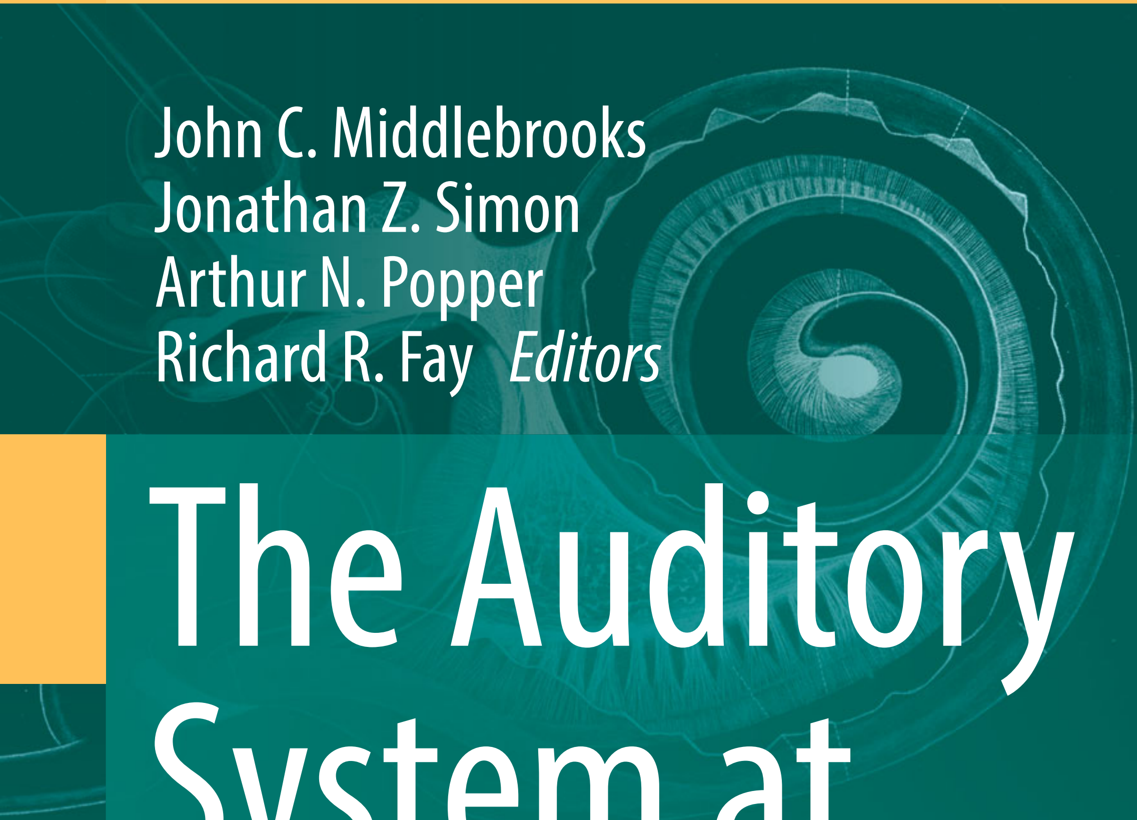
Springer Handbook of Auditory Research



John C. Middlebrooks
Jonathan Z. Simon
Arthur N. Popper
Richard R. Fay *Editors*

The Auditory System at the Cocktail Party

ASA Press

Springer

Springer Handbook of Auditory Research

Volume 60

More information about this series at http://www.springer.com/series/2506

John C. Middlebrooks · Jonathan Z. Simon
Arthur N. Popper · Richard R. Fay
Editors

The Auditory System at the Cocktail Party

With 41 Illustrations

ASA Press

Springer

Chapter 2
Auditory Object Formation and Selection

Barbara Shinn-Cunningham, Virginia Best, and Adrian K.C. Lee

Abstract Most normal-hearing listeners can understand a conversational partner in an everyday setting with an ease that is unmatched by any computational algorithm available today. This ability to reliably extract meaning from a sound source in a mixture of competing sources relies on the fact that natural, meaningful sounds have structure in both time and frequency. Such structure supports two processes that enable humans and animals to solve the cocktail party problem: auditory object formation and auditory object selection. These processes, which are closely intertwined and difficult to isolate, are linked to previous work on auditory scene analysis and auditory attention, respectively. This chapter considers how the brain may implement object formation and object selection. Specifically, the chapter focuses on how different regions of the brain cooperate to isolate the neural representation of sound coming from a source of interest and enhance it while suppressing the responses to distracting or unimportant sounds in a sound mixture.

Keywords Auditory grouping · Auditory streaming · Cocktail party · Energetic masking · Informational masking · Scene analysis · Selective attention

B. Shinn-Cunningham (✉)
Center for Research in Sensory Communication and Emerging Neural Technology,
Boston University, 677 Beacon St., Boston, MA 02215, USA
e-mail: shinn@bu.edu

V. Best
Department of Speech, Language and Hearing Sciences,
Boston University, 635 Commonwealth Ave., Boston, MA 02215, USA
e-mail: ginbest@bu.edu

A.K.C. Lee
Department of Speech and Hearing Sciences, Institute for Learning
and Brain Sciences (I-LABS), University of Washington,
1715 Columbia Road NE, Seattle, WA 98195-7988, USA
e-mail: akclee@uw.edu

J.C. Middlebrooks et al. (eds.), *The Auditory System*
at the Cocktail Party, Springer Handbook of Auditory Research 60,
DOI 10.1007/978-3-319-51662-2_2

2.1 Introduction

Most normal-hearing listeners can understand a conversational partner in everyday social settings, even when there are competing sounds from different talkers and from other ordinary sounds. Yet when one analyzes the signals reaching a listener's ears in such settings, this ability seems astonishing. In fact, despite the ubiquity of computational power today, even the most sophisticated machine listening algorithms cannot yet reliably extract meaning from everyday sound mixtures with the same skill as a toddler. Understanding how humans and other animals solve this "cocktail party problem" has interested auditory researchers for more than a half century (Cherry 1953).

This chapter reviews how different sound properties, operating on different time scales, support two specific processes that enable humans and animals to solve the cocktail party problem. Specifically, the chapter concentrates on the interrelated processes of auditory object formation and auditory object selection. A discussion of how the brain may implement these processes concludes the chapter.

2.1.1 *The Cocktail Party: Confusing Mixtures and Limited Processing Capacity*

To illustrate these ideas, consider Fig. 2.1, which presents a very simple auditory scene consisting of messages from two different talkers (see the spectrogram of the mixture in Fig. 2.1A, while the individual messages are shown in Fig. 2.1B and C, in blue and red, respectively). Many natural signals, such as speech, are relatively sparse in time and frequency. Luckily, this means that the time–frequency overlap of signals in a sound mixture is often modest (the signals do not fully mask each other "energetically"; see Culling and Stone, Chap. 3). For instance, in a mixture of two equally loud voices, the majority of each of the signals is audible. That can be seen in Fig. 2.1D, which labels each time–frequency point at which only one of the two sources has significant energy as either blue or red, depending on which source dominates. The points of overlap, where there is significant energy in both sources, are shown in green. To make sense of either one of the messages making up the mixture, one simply needs to know which energy is from that source. That is, either the red or blue time–frequency points in Fig. 2.1D represent enough of the respective message's information for it to be easily understood.

Unfortunately, there are many different "solutions" to the question of what produced any given sound mixture. For instance, in looking at Fig. 2.1A, where the mixture is not color labeled, one notes there are an infinite number of ways that the mixture could have come about. In fact, even knowing how many sound sources there are does not make it possible to determine what energy came from what source without making assumptions. The first broad burst of energy in Fig. 2.1C, representing the /ih/ sound in "It's" (see text annotation above the spectrogram)

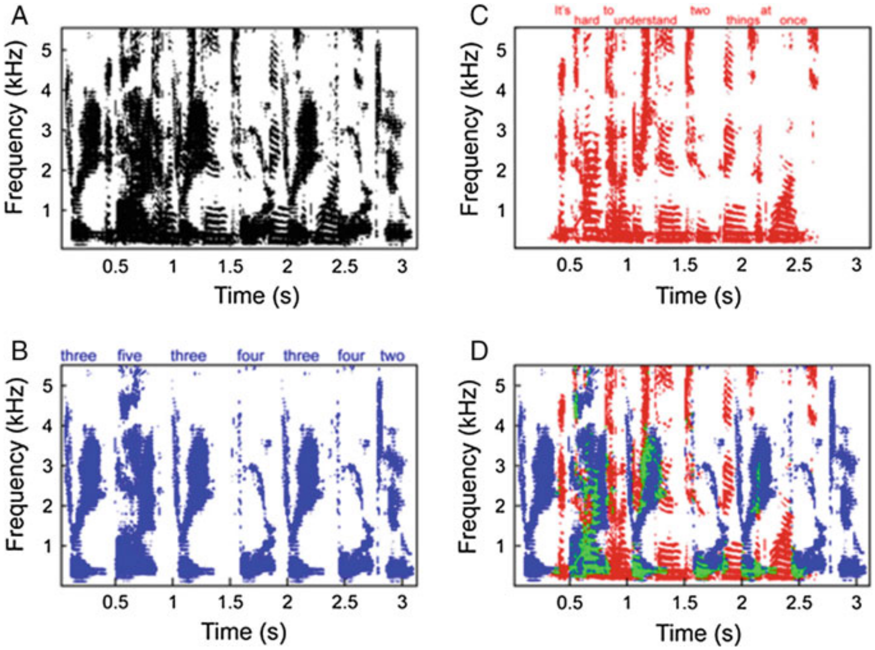


Fig. 2.1 Demonstration of how time–frequency sparseness leads to sound mixtures where clean "glimpses" of the component sounds are preserved, using two independent speech streams. (**A**) A thresholded spectrogram showing all time–frequency tiles with significant energy in a mixture of two sentences, added together. (**B, C**) Individual thresholded spectrograms of the two sentences making up the mixture shown in **A** and **D**. (**D**) A color-coded thresholded spectrogram, where each time–frequency tile is color coded depending on whether it is dominated by the sentence shown in **B** (blue), dominated by the sentence shown in **C** (red), or is a mixture of overlapping sound from the two sentences, leading to interference (green)

shows that there are three bands of energy visible that turn on and off together. Theoretically, each could have come from a different source (for that matter, portions of each could be from different sources); there is no way to determine unambiguously that they are from the same source. The brain seems to solve this mathematically underdetermined problem of estimating what mixture energy belongs to a particular external source by making educated guesses based on knowledge about the statistical properties of typical natural sounds. For instance, although it could have been a coincidence that all three bursts have a similar time course, that is unlikely—especially given that together, they sound like a voice making the vowel /ih/. In other words, to make sense of the acoustic world, the brain uses prior information about the spectrotemporal structure of natural sounds to group together acoustic energy that belongs together. As discussed further in Sect. 2.2, this process of *auditory object formation*, or estimating which components of a sound mixture came from the same external sound source, is an important part of solving the cocktail party problem.

Yet, even when auditory objects are easy to form from a sound mixture, listeners have difficulty understanding important sounds if they cannot select the proper object for analysis. This problem, of *auditory object selection*, is critical because listeners do not actually exhaustively analyze the content of every object in a multiobject scene. Although in theory one could imagine the brain recognizing the content of every source in a scene in parallel, there is a limit to the processing capacity of the brain. As a result, in most situations, listeners focus selective attention on one source for detailed analysis and suppress other competing sources (for a comprehensive review about auditory attention, see Fritz et al. 2007). The process of selecting what object to attend in a scene is another key aspect to listeners solving the cocktail party problem.

Together, the processes of forming and selecting auditory objects from a scene constitute different aspects of how auditory selective attention operates. These processes allow listeners to understand whatever auditory object seems most important at a given time, based jointly on the volitional goals of the listener on the statistics of the sound mixture, which can automatically guide attention to an unexpected event. For instance, when one is trying to listen to a dinner companion in a crowded restaurant, attention may nonetheless be drawn automatically to the crash of a dinner plate splintering as it hits the tile floor (Desimone and Duncan 1995). Many of the issues covered in this chapter are discussed in the literature in terms of these attentional processes.

