CHAPTER 2

Audition

JOSH H. MCDERMOTT

INTRODUCTION

Sound is created when matter in the world vibrates and takes the form of pressure waves that propagate through the air, containing clues about the environment around us. Audition is the process by which organisms utilize these clues to derive information about the world.

Audition is a crucial sense for most animals. Humans use sound to infer a vast number of important things—what someone said, their emotional state when they said it, and the whereabouts and nature of objects we cannot see, to name but a few. When hearing is impaired (via congenital conditions, noise exposure, or aging), the consequences can be devastating, such that a large industry is devoted to the design of prosthetic hearing devices.

As listeners, we are largely unaware of the computations underlying our auditory system's success, but they represent an impressive feat of engineering. The computational challenges of everyday audition are reflected in the gap between biological and machine hearing systems—machine systems for interpreting sound currently fall short of human abilities. At present, smart phones and other machine systems recognize speech reasonably well in quiet conditions, but in a noisy restaurant they are all but useless. Understanding the basis of our success in perceiving sound will hopefully help us to replicate it in machine systems and restore it in biological auditory systems when their function becomes impaired.

The goal of this chapter is to provide a bird's-eye view of contemporary hearing research. The chapter is an updated version of one I wrote a few years ago (McDermott, 2013). I provide brief overviews of classic areas of research as well as some central themes and advances from the past 10 years. The first section describes the sensory transduction of the cochlea. The second section discusses modulation and its measurement by subcortical and cortical regions of the auditory system, a key research focus of the last few decades. The third and fourth sections describe some of what is known about the primary and nonprimary auditory cortex, respectively. The fifth section discusses the perception of sound source properties,

This chapter is an updated version of a chapter written for *The Oxford Handbook of Cognitive Neuroscience*. I thank Dana Boebinger, Alex Kell, Wiktor Mlynarski, and Kevin Woods for helpful comments on earlier drafts of this chapter. Supported by a McDonnell Scholar Award and a National Science Foundation CAREER Award.

Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience, Fourth Edition, edited by John T. Wixted.

DOI: 10.1002/9781119170174.epcn202

focusing on location, loudness, and pitch. The sixth section presents an overview of auditory scene analysis. I conclude with a discussion of where hearing research is headed.

THE PROBLEM

Just by listening, we can routinely apprehend many aspects of the world around us: the size of a room in which we are talking, whether it is windy or raining outside, the speed of an approaching car, or whether the surface someone is walking on is gravel or marble. This ability is nontrivial because the properties of the world that are of interest to a listener are generally not explicit in the acoustic input—they cannot be easily recognized or discriminated from the sound waveform itself. The brain must process the sound entering the ear to generate representations in which the properties of interest are more evident. One of the main objectives of hearing science is to understand the nature of this transformation and its instantiation in the brain.

Like other senses, audition is further complicated by a second challenge—that of scene analysis. Although listeners are generally interested in the properties of individual objects or events, the ears are rarely presented with the sounds from isolated sources. Instead, the sound signal that reaches the ear is typically a mixture of sounds from different sources. Such mixtures of sound sources occur frequently in natural auditory environments, for example in social settings, where a single speaker of interest may be talking among many others, and in music. From the mixture it receives as input, the brain must derive representations of the individual sound sources of interest, as are needed to understand someone's speech, recognize a melody, or otherwise guide behavior. Known as the "cocktail party problem" (Cherry, 1953), or "auditory scene analysis" (Bregman, 1990), this problem has analogs in other sensory modalities, but the nature of sound presents the auditory system with unique challenges.

SOUND MEASUREMENT—THE PERIPHERAL AUDITORY SYSTEM

The transformation of the raw acoustic input into representations that are useful for behavior is apparently instantiated over many brain areas and stages of neural processing, spanning the cochlea, midbrain, thalamus, and cortex (Figure 2.1). The early stages of this cascade are particularly intricate in the auditory system relative to other sensory systems, with many processing stations occurring prior to the cortex. The sensory organ of the cochlea is itself a complex multicomponent system, whose investigation remains a considerable challenge—the mechanical nature of the cochlea renders it much more difficult to probe (e.g., with electrodes) than the retina or olfactory epithelium, for instance. Peripheral coding of sound is also unusual relative to that of other senses in its degree of clinical relevance. Unlike vision, for which the most common forms of dysfunction are optical in nature, and can be fixed with glasses, hearing impairment typically involves altered peripheral neural processing, and its treatment has benefited from a detailed understanding of the processes that are altered. Much of hearing research has accordingly been devoted to understanding the nature of the measurements made by the auditory periphery, and they provide a natural starting point for any discussion of how we hear.

Frequency Selectivity and the Cochlea

Hearing begins with the ear, where sound is transduced into action potentials that are sent to the brain via the auditory nerve.

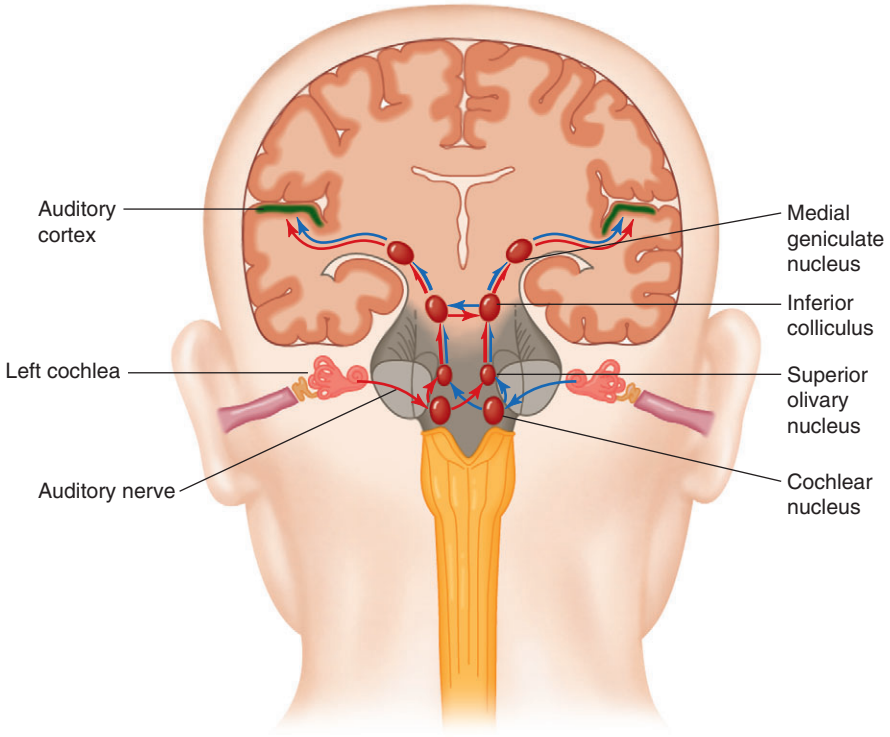

Figure 2.1 The auditory system. Sound is transduced by the cochlea, processed by an interconnected set of subcortical areas, and then fed into the core regions of auditory cortex.
SOURCE: From Goldstein (2007). © 2007 South-Western, a part of Cengage, Inc. Reproduced with permission. www.cengage/com/permissions

The transduction process is marked by several distinctive signal transformations, the most obvious of which is produced by frequency tuning.

The key components of sound transduction are depicted in Figure 2.2. Sound induces vibrations of the eardrum, which are then transmitted via the bones of the middle ear to the cochlea, the sensory organ of the auditory system. The cochlea is a coiled, fluid-filled tube. Several membranes extend through the tube and vibrate in response to sound. Transduction of this mechanical vibration into an electrical signal occurs in the organ of Corti, a mass of cells attached to the basilar membrane. The organ of Corti in particular contains what are known as hair cells, named for the stereocilia that protrude from them. The inner hair cells are responsible for sound transduction. When the section of membrane on which they lie vibrates, stereocilia shear against the membrane above, opening mechanically gated ion channels and inducing a voltage change within the body of the cell. Neurotransmitter release is triggered by the change in membrane potential, generating action potentials in the auditory nerve fibers that the hair cell synapses with. This electrical signal is carried by the auditory nerve fibers to the brain.

