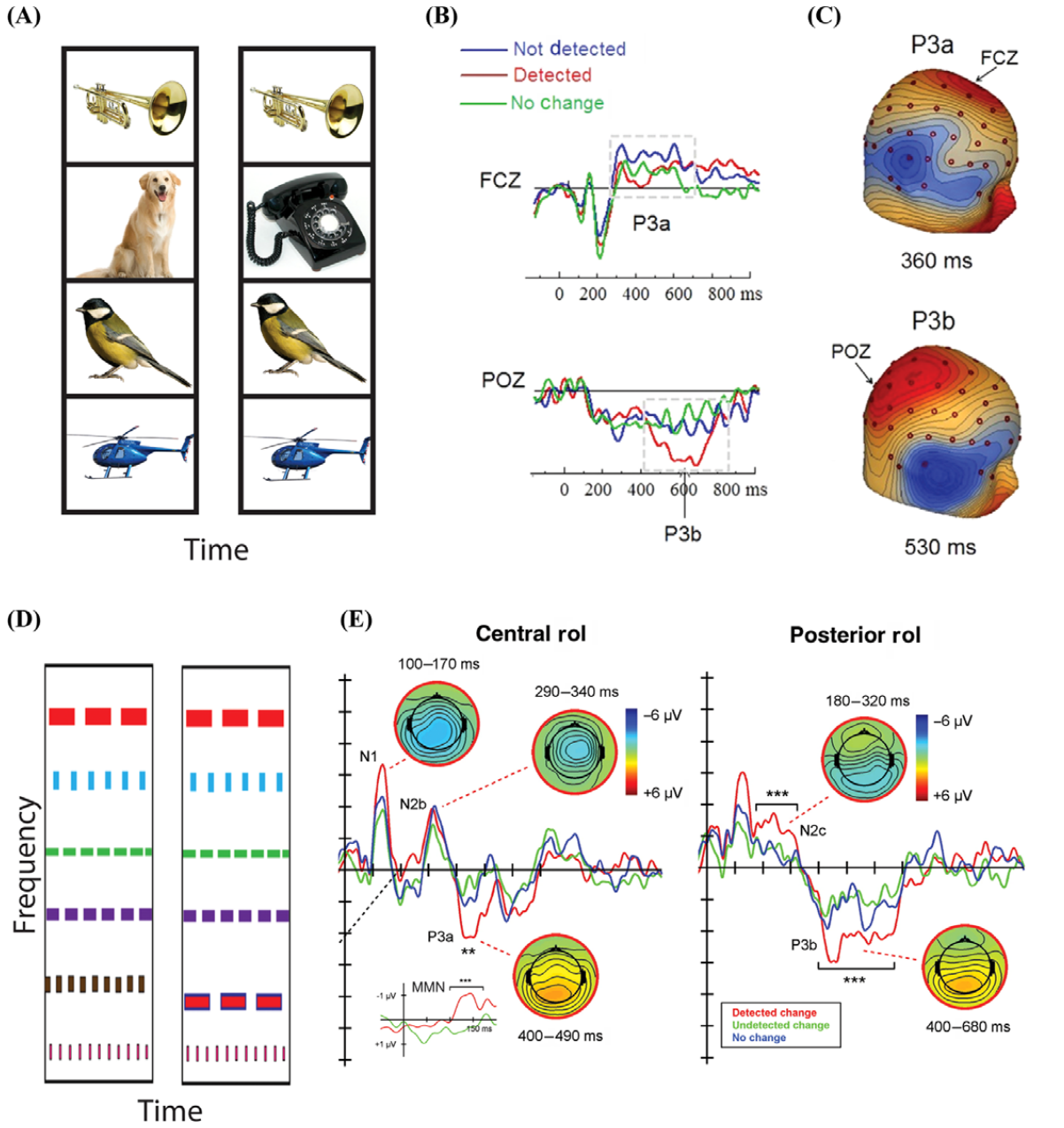

Figure 4. (A) Schematic of recognizable sounds used in change deafness experiments. A set of sounds are played at the same time and after a brief delay, the same set of sounds are played with no change, or one of the sounds is changed (as shown with dog turning into phone ringing). (B) Electrical brain responses (reprinted from Ref. 128 with permission from Elsevier) showing P3a and P3b responses that are enhanced on trials with a detected change, compared with trials with no change or a nondetected change (note positive voltage is plotted downward). (C) Topographies of the difference between detected and nondetected changes in Ref. 128 for the P3a and P3b. (D) Schematic of bandpass noise burst patterns used in change deafness experiments, with change in second-lowest frequency pattern. (E) Electrical brain responses (reprinted from Ref. 129 with permission from Elsevier), showing several enhanced components for detected changes.