2.1.2 *Object-Based Attention*

The ideas that central processing resources are limited and that attention determines what object the brain analyzes in a complex scene are not unique to the auditory system. In visual neuroscience, it is assumed that attention operates on *visual objects*. In her influential feature integration theory, Anne Treisman (see Treisman and Gelade 1980) proposed that visual stimulus features (color, shape, orientation, movement) are registered automatically and preattentively and are bound together into a coherent object (a perceptual rather than physical entity) when focused attention is directed to one or more of the elements of that object. If attention is focused on a location in space where the corner of a red triangle appears, the other corners, which together with the attended corner form the triangle, are also brought into attentional focus. It has since been argued that auditory objects are the “units” on which selective auditory attention operates (Shinn-Cunningham 2008; Shinn-Cunningham and Best 2008). Moreover, research suggests that inputs from different sensory modalities can bind together, creating objects comprising information from different modalities. For instance, when an observer focuses on some feature of a multisensory object in one sensory modality, there is a transfer of attention to the information in other, “task-irrelevant,” modalities (Molholm et al. 2007). This kind of obligatory enhancement of information that is not relevant to a particular task, but that “comes along for the ride” when an observer focuses on one aspect of a perceptual object, is a hallmark of object-based attention.

2.1.3 Heterarchical Rather Than Hierarchical Processing

When first faced with the ideas of object formation and object selection, it feels intuitive to assume that these two processes are distinct and that they occur sequentially, with segregation first parsing a complex scene into constituent auditory objects, and then selection operating to pull out an important sound to allow it to be analyzed in detail. The reality is more complex. Rather than a hierarchy in which object formation occurs first, followed by selection, processing of an auditory scene is more heterarchical: formation and selection influence one another, feed back upon each other, and are not easily separable in terms of either how they are implemented in the brain or how their effects are measured behaviorally. In line with this, it is currently impossible to pinpoint exactly what neural processing stages support object formation or where they occur. Indeed, it is unlikely that there is one particular site in the pathway where objects “first emerge;” instead, an object-based representation likely emerges gradually and imperfectly as one traverses up the auditory pathway from the auditory nerve through the brainstem and midbrain to the various divisions of the cortex. Similarly, attentional selection does not happen at any one particular processing stage, but instead occurs at every stage. A meta-analysis in the vision literature summarizes this phenomenon beautifully in that sensory system: in the periphery of the system, the representation is determined strongly by the pattern of light entering the retina and weakly by what information a listener is trying to process, but at each progressive stage of processing, the influence of attention becomes stronger and the influence of the input stimulus relatively weaker (Serences and Yantis 2006a). The same appears to be true in the auditory system (compare, for instance, the weak effects of attention on the representation in the midbrain, e.g., Varghese et al. 2015, to the strong effects in cortex, Choi et al. 2013).

Despite this complication, this chapter is organized around the two ideas of object formation and selection because there are clearly cases in which listening in a complex setting breaks down because of failures of one rather than the other of these processes. Understanding these two processes and how they can break down is crucial, as failures of either object formation or object selection can lead to catastrophic communication failures. That is, it is not uncommon for a listener to fail to “hear” a sound because of central limitations on perception, despite the sound being well represented on the auditory nerve; critical information that is perfectly audible can be misunderstood or can go unnoticed by a human operator in a complex scene.

2.1.4 A Historical Note

Auditory psychologists initially led in studies of human selective attention, with some of the earliest work in the area focused on auditory communication signals (Cherry 1953; Broadbent 1958; Treisman 1960). In the 1970s and 1980s, though, as visual studies on attention flourished, hearing research focused on how information is coded in the auditory periphery, with relatively little emphasis on how central processing capacity limits perception. During this time, seminal work by Albert Bregman (reviewed in Bregman 1990) described the challenge of "auditory scene analysis." In his work, Bregman articulated many of the rules governing the perceptual organization of sound mixtures (a concept that is nearly synonymous with the idea of auditory object formation, as used in this chapter). Bregman's efforts inspired a host of psychoacoustic studies that built on and quantified the principles he articulated (e.g., Culling and Darwin 1993a; Darwin and Carlyon 1995); however, most of these studies discussed how auditory scenes are parsed without any explicit discussion of the role of attention. Moreover, when auditory researchers did explore what happens when central bottlenecks, rather than sensory limitations, determined performance, the work was rarely related to modern theories of attention and memory. Instead, the term "informational masking" was coined to encompass any perceptual interference between sounds that was not explained by "energetic masking," which in turn was defined as interference explained by masking within the auditory nerve (for reviews, see Kidd et al. 2008; Kidd and Colburn, Chap. 4).

Whereas the field of hearing science largely ignored attentional studies, neuroimaging studies of auditory attention, typically using electroencephalography (EEG; Naatanen et al. 1992; Woldorff et al. 1993) or functional magnetic resonance imaging (fMRI; e.g., Pugh et al. 1996; Woodruff et al. 1996), were more common. These studies demonstrated the importance of attention in sculpting what auditory information is encoded in cortex and began to elucidate the cortical regions responsible for controlling attention (an issue we touch on in Sect. 2.6). Yet this work typically ignored how attentional performance depended on either early stages of sound encoding (e.g., in the cochlea, brainstem, and midbrain) or on auditory scene analysis. In short, historically, there was a great divide between hearing science and other aspects of neuroscience in understanding the cocktail party problem that has been gradually closing since the early 2000s.

A key realization that helped bridge this gap was that object formation and attention are best studied jointly (e.g., Shinn-Cunningham 2008). Interestingly, although the idea of object-based attention came from vision, there is relatively little discussion of the relationship between object formation and attention in that literature. It is not entirely clear why this is the case; historically, however, most visual attention studies use scenes consisting of very distinct, discrete objects (e.g., individual triangles and squares or individual letters), so that there is little ambiguity as to how to parse the inputs. In the auditory domain, failures of selective attention often arise because of failures to properly parse the acoustic scene into appropriate objects. Moreover, because auditory information (e.g., in speech) often unfolds over relatively long time scales (seconds), auditory selective attention depends on properly tracking auditory objects through time, a concept commonly referred to as "streaming." Given this, it may be that forming and streaming auditory objects is often inherently more challenging than forming visual objects.

A related omission in the visual literature on attention is a consideration of the time course of attention. Importantly, visual objects can often be defined without considering their temporal structure. Consider that a static two-dimensional picture of a natural scene generally contains enough information for visual objects to emerge without any further information. In contrast, auditory information is conveyed by changes in sounds as a function of time; it is the spectrotemporal content of sound that conveys a message's meaning. A "static" sound (such as stationary noise) has little informational content. Instead, basic temporal and spectral features and structure drive auditory stream formation. Because information in sound evolves through time, it takes time for listeners to make sense of what objects are in the scene, let alone to extract information about their content and meaning. Specifically, the perception of objects in a scene often emerges gradually. In turn, the ability to attend selectively to an object in the scene develops and becomes more specific over time. "Local" grouping features emerge over tens of milliseconds, but higher-order features and regularities can require on the order of seconds to be perceived (Cusack et al. 2004; Chait et al. 2010). Moreover, an auditory scene can be ambiguous, leading to an unstable percept (Hupe et al. 2008). For instance, over the course of seconds, a sequence of high and low tones may switch from being perceived as one stream to being perceived as two separate streams, and then switch back again. Current auditory theories deal directly with the fact that the percept of auditory objects evolves through time, and that this process may both influence and be influenced by attention (Elhilali et al., 2009a; Shamma et al. 2011).

2.2 Parsing the Acoustic Scene: Auditory Object Formation

All information in sound comes from its spectrotemporal structure. However, depending on the time scale being considered, this structure plays very different perceptual roles. For instance, we are sensitive to sound that has a frequency content ranging from 20 Hz to 20 kHz. Relatively rapid energy fluctuations in these acoustic signals determine perceptual attributes of an auditory object, such as its variation in loudness through time (for envelope fluctuations from about 5 Hz to 20 Hz), its "roughness" (for fluctuations between about 15 Hz and 75 Hz; e.g., see von Békésy 1960; Terhardt 1974), or its pitch (if there are regular fluctuations in the range from about 50 Hz to 4.5 kHz; e.g., see the review by Oxenham 2012). In contrast, object formation operates at two relatively long time scales: a "local" scale that helps bind together sound energy that is concurrent or spectrotemporally "connected" (discussed in Sect. 2.2.1), and a yet longer time scale that causes locally grouped energy bursts to connect into auditory objects that extend through time, forming what Bregman referred to as "streams" (discussed in Sect. 2.2.2).

2.2.1 Local Spectrotemporal Cues Support "Syllable-Level" Object Formation

Bregman noted several "local" features that cause sound elements to group together, perceptually, which he called "integration of simultaneous components" (see reviews by Carlyon 2004; Griffiths and Warren 2004). The rule of spectrotemporal proximity says that sounds that are close together and continuous in time and/or in frequency tend to be perceived as coming from the same source. Sounds that turn on and/or off together also tend to group together, even when they are far separated in frequency and "close together" only in time; more generally, sounds that have correlated fluctuations in amplitude modulation tend to group into the same perceptual object. Indeed, many of the studies of the psychoacoustic phenomenon of "co-modulation masking release" can be understood in terms of local grouping (Hall and Grose 1990; Oxenham and Dau 2001). The key modulations driving such object binding are slower than those that determine perceptual properties of sound (such as roughness and pitch), typically below about 7 Hz (e.g., Fujisaki and Nishida 2005; Maddox et al. 2015). Syllables in everyday spoken English have onset/offset envelopes whose fluctuations fall into this slow, below 10 Hz range, with durations typically between 100 and 450 ms (Greenberg et al. 2003). Note that although the word "syllable" often is used to refer exclusively to elements in human language, for the rest of this chapter, we use the term more generally to refer to distinct bursts of sound that cohere perceptually due to local spectrotemporal structure, even in the absence of linguistic structure.

Intuitively, it seems as if the spatial cues of concurrent sounds should impact auditory grouping strongly. However, instantaneous spatial cues actually are relatively weak cues for grouping at the syllabic level (Culling and Stone, Chap. 3). For instance, sound elements that turn on and off together tend to fuse together even if they have spatial cues that are inconsistent with one another (Darwin and Hukin 1997); conversely, spatial cues influence local grouping only weakly, with effects that may be observable only when other spectrotemporal cues are ambiguous (e.g., Shinn-Cunningham et al. 2007; Schwartz et al. 2012). This counterintuitive result may reflect the fact that spatial cues are derived, requiring a comparison of the inputs to the two ears, whereas amplitude and harmonic cues are inherent in the peripheral representation of sounds. The modest influence of spatial cues on object formation may also reflect the fact that in the real world, spatial cues are quite unreliable owing to effects of reverberation as well as interference from other sound sources (Palomaki et al. 2004; Ihlefeld and Shinn-Cunningham 2011). While such effects can distort interaural time and level differences quite significantly, their effects on amplitude modulation or harmonic structure are less pronounced; in line with this, moderate reverberant energy often degrades spatial cues significantly without interfering with perception of other sound properties, such as speech meaning (Culling et al. 1994; Ruggles et al. 2011). Although spatial cues have relatively weak effects on grouping at the syllabic level, when target and masker sources are at distinct locations, spatial cues can provide a strong basis for grouping of sequences of syllables into perceptual streams and for disentangling multiple interleaved sequences of sounds (Maddox and Shinn-Cunningham, 2012; Middlebrooks, Chap. 6).

Sounds that are harmonically related also tend to be perceived as having a common source, whereas inharmonicity can cause grouping to break down (Culling and Darwin 1993a; Culling and Stone, Chap. 3). Like spatial cues, though, harmonicity has less influence on local grouping than does common amplitude modulation (Darwin et al. 1995; Hukin and Darwin 1995).