The frequency tuning of the transduction process occurs because different parts of the basilar membrane vibrate maximally in response to different frequencies. This is

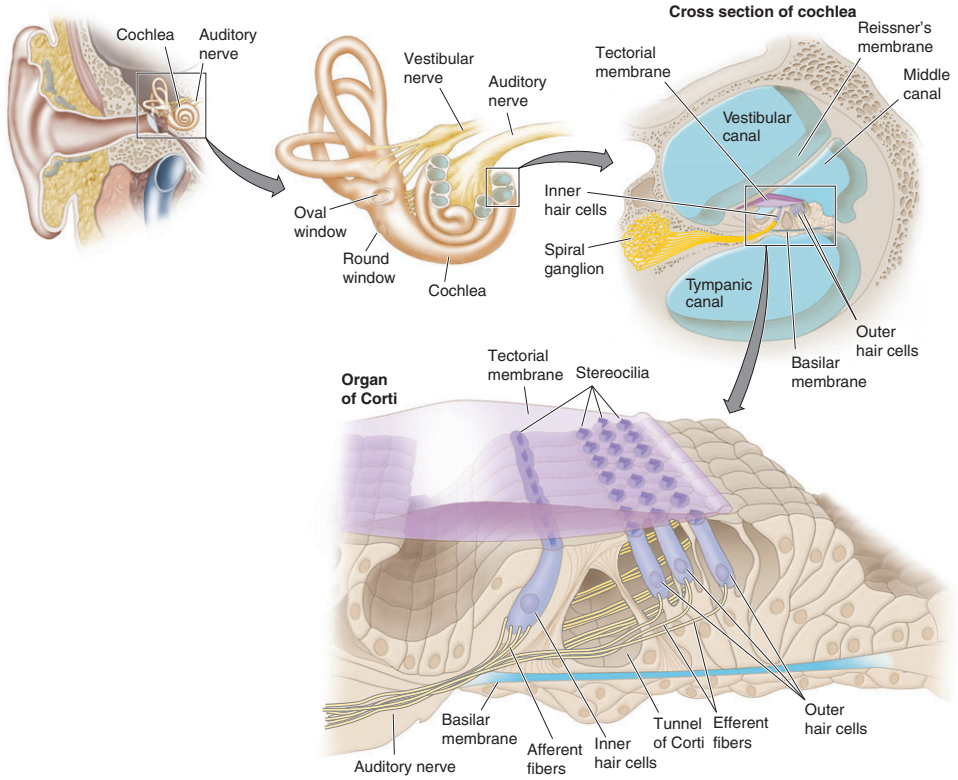

Figure 2.2 Structure of the peripheral auditory system. (Top left) Diagram of ear. The eardrum transmits sound to the cochlea via the middle ear bones (ossicles). (Top middle) Inner ear. The semicircular canals abut the cochlea. Sound enters the cochlea via the oval window and causes vibrations along the basilar membrane, which runs through the middle of the cochlea. (Top right) Cross section of cochlea. The organ of Corti, containing the hair cells that transduce sound into electrical potentials, sits on top of the basilar membrane. (Bottom) Schematic of section of the organ of Corti. The shearing that occurs between the basilar and tectorial membranes when they vibrate (in response to sound) causes the hair cell stereocilia to deform. The deformation causes a change in the membrane potential of the inner hair cells, transmitted to the brain via afferent auditory nerve fibers. The outer hair cells, which are 3 times more numerous than the inner hair cells, serve as a feedback system to alter the basilar membrane motion, tightening its tuning and amplifying the response to low-amplitude sounds.
SOURCE: From Wolfe (2006, Chapter 9). Reproduced with permission of Oxford University Press.

partly due to mechanical resonances—the thickness and stiffness of the membrane vary along its length, producing a different resonant frequency at each point. The mechanical resonances are actively enhanced via a feedback process, believed to be mediated largely by a second set of cells, called the outer hair cells. The outer hair cells abut the inner hair cells on the organ of Corti and serve to alter the basilar membrane vibration rather than transduce it. They expand and contract in response to sound (Ashmore, 2008; Dallos, 2008; Hudspeth, 2008). Their motion alters the passive mechanics of the basilar membrane, amplifying the response to low-intensity sounds and tightening the frequency tuning of the resonance. The upshot is that high frequencies produce vibrations at the basal end of the cochlea (close to the eardrum), while low frequencies produce

vibrations at the apical end (far from the eardrum), with frequencies in between stimulating intermediate regions. The auditory nerve fibers that synapse onto individual inner hair cells are thus frequency-tuned—they fire action potentials in response to a local range of frequencies, collectively providing the rest of the auditory system with a frequency decomposition of the incoming waveform. As a result of this behavior, the cochlea is often described functionally as a set of bandpass filters—filters that each pass frequencies within a particular range, and eliminate those outside of it. Collectively the filters span the audible spectrum.

The frequency decomposition of the cochlea is conceptually similar to the Fourier transform, but differs in important respects. Whereas the Fourier transform uses linearly spaced frequency bins, each separated by the same number of Hz, the tuning bandwidth of auditory nerve fibers increases with their preferred frequency. This characteristic is evident in Figure 2.3A, in which the frequency response of a set of auditory nerve fibers is plotted on a logarithmic frequency scale. Although the lowest frequency fibers are broader on a log scale than the high frequency fibers, in absolute terms their bandwidths are much lower—several hundred Hz instead of several thousand. The distribution of best frequency along the cochlea also follows a roughly logarithmic function, apparent in Figure 2.3B, which plots the best frequency of a large set of nerve fibers against the distance along the cochlea of the hair cell that they synapse with. These features of frequency selectivity are present in most biological auditory systems. It is partly for this reason that a log scale is commonly used for frequency.

Cochlear frequency selectivity has a host of perceptual consequences—for instance, our ability to detect a particular frequency is limited largely by the signal-to-noise ratio of the cochlear filter centered on the frequency. There are many treatments of frequency selectivity and perception (Moore, 2003), as it is perhaps the most studied aspect of hearing.

Although the frequency tuning of the cochlea is uncontroversial, the teleological question of *why* the cochlear transduction process is frequency tuned remains less settled. How does frequency tuning aid the brain's task of recovering useful information about the world from its acoustic input? Over the last two decades, a growing number of researchers have endeavored to explain properties of sensory systems as optimal for the task of encoding natural sensory stimuli, initially focusing on coding questions in vision, and using notions of efficiency as the optimality criterion (Field, 1987; Olshausen & Field, 1996). Lewicki and his colleagues have applied similar concepts to hearing, using algorithms that derive efficient and sparse representations of sounds (Lewicki, 2002; Smith & Lewicki, 2006), properties believed to be desirable of early sensory representations. They report that for speech, or sets of environmental sounds and animal vocalizations, efficient representations for sound look much like the representation produced by auditory nerve fiber responses—sounds are represented with filters whose tuning is localized in frequency. Interestingly, the resulting representations share the dependence of bandwidth and frequency found in biological hearing—bandwidths increase with frequency as they do in the ear. Moreover, representations derived in the same way for "unnatural" sets of sounds, such as samples of white noise, do not exhibit frequency tuning, indicating that the result is at least somewhat specific to the sorts of sounds commonly encountered in the world. These results suggest that frequency tuning of the sort found in the ear provides an efficient means to encode the sounds that were likely of importance when the

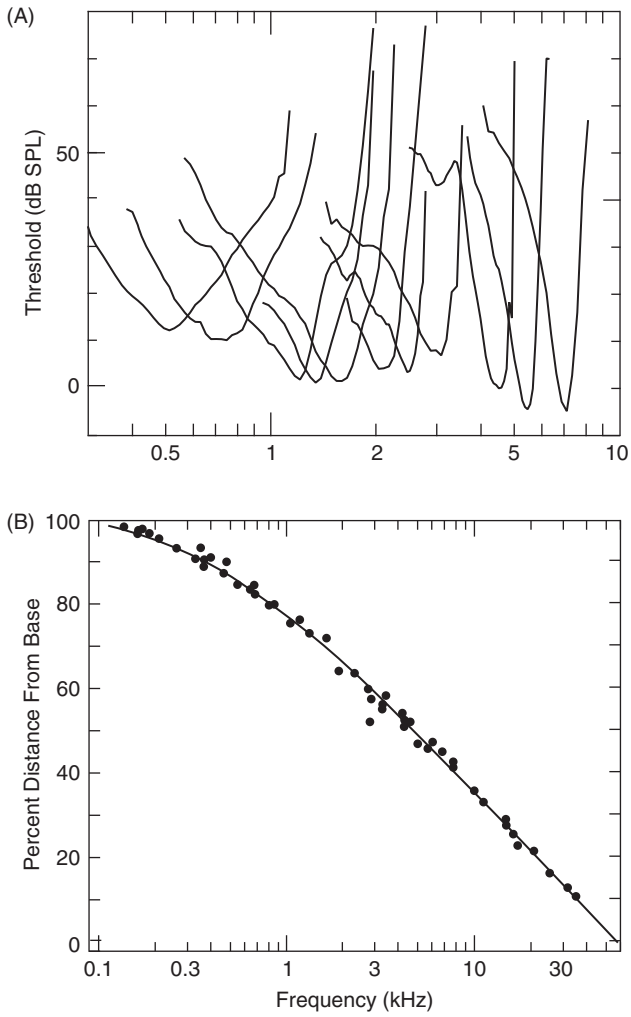

Figure 2.3 Frequency selectivity. (A) Threshold tuning curves of auditory nerve fibers from a cat ear, plotting the level that was necessary to evoke a criterion increase in firing rate for a given frequency (Miller, Schilling, Franck, & Young, 1997). (B) The tonotopy of the cochlea. The position along the basilar membrane at which auditory nerve fibers synapse with a hair cell (determined by dye injections) is plotted versus their best frequency (Liberman, 1982).
SOURCE: Both parts of this figure are courtesy of Eric Young (Young, 2010), who replotted data from the original sources. Reproduced with permission of Oxford University Press.

auditory system evolved, possibly explaining its ubiquitous presence in auditory systems as an optimal distribution of limited neural coding resources. It remains to be seen whether this framework can explain potential variation in frequency tuning bandwidths across species—humans have recently been claimed to possess narrower tuning than other species (Joris et al., 2011; Shera, Guinan, & Oxenham, 2002)—or the broadening of frequency tuning with increasing sound intensity (Rhode, 1978), but it provides one means by which to understand the origins of peripheral auditory processing.

