14 Brain Mechanisms of Auditory Scene Analysis

BARBARA G. SHINN-CUNNINGHAM

ABSTRACT When listening to a mixture of sound sources, the brain relies on prior knowledge of statistical regularities in natural sound to estimate what sound energy comes from which source in the external world. This process of perceptual *object formation*, also known as auditory scene analysis, has been a focus of research for decades, most recently for rich, natural sounds like speech. Because the brain can only analyze one perceptual object at a time, object formation influences how we perceive and interpret the sounds in our environment. This chapter reviews the acoustic features that drive object formation and then considers neural correlates of auditory scene analysis and attention.

Our ears usually receive a mixture of sound sources overlapping in time and frequency content. Yet we perceive distinct perceptual "objects," or estimates of the acoustic energy emitted by physical sound sources in the external world (Darwin, 1997; Griffiths & Warren, 2004; Shinn-Cunningham, 2008). The process of separating sound mixtures into perceptual objects is known as *auditory scene analysis*, or ASA. This chapter provides an introduction to both behavioral and neurophysiological studies of ASA in natural settings (see also Darwin, 1997; Kondo, van Loon, Kawahara, & Moore, 2017).

Determining what sound energy comes from what physical sound source is a mathematically ill-posed problem: there are an infinite number of sound mixtures that could have produced the signals reaching your left and right ears. For instance, imagine the word *boot*. It could be from a single source or from one source, *boo*, temporally abutting another source *oot*, or the low frequencies of *boot* could be from one source and the high frequencies from a different source. How does your brain estimate the sources in the scene given that the problem itself is underconstrained?

The process of ASA matters practically because how you parse the acoustic scene has a direct impact on your ability to understand the meaning of the sources around you (Shinn-Cunningham, 2008). Cognitive processes are limited; you cannot process everything in the scene in detail. Instead, you use attention to focus on one object and suppress the neural responses from competing objects. Think about what this means: your ability to make sense of the world depends on attention, which suppresses the representation of sounds you ignore, but for attention to be effective, you must successfully segregate the sources in the scene. Thus, your ability to understand sound sources depends on your ability to parse the scene into auditory objects that represent the physical sound sources around you.

At first blush, there seems to be a hierarchy of processing: the scene is first analyzed into objects and then one object is selected and attended. In truth, these processes form a heterarchy: the processes occur in parallel, feeding back upon one another. Your decision about what to attend in turn influences what sound features are emphasized, which influences how objects are formed (akin to the reverse hierarchy theory; see Nahum, Nelken, & Ahissar, 2008). Conversely, sounds that are particularly salient are more likely to grab attention even in the absence of a conscious decision to focus on them (see Kaya & Elhilali, 2014a; Kayser, Petkov, Lippert, & Logothetis, 2005). Once a sound is the focus of attention, it influences object formation—whether attentional selection was the result of volitional effort or of stimulus-driven salience.

Although there is a large body of research exploring ASA and attention, the brain mechanisms realizing these processes are still poorly understood. In part, this is because object formation and attention involve a wide range of brain areas, working together. Moreover, how sound is organized into objects and how a particular object rises to be the focus of attention are both accomplished gradually across stages of the auditory pathway and as sound content evolves through time. Thus, one cannot pinpoint a specific neural place or a time at which an object is formed or becomes the focus of attention.

This chapter first discusses how the brain forms objects from sound mixtures and briefly reviews how object formation and attention allow us to interpret complex sound mixtures. It then considers different neural correlates of object formation and attention.

Statistical Regularities in Natural Sounds Guide Auditory Object Formation

The intellectual father of ASA, Al Bregman, wrote a tome that lays out most of its fundamental principles (Bregman, 1990). In general, auditory objects, the output of ASA, are the brain's "best guess" of what belongs together, based on statistical regularity in the spectrotemporal structure of natural sounds. Much of Bregman's seminal work focused on artificial sounds, such as mixtures of simple tones combined in different sequences and rates. Indeed, to most psychoacousticians, the shorthand *A-B-A paradigm* immediately brings to mind a sequence of two tones with different frequencies (A and B) that are each repeated, with the A tones isochronous at one rate, the B tones isochronous at half that rate, and each B tone falling midway between a pair of A tones (see figure 14.1). Listeners may perceive either one stream (of A-B-A tones, together, forming a triplet rhythm) or two streams (one of fast-repeating A tones and another of slow-repeating B tones, each with an isochronous rhythm). Listeners are more likely to hear a galloping stream when the presentation rate is faster or the two frequencies are closer together. The similarity of the A and B tones in other perceptual dimensions also can influence how the scene is parsed.

While simple stimuli allow one to isolate how specific features of sound influence ASA, they are quite impoverished compared to natural soundscapes, where most sources are broadband, time varying, and statistically unrelated to every other source. This makes it hard to extrapolate from experimental results of ABA studies to understand what occurs for richer acoustic stimuli like those encountered in everyday settings. Below, we focus on how the brain copes with more natural sound mixtures, such as independent streams of speech or temporally uncorrelated melodies.