updating or other cognitive consequences of awareness, respectively.[125–127] Another study by the same group used a different set of recognizable sounds and unrecognizable versions of the sounds but found no N1 modulation for either type of sound.[128] This study did find the P3b to be modulated as in the prior study, and further found a P3a response modulation for detected unrecognizable sounds (Fig. 4B,C–c), possibly reflecting attention orienting.[126] Studies using simpler scenes composed

of multiple streams of bandpass noise bursts (Fig. 4D) found that detected changes were associated with modulations of a number of ERP components, including the N1, P3a, and P3b (Fig. 4E),[129] as found in the studies by Gregg and colleagues. Additional components that were larger for detected changes included the N2 and mismatch negativity. Another recent study used streams of pure tone sounds and recorded MEG responses that were modulated during successful change detection, starting around 100 ms after the onset of the change,[130] consistent with the studies just discussed that showed N1 modulations.

Finally, an fMRI study on change deafness found greater activity in the anterior cingulate cortex and right insula during successful change detection trials compared with unsuccessful trials.[131] These successful detection trials were also associated with stronger functional connectivity between the right auditory cortex and both the left insula and the left inferior frontal cortex regions, when compared with unsuccessful trials. In contrast, the right superior temporal sulcus showed stronger functional connectivity with the auditory cortex for unsuccessful trials compared with successful trials. While these findings need to be replicated, the anterior cingulate modulation is consistent with a neural orienting response.[132,133] This is suggested by the P3a modulations discussed above and the fact that P3a during oddball processing is in part generated in the anterior cingulate.[134] In contrast, insula activation could be related to interoceptive feelings related to conscious detection of changes or error detection.[135,136]

Conclusions

Recent studies of auditory scene perception have made considerable progress in advancing our understanding of auditory segregation and object formation. This is in part the result of using a wider variety of computational methods and experimental techniques in humans and different nonhuman animal species. This increasing diversity of approaches is offering a more complete picture of different phenomena related to auditory scene perception. As such, the discovery of convergent support for a particular quantitative theory using different techniques, species, and stimulus paradigms can provide more convincing evidence for that theory. Owing to the importance of computational theory development, we hope more researchers will begin publishing such work in the near future. This effort should include more neuromechanistic models that explain how particular computations are carried out in the brain using realistic cellular, synaptic, and circuit mechanisms.

While it is a well-established goal in the field of auditory scene analysis to study higher-level aspects of scene perception,[5] the use of relatively simple, artificial sounds has severely curtailed progress in this effort. Apart from attention,[68,69,137] other high-level aspects of auditory scene perception have largely been ignored, with a few exceptions.[138,139] However, more complex aspects of auditory scene perception have begun to garner interest in recent years. For example, scientists have developed paradigms using more intricate stimulus structures that can theoretically be parsed into more than just two objects or streams. Furthermore, IM paradigms impede the perception of a detectable target using a complex array of background sounds. Finally, the use of realistic or natural scenes, including concurrent speech utterances and challenging listening paradigms, is likely to give us a broader and perhaps deeper understanding of auditory scene perception. By using recognizable sounds (e.g., Refs. 12, 98, and 124), many studies have further highlighted the importance of studying real-world sounds, including speech, music, and other environmental sounds. Moreover, studies of change detection have taken the lead in uncovering the extent to which semantic knowledge[128,140,141] and object-based attention[142] influence perception. In the future, these behavioral findings can be leveraged to uncover the neural mechanisms of high-level processing of auditory scenes.

Breaking new ground in our understanding of auditory scene perception is also leading to new applications spanning engineering systems to medical technology. For example, a better understanding of neural processing of meaningful sounds has shown promise for neural decoding–based communication with severely brain-damaged individuals.[143] Along the same lines, computational models mimicking cocktail party listening (i.e., when multiple people are talking) are gradually increasing in complexity and providing more integrated architectures that span both low-level and high-level processing of realistic scenes. Indeed, models of auditory scene analysis are now extending beyond simple scenes composed of tones and sparse

sound patterns to more complex and challenging scenarios (e.g., concurrent speakers in noisy, natural environments (for review, see Ref. 144)). Many such models are now being compared, and some even outperform state-of-the-art systems developed using pure engineering principles that are tailored to specific applications using extensive training data.[29,145–147]

Acknowledgments

J.S.S. was supported by the Army Research Office (W9IINF-I2-I-0256) and the Office of Naval Research (N000141612879); M.E. was supported by the National Institutes of Health (R01HL133043) and the Office of Naval Research (N000141010278, N000141612045, and N000141210740).

Conflicts of interest

The authors declare no conflicts of interest.