Amplitude Compression

A second salient transformation that occurs in the cochlea is that of amplitude compression. Compression is reflected in the fact that the mechanical response of the cochlea to a soft sound (and thus the neural response that results) is larger than would be expected given the response to an intense sound. The response elicited by a sound is not proportional to the sound's amplitude (as it would be if the response were linear), but rather to a compressive nonlinear function of amplitude. The dynamic range of the response to sound is thus "compressed" relative to the dynamic range of the acoustic input. Whereas the range of audible sounds covers five orders of magnitude, or 100 dB, the range of cochlear response covers only one or two orders of magnitude (Ruggero, Rich, Recio, & Narayan, 1997).

Compression appears to serve to map the range of amplitudes that the listener needs to hear (i.e., those commonly encountered in the environment) onto the physical operating range of the cochlea. Without compression, it would have to be the case that either sounds low in level would be inaudible, or sounds high in level would be indiscriminable (for they would fall outside the range that could elicit a response change). Compression permits very soft sounds to produce a physical response that is (just barely) detectable, while maintaining some discriminability of higher levels.

The compressive nonlinearity is often approximated as a power function with an exponent of 0.3 or so. It is not obvious why the compressive nonlinearity should take the particular form that it does. Many different functions could in principle serve to compress the output response range. It remains to be seen whether compression can be explained in terms of optimizing the encoding of the input, as has been proposed for frequency tuning (though see Escabi, Miller, Read, and Schreiner (2003)). Most machine hearing applications also utilize amplitude compression prior to analyzing sound, however, and it is widely agreed to be useful to amplify low amplitudes relative to large when processing sound.

Amplitude compression was first noticed in measurements of the physical vibrations of the basilar membrane (Rhode, 1971; Ruggero, 1992), but it is also apparent in auditory nerve fiber responses (Yates, 1990) and is believed to account for a number of perceptual phenomena (Moore & Oxenham, 1998). The effects of compression are related to cochlear amplification, in that compression results from response amplification that is limited to low-intensity sounds. Compression is achieved in part via the outer hair cells, whose motility modifies the motion of the basilar membrane in response to sound (Ruggero & Rich, 1991). Outer hair cell function is frequently altered in hearing impairment, one consequence of which is a loss of compression, something that hearing aids attempt to mimic.

Neural Coding in the Auditory Nerve

Although frequency tuning and amplitude compression are at this point uncontroversial and relatively well understood, several other empirical questions about peripheral auditory coding remain unresolved. One important issue involves the means by which the auditory nerve encodes frequency information. As a result of the frequency tuning of the auditory nerve, the spike rate of a nerve fiber contains information about frequency (a large firing rate indicates that the sound input contains frequencies near the center of the range of the fiber's tuning). Collectively, the firing rates of all nerve fibers could thus be used to estimate the instantaneous spectrum of a sound. However, spike timings

also carry frequency information. At least for low frequencies, the spikes that are fired in response to sound do not occur randomly, but rather tend to occur at the peak displacements of the basilar membrane vibration. Because the motion of a particular section of the membrane mirrors the bandpass filtered sound waveform, the spikes occur at the waveform peaks (Rose, Brugge, Anderson, & Hind, 1967). If the input is a single frequency, spikes thus occur at a fixed phase of the frequency cycle (Figure 2.4A). This behavior is known as "phase locking" and produces spikes at regular intervals corresponding to the period of the frequency. The spike timings thus carry information that could potentially augment or supercede that conveyed by the rate of firing.

Phase locking degrades in accuracy as frequency is increased (Figure 2.4B) due to limitations in the temporal fidelity of the hair cell membrane potential (Palmer & Russell, 1986), and is believed to be largely absent for frequencies above 4 kHz in most mammals, though there is some variability across species (Johnson, 1980; Palmer & Russell, 1986; Sumner & Palmer, 2012). The appeal of phase locking as a code for sound frequency is partly due to features of rate-based frequency selectivity that are unappealing from an engineering standpoint. Although frequency tuning in the auditory system (as measured by auditory nerve spike rates or psychophysical masking experiments) is narrow at low stimulus levels, it broadens considerably as the level is raised (Glasberg & Moore, 1990; Rhode, 1978). Phase locking, by comparison, is robust to sound level—even though a nerve fiber responds to a broad range of frequencies when the level is high, the time intervals between spikes continue to convey frequency-specific information, as the peaks in the bandpass-filtered waveform tend to occur at integer multiples of the periods of the component frequencies.

Our ability to discriminate frequency is impressive, with thresholds on the order of 1% (Moore, 1973), and there has been longstanding interest in whether this ability in part depends on fine-grained spike timing information (Heinz, Colburn, & Carney, 2001).

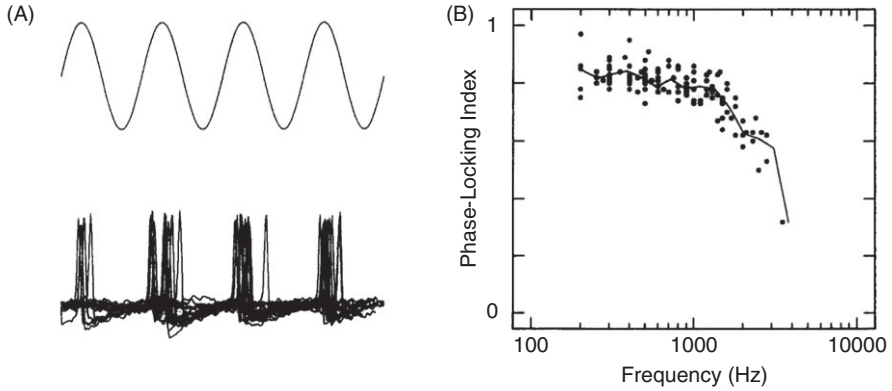

Figure 2.4 Phase locking. (A) A 200 Hz pure tone stimulus waveform aligned in time with several overlaid traces of an auditory nerve fiber's response to the tone. Note that the spikes are not uniformly distributed in time, but rather occur at particular phases of the sinusoidal input. (B) A measure of phase locking for each of a set of nerve fibers in response to different frequencies. Phase locking decreases at high frequencies.
SOURCE: Reprinted from Javel and Mott (1988). Reproduced with permission of Elsevier.

Although phase locking remains uncharacterized in humans due to the unavailability of human auditory nerve recordings, it is presumed to occur in much the same way as in nonhuman auditory systems. Moreover, several psychophysical phenomena are consistent with a role for phase locking in human hearing. For instance, frequency discrimination becomes much poorer for frequencies above 4 kHz (Moore, 1973), roughly the point at which phase locking declines in nonhuman animals. The fundamental frequency of the highest note on a piano is also approximately 4 kHz; this is also the point above which melodic intervals between pure tones (tones containing a single frequency) are also much less evident (Attneave & Olson, 1971; Demany & Semal, 1990). These findings provide some circumstantial evidence that phase locking is important for deriving precise estimates of frequency, but definitive evidence remains elusive. It remains possible that the perceptual degradations at high frequencies reflect a lack of experience with such frequencies, or their relative unimportance for typical behavioral judgments, rather than a physiological limitation.

The upper limit of phase locking is also known to decrease markedly at each successive stage of the auditory system (Wallace, Anderson, & Palmer, 2007). By primary auditory cortex, the upper cutoff is in the neighborhood of a few hundred Hz. It would thus seem that the phase locking that occurs robustly in the auditory nerve would need to be rapidly transformed into a spike rate code if it were to benefit processing throughout the auditory system. Adding to the puzzle is the fact that frequency tuning is not thought to be dramatically narrower at higher stages in the auditory system. Such tightening might be expected if the frequency information provided by phase-locked spikes was transformed to yield improved rate-based frequency tuning at subsequent stages (though see Bitterman, Mukamel, Malach, Fried, and Nelken (2008)).

Feedback

Like other sensory systems, the auditory system can be thought of as a processing cascade, extending from the sensory receptors to cortical areas believed to mediate auditory-based decisions. This "feed-forward" view of processing underlies much auditory research. As in other systems, however, feedback from later stages to earlier ones is ubiquitous and substantial, and in the auditory system is perhaps even more pronounced than elsewhere in the brain. Unlike the visual system, for instance, the auditory pathways contain feedback extending all the way back to the sensory receptors. The function of much of this feedback remains poorly understood, but one particular set of projections—the cochlear efferent system—has been the subject of much discussion.

Efferent connections to the cochlea originate primarily from the superior olivary nucleus, an area of the midbrain a few synapses removed from the cochlea (Figure 2.1, though the efferent pathways are not shown). The superior olive is divided into two subregions, medial and lateral, and to first order, these give rise to two efferent projections: one from the medial superior olive to the outer hair cells, called the medial olivocochlear (MOC) efferents, and one from the lateral superior olive to the inner hair cells (the LOC efferents) (Elgoyhen & Fuchs, 2010). The MOC efferents have been more thoroughly studied than their LOC counterparts. Their activation (by electrical stimulation, for instance) is known to reduce the basilar membrane response to low-intensity sounds and causes the frequency tuning of the response to broaden. This is probably because the MOC efferents inhibit the outer hair cells, which are crucial to

amplifying the response to low-intensity sounds, and to sharpening frequency tuning.