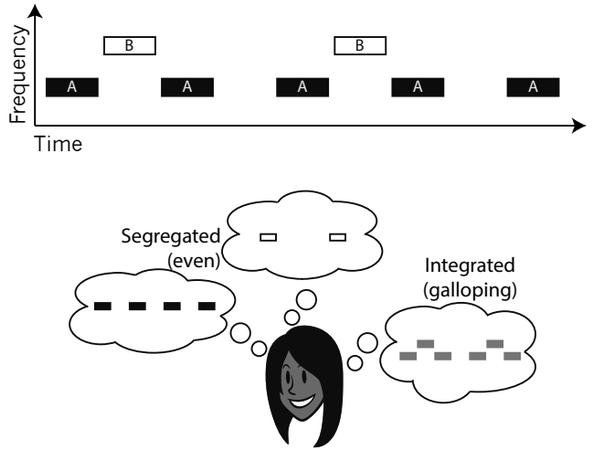

FIGURE 14.1 Schematic of the well tested A-B-A paradigm used to study auditory scene analysis. Two different frequency tones (denoted by *A* and *B*) are presented, each at a regular interval; however, the B tones repeat at half the rate of and in between the A tones. If listeners perceive the sound mixture as segregated, they report two separate streams, one of A tones and one of B tones, each of which has an "even" rhythm (*left side, bottom of figure*). If they perceive all of the tones as one single stream, they report a galloping rhythm. (See color plate 19.)

Bregman broke down ASA mechanisms into *simultaneous* grouping (rules governing whether tones that are present at the same moment are heard in the same object) and *sequential* grouping (rules governing whether events that occur in a sequence are heard as part of the same, or a different, object). While these are useful concepts, the words can be misleading when considering natural sounds. To extend these concepts to common, everyday sounds, we consider statistical regularities that arise at a "local" level to form *syllables* (akin to Bregman's simultaneous grouping) and higher-order statistical regularities that can perceptually link together syllables separated by silent gaps (generalizing Bregman's concept of sequential grouping).

Syllables form from spectrotemporally structured, contiguous sound Many real physical objects generate energy in bursts, often broadly spread across the audible spectrum. Such bursts are not typically static; their content changes continuously through time. For instance, imagine a person saying the syllable *oy*, which has a spectrum that changes smoothly from one vowel to another (see figure 14.2). These changes are generally structured in

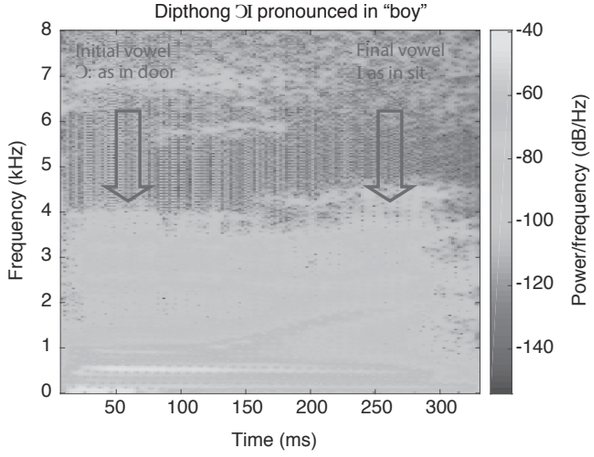

FIGURE 14.2 Spectrogram of the syllable *boy*, showing how dynamic changes in spectral content are nonetheless structured and continuous in time frequency. (See color plate 20.)

time–frequency; as a result, the entire syllable is perceived automatically as a single unit, despite the dynamic changes in content over time.

In speech the unit of a syllable (like *oy*) has spectrotemporal continuity and a predictable structure that allows its widespread spectral content to nonetheless be grouped together unambiguously. In many nonspeech sounds, the same kind of spectrotemporal structure also arises. Throughout this chapter we use the term *syllable* to refer to any continuous burst of sound whose "local" spectrotemporal structure groups it together, perceptually, such as a note from an oboe, the crash of a glass smashing against a tile floor, or the rattle of a coin rolling to a stop on a counter.

In each example there are multiple (often temporally contiguous) spectral components making up individual syllables. Various spectral components, however, group together not because they are contiguous in frequency but because of other shared structure. For example, the music note from an oboe comprises harmonically related spectral elements with a common fundamental frequency. For the less structured smash of a breaking glass, spectral elements are not harmonically related but have similar, correlated temporal envelopes, with a sudden common onset that decays rapidly back to nothing. Although spatial cues are relatively weak in influencing the formation of spectrotemporally local objects, they still have some perceptual weight, especially when there is ambiguity in a sound mixture, and one spectral element could otherwise belong to two competing objects (e.g., see Darwin, 2006; Middlebrooks, 2017; Schwartz & Shinn-Cunningham, 2010). In general, the statistical structure of the spectrotemporal content, including harmonicity, common amplitude modulation, and spatial cues like interaural time or level differences, influence how spectral elements are grouped into syllables.

Sequences of syllables group together into streams based on feature similarity Just as typical sources produce syllables made up of statistically structured spectral components, they tend to produce sequences of syllables with common attributes. Each syllable has higher-level perceptual features that link together the spectral elements making up that syllable, including pitch, timbre, and location. For instance, a sequence of oboe notes is heard as a single melody. For syllables produced by the same source, these high-level syllabic features tend to change slowly over time, allowing the brain to connect the syllables, perceptually. When perceived as a single sequence, the syllables are said to form a *stream*.