On the surface, these local spectrotemporal grouping cues, both strong and weak, seem fundamentally different from one another. However, in a more abstract sense, they are similar: all reflect statistical correlations in acoustic spectrotemporal structure (either monaurally or binaurally) that tend to arise when sound energy is generated by a common source. For instance, just as it is likely that a single source produced sound elements whose amplitude envelopes are correlated, it is likely that one object with a particular resonant frequency generated concurrent sounds sharing a common fundamental frequency. In general, then, one can think of syllabic grouping as being driven by correlations in short-term spectrotemporal content that are typically present in natural sounds.

Most of the early studies of local grouping used rather simple auditory stimuli. For example, many studies explored how simultaneous pure tone bursts of different frequencies are integrated, manipulating properties such as whether or not they turn on and off together, are harmonically related, or share spatial properties (Darwin and Sutherland 1984; Darwin and Ciocca 1992; de Cheveigne et al. 1997). Such studies are useful for demonstrating that particular spectrotemporal cues can influence syllabic grouping; however, they do not necessarily reflect what happens in everyday settings. In particular, in most laboratory studies, only one feature is manipulated. Yet most "interesting" sounds, such as speech, musical sounds, or collision sounds, have rich spectrotemporal structure. The sound components generated by a real-world source typically have correlated envelope structure, related harmonic structure, and related localization cues. In such situations, these multiple cues all support the same local grouping of sound, rather than being pitted against one another (as is common in many psychoacoustic studies). Moreover, even in the absence of strong grouping cues, repetition of complex acoustic structures in the context of different mixtures can allow them to emerge as objects (McDermott et al. 2011). What this means is that in most natural listening situations, local grouping is relatively robust—at least when sounds are audible (i.e., not masking each other energetically; see Culling and Stone, Chap. 3 and Kidd and Colburn, Chap. 4). For instance, when listening in a cocktail party mixture, individual syllables often are heard; the real challenge is tracking the stream of such syllables from a particular talker over time.

2.2.2 Higher-Order Features Link Syllables into "Streams"

Grouping also occurs across longer time scales to bind together syllables into coherent streams ("integration of sequential components," in Bregman's terms). For example, humans perceive ongoing speech as one stream even though there are often silent gaps between syllables, across which local spectrotemporal continuity cannot operate. To create an auditory stream (a perceptual object composed of multiple syllables), higher-order perceptual features are key. For instance, the continuity or similarity of cues including frequency (Dannenbring 1976; De Sanctis et al. 2008), pitch (Culling and Darwin 1993a; Vliegen et al. 1999), timbre (Culling and Darwin 1993b; Cusack and Roberts 2000), amplitude modulation rate (Grimault et al. 2002), and spatial location (Darwin 2006; Maddox and Shinn-Cunningham 2012) of syllables presented in a sequence all contribute to hearing them as a single ongoing source. Just as with simultaneous grouping, many of the early studies of sequential grouping were conducted using very simple stimuli, such as tone or noise bursts, that rather than which have carefully controlled—and somewhat impoverished—higher-order features. In contrast, a particular talker produces a stream of speech in which there are a myriad of cues to distinguish it from competing streams.

Streaming based on continuity depends on computing relevant feature values in each of the syllables. These computations themselves depend on integrating information in the constituent elements making up each syllable (Darwin 2005). Consider, for example, a number of sinusoidal components that are heard as a syllable because they turn on and off together. As noted in Sect. 2.2.1, spatial cues in each component may be inconsistent with one another, yet not break down the syllabic grouping driven by the shared temporal course of the components. The perceived location of the syllable depends on combining this spatial information across all of the components, typically weighting low-frequency (300–600 Hz) interaural time differences relatively strongly compared to other spatial cues in other frequencies (Heller and Trahiotis 1996; Heller and Richards 2010). Whereas the spatial cues of each component have a weak impact on syllabic grouping, the continuity of the locations of sequential syllables can influence streaming; in fact, at this time scale, location plays an important role in streaming (Darwin and Hukin 2000; Best et al. 2008). Similarly, the pitch and timbre of a syllable depend on the harmonic relationships among all of its components, whereas streaming of a syllable with its temporal neighbors is influenced by the perceived pitches of the individual syllables (Oxenham 2008).

Because various syllabic features, such as location or pitch, strongly influence streaming, they therefore influence how we focus attention (Maddox and

Shinn-Cunningham 2012; Bressler et al. 2014). For instance, when listeners are asked to report back target words that share one feature amid simultaneous distractor words that may share some other task-irrelevant feature, such as pitch, the pitch cues nonetheless influence performance. Specifically, listeners are more likely to fail on such a task when the irrelevant pitch of one target word matches that of a subsequent distractor word; they are led astray by the task-irrelevant feature's continuity (Maddox and Shinn-Cunningham 2012). Another aspect of the strength of syllabic feature continuity is that when listeners are asked to focus attention on one sound feature, such as location, their ability to filter out distractors improves through time (Best et al. 2008; Bressler et al. 2014). These are parallel effects: there are higher rates of failure of selective attention when feature continuity works against the formation of a perceptually continuous stream of target words, and there are improvements in selective attention through time when feature continuity supports hearing the target words as one perceptual stream. Despite this obligatory influence of feature continuity on selective attention, listeners are able to control which of the words they hear from such a mixture to some degree, based on task instructions. This is a demonstration of top-down selection, discussed in Sect. 2.3.

2.2.3 Open Questions

The role of attention in auditory object formation remains a subject of debate. Some argue that objects form only when a stream (an auditory object extending through time) is attended (Alain and Woods 1997; Cusack et al. 2004). However, other studies suggest that auditory streams form automatically and preattentively (Macken et al. 2003; Sussman et al. 2007). Most likely, both automatic and attention-driven processes influence stream formation. In cases in which low-level attributes are sufficiently distinct to define a stream unambiguously, the sound object will be segregated from a sound mixture even without attention. But sound mixtures are often ambiguous, in which case attention to a particular perceptual feature may help "pull out" the stream that is attended (Alain et al. 2001; Macken et al. 2003). Moreover, listeners weight different acoustic cues that influence streaming differently depending on whether the cues are task relevant or task irrelevant (Maddox and Shinn-Cunningham 2012). In general, the view that top-down factors influence object formation is supported by studies that show that perception of a complex auditory scene is refined through time (Carlyon et al. 2003; Teki et al. 2013).

A related question is whether the attentional "background," comprising those parts of an acoustic scene that are not the focus of attention, is organized into objects or whether it remains undifferentiated. This question, although of great theoretical interest, is difficult to test, given that the only direct way to probe listeners' percepts of "the background" is to ask them what they perceive; however, the very act of asking this question is likely to cause them to focus attention on the background, so that it flips to become the foreground. Studies of neural, rather than behavioral, responses may help shed light on this important question (e.g., Lepisto et al. 2009).

2.3 Focusing Attention: Selecting What to Process

Even when auditory object and stream formation takes place accurately on the basis of the principles described in Sect. 2.2, listeners faced with complex auditory mixtures must select which object or stream to process. In the context of the cocktail party situation, it is impossible to process everything being said by every talker as well as to analyze the background sounds in detail. Moreover, such a comprehensive analysis is rarely the goal in everyday communication. Instead, selective processing of one, or maybe a few, talkers is generally the goal. In vision, attention is argued to operate as a "biased competition" between the neural representations of perceptual objects (Desimone and Duncan 1995; Kastner and Ungerleider 2001). The biased-competition view argues that the focus of attention is determined by the interplay between the salience of stimuli (exogenously guided attention) and observer goals (endogenously guided attention). However, biased competition arises specifically between objects, each of which is a collection of attributes. At any one time, one object is the focus of attention and is processed in greater detail than other objects in the scene. Evidence for such effects in auditory processing has started to emerge from physiological studies (Chait et al. 2010; Mesgarani and Chang 2012).

2.3.1 Top-Down Control Guides Selection

Listeners can selectively listen to one source in a mixture by directing top-down attention to different acoustic dimensions, many of which also influence object and stream formation. There are numerous examples demonstrating that listeners can focus attention on a certain frequency region (Greenberg and Larkin 1968; Scharf et al. 1987) or a certain spatial location (e.g., Arbogast and Kidd 2000; Kidd et al., 2005b) to improve detection or discrimination at a particular locus. There are also examples demonstrating that attention can be directed to pitch (Maddox and Shinn-Cunningham 2012), level (e.g., attending to the softer of two voices; Brungart 2001; Kitterick et al. 2013), and talker characteristics such as timbre and gender (e.g., Culling et al. 2003; Darwin et al. 2003). Auditory attention can also be focused in time, such that sounds occurring at expected times are better detected than those occurring at unpredictable times (Wright and Fitzgerald 2004; Varghese et al. 2012). This idea has been elaborated to describe attention that is distributed in time to either enhance sensitivity to target sequences ("rhythmic attention"; e.g., Jones et al. 1981) or to cancel irrelevant sounds (Devergie et al. 2010).

2.3.2 Bottom-up Salience Influences Attention

It is generally agreed that many bottom-up factors affect the inherent salience of an auditory stimulus. These include unexpectedness (e.g., a sudden door slam) and uniqueness, in which a sound stands out from the other sounds in the scene because of its features or statistics (for a computational model realizing these ideas, see Kaya and Elhilali 2014, and Elhilali Chap. 5). In the context of the cocktail party problem, one very often cited example of salience is the sound of one's own name, which can capture a listener's attention even when it occurs in an otherwise "unattended" stream (Moray 1959). Subsequent experiments show that the strength of this effect varies across listeners; moreover, the stronger the effect is, the worse a listener is at listening selectively to the "attended" stream (Wood and Cowan 1995; Conway et al. 2001). In any case, although this kind of attentional capture is stimulus driven rather than voluntary, the salience comes from the "learned importance" of that stimulus; in other words, some aspects of the bottom-up salience of auditory stimuli are not "preprogrammed" in the auditory system, but instead develop through long-term learning. The true impact of bottom-up salience is difficult to measure, as its strong interactions with top-down factors make it very difficult to isolate experimentally (Best et al., 2007b; Shuai and Elhilali 2014).

2.3.3 Extracting Meaning from Imperfect Objects

The problem of how objects are formed in the complicated mixtures of sounds that we encounter every day is one that continues to intrigue researchers. However, many natural sounds, particularly interesting ones such as speech and other animal vocalizations, are relatively sparse in time and frequency. Thus mixtures are not uniformly "mixed," and in fact many time–frequency units offer clear looks of one or another component sound source in a mixture. This natural segregation starts to fail when there are too many sources or in the presence of continuous unstructured noise or strong reverberation, both of which act to energetically mask potentially clean glimpses of sounds of interest.

When clean glimpses are available, even if they represent only fragments of a sound, they can be sufficient to allow a listener to identify that sound (Cooke 2006; Culling and Stone, Chap. 3). Such glimpsing can also support perceptual completion of information that is energetically masked. For example, a tone that is interrupted by a brief, loud noise is perceived as continuous even though it is acoustically completely swamped during the noise; in fact, even if the tone is interrupted and not actually present during the noise, it is perceived as if it is ongoing (the "continuity illusion"; Warren et al. 1988). This effect also applies to speech. When speech is interrupted periodically by silent gaps, intelligibility suffers, but if the gaps are filled by a gated noise, the speech is both perceived as continuous and rendered more intelligible ("phonemic restoration"; Warren 1970;

Samuel 1981). Phonemic restoration appears to be based on top-down knowledge that is either learned or hard-wired or both, and as such is influenced by cognitive and linguistic skills (Benard et al. 2014).

2.4 Perceptual Consequences of Object-Based Auditory Selective Attention

Object-based auditory attention has proven to be a challenging concept to test and even discuss. In addition to the difficulty of defining what constitutes an auditory object, it can also be difficult to define which object a listener is attending, especially if there is a hierarchy of objects in the scene. Still, there are a number of perceptual phenomena consistent with the idea that complex auditory scenes are naturally, and somewhat automatically, parsed into constituent objects that vie to be the focus of attention.