The MOC efferents may serve a protective function by reducing the response to loud sounds (Rajan, 2000), but their most commonly proposed function is to enhance the response to transient sounds in noise (Guinan, 2006). When the MOC fibers are severed, for instance, performance on tasks involving discrimination of tones in noise is reduced (May & McQuone, 1995). Noise-related MOC effects are proposed to derive from its influence on adaptation, which when induced by background noise reduces the detectability of transient foreground sounds by decreasing the dynamic range of the auditory nerve's response. Because MOC activation reduces the response to ongoing sound, adaptation induced by continuous background noise is reduced, thus enhancing the response to transient tones that are too brief to trigger the MOC feedback themselves (Kawase, Delgutte, & Liberman, 1993; Winslow & Sachs, 1987). Another interesting but controversial proposal is that the MOC efferents play a role in auditory attention. One study, for instance, found that patients whose vestibular nerve (containing the MOC fibers) had been severed were better at detecting unexpected tones after the surgery, suggesting that selective attention had been altered so as to prevent the focusing of resources on expected frequencies (Scharf, Magnan, & Chays, 1997). See Guinan (2006) for a recent review of these and other ideas about MOC efferent function.

SOUND MEASUREMENT—MODULATION

Subcortical Auditory Pathways

The auditory nerve feeds into a cascade of interconnected subcortical regions that lead up to the auditory cortex, as shown in Figure 2.1. The subcortical auditory pathways have complex anatomy, only some of which is depicted in Figure 2.1. In contrast to the subcortical pathways of the visual system, which are less complex and largely preserve the representation generated in the retina, the subcortical auditory areas exhibit a panoply of interesting response properties not found in the auditory nerve, many of which remain active topics of investigation. Several subcortical regions will be referred to in the sections that follow in the context of other types of acoustic measurements or perceptual functions. One of the main features that emerges in subcortical auditory regions is tuning to amplitude modulation, the subject of the next section.

Amplitude Modulation and the Envelope

The cochlea decomposes the acoustic input into frequency channels, but much of the important information in sound is conveyed by the way that the output of these frequency channels is modulated in amplitude. Consider Figure 2.5A, which displays in blue the output of one such frequency channel for a short segment of a speech signal. The blue waveform oscillates at a rapid rate, but its amplitude waxes and wanes at a much lower rate (evident in the close-up view of Figure 2.5B). This waxing and waning is known as "amplitude modulation" and is a common feature of many modes of sound production (e.g., vocal articulation). The amplitude is captured by what is known as the "envelope" of a signal, shown in red for the signal of Figures 2.5A and B. The envelopes of a set of bandpass filters can be stacked vertically and displayed as an image, generating a spectrogram (referred to as a cochleogram when the filters mimic the frequency tuning of the cochlea, as in Figure 2.5C). Figure 2.5D shows the spectra of the signal and its envelope. The signal spectrum is bandpass (because it is the output of a

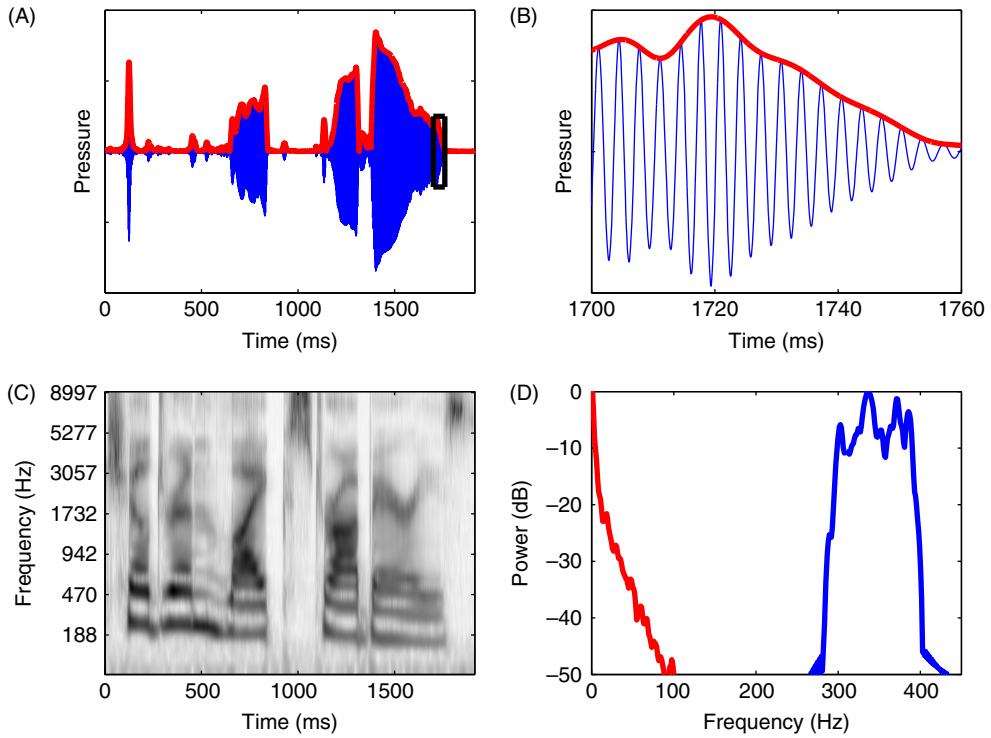

Figure 2.5 Amplitude modulation. (A) The output of a bandpass filter (centered at 340 Hz) for a recording of speech, plotted in blue, with its envelope plotted in red. (B) Close-up of part of (A) (corresponding to the black rectangle in (A)). Note that the filtered sound signal (like the unfiltered signal) fluctuates around zero at a high rate, whereas the envelope is positive valued, and fluctuates more slowly. (C) Cochleagram of the same speech signal formed from the envelopes of a set of filters mimicking the frequency tuning of the cochlea (one of which is plotted in (A)). The cochleagram is produced by plotting each envelope horizontally in grayscale. (D) Power spectra of the filtered speech signal in (A) and its envelope. Note that the envelope contains power only at low frequencies (modulation frequencies), whereas the filtered signal has power at a restricted range of high frequencies (audio frequencies).

bandpass filter), with energy at frequencies in the audible range. The envelope spectrum, in contrast, is lowpass, with most of the power below 10 Hz, corresponding to the slow rate at which the envelope changes. The frequencies that compose the envelope are typically termed "modulation frequencies," distinct from the "audio frequencies" that compose the signal that the envelope is derived from.

The information carried by a cochlear channel can thus be viewed as the product of an amplitude envelope—that varied slowly—and its "fine structure"—a waveform that varies rapidly, at a rate close to the center frequency of the channel (Rosen, 1992). The envelope and fine structure have a clear relation to common signal processing formulations in which the output of a bandpass filter is viewed as a single sinusoid varying in amplitude and frequency—the envelope describes the amplitude variation, and the fine structure describes the frequency variation. The envelope of a frequency channel is straightforward to extract from the auditory nerve—the envelope results from lowpass filtering a spike train, as the envelope is reflected in relatively slow changes in the

rectified sound signal. Despite the fact that envelope and fine structure are not completely independent (Ghitza, 2001), there has been much interest in the last decade in distinguishing their roles in different aspects of hearing (Smith, Delgutte, & Oxenham, 2002) and its impairment (Lorenzi, Gilbert, Carn, Garnier, & Moore, 2006).

Perhaps surprisingly, the temporal information contained in amplitude envelopes can be sufficient for speech comprehension even when spectral information is severely limited. In a classic paper, Shannon and colleagues isolated the information contained in the amplitude envelopes of speech signals with a stimulus known as "noise-vocoded speech" (Shannon et al., 1995). Noise-vocoded speech is generated by filtering a speech signal and a noise signal into frequency bands, multiplying the frequency bands of the noise by the envelopes of the speech, and then summing the modified noise bands to synthesize a new sound signal. By using a small number of broad frequency bands, spectral information can be greatly reduced, leaving amplitude variation over time (albeit smeared across a broader than normal range of frequencies) as the primary signal cue. Examples are shown in Figure 2.6 for two, four, and eight bands. Shannon and colleagues found that the resulting stimulus was intelligible even when just a few bands were used (i.e., with much broader frequency tuning than is present in the cochlea), indicating that the temporal modulation of the envelopes contains much information about speech content.

Modulation Tuning

Amplitude modulation has been proposed to be analyzed by dedicated banks of filters operating on the envelopes of cochlear filter outputs rather than the sound waveform itself (Dau, Kollmeier, & Kohlrausch, 1997). Early evidence for such a notion came from masking and adaptation experiments, which found that the detection of a modulated signal was impaired by a masker or adapting stimulus modulated at a similar frequency (Bacon & Grantham, 1989; Houtgast, 1989; Tansley & Suffield, 1983). There is now considerable evidence from neurophysiology that single neurons in the midbrain, thalamus, and cortex exhibit some degree of tuning to modulation (Depireux, Simon, Klein, & Shamma, 2001; Joris, Schreiner, & Rees, 2004; Miller, Escabi, Read, & Schreiner, 2001; Rodriguez, Chen, Read, & Escabi, 2010; Schreiner & Urbas, 1986; Schreiner & Urbas, 1988; Woolley, Fremouw, Hsu, & Theunissen, 2005), loosely consistent with the idea of a modulation filter bank (Figure 2.7A).