Expectations and predictions guide object formation Both syllable and stream formation depend upon expectations: statistical regularities in low-level acoustic features that indicate spectral components belong to the same syllable and in higher-level perceptual features that indicate that distinct syllables are part of the same stream. Some of these expectations are learned, while others may be "hardwired" through evolution. For instance, common onsets strongly drive syllabic-level grouping; octopus cells in the brain stem are hardwired to detect and respond to common onsets across broad frequency ranges. On the other hand, if an acoustic signal in a mixture repeats, the brain rapidly learns (over the course of two to three repetitions) to expect this sequence, increasing the likelihood that it is heard as a coherent stream, even in the absence of attention (e.g., Masutomi, Barascud, Kashino, McDermott, & Chait, 2016).

Attention and Object Formation Are Intricately Intertwined

The brain processes one sound object in detail at any one time; it is not capable of attending to two sounds simultaneously (e.g., see Ihlefeld & Shinn-Cunningham, 2008). In those moments when you believe you are "sharing" attention among simultaneous objects, you are probably switching rapidly between them—and may be able to extract the gist of each if each is predictable, allowing you to fill in the unattended information. However, attention is being multiplexed, rather than truly divided.

What object we attend at a given moment depends partially on how we direct attention. This volitional focus is effected through feedback from control areas in the frontal cortex that modulate auditory sensory responses, suppressing inputs that do not match what we wish to attend. However, salient auditory objects may "get through" in spite of volitional attention. For instance, if you are trying to attend to a woman on your right, top-down attention will suppress a man on your left; however, if a gunshot goes off behind your head, it will interrupt attention to allow your brain to decide how to act (Fight? Flight? Or is it your mischievous cousin with his cap gun, which you can ignore?; see Dalton & Fraenkel, 2012). In vision, this battle between top-down, conscious attention and bottom-up, salience-driven attention is described as a *biased-competition*, in which volitional focus biases the incoming sensory representation but where this biasing may be overcome by salient inputs (Desimone & Duncan, 1995). This same model of biased competition accounts well for how auditory attention operates, consistent with the two sensory modalities sharing either common or similar

attentional circuitry (Kaya & Elhilali, 2017; McDermott, 2009; Shinn-Cunningham, 2008).

Unambiguous objects form automatically, independent of attention In the natural world, auditory objects often are distinct, leading to strong interobject competition. The sounds of a phone ringing, music playing from your computer, the computer fan humming, and your colleague telling you about his weekend are mutually dissimilar; your brain has no trouble estimating the acoustic content of each source. However, because these objects are perceptually distinct, the automatic competition between them is strong, making it nearly impossible to analyze more than one (e.g., Best et al., 2006). Indeed, when distinct ongoing streams are present, an attended stream will tend to remain the focus of attention, biasing attention to remain on it as it continues (Best, Ozmeral, Kopco, & Shinn-Cunningham, 2008; Bressler, Masud, Bharadwaj, & Shinn-Cunningham, 2014; Woods & McDermott, 2015). This is not to say that top-down attentional focus (object selection) does not waver over time (e.g., oscillating between remaining focused on auditory target sounds versus *lapsing*, as shown recently in a cross-modal selective attention task in monkeys; Lakatos et al., 2016). Rather, while attentional selection may fail due to such top-down lapses of attentional control, top-down attention is not necessary to form streams from a sound mixture when the statistical structures of competing streams make each distinct. In these conditions, the objects are formed automatically, based on the statistical structure of the input sound mixture; top-down attentional failures may lead to selection of the wrong stream but will not lead to failures of object segregation (e.g., see Ruggles & Shinn-Cunningham, 2011).

Attention influences object formation when sound mixtures are ambiguous In some conditions, auditory object formation is ambiguous. Ambiguity may occur at the level of syllabic formation—such as when a flute and an oboe play a melody in unison—or at the level of stream formation—such as when two similar-sounding women speak at roughly the same intensity from roughly the same direction. In these cases, top-down focus affects what object will be attended and the formation of the objects themselves. If you focus on the sound of the flute within the flute-oboe mixture, your brain may, over time, isolate it and analyze it; however, if you focus on the flute and oboe as a single melody with an unusual timbre, you can perceive it as a single entity. As you focus on the woman who is slightly to the right and has a lower-pitched, raspier voice, you will be able to pull her speech out of the mixture; however, if you do not focus attention, you are likely to hear a gibberish of intermingled words spoken by both women (e.g., Best et al., 2006). Although syllables from each talker form automatically due to local structure, the stream corresponding to a speaker will not form without top-down, focused attention, and the specificity of attentional focus will tend to increase through time (e.g., see Best et al., 2008; Dai, Best, & Shinn-Cunningham, 2018). These examples illustrate two important principles of ambiguous mixtures. First, when objects are not particularly distinct, it takes time for objects to be segregated, perceptually. Second, in these cases attention can directly influence object formation, not just which object becomes the focus of detailed analysis.