References

1. Bregman, A.S. & J. Campbell. 1971. Primary auditory stream segregation and perception of order in rapid sequences of tones. *J. Exp. Psychol.* **89:** 244–249.
2. Miller, G.A. & G.A. Heise. 1950. The trill threshold. *J. Acoust. Soc. Am.* **22:** 637–638.
3. Van Noorden, L.P.A.S. 1975. Temporal coherence in the perception of tone sequences. Unpublished PhD dissertation. Eindhoven University of Technology, Eindhoven.
4. Warren, R.M., C.J. Obusek, R.M. Farmer, *et al.* 1969. Auditory sequence: confusion of patterns other than speech or music. *Science* **164:** 586–587.
5. Bregman, A.S. 1990. *Auditory Scene Analysis: The Perceptual Organization of Sound.* Cambridge, MA: MIT Press.
6. Gepshtein, S. & M. Kubovy. 2000. The emergence of visual objects in space–time. *Proc. Natl. Acad. Sci. U.S.A.* **97:** 8186–8191.
7. Wagemans, J., J.H. Elder, M. Kubovy, *et al.* 2012. A century of Gestalt psychology in visual perception: I. Perceptual grouping and figure-ground organization. *Psychol. Bull.* **138:** 1172–1217.
8. Wagemans, J., J. Feldman, S. Gepshtein, *et al.* 2012. A century of Gestalt psychology in visual perception: II. Conceptual and theoretical foundations. *Psychol. Bull.* **138:** 1218–1252.
9. Denham, S. & I. Winkler. 2015. Auditory perceptual organization. In *Oxford Handbook of Perceptual Organization.* Vol. 601–620. J. Wagemans, Ed. Oxford: Oxford University Press.
10. Kubovy, M. & D. Van Valkenburg. 2001. Auditory and visual objects. *Cognition* **80:** 97–126.
11. Fishman, Y.I., J.C. Arezzo & M. Steinschneider. 2004. Auditory stream segregation in monkey auditory cortex: effects of frequency separation, presentation rate, and tone duration. *J. Acoust. Soc. Am.* **116:** 1656–1670.
12. Hartmann, W.M. & D. Johnson. 1991. Stream segregation and peripheral channeling. *Music Percept.* **9:** 155–184.
13. Micheyl, C., B. Tian, R.P. Carlyon, *et al.* 2005. Perceptual organization of tone sequences in the auditory cortex of awake macaques. *Neuron* **48:** 139–148.
14. Moore, B.C.J. & H.E. Gockel. 2012. Properties of auditory stream formation. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **367:** 919–931.
15. Snyder, J.S. & C. Alain. 2007. Toward a neurophysiological theory of auditory stream segregation. *Psychol. Bull.* **133:** 780–799.
16. Bregman, A.S. 1978. Auditory streaming is cumulative. *J. Exp. Psychol. Hum. Percept. Perform.* **4:** 380–387.
17. Moore, B.C.J. & H. Gockel. 2002. Factors influencing sequential stream segregation. *Acta Acust. United Ac.* **88:** 320–333.
18. Beauvois, M.W. & R. Meddis. 1996. Computer simulation of auditory stream segregation in alternating-tone sequences. *J. Acoust. Soc. Am.* **99:** 2270–2280.
19. Fishman, Y.I., D.H. Reser, J.C. Arezzo, *et al.* 2001. Neural correlates of auditory stream segregation in primary auditory cortex of the awake monkey. *Hear. Res.* **151:** 167–187.
20. Pressnitzer, D., M. Sayles, C. Micheyl, *et al.* 2008. Perceptual organization of sound begins in the auditory periphery. *Curr. Biol.* **18:** 1124–1128.
21. Itatani, N. & G.M. Klump. 2011. Neural correlates of auditory streaming of harmonic complex sounds with different phase relations in the songbird forebrain. *J. Neurophysiol.* **105:** 188–199.
22. Micheyl, C., R.P. Carlyon, A. Gutschalk, *et al.* 2007. The role of auditory cortex in the formation of auditory streams. *Hear. Res.* **229:** 116–131.
23. Gutschalk, A. & A.R. Dykstra. 2014. Functional imaging of auditory scene analysis. *Hear. Res.* **307:** 98–110.
24. Simon, J.Z. 2017. Neurophysiology and neuroimaging of auditory stream segregation in humans. In *Springer Handbook of Auditory Research.* J. Middlebrooks, J.Z. Simon, A.N. Popper, & R.R. Fay, Eds. New York: Springer.
25. Shamma, S.A., M. Elhilali & C. Micheyl. 2011. Temporal coherence and attention in auditory scene analysis. *Trends Neurosci.* **34:** 114–123.
26. Darwin, C.J. & R.P. Carlyon. 1995. Auditory grouping. In *Hearing.* B.C.J. Moore, Ed.: 387–424. Orlando, FL: Academic Press.
27. Micheyl, C., H. Kreft, S. Shamma, *et al.* 2013. Temporal coherence versus harmonicity in auditory stream formation. *J. Acoust. Soc. Am.* **133:** EL188–EL194.
28. Elhilali, M., L. Ma, C. Micheyl, *et al.* 2009. Temporal coherence in the perceptual organization and cortical representation of auditory scenes. *Neuron* **61:** 317–329.
29. Krishnan, L., M. Elhilali & S. Shamma. 2014. Segregating complex sound sources through temporal coherence. *PLoS Comput. Biol.* **10:** e1003985.
30. Teki, S., M. Chait, S. Kumar, *et al.* 2011. Brain bases for auditory stimulus-driven figure–ground segregation. *J. Neurosci.* **31:** 164–171.
31. O'Sullivan, J.A., S.A. Shamma & E.C. Lalor. 2015. Evidence for neural computations of temporal coherence in an auditory scene and their enhancement during active listening. *J. Neurosci.* **35:** 7256–7263.