Modulation tuning in single neurons is often studied by measuring spectrotemporal receptive fields (STRFs) (Depireux et al., 2001), conventionally estimated using techniques such as spike-triggered averaging (Theunissen et al., 2001). To compute a STRF, neuronal responses to a long, random stimulus are recorded, after which the stimulus spectrogram segments preceding each spike are averaged to yield the STRF—the stimulus, described in terms of audio frequency content over time, that on average preceded a spike. Alternatively, a linear model can be fit to the neuronal response given the stimulus (Willmore & Smyth, 2003). In Figure 2.7B, for instance, the STRF consists of a decrease in power followed by an increase in power in the range of 10 kHz; the neuron would thus be likely to respond well to a rapidly modulated 10 kHz tone, and less so to a tone whose amplitude was constant. This STRF can be viewed as a filter that passes modulations in a certain range of rates, that is, modulation frequencies. Modulation tuning functions (e.g., those shown in Figure 2.7A) can be obtained via the Fourier transform of the STRF. Note, though, that the sample STRF in Figure 2.7B is also tuned in audio frequency (the dimension on the

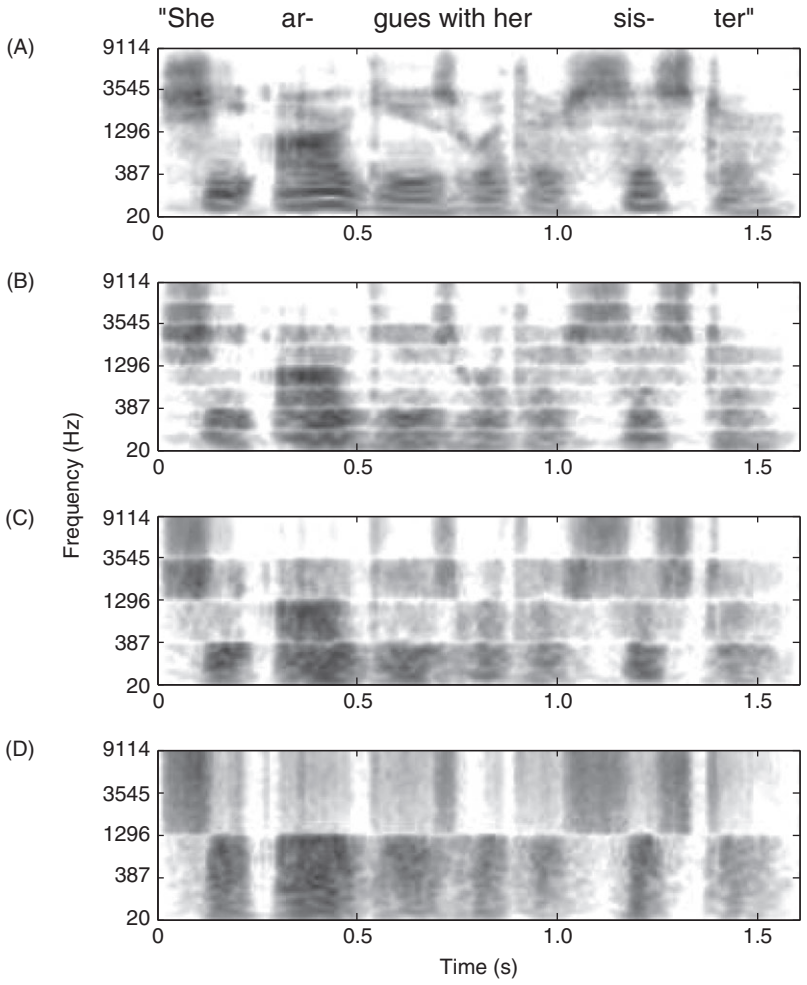

Figure 2.6 Noise-vocoded speech. (A) Cochleagram of a speech utterance, generated as in Figure 2.5C. (B–D) Cochleagrams of noise-vocoded versions of the utterance from (A), generated with eight (B), four (C), or two (D) channels. To generate the noise-vocoded speech, the amplitude envelope of the original speech signal was measured in each of the frequency bands in (B), (C), and (D). A white noise signal was then filtered into these same bands and the noise bands were multiplied by the corresponding speech envelopes. These modulated noise bands were then summed to generate a new sound signal. It is visually apparent that the sounds in (B)–(D) are spectrally coarser versions of the original utterance. Good speech intelligibility is usually obtained with only four channels, indicating that patterns of amplitude modulation can support speech recognition in the absence of fine spectral detail.

y-axis), responding only to modulations of fairly high audio frequencies. Such receptive fields are commonly observed in subcortical auditory regions such as the inferior colliculus and medial geniculate nucleus.

The signal processing effects of subcortical auditory circuitry are encapsulated in the modulation filter bank model, as shown in Figure 2.7C (Dau et al., 1997; McDermott & Simoncelli, 2011). The sound waveform is passed through a set of bandpass filters that simulate cochlear frequency selectivity. The envelopes of the filter outputs are extracted and passed through a compressive

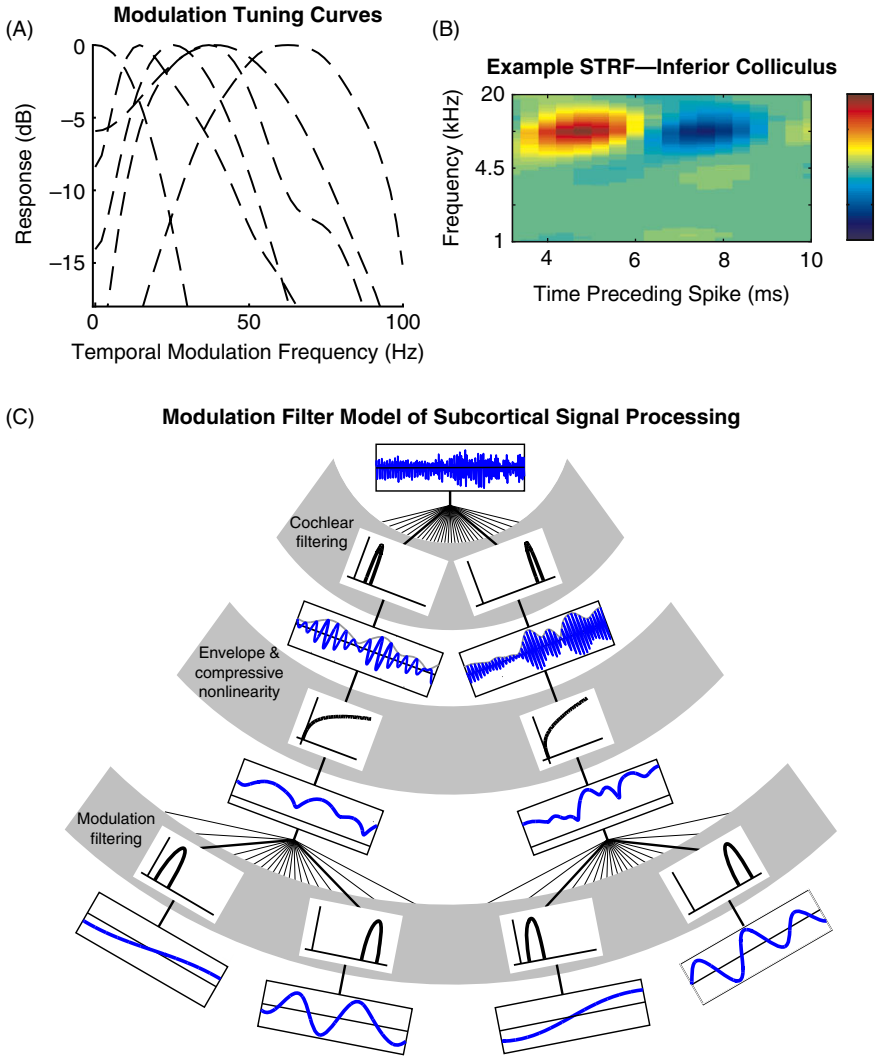

Figure 2.7 Modulation tuning. (A) Example temporal modulation tuning curves for neurons in the medial geniculate nucleus of the thalamus. (B) Example spectrotemporal receptive field (STRF) from a thalamic neuron. Note that the modulation in the STRF is predominantly along the temporal dimension, and that this neuron would thus be sensitive primarily to temporal modulation. (C) Diagram of modulation filter bank model of peripheral auditory processing. The sound waveform is filtered by a simulated cochlear filter bank, the envelopes of which are passed through a compressive nonlinearity before being filtered by a modulation filter bank.

SOURCE: From Miller, Escabi, Read, and Schreiner (2002). Reproduced with permission of the American Physiological Society. Diagram modified from McDermott and Simoncelli (2011). Reproduced with permission of Elsevier.

nonlinearity, simulating cochlear compression. These envelopes are then passed through a modulation filter bank. Because the modulation filters operate on the envelope of a particular cochlear channel, they are tuned both in audio frequency (courtesy of the cochlea) and modulation frequency, like the sample STRF in Figure 2.7B. It is important to note that the model discards the fine structure of each cochlear subband. The fine structure is reflected in the phase locking evident in auditory nerve fibers, but is neglected in envelope-based models of auditory processing (apart from being implicitly captured to some extent by the envelopes of adjacent filters, which jointly constrain their fine structure). This model is often conceived as capturing the signal processing that occurs between the ear and the thalamus (McDermott & Simoncelli, 2011), although it is clearly only a first-pass approximation.