2.4.1 Failure to Divide Attention

There is evidence that listeners cannot actually divide attention between multiple simultaneous auditory objects. In fact this idea forms the basis of one of the paradigms used to measure stream segregation objectively: when presented with a sequence of interleaved tones of two different frequencies (A and B), judgments about the timing between neighboring A and B tones are impaired as the frequency separation is increased (i.e., as the sequence segregates into two distinct streams). The "change deafness" paradigm has been used to examine the role of selective and divided attention in busy, natural listening scenarios (Eramudugolla et al. 2005). Listeners are remarkably good at monitoring one object in a scene consisting of multiple, spatially separated natural sounds, and detecting its disappearance in a subsequent exposure to the scene, as long as selective attention is directed in advance to the object. In the absence of directed attention (i.e., when relying on divided attention) listeners are unable to detect the disappearance of one of the objects reliably: if the object that disappears is not in the focus of attention when it stops, listeners do not readily notice the change. Conversely, when listeners do focus attention selectively within a complex scene, it can leave them completely unaware of unusual or unexpected auditory events ("inattentional deafness"; e.g., see Dalton and Fraenkel 2012; Koreimann et al. 2014). There is also some evidence of an asymmetry when comparing the ability to detect a sudden disappearance versus detecting the sudden appearance of an object; when a sound suddenly appears, listeners are slightly better at detecting the change than when a sound suddenly disappears (Pavani and Turatto 2008). To the extent such asymmetry exists, it suggests that the appearance of a new event draws attention exogenously, whereas the disappearance of an unattended object does not.

In the case of speech, when listeners attend to one talker, they can recall little about unattended talkers (Cherry 1953). When instructed in advance to report back both of two brief competing messages, listeners can perform relatively well (Broadbent 1954; Best et al. 2006); however, it is not clear that this good performance indicates a true sharing of attention across streams. One possibility is that attention can be divided to a point, when the stimuli are brief, when the two tasks are not demanding, and/or when the two tasks do not compete for a limited pool of processing resources (Gallun et al. 2007; McCloy and Lee 2015). Another possibility is that simultaneous sensory inputs are stored temporarily via immediate auditory memory and then processed serially by a limited-capacity mechanism, which works reasonably well when recalling brief messages (Broadbent 1957; Lachter et al. 2004).

2.4.2 Obligatory Interactions Between Formation and Selection

Some of the strongest evidence that auditory objects are the units of auditory attention is in what information listeners can access and how grouping influences perception in an obligatory way. For instance, listeners have difficulty making judgments about individual frequency components within a complex tone or vowel; instead, they are obliged to make global judgments about the unitary auditory object. Importantly, by changing the surrounding context, this kind of obligatory integration of information can be dramatically reduced, demonstrating that it is likely because a component is a part of an object that its information is hard to analyze. For instance, the contribution of a mistuned harmonic to the pitch of a complex tone is reduced when the tone is perceived as a separate event, such as when it has a different onset from the other components or when it is “captured” into a different, sequential object (Darwin and Ciocca 1992; Darwin et al. 1995). Similarly, listeners can have difficulty judging the interaural cues of a high-frequency sound that is gated on and off with a low-frequency sound. However, if the low-frequency sound is preceded by a stream of identical low sounds, causing them to form one stream, the high-frequency element is “released,” its spatial cues dominate the perceived location of the now-separate high-frequency object, and discrimination of the high-frequency interaural cue becomes easy (Best et al., 2007a).

The influence of feature continuity on perception also supports the idea that objects are the focus of attention. As mentioned in Sect. 2.2.2, even when listeners try to ignore some task-irrelevant feature, the perceptual continuity of that feature influences the ability to extract information from a sound mixture. In particular, once a listener attends to one word, a subsequent word that shares some perceptual feature with the attended word is automatically more likely to be the focus of attention than a word that does not match the preceding word (Bressler et al. 2014). This result supports the idea that auditory objects extend through time, and that the resulting stream is the unit of attention.

Although these phenomena support the idea that selective auditory attention operates on perceptual objects, one of the complications is that object formation is *not* all or nothing. Take, for example, the distinction between attending to one instrument (or object) in an orchestra versus attending to the whole orchestra (itself also an object). Object formation can be thought of as a hierarchical structure in which objects form at different levels depending on contextual factors and listener goals (see Feldman 2003 for a similar argument about visual objects).

2.4.3 Costs of Switching Attention

A question that has interested researchers for many decades is how easily and rapidly selective attention can be switched from one object to another when the focus of interest changes. There are many examples showing that there is a cost associated with switching auditory attention. Early experiments demonstrated deficits in recall of speech items when presented alternately to the two ears (Cherry and Taylor 1954; Broadbent 1956). This cost is also apparent in more complex scenarios in which listeners must switch attention on cue between multiple simultaneous streams of speech (e.g., Best et al. 2008) or from one voice to another (Larson and Lee 2013; Lawo and Koch 2014). The cost of switching attention is associated with the time required to disengage and reengage attention, but may also come from an improvement in performance over time when listeners are able to hone the attentional filter more finely when they maintain focus on a single stream (Best et al. 2008; Bressler et al. 2014).

2.5 Neural Mechanisms Supporting Object Formation

There are a multitude of hypotheses and models concerning the neural underpinnings of auditory object formation. One hypothesis postulates that sound elements segregate into separate streams whenever they activate well-separated populations of auditory neurons, such as when the streams do not overlap in frequency (Micheyl et al. 2005). However, sounds can bind together into one perceptual group even if they excite distinct neural populations (Elhilali et al., 2009b). The temporal coherence theory (TCT) of object formation accounts for these results by assuming that when neurons encoding various sound features have responses that modulate coherently through time, the features are bound together, perceptually (Shamma et al. 2011; O'Sullivan et al. 2015). A multifeature representation such as that proposed in TCT provides a general and flexible framework for explaining how perceptual objects can emerge from a distributed neural code. The proposal that temporal coherence between different feature-selective neurons drives perceptual binding leverages two statistical aspects of a natural auditory scene: (1) In general, the strength of the response to a feature of a particular sound source will be proportional to the intensity of the source at a given moment, (2) The intensity of distinct sound sources, and thus the response to any associated features of the two sources, will be statistically independent over time. Attention has been hypothesized to influence object formation by modulating the temporal coherence of neural populations (O'Sullivan et al. 2015; see Gregoriou et al., 2009, for an example from the vision literature). When a listener selectively attends to a feature, this attentional focus is thought to up-regulate activity, which strengthens the binding of features that are temporally coherent with the attended feature.

Although this kind of theory is plausible, it does not address how an “object” is represented in a neural population. For instance, for selective attention to operate, the attended object and the competition must be separable in the neural code. Neural oscillations may help separate competing neural representations of different objects (Engel et al. 2001; Engel and Singer 2001). Growing evidence suggests that slow oscillations in the brain entrain to the syllabic structure of attended sound (Ding and Simon 2012a; Mesgarani and Chang 2012), and also that these oscillations gate information flow (i.e., that enhancement and suppression of sensory events occur, depending on the phase of these slow oscillations; Lakatos et al. 2013; Zion-Golumbic et al. 2013). Thus, slow neural oscillations are both driven by selectional focus (entraining to the syllabic rhythms of an attended stream) and support segregation by passing through information whose temporal information correlates with syllabic structure of the attended source. Just as effects of selection and segregation are intertwined perceptually, slow neural oscillations are driven by attention while at the same time supporting segregation. Such a selection–segregation mechanism could enable a type of temporal multiplexing of information, an idea with real appeal in the auditory realm, where competing signals often excite the same peripheral channels but with different time courses. Although such theories have some support, there remains a great deal to discover about where and how in the neural pathway an object-based representation of an attended sound emerges.

2.6 Neural Mechanisms Supporting Object Selection

In the past two decades, the field of cognitive neuroscience has witnessed a growing interest in understanding the mechanisms controlling attentional selection. This may be partly due to the rapid advancement of recording techniques that have enabled scientists to study the brain while an observer is engaged in attentionally demanding tasks. Both noninvasive techniques, such as fMRI, EEG, and magnetoencephalography (MEG), and invasive electrocorticography (intracranial recording from the exposed surface of the brain, typically done in conjunction with presurgery testing of epileptic patients) provide important, complementary information about how the human cortical response is modulated by attention. To a large degree, vision scientists have led the search for neural mechanisms underpinning attention. Given that the networks controlling attention seem at least partially to be shared across the senses (e.g., see Tark and Curtis 2009), understanding the attentional networks found by vision scientists is helpful for understanding the control of auditory attention. Thus, evidence about networks defined from visual studies is reviewed before returning to audition.

2.6.1 Visual Cognitive Networks Controlling Attention

Early work based on behavioral and lesion studies identified three different functional brain networks associated with different aspects of attentional control: the alerting, orienting, and executive networks (originally proposed by Posner and Petersen 1990). These basic ideas have since been expanded and refined (e.g., see Corbetta and Shulman 2002 and Petersen and Posner 2012).

The alerting network, which has been linked to the neuromodulator norepinephrine (NE), maintains vigilance throughout task performance. For instance, when a warning signal precedes a target event, there is a phasic change in alertness that leads to faster reaction times; the alerting network governs this sort of increase in responsiveness. Warning signals evoke activity in the locus coeruleus, which is the origin of an NE-containing neurochemical pathway that includes major nodes in the frontal cortex and in the parietal areas (Marrocco and Davidson 1998). Alerting is not closely linked to sensory modality, and is likely to affect auditory and visual processing similarly.

Orienting, originally associated with a single visual control network, appears instead to be controlled by at least two distinct networks relevant to auditory attention, one associated with spatial orienting of attention and the other with reorienting attention (Corbetta and Shulman 2002; Petersen and Posner 2012). The dorsal frontoparietal network (including the superior parietal lobe and the frontal eye fields [FEFs]) enables volitional focusing of attention to events at particular locations (e.g., see Bressler et al. 2008). In vision, there have been efforts to tease apart which parts of this spatial attention network are specifically controlling attention and which are controlling eye gaze, independent of spatial attention; however, this has proven difficult. Specifically, FEF, located in the premotor cortex, not only controls eye gaze but also participates in orienting attention independent of eye movements (i.e., directing "covert attention"; e.g., see Goldberg and Bruce 1985; Wardak et al. 2006). Indeed, it may be artificial to try to separate these functions. Moving the eyes changes the focus of spatial attention, and attending to an object makes one want to move one's eyes to an object's location, even if these eye movements are suppressed. Regardless, the dorsal frontoparietal network, which includes the FEF, is intimately involved in volitional focusing of visuospatial attention. As discussed further in Sect. 2.6.2, there is clear support for the idea that this orienting network is engaged during auditory spatial processing (Tark and Curtis 2009; Michalka et al. 2015).

A second, separate network, which runs more ventrally and includes the temporoparietal junction (TPJ), “interrupts” sustained, focused attention to allow observers to orient to new events (Corbetta et al. 2008). Interestingly, in the vision literature, this “reorienting” network has been associated primarily with bottom-up, stimulus-driven interruptions, such as from particularly salient or unexpected stimuli (e.g., see Serences and Yantis 2006b); however, many of the paradigms used to explore the role of “reorienting” in the vision literature do not test whether the reorienting network can be engaged by endogenous control (i.e., whether volitionally interrupting sustained attention also deploys the reorienting network). Moreover, there is support for the idea that volitional and stimulus-driven reorienting activates this more ventral attention network. Most current theories about the orienting and reorienting networks acknowledge that, although distinct, the two networks typically work together, dynamically, to direct visual attention (see Vossel et al. 2014). Note that the ventral reorienting network is distinct from an even more ventral network, known variously as the “what” or “action” pathway, which appears to be devoted almost exclusively to processing of visual form and visual features (Ungerleider and Mishkin 1982; Goodale and Milner 1992). Importantly, in visual studies, attention to a nonspatial feature (which one might expect to engage this ventral “what” pathway) may also cause activity in the more dorsal, “where” pathway (for review, see Ptak 2012). However, this engagement of the visuospatial attention network during “feature-based” attention may also be a consequence of how information is encoded; specifically, all of visual inputs are represented spatially, from the moment light hits the retina, and thus may always have “where” information associated with them.