In the above examples, object ambiguity arises because distinct physical sound sources produce sounds that are perceptually similar, either at the syllabic or the stream level. However, sometimes, ambiguity in a sound mixture arises because the scene is noisy. Imagine trying to listen to a talker in a cathedral. Echoes and reverberation smear out the sound, muddying both low-level attributes (temporal envelope, harmonic structure, location cues, and more) and higher-order perceptual features (pitch, location, and more), thereby interfering with object formation. In such situations, attention helps to isolate a desired sound, influencing both object selection and formation.

For listeners with peripheral hearing deficits, object formation may be ambiguous even in conditions for which a normal-hearing listener has no difficulty. Hearing impairment reduces spectral and temporal resolution, leading to less distinct acoustic features. The reduced spectral resolution of the hearing-impaired listener also increases the amount of acoustic overlap of competing, simultaneous sources. In a normal-hearing listener, peripheral neural channels are frequency-specific; therefore, each individual frequency channel is likely to be dominated by a single source at any one point in time. However, for a hearing-impaired listener with broader than normal peripheral filtering, each frequency channel is more likely to contain a mixture of sound sources, making it impossible to separate which energy is coming from which source (Best, Mason, & Kidd, 2011; Shinn-Cunningham & Best, 2008). These effects help explain the subjective reports of many hearing-impaired listeners (Dai, Best, & Shinn-Cunningham, 2018; Roverud, Best, Mason, Swaminathan, & Kidd, 2016). For instance, hearing aids can make perceptible sound that would otherwise be inaudible; however, many users report that in a noisy setting, improving audibility increases perceptual interference ("When I put in my hearing aid, it just makes the clattering dishes louder!"). These symptoms are consistent

with failures of object formation: if the peripheral representation is degraded to the point that objects cannot form, then attentional selection will fail.

Neuroimaging Reveals Neural Correlates of Auditory Scene Analysis

The majority of neurophysiological studies of neural responses to sound mixtures have employed Bregman-like, A-B-A tone sequences (e.g., Itatani & Klump, 2018; Mehta, Jacoby, Yasin, Oxenham, & Shamma, 2017). More recent noninvasive studies in humans, however, have embraced the use of richer sound mixtures made up of speech (e.g., Ding & Simon, 2012), competing melodies (e.g., Xiang, Simon, & Elhilali, 2010), or complex "textures" of ongoing random events (e.g., Barascud, Pearce, Griffiths, Friston, & Chait, 2016). In exploring neural correlates of ASA, here we focus on neuroelectric (electro- and magnetoencephalography, or EEG and MEG) and functional magnetic resonance (fMRI) imaging studies in human listeners.

These two "views" of neural processing afford different strengths and weaknesses (Lee, Larson, Maddox, & Shinn-Cunningham, 2014). Whereas neuroelectric imaging provides exceptional temporal resolution, the neural locus of activity is difficult to determine. Conversely, fMRI provides good spatial precision in defining the brain regions involved in neural processing; however, its temporal resolution is too poor to track responses to individual sound events. While new approaches have been successfully used to wed the temporal resolution of neuroelectric imaging with the spatial resolution of fMRI in visual studies, these approaches have yet to be applied widely in studies of auditory processing (see Cichy & Teng, 2017). Still, the results of neuroelectric and fMRI studies converge to reveal that ASA and attention are controlled by interactions across disperse brain networks.

Neuroelectric imaging tracks neural responses to objects and streams EEG and MEG can both be used to measure the mismatch negativity (MMN), a larger than normal negative response to a syllable that is unexpected in a given stimulus context compared to when it is expected (e.g., Horvath, Czigler, Sussman, & Winkler, 2001; Näätänen, Paavilainen, Rinne, & Alho, 2007). For instance, in a sequence of repeated syllables ("dah, dah, dah"), a different syllable ("bah," known as a *deviant*) will elicit a stronger response than if it occurs in a sequence of "bahs." MMN responses occur relatively early in the neural cascade, typically within 200 ms of the deviant, consistent with a relatively early sensory response in the cortex. Importantly, MMN responses are preattentive (they can be elicited in a listener who is asleep) but tend to be enhanced when a listener is attending to the stream containing the deviant (Horvath et al., 2001; Näätänen et al., 2007).

The MMN provides evidence that syllables form coherent streams when they are expected and predictable and fit with the ongoing stream context. Deviants are inherently more salient than when the same syllable is expected. By deviating from built-up expectations about a stream's content, a deviant is more likely to be heard as a new object and to interrupt top-down processing in the biased competition governing attention. Thus, although rarely discussed in this way, the MMN is an early sensory marker of auditory stream formation, signaling an automatically detected deviation that may constitute a new object (see, for instance, Kaya & Elhilali, 2014b; Sohoglu & Chait, 2016; Southwell et al., 2017).