One inadequacy of the modulation filter bank model of auditory processing is that the full range of modulation frequencies that are perceptually relevant does not appear to be represented at any single stage of auditory processing. Neurophysiological studies in nonhuman animals have generally found subcortical neurons to prefer relatively high modulation rates (up to 100–200 Hz) (Miller et al., 2002), with lower modulation rates being represented preferentially in the cortex (Schreiner & Urbas 1986; Schreiner & Urbas, 1988). Neuroimaging results in humans have similarly found that the auditory cortex responds preferentially to low modulation frequencies (in the range of 4–8 Hz) (Boemio, Fromm, Braun, & Poeppel, 2005; Giraud et al., 2000; Schonwiesner & Zatorre, 2009). It seems that the range of preferred modulation frequencies decreases as one ascends the auditory pathway.

Based on this, it is intriguing to speculate that successive stages of the auditory system might process structure at progressively longer (slower) timescales, analogous to the progressive increase in receptive field size that occurs in the visual system from V1 to inferotemporal cortex (Lerner, Honey, Silbert, & Hasson, 2011). Within the cortex, however, no hierarchy is clearly evident as of yet, at least in the response to simple patterns of modulation (Giraud et al., 2000; Boemio et al., 2005). Moreover, there is considerable variation within each stage of the pathway in the preferred modulation frequency of individual neurons (Miller et al., 2001; Rodriguez et al., 2010). There are several reports of topographic organization for modulation frequency in the inferior colliculus, in which a gradient of preferred modulation frequency is observed orthogonal to the gradient of preferred audio frequency (Baumann et al., 2011; Langner, Sams, Heil, & Schulze, 1997). Similar topographic organization has been proposed to exist in the cortex, though the issue remains unsettled (Barton, Venezia, Saberi, Hickok, & Brewer, 2012; Herdener et al., 2013; Nelken et al., 2008).

As with the frequency tuning of the auditory nerve (Lewicki, 2002; Smith & Lewicki, 2006), modulation tuning has been proposed to be consistent with an efficient coding strategy. Modulation tuning bandwidths in the inferior colliculus tend to increase with preferred modulation frequency (Rodriguez et al., 2010), as would be predicted if the lowpass modulation spectra of most natural sounds (Attias & Schreiner, 1997; McDermott, Wrobleski, & Oxenham, 2011; Singh & Theunissen, 2003) were to be divided into channels conveying equal power. Auditory neurons have also been found to convey more information about sounds whose amplitude distribution follows that of natural sounds rather than that of white noise (Escabi et al., 2003). Studies of STRFs in the bird auditory system also indicate that neurons are tuned to the properties of bird song and other natural sounds, maximizing discriminability

of behaviorally important sounds (Hsu, Woolley, Fremouw, Theunissen, 2004; Woolley et al., 2005). Similar arguments have been made about the coding of binaural cues to sound localization (Harper & McAlpine, 2004).

PRIMARY AUDITORY CORTEX

The auditory nucleus of the thalamus directs most of its projections to one region of the auditory cortex, defined on this basis as primary auditory cortex. Other cortical regions also receive thalamic projections, but they are substantially sparser. Primary auditory cortex is also often referred to as the "core" auditory cortex. In humans, primary auditory cortex occupies Heschl's gyrus, also known as the transverse temporal gyrus, located within the lateral sulcus. The rare cases in which humans have bilateral lesions of primary auditory cortex produce profound hearing impairment, termed "cortical deafness" (Hood, Berlin, & Allen, 1994). The structure and functional properties of the PAC are relatively well established compared to the rest of the auditory cortex, and it is the last stage of auditory processing for which computational models exist.

Spectrotemporal Modulation Tuning

Particularly in the auditory cortex, neurons often exhibit tuning for spectral modulation in addition to the tuning for temporal modulation discussed in the previous section. Spectral modulation is variation in power that occurs along the frequency axis. Spectral modulation is frequently evident in natural sounds such as speech, both from individual frequency components, and from formants—the broad peaks in the instantaneous spectra produced by vocal tract resonances that characterize vowel sounds (e.g., Figure 2.5C). Tuning to spectral modulation is generally less pronounced than to amplitude modulation, but is an important feature of cortical responses (Barbour & Wang, 2003). Examples of cortical STRFs with spectral modulation sensitivity are shown in Figure 2.8A. Observations of complex spectrotemporal modulation tuning in cortical neurons underlie what is arguably the standard model of cortical auditory processing (Figure 2.8B), in which a cochleogram-like representation is passed through a bank of filters tuned to temporal and spectral modulations of various rates (Chi, Ru, & Shamma, 2005).

The STRF approximates a neuron's output as a linear function of the cochlear input—the result of convolving the spectrogram of the acoustic input with the STRF. However, particularly in the cortex, it is clear that linear models are inadequate to explain neuronal responses (Christianson, Sahani, & Linden, 2008; Machens, Wehr, & Zador, 2004; Rotman, Bar Yosef, & Nelken, 2001; Theunissen, Sen, & Doupe, 2000). Understanding the nonlinear contributions is an important direction of future research (Ahrens, Linden, & Sahani, 2008; David, Mesgarani, Fritz, & Shamma, 2009), but at present much analysis is restricted to linear receptive field estimates. There are established methods for computing STRFs, and they exhibit many interesting properties even though it is clear that they are not the whole story.

Tonotopy

Although many of the functional properties of cortical neurons are distinct from what is found in auditory nerve responses, frequency tuning persists. Many cortical neurons have a preferred frequency, although they are often less responsive to pure tones (relative to sounds with more complex spectra) and often have broader tuning than neurons in

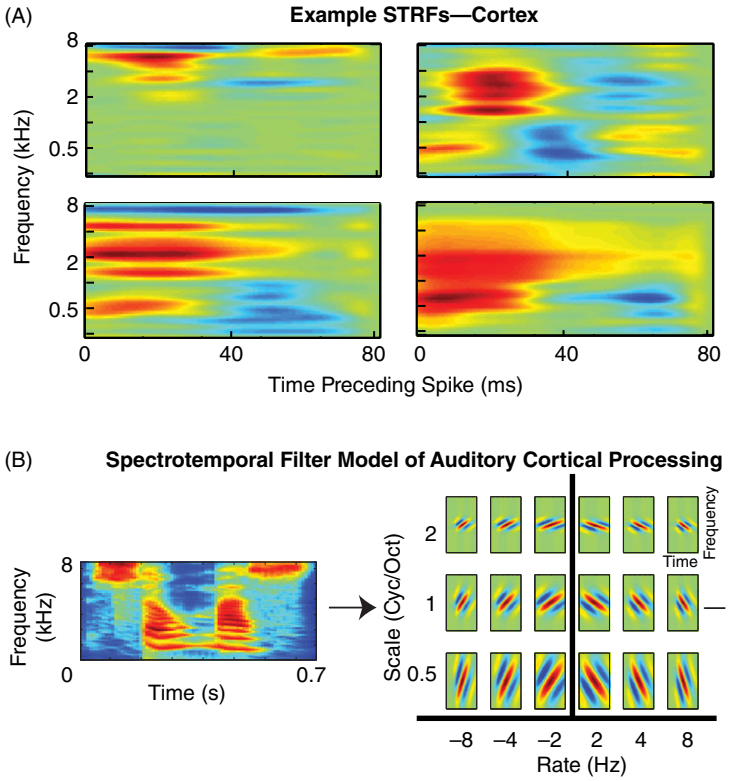

Figure 2.8 STRFs. (A) Example STRFs from cortical neurons. Note that the STRFs feature spectral modulation in addition to temporal modulation, and as such are selective for more complex acoustic features. Cortical neurons typically have longer latencies than subcortical neurons, but this is not evident in the STRFs, probably because of nonlinearities in the cortical neurons that produce small artifacts in the STRFs (Stephen David, personal communication, 2011). (B) Spectrotemporal filter model of auditory cortical processing, in which a cochleogram-like representation of sound is filtered by a set of linear spectrotemporal filters tuned to scale (spectral modulation) and rate (temporal modulation).
SOURCE: (A): Mesgarani, David, Fritz, and Shamma (2008). Reproduced with permission of AIP Publishing LLC. (B): Chi, Ru, and Shamma (2005); Mesgarani and Shamma (2011). Reproduced with permission of AIP Publishing LLC and IEEE.

peripheral stages (Moshitch, Las, Ulanovsky, Bar Yosef, & Nelken, 2006). Moreover, neurons tend to be spatially organized to some extent according to their best frequency, forming "tonotopic" maps. Cortical frequency maps were one of the first reported findings in single-unit neurophysiology studies of the auditory cortex in animals, and have since been found using functional magnetic resonance imaging (fMRI) in humans (Formisano et al., 2003; Humphries, Liebenthal, & Binder, 2010; Talavage et al., 2004) as well as monkeys (Petkov, Kayser, Augath, & Logothetis, 2006). Tonotopic maps are also present in subcortical auditory regions. Although never formally quantified, it seems that tonotopy is less robust than the retinotopy found in the visual system (evident, for instance, in two-photon imaging studies [Bandyopadhyay, Shamma, & Kanold, 2010; Rothschild, Nelken, & Mizrahi, 2010]).