32. Teki, S., N. Barascud, S. Picard, *et al.* 2016. Neural correlates of auditory figure–ground segregation based on temporal coherence. *Cereb. Cortex* **26:** 3669–3680.
33. Toth, B., Z. Kocsis, G.P. Haden, *et al.* 2016. EEG signatures accompanying auditory figure–ground segregation. *Neuroimage* **141:** 108–119.
34. Shamma, S., M. Elhilali, L. Ma, *et al.* 2013. Temporal coherence and the streaming of complex sounds. In *Basic Aspects of Hearing: Physiology and Perception.* Vol. 787. B.C.J. Moore, R.D. Patterson, I.M. Winter, *et al.*, Eds.: 535–543. New York: Springer.
35. Alain, C. 2007. Breaking the wave: effects of attention and learning on concurrent sound perception. *Hear. Res.* **229:** 225–236.
36. Carlyon, R.P. 2004. How the brain separates sounds. *Trends Cogn. Sci.* **8:** 465–471.
37. Helmholtz, H.V. & A.J. Ellis. 1885. *On the Sensations of Tone as a Physiological Basis for the Theory Of Music.* London: Longmans, Green.
38. Yost, W.A. 2009. Pitch perception. *Atten. Percept. Psychophys.* **71:** 1701–1715.
39. Hartmann, W.M., S. McAdams & B.K. Smith. 1990. Hearing a mistuned harmonic in an otherwise periodic complex tone. *J. Acoust. Soc. Am.* **88:** 1712–1724.
40. Moore, B.C.J., B.R. Glasberg & R.W. Peters. 1986. Thresholds for hearing mistuned partials as separate tones in harmonic complexes. *J. Acoust. Soc. Am.* **80:** 479–483.
41. Micheyl, C. & A.J. Oxenham. 2010. Pitch, harmonicity and concurrent sound segregation: psychoacoustical and neurophysiological findings. *Hear. Res.* **266:** 36–51.
42. Assmann, P.F. & Q. Summerfield. 1990. Modeling the perception of concurrent vowels: vowels with different fundamental frequencies. *J. Acoust. Soc. Am.* **88:** 680–697.
43. Assmann, P.F. & Q. Summerfield. 1994. The contribution of wave-form interactions to the perception of concurrent vowels. *J. Acoust. Soc. Am.* **95:** 471–484.
44. Culling, J.F. & C.J. Darwin. 1993. Perceptual separation of simultaneous vowels: within and across-formant grouping by F_0. *J. Acoust. Soc. Am.* **93:** 3454–3467.
45. deCheveigne, A. 1997. Concurrent vowel identification. III. A neural model of harmonic interference cancellation. *J. Acoust. Soc. Am.* **101:** 2857–2865.
46. Meddis, R. & M.J. Hewitt. 1992. Modeling the identification of concurrent vowels with different fundamental frequencies. *J. Acoust. Soc. Am.* **91:** 233–245.
47. deCheveigne, A. 1997. Harmonic fusion and pitch shifts of mistuned partials. *J. Acoust. Soc. Am.* **102:** 1083–1087.
48. Roberts, B. & J.M. Brunstrom. 1998. Perceptual segregation and pitch shifts of mistuned components in harmonic complexes and in regular inharmonic complexes. *J. Acoust. Soc. Am.* **104:** 2326–2338.
49. Roberts, B. & J.M. Brunstrom. 2001. Perceptual fusion and fragmentation of complex tones made inharmonic by applying different degrees of frequency shift and spectral stretch. *J. Acoust. Soc. Am.* **110:** 2479–2490.
50. Palmer, A.R. 1990. The representation of the spectra and fundamental frequencies of steady-state single- and double-vowel sounds in the temporal discharge patterns of guinea pig cochlear-nerve fibers. *J. Acoust. Soc. Am.* **88:** 1412–1426.