Early neuroelectric responses evoked by competing streams in a mixture are strongly modulated by attentional focus. In particular, when competing streams contain isolated syllables, the auditory N1 (a negativity in frontocentral electrodes, evoked by early auditory sensory responses), in response to a syllable of greater amplitude when the stream containing the syllable is attended compared to when it is ignored, provides a quantitative index of the strength of top-down attention (Choi, Rajaram, Varghese, & Shinn-Cunningham, 2013). Indeed, the strength of attentional modulation of the N1 response often correlates with how well listeners perform in challenging selective auditory attention tasks (Dai, Best, & Shinn-Cunningham, 2018). Attentional modulation also alters the strength of the correlation between the low-frequency portion of the measured neuroelectric waveform and the envelope of an attended speech stream (Ding, Melloni, Zhang, Tian, & Poeppel, 2016; Ding & Simon, 2013; O'Sullivan et al., 2014; Zion Golumbic et al., 2013). Consistent with behavioral studies demonstrating that it can take time to segregate an attended stream from competing sounds, these neural signatures of attentional focus show a buildup in the brain's ability to lock onto an attended sound stream amid competing sounds (Best, & Shinn-Cunningham, 2018; Riecke, Sack, & Schroeder, 2015; Dai, Best, & Shinn-Cunningham, 2018).

Neuroelectric studies comparing different forms of attentional control reveal that spatial attention tends to recruit activity in the superior prefrontal sulcus (sPCS; see Bharadwaj, Lee, & Shinn-Cunningham, 2014; Hill & Miller, 2010; Lee et al., 2013), while focusing attention on nonspatial features evokes activity in the temporal lobe (Hill & Miller, 2010; Lee et al., 2013). These results show that task demands change how control networks in

the brain are activated. Neuroelectric studies of induced responses (looking at changes in the power of neural energy in different frequency bands that is not phase locked) also reveal differences in neural processing that depend on task demands, such as changes in energy in the alpha band (8–12 Hz) over the parietal cortex (see chapter 15). These studies suggest that IPS is engaged by auditory spatial selective attention, presumably by top-down signaling from the prefrontal control regions of the cortex (e.g., see Buschman & Miller, 2007).

Functional magnetic resonance imaging defines brain networks involved in auditory perception fMRI studies have been used to explore which brain regions are involved in different aspects of ASA and the control of auditory cognitive processes. These studies support the view that preattentive, primitive processes, largely in auditory sensory and multisensory regions of the brain, represent the fundamental features that drive object formation (see Nelken, 2004, for a review). Meanwhile, top-down cognitive control networks involving sensory, multisensory, and frontal regions modulate how auditory information is processed (Michalka, Rosen, Kong, Shinn-Cunningham, & Somers, 2016; Noyce, Cestero, Michalka, Shinn-Cunningham, & Somers, 2017; Osher, Tobyne, Congden, Michalka, & Somers, 2015; Shomstein & Yantis, 2004, 2006; Tobyne, Osher, Michalka, & Somers, 2017; Westerhausen et al., 2010). IPS is known to contain multisensory representations. In ASA experiments, IPS is active in the automatic, bottom-up processing that causes objects to form from complex acoustic scenes (Sohoglu & Chait, 2016; Teki et al., 2016).

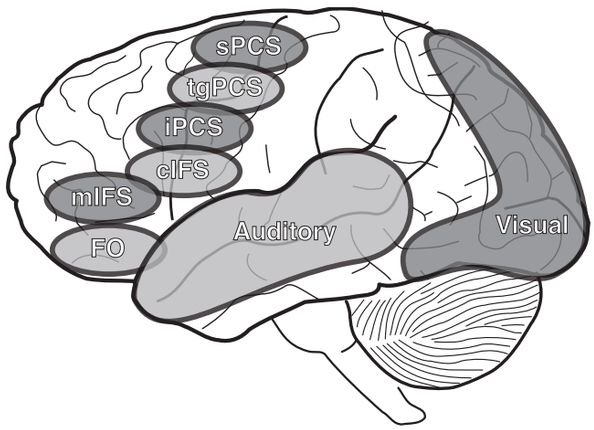

FIGURE 14.3 Schematic illustrating interleaved, discrete executive control regions that are preferentially recruited for visual (*dark gray*) or auditory (*light gray*) attention and working memory. Provided by Dr. Abigail Noyce, Boston University. The regions span from the superior precentral sulcus (*sPCS*, a visually biased region) through to the frontorbital cortex (*FO*, an auditory biased region). Other visually biased regions include the inferior precentral sulcus (*iPCS*) and the medial inferior frontal sulcus (*mIFS*), while auditory biased regions include the transverse gyrus of the precentral sulcus (*tgPCS*) and the caudo inferior frontal sulcus (*clFS*).

fMRI studies show that auditory spatial selective attention engages areas in the frontoparietal network associated with visual spatial attention (Hill & Miller, 2010; Kong et al., 2014; Michalka et al., 2016). While these results hint that visual and auditory tasks use the same executive function regions, within-subject contrasts of activation in auditory versus in visual tasks reveal alternating sensory-biased regions (see schematic in figure 14.3; Michalka et al., 2016; Noyce et al., 2017). Importantly, if typical methods are used to coregister and average the activity of these caudolateral frontal cortex regions across subjects, the result is a large swath of multisensory control regions, smearing out the consistent alternating pattern (Noyce et al., 2017). Moreover, although these regions are biased toward either visual (figure 14.3, *blue regions*) or auditory (14.3, *orange regions*) inputs, the "more visual" areas are engaged when listeners process spatial, rather than temporal, aspects of an auditory input; conversely, the "more auditory" regions are more active when observers process temporal, rather than spatial, aspects of a visual input. Thus, while previous studies have shown that the caudolateral frontal cortex is part of a multiple-demand network (e.g., see Duncan, 2010; Fedorenko, Duncan, & Kanwisher, 2013), we see more structured, differentiated activity. Together, these studies suggest that multiple, broad brain networks with different processing specializations influence ASA and attention.