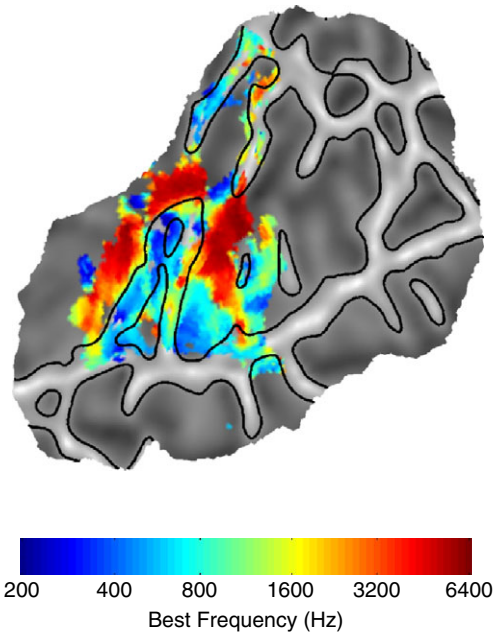

Figure 2.9 Tonotopy. Best frequency of voxels in the human auditory cortex, measured with fMRI, plotted on the flattened cortical surface. Note that the best frequency varies quasi-smoothly over the cortical surface, and is suggestive of two maps that are approximately mirror images of each other. SOURCE: From Humphries, Liebenthal, and Binder (2010). Reproduced with courtesy of Elsevier.

Although the presence of some degree of tonotopy in the cortex is beyond question, its functional importance remains unclear. Frequency selectivity is not the end goal of the auditory system, and it does not obviously bear much relevance to behavior, so it is unclear why tonotopy would be a dominant principle of organization throughout the auditory system. At present, however, tonotopy remains a staple of textbooks and review chapters such as this. Practically, tonotopy is useful to auditory neuroscientists because it provides a simple functional signature of the primary auditory cortex. Figure 2.9 shows an example tonotopic map obtained in a human listener with fMRI. Humans exhibit a stereotyped high-low-high gradient of preferred frequency, typically interpreted as two mirror-reversed maps. These two maps are sometimes referred to as Te 1.0 and Te 1.2 in humans (Morosan et al., 2001). The macaque exhibits similar organization, although additional fields are typically evident (Baumann, Petkov, & Griffiths, 2013). Tonotopy remains the primary functional criterion by which auditory cortical regions are distinguished.

NONPRIMARY AUDITORY CORTEX

Largely on grounds of anatomy and connectivity, the mammalian auditory cortex is standardly divided into three sets of regions (Figure 2.10): a core region receiving direct input from the thalamus, a "belt" region surrounding it, and a "parabelt" region beyond that (Kaas & Hackett 2000; Sweet, Dorph-Petersen, & Lewis, 2005). Within these areas tonotopy is often used to delineate distinct "fields." The core region is divided in this way into areas A1, R (for rostral), and RT (for rostrotemporal) in nonhuman primates,

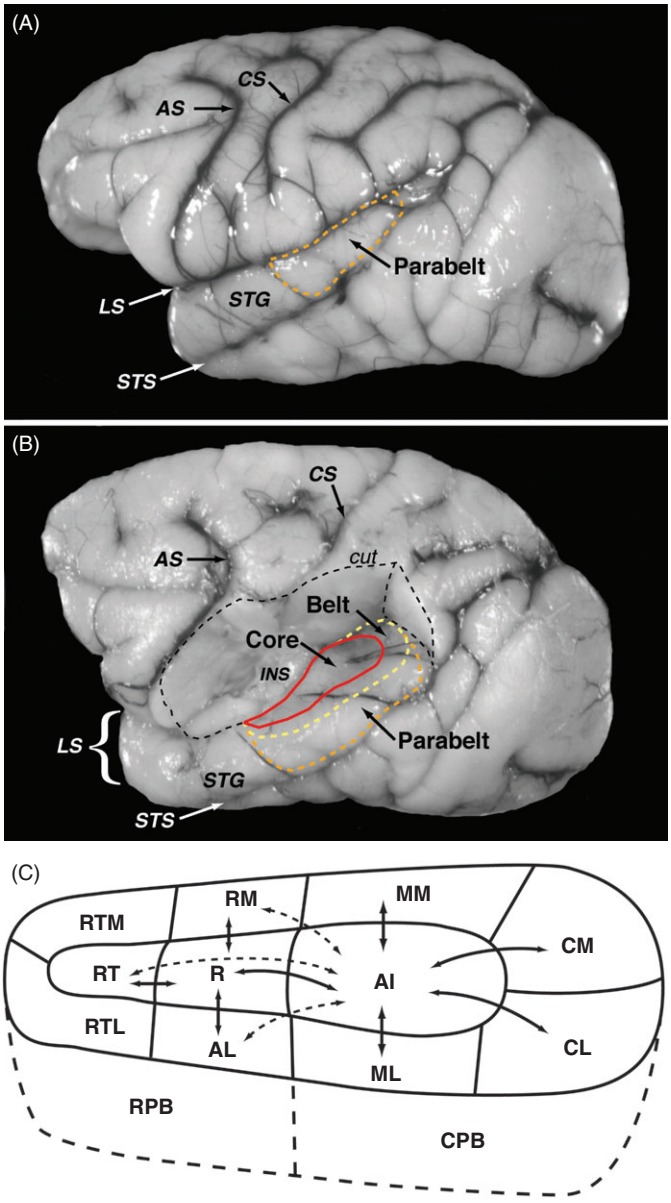

Figure 2.10 Anatomy of the auditory cortex. (A) Lateral view of a macaque's cortex. The approximate location of the parabelt region is indicated with dashed red lines. (B) View of the brain from (A) after removal of the overlying parietal cortex. Approximate locations of the core (solid red line), belt (dashed yellow line), and parabelt (dashed orange line) regions are shown. Abbreviations: superior temporal gyrus (STG), superior temporal sulcus (STS), lateral sulcus (LS), central sulcus (CS), arcuate sulcus (AS), insula (INS). (C) Connectivity between A1 and other auditory cortical areas. Solid lines with arrows denote dense connections; dashed lines with arrows denote less dense connections. RT (the rostrotemporal field), R (the rostral field), and A1 comprise the core; all three subregions receive input from the thalamus. The areas surrounding the core comprise the belt, and the two regions outlined with dashed lines comprise the parabelt. The core has few direct connections with the parabelt or more distant cortical areas.
SOURCE: From Kaas and Hackett (2000).

with A1 and R receiving direct input from the medial geniculate nucleus of the thalamus. There are also multiple belt areas (Petkov et al., 2006), each receiving input from the core areas. Functional imaging reveals many additional areas that respond to sound in the awake primate, including parts of the parietal and frontal cortex (Poremba et al., 2003). There are some indications that the three core regions have different properties (Bendor & Wang, 2008), and that stimulus selectivity increases in complexity from the core to surrounding areas (Kikuchi, Horwitz, & Mishkin, 2010; Rauschecker & Tian, 2004; Tian & Rauschecker, 2004), suggestive of a hierarchy of processing. However, at present there is not a single widely accepted framework for auditory cortical organization. Several principles of organization have been proposed with varying degrees of empirical support.

Some of the proposed organizational principles clearly derive inspiration from the visual system. For instance, selectivity for vocalizations and selectivity for spatial location have been found to be partially segregated, each being most pronounced in a different part of the lateral belt (Tian, Reser, Durham, Kustove, & Rauschecker, 2001; Woods, Lopez, Long, Rahman, & Recanzone, 2006). These regions have thus been proposed to constitute the beginning of ventral "what" and dorsal "where" pathways analogous to those in the visual system, perhaps culminating in the same parts of the prefrontal cortex as the analogous visual pathways (Cohen et al., 2009; Romanski et al., 1999). Functional imaging results in humans have also been viewed as supportive of this framework (Ahveninen et al., 2006; Alain, Arnott, Hevenor, Graham, & Grady, 2001; Warren, Zielinski, Green, Rauschecker, & Griffiths, 2002). Additional evidence for a "what"/"where" dissociation comes from a recent study in which sound localization and temporal pattern discrimination in cats were selectively impaired by reversibly deactivating different regions of nonprimary auditory cortex (Lomber & Malhotra, 2008). However, other studies have found less evidence for segregation of tuning properties in early auditory cortex (Bizley, Walker, Silverman, King, & Schnupp, 2009). Moreover, the properties of the "what" stream remain relatively undefined (Recanzone, 2008); at this point it has been defined mainly by reduced selectivity to location.

There have been further attempts to extend the characterization of a ventral auditory pathway by testing for specialization for the analysis of particular types of sounds, potentially analogous to what has been found in the ventral visual system (Kanwisher, 2010). The most widely proposed specialization is for speech and/or for vocalizations more generally. Responses to speech have been reported in the superior temporal gyrus (STG) of humans for over a decade (Binder et al., 2000; Hickok & Poeppel, 2007; Liebenthal, Binder, Spitzer, Possing, & Medler, 2005; Obleser, Zimmermann, Van Meter, & Rauschecker, 2007; Scott, Blank, Rosen, & Wise, 2000). Recent fMRI results indicate that the STG is involved in an analysis of speech that is at least partly distinct from linguistic processing, in that its response is driven by speech structure even when the speech is foreign and thus unintelligible (Overath, McDermott, Zarate, & Poeppel, 2015). The extent of naturalistic speech structure was manipulated using "quilts" that concatenate speech segments of some length in random order. As the quilt segment length increases, the stimulus becomes increasingly similar to natural speech. The response of Heschl's gyrus (primary auditory cortex in humans) was found to be similar irrespective of quilt segment length. By contrast, the response of regions of the STG increased with segment length, indicating sensitivity

to the temporal structure of speech (Overath et al., 2015). These results complement recent findings of phonemic selectivity in the STG, measured using recordings from the surface of the cortex in epilepsy patients (Mesgarani, Cheung, Johnson, & Chang, 2014). The relationship of these speech-selective responses to the representation of voice identity (Belin, Zatorre, Lafaille, Ahad, & Pike, 2000) remains unclear. Voice-selective regions appear to be present in macaque nonprimary auditory cortex (Petkov et al., 2008) and could be homologous to voice- or speech-selective regions in humans.