51. Sinex, D.G., J.H. Sabes & H. Li. 2002. Responses of inferior colliculus neurons to harmonic and mistuned complex tones. *Hear. Res.* **168:** 150–162.
52. Sinex, D.G., H. Guzik, H.Z. Li, *et al.* 2003. Responses of auditory nerve fibers to harmonic and mistuned complex tones. *Hear. Res.* **182:** 130–139.
53. Sinex, D.G. & H. Li. 2007. Responses of inferior colliculus neurons to double harmonic tones. *J. Neurophysiol.* **98:** 3171–3184.
54. Sinex, D.G. 2008. Responses of cochlear nucleus neurons to harmonic and mistuned complex tones. *Hear. Res.* **238:** 39–48.
55. Tramo, M.J., P.A. Cariani, B. Delgutte, *et al.* 2001. Neurobiological foundations for the theory of harmony in western tonal music. *Ann. N.Y. Acad. Sci.* **930:** 92–116.
56. Alain, C., B.M. Schuler & K.L. McDonald. 2002. Neural activity associated with distinguishing concurrent auditory objects. *J. Acoust. Soc. Am.* **111:** 990–995.
57. Alain, C., S.R. Arnott & T.W. Picton. 2001. Bottom-up and top-down influences on auditory scene analysis: evidence from event-related brain potentials. *J. Exp. Psychol. Hum. Percept. Perform.* **27:** 1072–1089.
58. Fishman, Y.I., I.O. Volkov, M.D. Noh, *et al.* 2001. Consonance and dissonance of musical chords: neural correlates in auditory cortex of monkeys and humans. *J. Neurophysiol.* **86:** 2761–2788.
59. Fishman, Y.I. & M. Steinschneider. 2010. Neural correlates of auditory scene analysis based on inharmonicity in monkey primary auditory cortex. *J. Neurosci.* **30:** 12480–12494.
60. Hautus, M.J. & B.W. Johnson. 2005. Object-related brain potentials associated with the perceptual segregation of a dichotically embedded pitch. *J. Acoust. Soc. Am.* **117:** 275–280.
61. Johnson, B.W., M. Hautus & W.C. Clapp. 2003. Neural activity associated with binaural processes for the perceptual segregation of pitch. *Clin. Neurophysiol.* **114:** 2245–2250.
62. Johnson, B.W. & M.J. Hautus. 2010. Processing of binaural spatial information in human auditory cortex: neuromagnetic responses to interaural timing and level differences. *Neuropsychologia* **48:** 2610–2619.
63. Lipp, R., P. Kitterick, Q. Summerfield, *et al.* 2010. Concurrent sound segregation based on inharmonicity and onset asynchrony. *Neuropsychologia* **48:** 1417–1425.
64. Sanders, L.D., A.S. Joh, R.E. Keen, *et al.* 2008. One sound or two? Object-related negativity indexes echo perception. *Percept. Psychophys.* **70:** 1558–1570.
65. Fishman, Y.I., M. Steinschneider & C. Micheyl. 2014. Neural representation of concurrent harmonic sounds in monkey primary auditory cortex: implications for models of auditory scene analysis. *J. Neurosci.* **34:** 12425–12443.
66. Bidelman, G.M. & C. Alain. 2015. Hierarchical neurocomputations underlying concurrent sound segregation: connecting periphery to percept. *Neuropsychologia* **68:** 38–50.
67. Chandrasekaran, B. & N. Kraus. 2010. The scalp-recorded brainstem response to speech: neural origins and plasticity. *Psychophysiology* **47:** 236–246.
68. Alain, C. & L.J. Bernstein. 2008. From sounds to meaning: the role of attention during auditory scene analysis. *Curr. Opin. Otolaryngol. Head Neck Surg.* **16:** 485–489.