Finally, executive control, which is associated with activity in the anterior cingulate and dorsal lateral prefrontal cortex (DLPFC), serves in decision making. For instance, the executive control network resolves conflict among potential responses (e.g., press a button with the right finger when there is a tone on the left and vice versa; Bush et al. 2000; Botvinick et al. 2001). Associated with processing of high-level, abstract concepts, executive control regions are likely engaged during judgments about various sensory inputs, regardless of modality.

2.6.2 Auditory Spatial Attention Engages Visual Orienting and Reorienting Networks

In audition research, more effort has been devoted to understanding how we direct attention (i.e., select what sound source to attend) than to the alerting or executive function. This perhaps reflects the fundamental question at the heart of the cocktail party problem: How does one recognize what another person is saying when there are multiple people speaking at the same time? As discussed in Sect. 2.3, many psychophysical studies have addressed how people orient attention or selectively attend to a particular sound object in a mixture.

A number of studies provide evidence that auditory spatial attention engages the frontoparietal spatial attention network documented in the vision literature. For instance, areas in this network are more active during spatial auditory tasks compared to when not performing a task, both in FEF (Tark and Curtis 2009; Michalka et al. 2015) and the intraparietal sulcus (IPS; Kong et al. 2014; Michalka et al. 2016). Moreover, the dorsal visuospatial network shows greater activation when listeners deploy spatial auditory processing compared to when they are attending some other acoustic feature, based on both MEG (Lee et al. 2013) and fMRI studies (Hill and Miller 2010; Michalka et al. 2015); interestingly, in some of these auditory studies, activity was asymmetrical, and greater in the left than in the right hemifield. Yet another MEG study showed that when listeners direct spatial attention to one of two sound streams, regions of the left precentral sulcus area (left PCS, most likely containing left FEF) phase lock to the temporal content of the attended, but not the unattended stream (Bharadwaj et al. 2014). These results show that auditory spatial processing engages many of the same brain regions as visual orienting, albeit with hints of a left hemisphere favoring asymmetry. Such an asymmetry is consistent with the view that left FEF may be part of a dorsal network controlling top-down attention, while right FEF may be more engaged during exogenous attention and attention shifting (Corbetta et al., 2008).

Similarly, dynamically switching spatial attention from one object to another in an auditory scene engages cortical regions such as those that are active when switching visual attention. In an imaging study combining MEG, EEG, and MRI anatomical information, listeners either maintained attention on one stream of letters throughout a trial or switched attention to a competing stream of letters after a brief gap (Larson and Lee 2014). The two competing streams were either separated spatially or differed in their pitch; therefore listeners either had to switch or maintain attention based on spatial or nonspatial cues. When listeners switched attention based on spatial features, the right TPJ (part of the reorienting network identified in visual studies) was significantly more active than when they switched focus based on pitch features. An fMRI study found that switching auditory attention from one auditory stream to another either voluntarily (based on a visual cue) or involuntarily (based on an unexpected, rare loud tone) evoked activity that overlapped substantially, and included areas associated with both the dorsal frontoparietal network (including FEF) and the reorienting network (including TPJ; see Alho et al., 2015). These results support the idea that auditory attention is focused by cooperative activity from the orienting and reorienting networks, and highlights the fact that even top-down, volitional switches of attention can evoke activity in the reorienting network.

2.6.3 Nonspatial Auditory Attention Differentially Engages Auditory-Specific Networks

While the visuospatial orienting and reorienting networks appear to be engaged by auditory tasks, direct contrasts between spatial and nonspatial auditory attention reveal activity in more auditory-specific processing regions. For instance, when listeners had to attend to one of two simultaneously presented syllables based on either location (left vs. right) or on pitch (high vs. low), network activity depended on how attention was deployed (Lee et al. 2013). Specifically, left (but not right) FEF, in the frontoparietal network, was significantly more active once a listener knew *where* a target sound would be located (even before it started), and stayed active throughout the spatial-based attention task; in contrast, when performing the same task based on the pitch of a syllable, the left posterior superior temporal sulcus (which has previously been associated with pitch categorization) showed enhanced activity (Lee et al. 2013). Similarly, in the switching study mentioned in Sect. 2.6.2, greater activity was found in the left inferior parietal supramarginal cortex (an area associated with memory processes in audition; see Vines et al. 2006; Schaal et al. 2013) when listeners switched attention based on pitch compared to when they switched attention based on location cues (Larson and Lee 2014). These results align with a previous fMRI study that contrasted spatial- and pitch-based auditory attention, which showed greater engagement of the dorsal frontoparietal network during spatial attention and greater engagement of auditory processing areas (in the inferior frontal gyrus) during pitch-based attention (Hill and Miller 2010). Thus, top-down attention to nonspatial auditory features differentially engages areas associated with auditory-specific processing, and causes less activity in the visuospatial orienting network.

2.6.4 Both Sensory Modality and Task Demands Affect Network Activity

A few studies underscore the emerging idea that which control networks are engaged by attention depends jointly on both the sensory modality of the input stimulus and the attributes to which attention is focused. In one fMRI study, directly contrasting activity during processing of auditory versus visual targets reveals interdigitated regions in lateral frontal cortex (LFC) that either favor visual attentional processing (superior precentral sulcus and inferior precentral sulcus) or auditory attentional processing (transverse gyrus intersecting precentral sulcus and caudal inferior frontal sulcus; see Michalka et al. 2015). These modality biases both are consistent with resting state analysis in individual subjects (i.e., the visual-biased LFC regions show intrinsic connectivity with visual sensory regions, whereas the auditory-biased LFC regions show intrinsic connectivity with auditory sensory regions; Michalka et al. 2015), and are also supported by analysis of anatomical connectivity using data taken from the Human Connectome Project (Osher et al. 2015). These new findings can be resolved with previous reports that suggest broad, cross-modal control regions in LFC (e.g., see the review by Duncan 2010), in part by understanding that averaging brain regions across subjects (the approach normally taken) blurs away important distinctions in these regions because of the challenge of co-registration of activity in frontal cortex, where individual variations in anatomical and function patterns can be significant.

Importantly, the kind of information that listeners had to extract from auditory and visual stimuli interacted with the modality of presentation in determining how LFC was engaged. Specifically, auditory LFC regions were active when either spatial or temporal information was extracted from sound; however, when spatial auditory information was processed, the visually biased LFC regions were also strongly recruited (Michalka et al. 2015). Conversely, visual LFC regions were active when either spatial or temporal information was extracted from visual inputs. When temporal visual information was processed, auditory LFC regions were significantly engaged, but when spatial visual information was processed, neither of the auditory LFC regions was significantly active. Similarly, parietal regions associated with the dorsal frontoparietal control network were engaged during auditory spatial tasks, but not during auditory temporal tasks (Michalka et al. 2016).

Figure 2.2 summarizes these findings: there seem to be two cooperating networks governing volitional control of auditory and visual attention. The "traditional" frontoparietal attention network appears to be engaged during visual tasks, regardless of task demands, as well as during spatial tasks, regardless of stimulus modality. In addition, there is a second "auditory–temporal" control network that is engaged during auditory tasks, regardless of task demands, as well as doing tasks

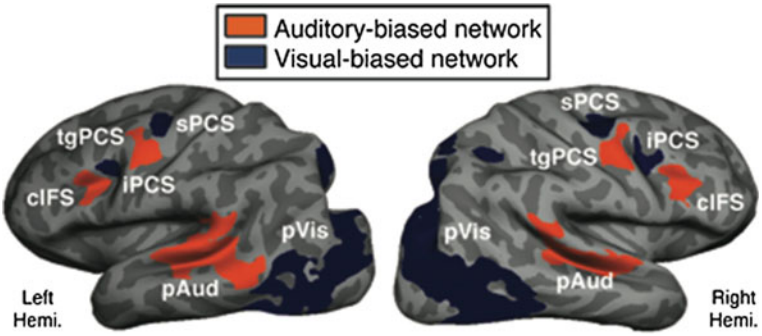


Fig. 2.2 Illustration of the brain regions making up auditory-biased (red) and vision-biased (blue) attentional control networks (derived from data reported in Michalka et al. 2015; figure provided by S. Michalka), shown on a "semi-inflated" map of the cortical surface (gyri shown in light gray; sulci shown in dark gray). The auditory-biased network includes two areas of lateral prefrontal cortex (LPC), the transverse gyrus intersecting precentral sulcus (tgPCS), and the caudal inferior frontal sulcus (cIFS), as well as sensory auditory regions (pAud). The visual-biased network includes two areas of lateral prefrontal cortex (LPC), the superior precentral sulcus (sPCS), and the inferior precentral sulcus (iPCS), as well as sensory visual regions (pVis)

that require judgments about temporal structure of inputs, regardless of stimulus modality. These results are consistent with the idea that vision excels at coding spatial information, while audition is a strongly temporal modality (Welch and Warren 1980); recruitment of the control network associated with the "other" modality may be the natural way to code information that does not match the natural strengths of a given sensory system (e.g., see Noyce et al. 2016).

2.6.5 Entrainment of Neural Responses to Attended Speech

Auditory streams evoke cortical responses that naturally reflect syllabic temporal structure. This structure can be captured using MEG and EEG, which have appropriate temporal resolution to reveal this activity (Simon, Chap. 7). For instance, for auditory stimuli with irregular rhythms, such as speech with its strong syllabic structure, one can find a linear kernel that predicts how the electric signals measured using MEG or EEG are related to the amplitude envelope of the input speech stream (Lalor et al. 2009; Lalor and Foxe 2010). In addition, because attention strongly modulates the strength of cortical responses, the temporal structure of neural MEG and EEG responses reflects the modulatory effects of attention. If a listener attends to one stream in a mixture of streams whose amplitude envelopes are uncorrelated, one can estimate which of the sources is being attended from MEG or EEG responses. For example, when listeners try to detect a rhythmic deviant in one of two isochronous tone sequences (repeating at 4 and 7 Hz, respectively), the neural power at the repetition rate of the attended stream is enhanced in MEG responses (Xiang et al. 2010). Similarly, when listeners selectively attend to one of two spoken stories, similar attentional modulation effects are seen in both EEG (Power et al. 2012) and MEG (Ding and Simon 2012b; Simon, Chap. 7). The attentional modulation of cortical responses is so strong that neural signals on single trials obtained from MEG and EEG can be used to decode which stream a listener is attending to in a mixture of melodies (Choi et al. 2013) or speech streams (Ding and Simon 2012b; O'Sullivan et al. 2014). These effects seem to be driven by responses in secondary sensory processing regions in the temporal lobe (e.g., planum temporale), but not in primary auditory cortex (Ding and Simon 2012b).

Patients undergoing medical procedures that require implantation of electrodes into the brain (for instance, to discover the focal source of epileptic seizures for surgical planning) now often agree to participate in studies of brain function (producing what is known as electrocorticography [ECoG], measured from penetrating or surface electrodes on the brain). A number of such patients have participated in studies of auditory attention. Signals from these studies have provided further insight into the neural encoding of attended and unattended auditory signals. Whereas the cortical coverage of ECoG is driven exclusively by clinical needs, and thus provides only a limited window on cortical activity, ECoG yields exquisite temporal and spatial resolution. In particular, the signal-to-noise ratio for high-frequency neural signals (especially in the high-gamma range of 80–150 Hz, which correlates with spiking activity in the underlying neural populations) is much greater in ECoG than with EEG or MEG.

