Springer Handbook of Auditory Research

V 60

M a ab a :// . .c / /2506

J C. M dd b J a a Z. S
A N. P R c a d R. Fa
Ed

T A d S
a C c. a Pa

W 41 I a

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

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 Problem: 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)

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,w, tous

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

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

"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

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

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.

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 inßuence 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.

background, so that it ßips 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

2.3.2 Bton-pSaliece In ßencesAtetto

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"

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 inßuenced by cognitive and linguistic skills (Benard et al. 2014).

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

feature with the attended word is automatically more likely to be the focus of

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 inßuence 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.

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 Cortical 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

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

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

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

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 to 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 Activity/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.

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

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

Larson, E., & Lee, A. K. C. (2013). Inß

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

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 Ampli cation, 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.

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