69. Snyder, J.S., M.K. Gregg, D.M. Weintraub, *et al.* 2012. Attention, awareness, and the perception of auditory scenes. *Front. Psychol.* **3:** 15.
70. Macken, W.J., S. Tremblay, R.J. Houghton, *et al.* 2003. Does auditory streaming require attention? Evidence from attentional selectivity in short-term memory. *J. Exp. Psychol. Hum. Percept. Perform.* **29:** 43–51.
71. Carlyon, R.P., R. Cusack, J.M. Foxton, *et al.* 2001. Effects of attention and unilateral neglect on auditory stream segregation. *J. Exp. Psychol. Hum. Percept. Perform.* **27:** 115–127.
72. Snyder, J.S., C. Alain & T.W. Picton. 2006. Effects of attention on neuroelectric correlates of auditory stream segregation. *J. Cogn. Neurosci.* **18:** 1–13.
73. Sussman, E.S., J. Horvath, I. Winkler, *et al.* 2007. The role of attention in the formation of auditory streams. *Percept. Psychophys.* **69:** 136–152.
74. Fritz, J.B., S.V. David, S. Radtke-Schuller, *et al.* 2010. Adaptive, behaviorally gated, persistent encoding of task-relevant auditory information in ferret frontal cortex. *Nat. Neurosci.* **13:** 1011–1019.
75. Fritz, J., S. Shamma, M. Elhilali, *et al.* 2003. Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex. *Nat. Neurosci.* **6:** 1216–1223.
76. Mesgarani, N. & E.F. Chang. 2012. Selective cortical representation of attended speaker in multi-talker speech perception. *Nature* **485:** 233–236.
77. O'Sullivan, J.A., A.J. Power, N. Mesgarani, *et al.* 2015. Attentional selection in a cocktail party environment can be decoded from single-trial EEG. *Cereb. Cortex* **25:** 1697–1706.
78. Pasley, B.N., S.V. David, N. Mesgarani, *et al.* 2012. Reconstructing speech from human auditory cortex. *PLoS Biol.* **10:** e1001251.
79. Stanley, G.B., F.F. Li & Y. Dan. 1999. Reconstruction of natural scenes from ensemble responses in the lateral geniculate nucleus. *J. Neurosci.* **19:** 8036–8042.
80. Zion Golumbic, E.M., N. Ding, S. Bickel, *et al.* 2013. Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party." *Neuron* **77:** 980–991.
81. Christison-Lagay, K.L. & Y.E. Cohen. 2014. Behavioral correlates of auditory streaming in rhesus macaques. *Hear. Res.* **309:** 17–25.
82. Ma, L., C. Micheyl, P. Yin, *et al.* 2010. Behavioral measures of auditory streaming in ferrets (*Mustela putorius*). *J. Comp. Psychol.* **124:** 317–330.
83. MacDougall-Shackleton, S.A., S.H. Hulse, T.Q. Gentner, *et al.* 1998. Auditory scene analysis by European starlings (*Sturnus vulgaris*): perceptual segregation of tone sequences. *J. Acoust. Soc. Am.* **103:** 3581–3587.
84. Bee, M.A. & C. Micheyl. 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.
85. Cusack, R. 2005. The intraparietal sulcus and perceptual organization. *J. Cogn. Neurosci.* **17:** 641–651.
86. Gutschalk, A., C. Micheyl, J.R. Melcher, *et al.* 2005. Neuromagnetic correlates of streaming in human auditory cortex. *J. Neurosci.* **25:** 5382–5388.
87. Hill, K.T., C.W. Bishop, D. Yadav, *et al.* 2011. Pattern of BOLD signal in auditory cortex relates acoustic response to perceptual streaming. *BMC Neurosci.* **12:** 85.
88. Kondo, H.M. & M. Kashino. 2009. Involvement of the thalamocortical loop in the spontaneous switching of percepts in auditory streaming. *J. Neurosci.* **29:** 12695–12701.
89. Kondo, H.M. & M. Kashino. 2007. Neural mechanisms of auditory awareness underlying verbal transformations. *Neuroimage* **36:** 123–130.
90. Schadwinkel, S. & A. Gutschalk. 2011. Transient bold activity locked to perceptual reversals of auditory streaming in human auditory cortex and inferior colliculus. *J. Neurophysiol.* **105:** 1977–1983.
91. Yamagishi, S., S. Otsuka, S. Furukawa, *et al.* 2016. Subcortical correlates of auditory perceptual organization in humans. *Hear. Res.* **339:** 104–111.
92. Elhilali, M., J. Xiang, S.A. Shamma, *et al.* 2009. Interaction between attention and bottom-up saliency mediates the representation of foreground and background in an auditory scene. *PLoS Biol.* **7:** e1000129.
93. Bullier, J. 2001. Feedback connections and conscious vision. *Trends Cogn. Sci.* **5:** 369–370.
94. Dehaene, S. & J.P. Changeux. 2011. Experimental and theoretical approaches to conscious processing. *Neuron* **70:** 200–227.
95. Hochstein, S. & M. Ahissar. 2002. View from the top: hierarchies and reverse hierarchies in the visual system. *Neuron* **36:** 791–804.
96. Tononi, G., M. Boly, M. Massimini, *et al.* 2016. Integrated information theory: from consciousness to its physical substrate. *Nat. Rev. Neurosci.* **17:** 450–461.
97. Nahum, M., I. Nelken & M. Ahissar. 2008. Low-level information and high-level perception: the case of speech in noise. *PLoS Biol.* **6:** e126.
98. Ding, N. & J.Z. Simon. 2012. Emergence of neural encoding of auditory objects while listening to competing speakers. *Proc. Natl. Acad. Sci. U.S.A.* **109:** 11854–11859.
99. Griffiths, T.D. & J.D. Warren. 2004. What is an auditory object? *Nat. Rev. Neurosci.* **5:** 887–892.
100. Ding, N. & J.Z. Simon. 2013. Adaptive temporal encoding leads to a background-insensitive cortical representation of speech. *J. Neurosci.* **33:** 5728–5735.
101. Xiang, J., J. Simon & M. Elhilali. 2010. Competing streams at the cocktail party: exploring the mechanisms of attention and temporal integration. *J. Neurosci.* **30:** 12084–12093.
102. Riecke, L., A.T. Sack & C.E. Schroeder. 2015. Endogenous delta/theta sound-brain phase entrainment accelerates the buildup of auditory streaming. *Curr. Biol.* **25:** 3196–3201.
103. Kondo, H.M., N. Kitagawa, M.S. Kitamura, *et al.* 2012. Separability and commonality of auditory and visual bistable perception. *Cereb. Cortex* **22:** 1915–1922.
104. Pressnitzer, D. & J.M. Hupé. 2006. Temporal dynamics of auditory and visual bistability reveal common principles of perceptual organization. *Curr. Biol.* **16:** 1351–1357.
105. Denham, S.L. & I. Winkler. 2006. The role of predictive models in the formation of auditory streams. *J. Physiol. Paris* **100:** 154–170.
106. Rankin, J., E. Sussman & J. Rinzel. 2015. Neuromechanistic model of auditory bistability. *PLoS Comput. Biol.* **11:** e1004555.
107. Mill, R.W., T.M. Bohm, A. Bendixen, *et al.* 2013. Modelling the emergence and dynamics of perceptual organisation in auditory streaming. *PLoS Comput. Biol.* **9:** e1002925.