One ECoG study analyzed the high gamma (75–150 Hz) local field potentials recorded directly from human posterior superior temporal gyrus (Mesgarani and Chang 2012), which provided an opportunity to estimate the speech spectrogram represented by the population neural response using a stimulus reconstruction method (Pasley et al. 2012). Subjects listened to a sentence presented either alone or simultaneously with another similar sentence spoken by a talker of the opposite gender. When an individual listned to a single sentence, the reconstructed spectrogram corresponded well to the spectrotemporal features of the original acoustic spectrogram. Importantly, the spectrotemporal encoding of the attended speaker in a two-speaker mixture also mirrored the neural response encoding that single speaker alone. A regularized linear classifier, trained on neural responses to an isolated speaker, was able to decode keywords of attended speech presented in the speech mixture. In trials in which the listener was able to report back the attended stream content, keywords from the attended sentence were decoded with high accuracy (around 80%). Equally telling, on trials in which the subject failed to correctly report back the target stream, decoding performance was significantly below chance, suggesting that the decoded signal was encoding the wrong sound, rather than that the encoded signal was too weak. In other words, it appeared that the errors were a consequence of improper *selection* by the subject, mirroring findings from psychoacoustic studies (e.g., Kidd et al., 2005a).

The aforementioned studies show that both low-frequency envelope-frequency oscillations and high-frequency gamma oscillations entrain to attended speech, consistent with the “selective entrainment hypothesis” (Giraud and Poeppel 2012; Zion-Golumbic and Schroeder 2012). Another ECoG study designed to characterize and compare speech-tracking effects in both low-frequency phase and high gamma power found that there were different spatial distributions and response time courses for these two frequency bands, suggesting that they reflect distinct aspects of attentional modulation in a cocktail party setting (Zion-Golumbic et al. 2013). Specifically, high-frequency gamma entrainment was found primarily in the superior temporal lobe (auditory sensory regions). In contrast, low-frequency (delta–theta rhythms, at syllabic rates of 1–7 Hz) had a wider topographic distribution that included not only low-level auditory areas but also higher-order language processing and attentional control regions such as inferior frontal cortex, anterior and inferior temporal cortex, and inferior parietal lobule. These results are consistent with growing evidence that neural encoding of complex stimuli relies on the combination of local processing, manifest in single-unit and multiunit activity (encoded by high-frequency gamma activity), and slow fluctuations that reflect modulatory control signals that regulate the phase of population excitability (e.g., Kayser et al. 2009; Whittingstall and Logothetis 2009).

2.6.6 *Other Neural Signatures of Focused Auditory Attention*

Attention not only causes portions of the brain to entrain to the attended input stimulus, but also affects neural oscillations that are not phase locked to the input. These changes are thought to reflect changes in the state of neural regions that encode and process inputs, such as changes in effort or load, or suppression of sensory information that is not the focus of attention.

One key example of such oscillatory effects is seen in the alpha oscillation band (roughly 8–12 Hz). In the visual occipital lobe, alpha oscillations that are not phase locked to any particular visual input are associated with suppression of visual processing (e.g., see Toscani et al. 2010). As discussed in Sect. 2.6.2, spatial processing of both visual and auditory stimuli is associated with the frontoparietal network, which is thought to have a lateralized encoding bias (e.g., sources on the left are coding strongly in right parietal regions). Consistent with this, spatial attention modulates the magnitude of alpha oscillations in parietal regions; the parietal region that is contralateral to a stimulus to be ignored typically has larger alpha oscillations, across modalities (see the review by Foxe and Snyder 2011). A growing number of auditory attention studies find that when spatial auditory attention is deployed, alpha activity is enhanced in parietal regions ipsilateral to the attended stimulus (consistent with suppression of sources that are contralateral to the target; e.g., see Kerlin et al. 2010; Strauss et al. 2014).

Studies of oscillations (such as alpha) that are not phase locked to input stimuli provide yet another way to measure neural activity associated with attentional selection. However, the mechanisms that produce such activity are still not understood. Future work exploring the circumstances that lead to these invoked oscillations and the time course of the activity and its generators will undoubtedly lead to even more insights into the processes governing auditory attentional control.

2.7 Summary Comments

As noted in the Introduction, it is amazing that humans communicate as well as they do, given the complexity of the problem of making sense of an acoustic signal in a crowded, noisy setting. In reality, though, the brain does not really “solve” the cocktail party problem. Instead, the brain assumes that the sources in today’s cocktail party are just like all the other sources in all past cocktail parties (both on an evolutionary time scale and over a lifetime of experience). Expectations constrain what we hear and perceive, helping us to form auditory objects out of a cacophony of competing sources. Although many aspects of object formation (at the levels of both the syllable and stream) appear to be automatic, they also influence and are influenced by object selection. Together, object formation and selection bring one perceived sound source into attentional focus, allowing the listener to analyze that object in detail.

Understanding these processes in the typically developing, healthy listener is of interest not only on theoretical grounds, but also because failures of these processes can have a crippling impact on the ability to communicate and interact in everyday settings. Because both object formation and object selection require a high-fidelity representation of spectrotemporal sound features, hearing impairment can lead to real difficulties in settings with competing sounds, even in listeners whose impairment allows them to communicate well in one-on-one settings (see discussion in Shinn-Cunningham and Best 2008; Litovsky, Goupell, Misurelli, and Kay, Chap. 10). Problems in the cocktail party are pronounced in cochlear implant users, who receive degraded spectrotemporal cues (e.g., see Loizou et al. 2009 and Litovsky et al., Chap. 10). In subclinical "hidden hearing loss," which is gaining increased attention in the field of hearing science, problems understanding sound in mixtures (but not in quiet settings) are often found (Plack et al. 2014; Bharadwaj et al. 2015). Other special populations, from listeners with attention-deficit disorder to veterans with mild traumatic brain injury to young adults with autism, struggle to communicate in complex settings owing to failures of executive control. Understanding how sensory and central factors interact to enable communication in everyday settings is a key step toward finding ways to ameliorate such communication disorders and improve the quality of life for listeners struggling at the cocktail party.

Compliance with Ethics Requirements
Barbara Shinn-Cunningham has no conflicts of interest.
Virginia Best has no conflicts of interest.
Adrian K. C. Lee has no conflicts of interest.

References

Alain, C., Arnott, S. R., & Picton, T. W. (2001). Bottom-up and top-down influences on auditory scene analysis: Evidence from event-related brain potentials. *Journal of Experimental Psychology: Human Perception and Performance, 27*(5), 1072–1089.

Alain, C., & Woods, D. L. (1997). Attention modulates auditory pattern memory as indexed by event-related brain potentials. *Psychophysiology, 34*(5), 534–546.

Alho, K., Salmi, J., Koistinen, S., Salonen, O., & Rinne, T. (2015). Top-down controlled and bottom-up triggered orienting of auditory attention to pitch activate overlapping brain networks. *Brain Research, 1626*, 136–145.

Arbogast, T. L., & Kidd, G., Jr. (2000). Evidence for spatial tuning in informational masking using the probe-signal method. *The Journal of the Acoustical Society of America, 108*(4), 1803–1810.

Benard, M. R., Mensink, J. S., & Başkent, D. (2014). Individual differences in top-down restoration of interrupted speech: Links to linguistic and cognitive abilities. *The Journal of the Acoustical Society of America*, 135, EL88–94.

Best, V., Gallun, F. J., Carlile, S., & Shinn-Cunningham, B. G. (2007a). Binaural interference and auditory grouping. *The Journal of the Acoustical Society of America, 121*(2), 1070–1076.

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

Best, V., Ozmeral, E. J., Kopco, N., & Shinn-Cunningham, B. G. (2008). Object continuity enhances selective auditory attention. *Proceedings of the National Academy of Sciences of the USA, 105*(35), 13174–13178.

Best, V., Ozmeral, E. J., & Shinn-Cunningham, B. G. (2007b). Visually-guided attention enhances target identification in a complex auditory scene. *Journal of the Association for Research in Otolaryngology, 8*(2), 294–304.

Bharadwaj, H. M., Lee, A. K. C., & Shinn-Cunningham, B. G. (2014). Measuring auditory selective attention using frequency tagging. *Frontiers in Integrative Neuroscience, 8*, 6.

Bharadwaj, H. M., Masud, S., Mehraei, G., Verhulst, S., & Shinn-Cunningham, B. G. (2015). Individual differences reveal correlates of hidden hearing deficits. *The Journal of Neuroscience, 35*(5), 2161–2172.

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review, 108*(3), 624–652.

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.

Bressler, S. L., Tang, W., Sylvester, C. M., Shulman, G. L., & Corbetta, M. (2008). Top-down control of human visual cortex by frontal and parietal cortex in anticipatory visual spatial attention. *The Journal of Neuroscience, 28*(40), 10056–10061.

Broadbent, D. E. (1954). The role of auditory localization in attention and memory span. *Journal of Experimental Psychology, 47*(3), 191–196.

Broadbent, D. E. (1956). Successive responses to simultaneous stimuli. *Quarterly Journal of Experimental Psychology*, 145–152.

Broadbent, D. E. (1957). Immediate memory and simultaneous stimuli. *Quarterly Journal of Experimental Psychology, 9*, 1–11.

Broadbent, D. E. (1958). *Perception and communication*. New York: Pergamon Press.

Brungart, D. S. (2001). Informational and energetic masking effects in the perception of two simultaneous talkers. *The Journal of the Acoustical Society of America, 109*(3), 1101–1109.

Bush, G., Luu, P., & Posner, M. I. (2000). Cognitive and emotional influences in anterior cingulate cortex. *Trends in Cognitive Sciences, 4*(6), 215–222.

Carlyon, R. P. (2004). How the brain separates sounds. *Trends in Cognitive Sciences, 8*(10), 465–471.

Carlyon, R. P., Plack, C. J., Fantini, D. A., & Cusack, R. (2003). Cross-modal and non-sensory influences on auditory streaming. *Perception, 32*(11), 1393–1402.

Chait, M., de Cheveigne, A., Poeppel, D., & Simon, J. Z. (2010). Neural dynamics of attending and ignoring in human auditory cortex. *Neuropsychologia, 48*(11), 3262–3271.

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *The Journal of the Acoustical Society of America, 25*, 975–979.

Cherry, E. C., & Taylor, W. K. (1954). Some further experiments upon the recognition of speech, with one and with two ears. *The Journal of the Acoustical Society of America, 26*, 554–559.

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.

Conway, A. R., Cowan, N., & Bunting, M. F. (2001). The cocktail party phenomenon revisited: The importance of working memory capacity. *Psychonomic Bulletin Review, 8*(2), 331–335.

Cooke, M. (2006). A glimpsing model of speech perception in noise. *The Journal of the Acoustical Society of America, 119*(3), 1562–1573.

Corbetta, M., Patel, G., & Shulman, G. L. (2008). The reorienting system of the human brain: From environment to theory of mind. *Neuron, 58*(3), 306–324.

Corbetta, M., & Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience, 3*(3), 201–215.

Culling, J. F., & Darwin, C. J. (1993a). Perceptual separation of simultaneous vowels: Within and across-formant grouping by F0. *The Journal of the Acoustical Society of America, 93*(6), 3454–3467.

Culling, J. F., & Darwin, C. J. (1993b). The role of timbre in the segregation of simultaneous voices with intersecting F0 contours. *Perception and Psychophysics, 54*(3), 303–309.

Culling, J. F., Hodder, K. I., & Toh, C. Y. (2003). Effects of reverberation on perceptual segregation of competing voices. *The Journal of the Acoustical Society of America, 114*(5), 2871–2876.

Culling, J. F., Summerfield, Q., & Marshall, D. H. (1994). Effects of simulated reverberation on the use of binaural cues and fundamental-frequency differences for separating concurrent vowels. *Speech Communication, 14*, 71–95.

Cusack, R., Deeks, J., Aikman, G., & Carlyon, R. P. (2004). Effects of location, frequency region, and time course of selective attention on auditory scene analysis. *Journal of Experimental Psychology: Human Perception and Performance, 30*(4), 643–656.

Cusack, R., & Roberts, B. (2000). Effects of differences in timbre on sequential grouping. *Perception and Psychophysics, 62*(5), 1112–1120.

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

Dannenbring, G. L. (1976). Perceived auditory continuity with alternately rising and falling frequency transitions. *Canadian Journal of Psychology, 30*(2), 99–114.

Darwin, C. J. (2005). Simultaneous grouping and auditory continuity. *Perception and Psychophysics, 67*(8), 1384–1390.