Summary

Behavioral and neuroimaging results show that ASA depends on both automatic and attentionally driven processes. Simple, unambiguous mixtures are parsed automatically into distinct perceptual objects, whereas noisier scenes may require both time and focused attention for auditory objects to emerge. In forming auditory objects, the brain relies not only on a prior knowledge of the statistical regularities of natural sounds but on an iterative estimation of ongoing sources, effected through multiple, distributed brain networks.

REFERENCES

Barascud, N., Pearce, M. T., Griffiths, T. D., Friston, K. J., & Chait, M. (2016). Brain responses in humans reveal ideal observer-like sensitivity to complex acoustic patterns. *Proceedings of the National Academy of Sciences, 113*(5), E616–E625.

Best, V., Gallun, F. J., Ihlefeld, A., & Shinn-Cunningham, B. G. (2006). The influence of spatial separation on divided listening. *Journal of the Acoustical Society of America, 120*(3), 1506–1516.

Best, V., Mason, C. R., & Kidd Jr., G. (2011). Spatial release from masking in normally hearing and hearing-impaired listeners as a function of the temporal overlap of competing talkers. *Journal of the Acoustical Society of America, 129*(3), 1616–1625. doi:10.1121/1.3533733.

Best, V., Ozmeral, E. J., Kopco, N., & Shinn-Cunningham, B. G. (2008). Object continuity enhances selective auditory attention. *Proceedings of the National Academy of Science, 105*(35), 13174–13178. doi:10.1073/pnas.0803718105.

Bharadwaj, H. M., Lee, A. K. C., & Shinn-Cunningham, B. G. (2014). Measuring auditory selective attention using frequency tagging. *Frontiers in Integrative Neuroscience, 8*. doi:10.3389/fnint.2014.00006.

Bregman, A. S. (1990). *Auditory Scene Analysis: The Perceptual Organization of Sound*. Cambridge, MA: MIT Press.

Bressler, S., Masud, S., Bharadwaj, H., & Shinn-Cunningham, B. (2014). Bottom-up influences of voice continuity in focusing selective auditory attention. *Psychological Research, 78*(3), 349–360. doi:10.1007/s00426-014-0555-7.

Buschman, T. J., & Miller, E. K. (2007). Top-down versus bottom-up control of attention in the prefrontal and posterior parietal cortices. *Science, 315*(5820), 1860–1862.

Choi, I., Rajaram, S., Varghese, L. A., & Shinn-Cunningham, B. G. (2013). Quantifying attentional modulation of auditory-evoked cortical responses from single-trial electroencephalography. *Frontiers in Human Neuroscience, 7*, 115. doi:10.3389/fnhum.2013.00115.

Cichy, R. M., & Teng, S. (2017). Resolving the neural dynamics of visual and auditory scene processing in the human brain: A methodological approach. *Philosophical Transactions of the Royal Society B, 372*(1714), 20160108.

Dai, L., Best, V., & Shinn-Cunningham, B. G. (2018). Sensorineural hearing loss degrades behavioral and physiological measures of human spatial selective auditory attention. *Proceedings of the National Academy of Science, 115*(14), E3286–E3295. doi:10.1073/pnas.1721226115.

Dalton, P., & Fraenkel, N. (2012). Gorillas we have missed: Sustained inattentional deafness for dynamic events. *Cognition, 124*(3), 367–372.

Darwin, C. J. (1997). Auditory grouping. *Trends in Cognitive Sciences, 1*(9), 327–333.

Darwin, C. J. (2006). Contributions of binaural information to the separation of different sound sources. *International Journal of Audiology, 45*(Suppl 1), S20–24.

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience, 18*, 193–222.

Ding, N., Melloni, L., Zhang, H., Tian, X., & Poeppel, D. (2016). Cortical tracking of hierarchical linguistic structures in connected speech. *Nature Neuroscience, 19*(1), 158–164. doi:10.1038/nn.4186.

Ding, N., & Simon, J. Z. (2012). Emergence of neural encoding of auditory objects while listening to competing speakers. *Proceedings of the National Academy of Science, 109*(29), 11854–11859. doi:10.1073/pnas.1205381109.

Ding, N., & Simon, J. Z. (2013). Adaptive temporal encoding leads to a background-insensitive cortical representation of speech. *Journal of Neuroscience, 33*(13), 5728–5735. doi:10.1523/JNEUROSCI.5297-12.2013.

Duncan, J. (2010). The multiple-demand (MD) system of the primate brain: Mental programs for intelligent behaviour. *Trends in Cognitive Sciences, 14*(4), 172–179.

Fedorenko, E., Duncan, J., & Kanwisher, N. (2013). Broad domain generality in focal regions of frontal and parietal cortex. *Proceedings of the National Academy of Sciences, 110*(41), 16616–16621.