108. Barniv, D. & I. Nelken. 2015. Auditory streaming as an online classification process with evidence accumulation. *PLoS One* **10:** e0144788.
109. Durlach, N.I., C.R. Mason, G. Kidd, Jr., *et al.* 2003. Note on informational masking. *J. Acoust. Soc. Am.* **113:** 2984–2987.
110. Kidd, G., Jr., C.R. Mason & T.L. Arbogast. 2002. Similarity, uncertainty, and masking in the identification of nonspeech auditory patterns. *J. Acoust. Soc. Am.* **111:** 1367–1376.
111. Watson, C.S. 2005. Some comments on informational masking. *Acta Acust. United Ac.* **91:** 502–512.
112. Wiegand, K. & A. Gutschalk. 2012. Correlates of perceptual awareness in human primary auditory cortex revealed by an informational masking experiment. *Neuroimage* **61:** 62–69.
113. Gutschalk, A., C. Micheyl & A.J. Oxenham. 2008. Neural correlates of auditory perceptual awareness under informational masking. *PLoS Biol.* **6:** e138.
114. Nelken, I., A. Fishbach, L. Las, *et al.* 2003. Primary auditory cortex of cats: feature detection or something else? *Biol. Cybern.* **89:** 397–406.
115. Crick, F.C. & C. Koch. 2005. What is the function of the claustrum? *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **360:** 1271–1279.
116. Remedios, R., N.K. Logothetis & C. Kayser. 2014. A role of the claustrum in auditory scene analysis by reflecting sensory change. *Front. Syst. Neurosci.* **8:** 44.
117. Giani, A.S., P. Belardinelli, E. Ortiz, *et al.* 2015. Detecting tones in complex auditory scenes. *Neuroimage* **122:** 203–213.
118. Dykstra, A.R., E. Halgren, A. Gutschalk, *et al.* 2016. Neural correlates of auditory perceptual awareness and release from informational masking recorded directly from human cortex: a case study. *Front. Neurosci.* **10:** 472.
119. Eramudugolla, R., D.R. Irvine, K.I. McAnally, *et al.* 2005. Directed attention eliminates 'change deafness' in complex auditory scenes. *Curr. Biol.* **15:** 1108–1113.
120. Gygi, B. & V. Shafiro. 2011. The incongruency advantage for environmental sounds presented in natural auditory scenes. *J. Exp. Psychol. Hum. Percept. Perform.* **37:** 551–565.
121. McDermott, J.H., D. Wrobleski & A.J. Oxenham. 2011. Recovering sound sources from embedded repetition. *Proc. Natl. Acad. Sci. U.S.A.* **108:** 1188–1193.
122. Dickerson, K. & J.R. Gaston. 2014. Did you hear that? The role of stimulus similarity and uncertainty in auditory change deafness. *Front. Psychol.* **5:** 1125.
123. Snyder, J.S. & M.K. Gregg. 2011. Memory for sound, with an ear toward hearing in complex auditory scenes. *Atten. Percept. Psychophys.* **73:** 1993–2007.
124. Gregg, M.K. & J.S. Snyder. 2012. Enhanced sensory processing accompanies successful detection of change for real-world sounds. *Neuroimage* **62:** 113–119.
125. Aru, J., T. Bachmann, W. Singer, *et al.* 2012. Distilling the neural correlates of consciousness. *Neurosci. Biobehav. Rev.* **36:** 737–746.
126. Polich, J. 2007. Updating P300: an integrative theory of P3a and P3b. *Clin. Neurophysiol.* **118:** 2128–2148.
127. Snyder, J.S., B.D. Yerkes & M.A. Pitts. 2015. Testing domain-general theories of perceptual awareness with auditory brain responses. *Trends Cogn. Sci.* **19:** 295–297.
128. Gregg, M.K., V.C. Irsik & J.S. Snyder. 2014. Change deafness and object encoding with recognizable and unrecognizable sounds. *Neuropsychologia* **61:** 19–30.