Darwin, C. J. (2006). Contributions of binaural information to the separation of different sound sources. *International Journal of Audiology, 45*(Supplement 1), S20–S24.

Darwin, C. J., Brungart, D. S., & Simpson, B. D. (2003). Effects of fundamental frequency and vocal-tract length changes on attention to one of two simultaneous talkers. *The Journal of the Acoustical Society of America, 114*(5), 2913–2922.

Darwin, C. J., & Carlyon, R. P. (1995). Auditory grouping. In B. C. J. Moore (Ed.), *Hearing* (pp. 387–424). San Diego: Academic Press.

Darwin, C. J., & Ciocca, V. (1992). Grouping in pitch perception: Effects of onset asynchrony and ear of presentation of a mistuned component. *The Journal of the Acoustical Society of America, 91*(6), 3381–3390.

Darwin, C. J., & Hukin, R. W. (1997). Perceptual segregation of a harmonic from a vowel by interaural time difference and frequency proximity. *The Journal of the Acoustical Society of America, 102*(4), 2316–2324.

Darwin, C. J., & Hukin, R. W. (2000). Effects of reverberation on spatial, prosodic, and vocal-tract size cues to selective attention. *The Journal of the Acoustical Society of America, 108*(1), 335–342.

Darwin, C. J., Hukin, R. W., & al-Khatib, B. Y. (1995). Grouping in pitch perception: Evidence for sequential constraints. *The Journal of the Acoustical Society of America, 98*(2 Pt 1), 880–885.

Darwin, C. J., & Sutherland, N. S. (1984). Grouping frequency components of vowels: When is a harmonic not a harmonic? *Quarterly Journal of Experimental Psychology, 36A,* 193–208.

de Cheveigne, A., McAdams, S., & Marin, C. M. H. (1997). Concurrent vowel identification. II. Effects of phase, harmonicity, and task. *The Journal of the Acoustical Society of America, 101,* 2848–2856.

De Sanctis, P., Ritter, W., Molholm, S., Kelly, S. P., & Foxe, J. J. (2008). Auditory scene analysis: The interaction of stimulation rate and frequency separation on pre-attentive grouping. *European Journal of Neuroscience, 27*(5), 1271–1276.

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

Devergie, A., Grimault, N., Tillmann, B., & Berthommier, F. (2010). Effect of rhythmic attention on the segregation of interleaved melodies. *The Journal of the Acoustical Society of America, 128*(1), EL1–7.

Ding, N., & Simon, J. Z. (2012a). Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. *Journal of Neurophysiology, 107*(1), 78–89.

Ding, N., & Simon, J. Z. (2012b). Emergence of neural encoding of auditory objects while listening to competing speakers. *Proceedings of the National Academy of Sciences of the USA, 109*(29), 11854–11859.

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.

Elhilali, M., Ma, L., Micheyl, C., Oxenham, A. J., & Shamma, S. A. (2009a). Temporal coherence in the perceptual organization and cortical representation of auditory scenes. *Neuron, 61*(2), 317–329.

Elhilali, M., Xiang, J., Shamma, S. A., & Simon, J. Z. (2009b). Interaction between attention and bottom-up saliency mediates the representation of foreground and background in an auditory scene. *PLoS Biology, 7*(6), e1000129.

Engel, A. K., Fries, P., & Singer, W. (2001). Dynamic predictions: oscillations and synchrony in top-down processing. *Nature Reviews Neuroscience, 2*(10), 704–716.

Engel, A. K., & Singer, W. (2001). Temporal binding and the neural correlates of sensory awareness. *Trends in Cognitive Sciences, 5*(1), 16–25.

Eramudugolla, R., Irvine, D. R., McAnally, K. I., Martin, R. L., & Mattingley, J. B. (2005). Directed attention eliminates 'change deafness' in complex auditory scenes. *Current Biology, 15*(12), 1108–1113.

Feldman, J. (2003). What is a visual object? *Trends in Cognitive Sciences, 7*(6), 252–256.

Foxe, J. J., & Snyder, A. C. (2011). The role of alpha-band brain oscillations as a sensory suppression mechanism during selective attention. *Frontiers of Psychology, 2,* 154.

Fritz, J. B., Elhilali, M., David, S. V., & Shamma, S. A. (2007). Auditory attention: Focusing the searchlight on sound. *Current Opinion in Neurobiology, 17*(4), 437–455.

Fujisaki, W., & Nishida, S. (2005). Temporal frequency characteristics of synchrony-asynchrony discrimination of audio-visual signals. *Experimental Brain Research, 166*(3–4), 455–464.

Gallun, F. J., Mason, C. R., & Kidd, G., Jr. (2007). The ability to listen with independent ears. *The Journal of the Acoustical Society of America, 122*(5), 2814–2825.

Giraud, A. L., & Poeppel, D. (2012). Cortical oscillations and speech processing: Emerging computational principles and operations. *Nature Neuroscience, 15*(4), 511–517.

Goldberg, M. E., & Bruce, C. J. (1985). Cerebral cortical activity associated with the orientation of visual attention in the rhesus monkey. *Vision Research, 25*(3), 471–481.

Goodale, M. A., & Milner, A. D. (1992). Separate visual pathways for perception and action. *Trends in Neurosciences, 15*(1), 20–25.

Greenberg, S., Carvey, H., Hitchcock, L., & Chang, S. (2003). Temporal properties of spontaneous speech—A syllable-centric perspective. *Journal of Phonetics, 31*(3–4), 465–485.

Greenberg, G. Z., & Larkin, W. D. (1968). Frequency-response characteristic of auditory observers detecting signals of a single frequency in noise: The probe-signal method. *The Journal of the Acoustical Society of America, 44*(6), 1513–1523.

Gregoriou, G. G., Gotts, S. J., Zhou, H., & Desimone, R. (2009). High-frequency, long-range coupling between prefrontal and visual cortex during attention. *Science, 324*(5931), 1207–1210.

Griffiths, T. D., & Warren, J. D. (2004). What is an auditory object? *Nature Reviews Neuroscience, 5*(11), 887–892.

Grimault, N., Bacon, S. P., & Micheyl, C. (2002). Auditory stream segregation on the basis of amplitude-modulation rate. *The Journal of the Acoustical Society of America, 111*(3), 1340–1348.

Hall, J. W., 3rd, & Grose, J. H. (1990). Comodulation masking release and auditory grouping. *The Journal of the Acoustical Society of America, 88*(1), 119–125.

Heller, L. M., & Richards, V. M. (2010). Binaural interference in lateralization thresholds for interaural time and level differences. *The Journal of the Acoustical Society of America, 128*(1), 310–319.

Heller, L. M., & Trahiotis, C. (1996). Extents of laterality and binaural interference effects. *The Journal of the Acoustical Society of America, 99*(6), 3632–3637.

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

Hukin, R. W., & Darwin, C. J. (1995). Comparison of the effect of onset asynchrony on auditory grouping in pitch matching and vowel identification. *Perception and Psychophysics, 57*(2), 191–196.

Hupe, J. M., Joffo, L. M., & Pressnitzer, D. (2008). Bistability for audiovisual stimuli: Perceptual decision is modality specific. *Journal of Vision, 8*(7), 11–15.

Ihlefeld, A., & Shinn-Cunningham, B. G. (2011). Effect of source spectrum on sound localization in an everyday reverberant room. *The Journal of the Acoustical Society of America, 130*(1), 324–333.

Jones, M. R., Kidd, G., & Wetzel, R. (1981). Evidence for rhythmic attention. *Journal of Experimental Psychology: Human Perception and Performance, 7*(5), 1059–1073.

Kastner, S., & Ungerleider, L. G. (2001). The neural basis of biased competition in human visual cortex. *Neuropsychologia, 39*(12), 1263–1276.

Kaya, E. M., & Elhilali, M. (2014). Investigating bottom-up auditory attention. *Frontiers in Human Neuroscience, 8*(327), 1–12.

Kayser, C., Montemurro, M. A., Logothetis, N. K., & Panzeri, S. (2009). Spike-phase coding boosts and stabilizes information carried by spatial and temporal spike patterns. *Neuron, 61*(4), 597–608.

Kerlin, J. R., Shahin, A. J., & Miller, L. M. (2010). Attentional gain control of ongoing cortical speech representations in a "cocktail party". *The Journal of Neuroscience, 30*(2), 620–628.

Kidd, G., Jr., Arbogast, T. L., Mason, C. R., & Gallun, F. J. (2005a). The advantage of knowing where to listen. *The Journal of the Acoustical Society of America, 118*(6), 3804–3815.

Kidd, G., Mason, C. R., Brughera, A., & Hartmann, W. M. (2005b). The role of reverberation in release from masking due to spatial separation of sources for speech identification. *Acta Acustica united with Acustica, 91*(3), 526–536.

Kidd, G., Jr., Mason, C. R., Richards, V. M., Gallun, F. J., & Durlach, N. I. (2008). Informational Masking. In W. Yost, A. Popper, & R. Fay (Eds.), *Auditory perception of sound sources* (pp. 143–189). New York: Springer Science+Business Media.

Kitterick, P. T., Clarke, E., O'Shea, C., Seymour, J., & Summerfield, A. Q. (2013). Target identification using relative level in multi-talker listening. *The Journal of the Acoustical Society of America, 133*(5), 2899–2909.

Kong, L., Michalka, S. W., Rosen, M. L., Sheremata, S. L., et al. (2014). Auditory spatial attention representations in the human cerebral cortex. *Cerebral Cortex, 24*(3), 773–784.

Koreimann, S., Gula, B., & Vitouch, O. (2014). Inattentional deafness in music. *Psychological Research, 78,* 304–312.

Lachter, J., Forster, K. I., & Ruthruff, E. (2004). Forty-five years after Broadbent (1958): Still no identification without attention. *Psychological Review, 111*(4), 880–913.

Lakatos, P., Musacchia, G., O'Connel, M. N., Falchier, A. Y., Javitt, D. C., & Schroeder, C. E. (2013). The spectrotemporal filter mechanism of auditory selective attention. *Neuron, 77*(4), 750–761.

Lalor, E. C., & Foxe, J. J. (2010). Neural responses to uninterrupted natural speech can be extracted with precise temporal resolution. *European Journal of Neuroscience, 31*(1), 189–193.

Lalor, E. C., Power, A. J., Reilly, R. B., & Foxe, J. J. (2009). Resolving precise temporal processing properties of the auditory system using continuous stimuli. *Journal of Neurophysiology, 102*(1), 349–359.

Larson, E., & Lee, A. K. C. (2013). Influence of preparation time and pitch separation in switching of auditory attention between streams. *The Journal of the Acoustical Society of America, 134* (2), EL165–171.

Larson, E., & Lee, A. K. C. (2014). Switching auditory attention using spatial and non-spatial features recruits different cortical networks. *NeuroImage, 84,* 681–687.

Lawo, V., & Koch, I. (2014). Dissociable effects of auditory attention switching and stimulus–response compatibility. *Psychological Research, 78,* 379–386.

Lee, A. K. C., Rajaram, S., Xia, J., Bharadwaj, H., et al. (2013). Auditory selective attention reveals preparatory activity in different cortical regions for selection based on source location and source pitch. *Frontiers in Neuroscience, 6*, 190.

Lepisto, T., Kuitunen, A., Sussman, E., Saalasti, S., et al. (2009). Auditory stream segregation in children with Asperger syndrome. *Biological Psychology, 82*(3), 301–307.

Loizou, P. C., Hu, Y., Litovsky, R., Yu, G., et al. (2009). Speech recognition by bilateral cochlear implant users in a cocktail-party setting. *The Journal of the Acoustical Society of America, 125* (1), 372–383.

Macken, W. J., Tremblay, S., Houghton, R. J., Nicholls, A. P., & Jones, D. M. (2003). Does auditory streaming require attention? Evidence from attentional selectivity in short-term memory. *Journal of Experimental Psychology: Human Perception and Performance, 29*(1), 43–51.