Griffiths, T. D., & Warren, J. D. (2004). What is an auditory object? *Nature Review Neuroscience, 5*(11), 887–892. doi:10.1038/nrn1538.

Hill, K. T., & Miller, L. M. (2010). Auditory attentional control and selection during cocktail party listening. *Cerebral Cortex, 20*(3), 583–590. doi:10.1093/cercor/bhp124.

Horvath, J., Czigler, I., Sussman, E., & Winkler, I. (2001). Simultaneously active pre-attentive representations of local and global rules for sound sequences in the human brain. *Brain Research. Cognitive Brain Research, 12*(1), 131–144. doi:S0926-6410(01)00038-6.

Ihlefeld, A., & Shinn-Cunningham, B. (2008). Spatial release from energetic and informational masking in a divided speech identification task. *Journal of the Acoustical Society of America, 123*(6), 4380–4392. doi:10.1121/1.2904825.

Itatani, N., & Klump, G. M. (2018). Interaction of spatial and non-spatial cues in auditory stream segregation in the European starling. *European Journal of Neuroscience*. doi:10.1111/ejn.13716.

Kaya, E. M., & Elhilali, M. (2014a). Investigating bottom-up auditory attention. *Frontiers in Human Neuroscience, 8*, 327. doi:10.3389/fnhum.2014.00327.

Kaya, E. M., & Elhilali, M. (2014b). Investigating bottom-up auditory attention. *Frontiers in Human Neuroscience, 8*, 327.

Kaya, E. M., & Elhilali, M. (2017). Modelling auditory attention. *Philosophical Transactions of the Royal Society B, 372*(1714), 20160101.

Kayser, C., Petkov, C. I., Lippert, M., & Logothetis, N. K. (2005). Mechanisms for allocating auditory attention: An auditory saliency map. *Current Biology, 15*(21), 1943–1947.

Kondo, H. M., van Loon, A. M., Kawahara, J.-I., & Moore, B. C. (2017). Auditory and visual scene analysis: An overview. *Philosophical Transactions of the Royal Society B, 372* (1714), 20160099.

Kong, L., Michalka, S. W., Rosen, M. L., Sheremata, S. L., Swisher, J. D., Shinn-Cunningham, B. G., & Somers, D. C. (2014). Auditory spatial attention representations in the human cerebral cortex. *Cerebral Cortex, 24*(3), 773–784. doi:10.1093/cercor/bhs359.

Lakatos, P., Barczak, A., Neymotin, S. A., McGinnis, T., Ross, D., Javitt, D. C., & O'Connell, M. N. (2016). Global dynamics of selective attention and its lapses in primary auditory cortex. *Nature Neuroscience, 19*(12), 1707–1717. doi:10.1038/nn.4386.

Lee, A. K. C., Larson, E., Maddox, R. K., & Shinn-Cunningham, B. G. (2014). Using neuroimaging to understand the cortical mechanisms of auditory selective attention. *Hearing Research, 307*, 111–120. doi:10.1016/J.Heares.2013.06.010.

Lee, A. K. C., Rajaram, S., Xia, J., Bharadwaj, H., Larson, E., Hamalainen, M., & Shinn-Cunningham, B. G. (2013). Auditory selective attention reveals preparatory activity in different cortical regions for selection based on source location and source pitch. *Frontiers in Neuroscience, 6*. doi:10.3389/fnins.2012.00190.

—-1
—0
—+1

Masutomi, K., Barascud, N., Kashino, M., McDermott, J. H., & Chait, M. (2016). Sound segregation via embedded repetition is robust to inattention. *Journal of Experimental Psychology: Human Perception and Performance, 42*(3), 386.

McDermott, J. H. (2009). The cocktail party problem. *Current Biology, 19*(22), R1024–R1027.

Mehta, A. H., Jacoby, N., Yasin, I., Oxenham, A. J., & Shamma, S. A. (2017). An auditory illusion reveals the role of streaming in the temporal misallocation of perceptual objects. *Philosophical Transactions of the Royal Society B, 372*(1714), 20160114.

Michalka, S. W., Kong, L., Rosen, M. L., Shinn-Cunningham, B. G., & Somers, D. C. (2015). Short-term memory for space and time flexibly recruit complementary sensory-biased frontal lobe attention networks. *Neuron, 87*(4), 882–892. doi:10.1016/j.neuron.2015.07.028.

Michalka, S. W., Rosen, M. L., Kong, L., Shinn-Cunningham, B. G., & Somers, D. C. (2016). Auditory spatial coding flexibly recruits anterior, but not posterior, visuotopic parietal cortex. *Cerebral Cortex, 26*(3), 1302–1308. doi:10.1093/cercor/bhv303

Middlebrooks, J. C. (2017). Spatial stream segregation. In J. C. Middlebrooks, J. Z. Simon, A. N. Popper, & R. R. Fay (Eds.), *The auditory system at the cocktail party* (pp. 137–168). New York: Springer.

Näätänen, R., Paavilainen, P., Rinne, T., & Alho, K. (2007). The mismatch negativity (MMN) in basic research of central auditory processing: A review. *Clinical Neurophysiology, 118*(12), 2544–2590.