129. Puschmann, S., P. Sandmann, J. Ahrens, *et al.* 2013. Electrophysiological correlates of auditory change detection and change deafness in complex auditory scenes. *Neuroimage* **75:** 155–164.
130. Sohoglu, E. & M. Chait. 2016. Neural dynamics of change detection in crowded acoustic scenes. *Neuroimage* **126:** 164–172.
131. Puschmann, S., R. Weerda, G. Klump, *et al.* 2013. Segregating the neural correlates of physical and perceived change in auditory input using the change deafness effect. *J. Cogn. Neurosci.* **25:** 730–742.
132. Crottaz-Herbette, S. & V. Menon. 2006. Where and when the anterior cingulate cortex modulates attentional response: combined fMRI and ERP evidence. *J. Cogn. Neurosci.* **18:** 766–780.
133. Petersen, S.E. & M.I. Posner. 2012. The attention system of the human brain: 20 years after. *Annu. Rev. Neurosci.* **35:** 73–89.
134. Halgren, E., K. Marinkovic & P. Chauvel. 1998. Generators of the late cognitive potentials in auditory and visual oddball tasks. *Electroencephalogr. Clin. Neurophysiol.* **106:** 156–164.
135. Ullsperger, M., H.A. Harsay, J.R. Wessel, *et al.* 2010. Conscious perception of errors and its relation to the anterior insula. *Brain Struct. Funct.* **214:** 629–643.
136. Klein, T.A., M. Ullsperger & C. Danielmeier. 2013. Error awareness and the insula: links to neurological and psychiatric diseases. *Front. Hum. Neurosci.* **7:** 14.
137. Shinn-Cunningham, B.G. 2008. Object-based auditory and visual attention. *Trends Cogn. Sci.* **12:** 182–186.
138. Bey, C. & S. McAdams. 2002. Schema-based processing in auditory scene analysis. *Percept. Psychophys.* **64:** 844–854.
139. Dowling, W.J. 1973. Perception of interleaved melodies. *Cogn. Psychol.* **5:** 322–337.
140. Gregg, M.K. & A.G. Samuel. 2009. The importance of semantics in auditory representations. *Atten. Percept. Psychophys.* **71:** 607–619.
141. Vanden Bosch der Nederlanden, C.M., J.S. Snyder & E.E. Hannon. 2016. Children use object-level category knowledge to detect changes in complex auditory scenes. *Dev. Psychol.* **52:** 1867–1877.
142. Irsik, V.C., C.M. Vanden Bosch der Nederlanden & J.S. Snyder. 2016. Broad attention to multiple individual objects may facilitate change detection with complex auditory scenes. *J. Exp. Psychol. Hum. Percept. Perform.* **42:** 1806–1817.
143. Naci, L., R. Cusack, V.Z. Jia, *et al.* 2013. The brain's silent messenger: using selective attention to decode human thought for brain-based communication. *J. Neurosci.* **33:** 9385–9393.
144. Elhilali, M. 2017. Modeling the cocktail party problem. In *The Auditory System at the Cocktail Party.* J.C. Middlebrooks, J.Z. Simon, A.N. Popper & R.R. Fay, Eds. New York: Springer.

145. Alinaghi, A.Y., P.J. Jackson, Q.J. Liu, *et al.* 2014. Joint mixing vector and binaural model based stereo source separation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **22:** 1434–1448.
146. Bellur, A. & M. Elhilali. 2017 Feedback driven sensory mapping adaptation for robust speech activity detection. *IEEE/ACM Trans. Audio Speech Lang. Process.* **25:** 481–492.
147. Hu, K. & D.L. Wang. 2013. An unsupervised spproach to cochannel speech separation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **21:** 120–129.
148. Akram, S., A. Presacco, J.Z. Simon, *et al.* 2016. Robust decoding of selective auditory attention from MEG in a competing-speaker environment via state-space modeling. *Neuroimage* **124:** 906–917.