Maddox, R. K., Atilgan, H., Bizley, J. K., & Lee, A. K. (2015). Auditory selective attention is enhanced by a task-irrelevant temporally coherent visual stimulus in human listeners. *Elife, 4*. doi:10.7554/eLife.04995

Maddox, R. K., & Shinn-Cunningham, B. G. (2012). Influence of task-relevant and task-irrelevant feature continuity on selective auditory attention. *Journal of the Association for Research in Otolaryngology, 13*(1), 119–129.

Marrocco, R. T., & Davidson, M. C. (1998). Neurochemistry of attention. In R. Parasuraman (Ed.), *The attentive brain* (Vol. xii, pp. 35–50). Cambridge, MA: MIT Press.

McCloy, D. R., & Lee, A. K. (2015). Auditory attention strategy depends on target linguistic properties and spatial configuration. *The Journal of the Acoustical Society of America, 138*(1), 97–114.

McDermott, J. H., Wrobleski, D., & Oxenham, A. J. (2011). Recovering sound sources from embedded repetition. *Proceedings of the National Academy of Sciences of the USA, 108,* 1188–1193.

Mesgarani, N., & Chang, E. F. (2012). Selective cortical representation of attended speaker in multi-talker speech perception. *Nature, 485*(7397), 233–236.

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.

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.

Micheyl, C., Tian, B., Carlyon, R. P., & Rauschecker, J. P. (2005). Perceptual organization of tone sequences in the auditory cortex of awake macaques. *Neuron, 48*(1), 139–148.

Molholm, S., Martinez, A., Shpaner, M., & Foxe, J. J. (2007). Object-based attention is multisensory: Co-activation of an object's representations in ignored sensory modalities. *European Journal of Neuroscience, 26*(2), 499–509.

Moray, N. (1959). Attention in dichotic listening: Affective cues and the influence of instructions. *Quarterly Journal of Experimental Psychology, 11,* 56–60.

Naatanen, R., Teder, W., Alho, K., & Lavikainen, J. (1992). Auditory attention and selective input modulation: A topographical ERP study. *NeuroReport, 3*(6), 493–496.

Noyce, A. L., Cestero, N., Shinn-Cunningham, B. G., & Somers, D. C. (2016). Short-term memory stores organized by information domain. *Attention, Perception, & Psychophysics, 78* (30), 960–970.

O'Sullivan, J. A., Power, A. J., Mesgarani, N., Rajaram, S., et al. (2014). Attentional selection in a cocktail party environment can be decoded from single-trial EEG. *Cerebral Cortex, 25*(7), 1697–1706.

O'Sullivan, J. A., Shamma, S. A., & Lalor, E. C. (2015). Evidence for neural computations of temporal coherence in an auditory scene and their enhancement during active listening. *The Journal of Neuroscience, 35*(18), 7256–7263.

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.

Oxenham, A. J. (2008). Pitch perception and auditory stream segregation: Implications for hearing loss and cochlear implants. *Trends in Amplification, 12*(4), 316–331.

Oxenham, A. J. (2012). Pitch perception. *The Journal of Neuroscience, 32*(39), 13335–13338.

Oxenham, A. J., & Dau, T. (2001). Modulation detection interference: Effects of concurrent and sequential streaming. *The Journal of the Acoustical Society of America, 110*(1), 402–408.

Palomaki, K. J., Brown, G. J., & Wang, D. L. (2004). A binaural processor for missing data speech recognition in the presence of noise and small-room reverberation. *Speech Communication, 43* (4), 361–378.

Pasley, B. N., David, S. V., Mesgarani, N., Flinker, A., et al. (2012). Reconstructing speech from human auditory cortex. *PLoS Biology, 10*(1), e1001251.

Pavani, F., & Turatto, M. (2008). Change perception in complex auditory scenes. *Perception and Psychophysics, 70*(4), 619–629.

Petersen, S. E., & Posner, M. I. (2012). The attention system of the human brain: 20 years after. *Annual Review of Neuroscience, 35,* 73–89.

Plack, C. J., Barker, D., & Prendergast, G. (2014). Perceptual consequences of "hidden" hearing loss. *Trends in Hearing, 18*. doi:10.1177/2331216514550621

Posner, M. I., & Petersen, S. E. (1990). The attention system of the human brain. *Annual Review Neuroscience, 13,* 25–42.

Power, A. J., Foxe, J. J., Forde, E. J., Reilly, R. B., & Lalor, E. C. (2012). At what time is the cocktail party? A late locus of selective attention to natural speech. *European Journal of Neuroscience, 35*(9), 1497–1503.

Ptak, R. (2012). The frontoparietal attention network of the human brain: Action, saliency, and a priority map of the environment. *Neuroscientist, 18*(5), 502–515.

Pugh, K. R., Offywitz, B. A., Shaywitz, S. E., Fulbright, R. K., et al. (1996). Auditory selective attention: An fMRI investigation. *NeuroImage, 4*(3 Pt 1), 159–173.

Ruggles, D., Bharadwaj, H., & Shinn-Cunningham, B. G. (2011). Normal hearing is not enough to guarantee robust encoding of suprathreshold features important in everyday communication. *Proceedings of the National Academy of Sciences of the USA, 108*(37), 15516–15521.

Samuel, A. G. (1981). The role of bottom-up confirmation in the phonemic restoration illusion. *Journal of Experimental Psychology: Human Perception and Performance, 7*(5), 1124–1131.

Schaal, N. K., Williamson, V. J., & Banissy, M. J. (2013). Anodal transcranial direct current stimulation over the supramarginal gyrus facilitates pitch memory. *European Journal of Neuroscience, 38*(10), 3513–3518.

Scharf, B., Quigley, S., Aoki, C., Peachey, N., & Reeves, A. (1987). Focused auditory attention and frequency selectivity. *Perception and Psychophysics, 42*(3), 215–223.

Schwartz, A., McDermott, J. H., & Shinn-Cunningham, B. (2012). Spatial cues alone produce inaccurate sound segregation: The effect of interaural time differences. *The Journal of the Acoustical Society of America, 132*(1), 357–368.

Serences, J. T., & Yantis, S. (2006a). Selective visual attention and perceptual coherence. *Trends in Cognitive Sciences, 10*(1), 38–45.

Serences, J. T., & Yantis, S. (2006b). Spatially selective representations of voluntary and stimulus-driven attentional priority in human occipital, parietal, and frontal cortex. *Cerebral Cortex, 17*(2), 284–293.

Shamma, S. A., Elhilali, M., & Micheyl, C. (2011). Temporal coherence and attention in auditory scene analysis. *Trends in Neurosciences, 34*(3), 114–123.

Shinn-Cunningham, B. G. (2008). Object-based auditory and visual attention. *Trends in Cognitive Sciences, 12*(5), 182–186.

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

Shinn-Cunningham, B. G., Lee, A. K. C., & Oxenham, A. J. (2007). A sound element gets lost in perceptual competition. *Proceedings of the National Academy of Sciences of the USA, 104*(29), 12223–12227.

Shuai, L., & Elhilali, M. (2014). Task-dependent neural representations of salient events in dynamic auditory scenes. *Frontiers in Neuroscience, 8*(203), 1–11.

Strauss, A., Wostmann, M., & Obleser, J. (2014). Cortical alpha oscillations as a tool for auditory selective inhibition. *Frontiers in Human Neuroscience, 8*, 350.

Sussman, E. S., Horvath, J., Winkler, I., & Orr, M. (2007). The role of attention in the formation of auditory streams. *Perception and Psychophysics, 69*(1), 136–152.

Tark, K. J., & Curtis, C. E. (2009). Persistent neural activity in the human frontal cortex when maintaining space that is off the map. *Nature Neuroscience, 12*(11), 1463–1468.

Teki, S., Chait, M., Kumar, S., Shamma, S., & Griffiths, T. D. (2013). Segregation of complex acoustic scenes based on temporal coherence. *Elife, 2,* e00699.

Terhardt, E. (1974). Pitch, consonance, and harmony. *The Journal of the Acoustical Society of America, 55*(5), 1061–1069.

Toscani, M., Marzi, T., Righi, S., Viggiano, M. P., & Baldassi, S. (2010). Alpha waves: A neural signature of visual suppression. *Experimental Brain Research, 207*(3–4), 213–219.

Treisman, A. M. (1960). Contextual cues in selective listening. *Quarterly Journal of Experimental Psychology, 12,* 157–167.

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology, 12*(1), 97–136.

Ungerleider, L. G., & Mishkin, M. (1982). Two cortical visual systems. In D. J. Ingle, M. A. Goodale, & R. J. W. Mansfield (Eds.), *Analysis of visual behaviour* (pp. 549–586). Cambridge, MA: MIT Press.

Varghese, L., Bharadwaj, H. M., & Shinn-Cunningham, B. G. (2015). Evidence against attentional state modulating scalp-recorded auditory brainstem steady-state responses. *Brain Research, 1626,* 146–164.

Varghese, L. A., Ozmeral, E. J., Best, V., & Shinn-Cunningham, B. G. (2012). How visual cues for when to listen aid selective auditory attention. *Journal of the Association for Research in Otolaryngology, 13*(3), 359–368.

Vines, B. W., Schnider, N. M., & Schlaug, G. (2006). Testing for causality with transcranial direct current stimulation: Pitch memory and the left supramarginal gyrus. *NeuroReport, 17*(10), 1047–1050.

Vliegen, J., Moore, B. C., & Oxenham, A. J. (1999). The role of spectral and periodicity cues in auditory stream segregation, measured using a temporal discrimination task. *The Journal of the Acoustical Society of America, 106*(2), 938–945.

von Békésy, G. (1960). *Experiments in hearing* (1989th ed.). New York: Acoustical Society of America Press.

Vossel, S., Geng, J. J., & Fink, G. R. (2014). Dorsal and ventral attention systems: Distinct neural circuits but collaborative roles. *Neuroscientist, 20*(2), 150–159.

Wardak, C., Ibos, G., Duhamel, J. R., & Olivier, E. (2006). Contribution of the monkey frontal eye field to covert visual attention. *The Journal of Neuroscience, 26*(16), 4228–4235.

Warren, R. M. (1970). Perceptual restoration of missing speech sounds. *Science, 167*(917), 392–393.

Warren, R. M., Wrightson, J. M., & Puretz, J. (1988). Illusory continuity of tonal and infratonal periodic sounds. *The Journal of the Acoustical Society of America*, *84*(4), 1338–1342.

Welch, R. B., & Warren, D. H. (1980). Immediate perceptual response to intersensory discrepancy. *Psychological Bulletin, 88,* 638–667.

Whittingstall, K., & Logothetis, N. K. (2009). Frequency-band coupling in surface EEG reflects spiking activity in monkey visual cortex. *Neuron, 64*(2), 281–289.

Woldorff, M. G., Gallen, C. C., Hampson, S. A., Hillyard, S. A., et al. (1993). Modulation of early sensory processing in human auditory-cortex during auditory selective attention. *Proceedings of the National Academy of Sciences of the USA, 90*(18), 8722–8726.

Wood, N., & Cowan, N. (1995). The cocktail party phenomenon revisited: How frequent are attention shifts to one's name in an irrelevant auditory channel? *Journal of Experimental Psychology. Learning, Memory, and Cognition, 21*(1), 255–260.

Woodruff, P. W., Benson, R. R., Bandettini, P. A., Kwong, K. K., et al. (1996). Modulation of auditory and visual cortex by selective attention is modality-dependent. *NeuroReport, 7*(12), 1909–1913.

Wright, B. A., & Fitzgerald, M. B. (2004). The time course of attention in a simple auditory detection task. *Perception and Psychophysics, 66*(3), 508–516.

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

Zion-Golumbic, E. M., Ding, N., Bickel, S., Lakatos, P., et al. (2013). Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party". *Neuron, 77*(5), 980–991.

Zion-Golumbic, E., & Schroeder, C. E. (2012). Attention modulates 'speech-tracking' at a cocktail party. *Trends in Cognitive Sciences, 16*(7), 363–364.