Nahum, M., Nelken, I., & Ahissar, M. (2008). Low-level information and high-level perception: The case of speech in noise. *PLoS Biology, 6*(5), 978–991. doi:e12610.1371/journal.pbio.0060126.

Nelken, I. (2004). Processing of complex stimuli and natural scenes in the auditory cortex. *Current Opinion in Neurobiology, 14*(4), 474–480.

Noyce, A. L., Cestero, N., Michalka, S. W., Shinn-Cunningham, B. G., & Somers, D. C. (2017). Sensory-biased and multiple-demand processing in human lateral frontal cortex. *Journal of Neuroscience, 37*(36), 8755–8766.

Osher, D., Tobyne, S., Congden, K., Michalka, S., & Somers, D. (2015). Structural and functional connectivity of visual and auditory attentional networks: Insights from the Human Connectome Project. *Journal of Vision, 15*(12), 223. doi:10.1167/15.12.223.

O'Sullivan, J. A., Power, A. J., Mesgarani, N., Rajaram, S., Foxe, J. J., Shinn-Cunningham, B. G., Slaney, M., Shamma, S. A., & Lalor, E. C. (2014). Attentional selection in a cocktail party environment can be decoded from single-trial EEG. *Cerebral Cortex, 25*(7), 1697–1706. doi:10.1093/cercor/bht355.

Riecke, L., Sack, A. T., & Schroeder, C. E. (2015). Endogenous delta/theta sound-brain phase entrainment accelerates the buildup of auditory streaming. *Current Biology, 25*(24), 3196–3201. doi:10.1016/j.cub.2015.10.045.

Roverud, E., Best, V., Mason, C. R., Swaminathan, J., & Kidd Jr., G. (2016). Informational masking in normal-hearing and hearing-impaired listeners measured in a nonspeech pattern identification task. *Trends in Hearing, 20*. doi:10.1177/2331216516638516.

Ruggles, D., & Shinn-Cunningham, B. (2011). Spatial selective auditory attention in the presence of reverberant energy: Individual differences in normal-hearing listeners. *Journal of the Association for Research in Otolaryngology, 12*(3), 395–405. doi:10.1007/s10162-010-0254-z.

Schwartz, A. H., & Shinn-Cunningham, B. G. (2010). Dissociation of perceptual judgments of "what" and "where" in an ambiguous auditory scene. *Journal of the Acoustical Society of America, 128*(5), 3041–3051. doi:10.1121/1.3495942.

Shinn-Cunningham, B. G. (2008). Object-based auditory and visual attention. *Trends in Cognitive Sciences, 12*(5), 182–186. doi:10.1016/j.tics.2008.02.003.

Shinn-Cunningham, B. G., & Best, V. (2008). Selective attention in normal and impaired hearing. *Trends in Amplification, 12*(4), 283–299. doi:10.1177/1084713808325306.

Shomstein, S., & Yantis, S. (2004). Control of attention shifts between vision and audition in human cortex. *Journal of Neuroscience, 24*(47), 10702–10706.

Shomstein, S., & Yantis, S. (2006). Parietal cortex mediates voluntary control of spatial and nonspatial auditory attention. *Journal of Neuroscience, 26*(2), 435–439.

Sohoglu, E., & Chait, M. (2016). Detecting and representing predictable structure during auditory scene analysis. *eLife, 5*.

Southwell, R., Baumann, A., Gal, C., Barascud, N., Friston, K., & Chait, M. (2017). Is predictability salient? A study of attentional capture by auditory patterns. *Philosophical Transactions of the Royal Society B, 372*(1714), 20160105.

Teki, S., Barascud, N., Picard, S., Payne, C., Griffiths, T. D., & Chait, M. (2016). Neural correlates of auditory figure-ground segregation based on temporal coherence. *Cerebral Cortex, 26*(9), 3669–3680.

Tobyne, S. M., Osher, D. E., Michalka, S. W., & Somers, D. C. (2017). Sensory-biased attention networks in human lateral frontal cortex revealed by intrinsic functional connectivity. *NeuroImage, 162*, 362–372.

Westerhausen, R., Moosmann, M., Alho, K., Belsby, S.-O., Hämäläinen, H., Medvedev, S., Specht, K., & Hugdahl, K. (2010). Identification of attention and cognitive control networks in a parametric auditory fMRI study. *Neuropsychologia, 48*(7), 2075–2081.

Woods, K. J., & McDermott, J. H. (2015). Attentive tracking of sound sources. *Current Biology, 25*(17), 2238–2246. doi:10.1016/j.cub.2015.07.043.

Xiang, J., Simon, J., & Elhilali, M. (2010). Competing streams at the cocktail party: Exploring the mechanisms of attention and temporal integration. *Journal of Neuroscience, 30*(36), 12084–12093. doi:10.1523/JNEUROSCI.0827-10.2010.

Zion Golumbic, E. M., Ding, N., Bickel, S., Lakatos, P., Schevon, C. A., McKhann, G. M., Goodman, R. R., Emerson, R., Mehta, A. D., Simon, J. Z., Poeppel, D., & Schroeder, C. E. (2013). Mechanisms underlying selective-neuronal tracking of attended speech at a "cocktail party." *Neuron, 77*(5), 980–991. doi:10.1016/j.neuron.2012.12.037.