Traditionally, segregation of function has been explored by testing whether particular brain regions respond more to one class of sound than to a small set of other classes, with the sound classes typically linked to particular prior hypotheses. The approach is limited by the ability of the experimenter to construct relevant hypotheses, and by the small sets of stimulus conditions used to establish selectivity. One recent study from my lab has attempted to circumvent these limitations by measuring responses throughout the auditory cortex to a very large set of natural sounds (Norman-Haignere, Kanwisher, & McDermott, 2015). We measured fMRI responses of "voxels" (small volumes of brain tissue) to 165 natural sounds intended to be representative of the sounds we encounter in daily life, including speech, music, and many types of environmental sounds. We then inferred tuning functions across this stimulus set whose linear combination could best explain the voxel responses. This "voxel decomposition analysis" yielded six components, each characterized by a response profile across the stimulus set and a weight for each voxel in auditory cortex. Four of the components had responses that were largely explained by frequency and modulation tuning, and thus were not strongly selective for the category of the sounds (Figures 2.11A and B). The most frequency-selective components (numbered 1 and 2 in Figure 2.11) had weights that were strongest in the low- and high-frequency portions of the tonotopic map, respectively (Figure 2.11C), as one would expect. The last two components were strongly selective for speech and music, responding strongly to every speech or music sound, respectively, and much less to other types of sounds. The speech-selective component localized lateral to primary auditory cortex, in the STG, consistent with other recent work on speech selectivity (Overath et al., 2015). By contrast, the music-selective component was largely localized anterior to primary auditory cortex. The results thus provide evidence for distinct pathways for music and speech processing in nonprimary auditory cortex. This apparent functional segregation raises many questions about the role of these regions in speech and music perception, about their evolutionary history, and about their dependence on auditory experience and expertise.

One obvious feature of the component weight maps in Figure 2.11C is a strong degree of bilaterality. This symmetry contrasts with several prior proposals for functional segregation between hemispheres. One proposal is that the left and right auditory cortices are specialized for different aspects of signal processing, with the left optimized for temporal resolution and the right for frequency resolution (Zatorre, Belin, & Penhune, 2002). The evidence for hemispheric differences comes mainly from functional imaging studies that manipulate spectral and temporal stimulus characteristics (Samson, Zeffiro, Toussaint, & Belin, 2011; Zatorre & Belin, 2001) and neuropsychology studies that find pitch perception deficits associated with right temporal lesions (Johnsrude, Penhune, & Zatorre, 2000; Zatorre, 1985). A related alternative idea is that the two hemispheres are specialized to analyze distinct timescales, with the left hemisphere more

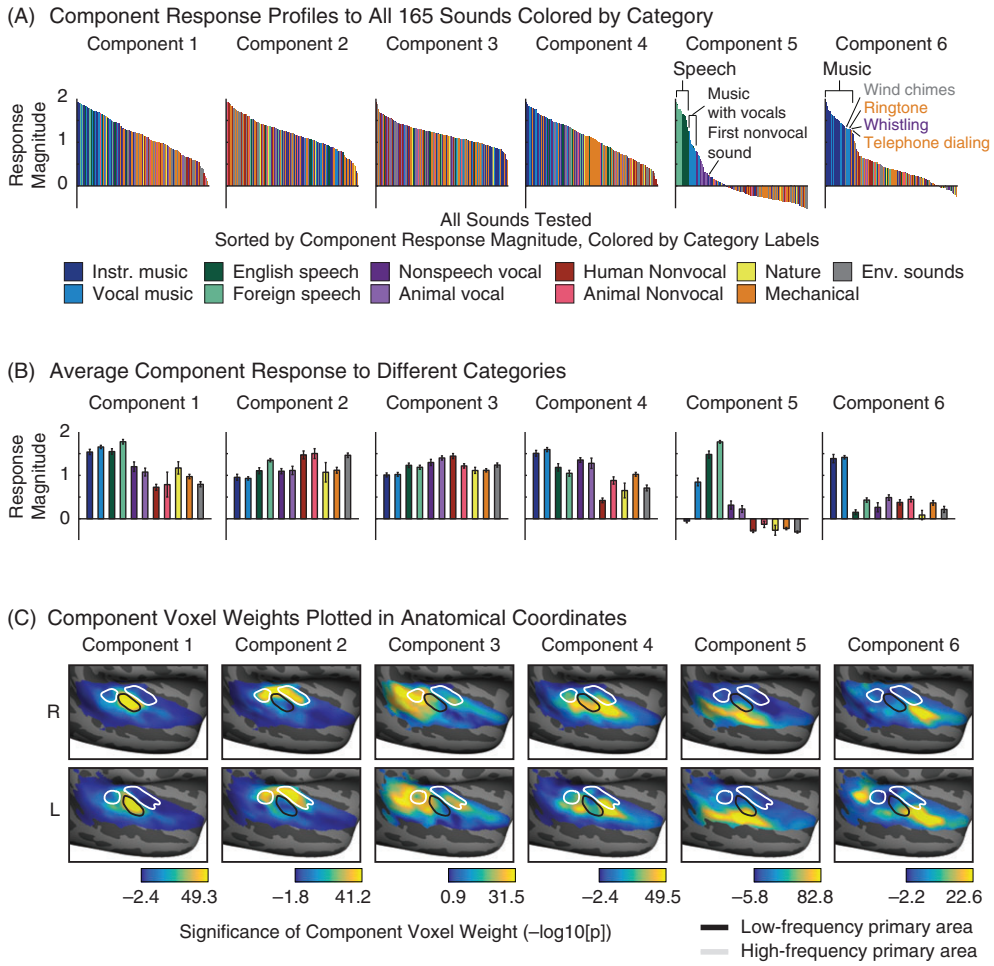

Figure 2.11 Functional organization of the nonprimary auditory cortex. (A) Results of decomposing voxel responses to 165 natural sounds into six components. Each component is described by its response to each of the sounds, here ordered by the response magnitude and color coded with the sound category. The first four components are well described by selectivity to established acoustic properties, and are not strongly category selective. By contrast, the last two components are selective for speech and music, respectively. (B) The average response of each of the six components to each category of sound. (C) The weights for each component plotted on an inflated brain. White and black outlines mark the high- and low-frequency fields of the tonotopic map, commonly identified with primary auditory cortex. Components 1–4 are primarily localized in and around primary auditory cortex, whereas the speech- and music-selective components localize to distinct regions of nonprimary auditory cortex.
SOURCE: From Norman-Haignere, Kanwisher, and McDermott (2015). Reprinted with permission of Elsevier.

responsive to short-scale temporal variation (e.g., tens of milliseconds) and the right hemisphere more responsive to long-scale variation (e.g., hundreds of milliseconds) (Boemio et al., 2005; Poeppel, 2003). Such asymmetries are not obvious in our fMRI results, but might become evident with measurements that have better temporal resolution.

SOUND SOURCE PERCEPTION

Ultimately, we wish to understand not only what acoustic measurements are made by the auditory system, as were characterized in the previous sections, but also how they give rise to perception—what we hear when we listen to sound. Following Helmholtz, we might suppose that the purpose of audition is to infer something about the events in the world that produce sound. We can often identify sound sources with a verbal label, for instance, and realize that we heard a finger snap, a flock of birds, or construction noise. Even if we cannot determine the object(s) that caused the sound, we may nonetheless know something about what happened: that something fell onto a hard floor, or into water (Gaver, 1993). Despite the richness of these aspects of auditory recognition, remarkably little is known at present about them (speech recognition stands alone as an exception), mainly because they are rarely studied (though see Gygi, Kidd, & Watson, 2004; Lutfi, 2008; and McDermott & Simoncelli, 2011).

Perhaps because they are more easily linked to peripheral processing than are our recognition abilities, researchers have been more inclined to instead study the perception of isolated properties of sounds or their sources (e.g., location, intensity, rate of vibration, or temporal pattern). Much research has concentrated in particular on three well-known properties of sound: spatial location, pitch, and loudness. This focus is on the one hand unfortunate, as auditory perception is much richer than the hegemony of these three attributes in hearing science would indicate. However, their study has given rise to rich lines of research that have yielded many useful insights about hearing.

Localization

Localization is less precise in hearing than in vision, but enables us to localize objects that we may not be able to see. Human observers can judge the location of a source to within a few degrees if conditions are optimal. The processes by which this occurs are among the best understood in hearing.

Spatial location is not made explicit on the cochlea, which provides a map of frequency rather than of space, and instead must be derived from three primary sources of information. Two of these are binaural, resulting from differences in the acoustic input to the two ears. Sounds to one side of the vertical meridian reach the two ears at different times and with different intensities. This is due to the difference in path length from the source to the ears, and to the acoustic shadowing effect of the head. These interaural time and level differences vary with direction and thus provide a cue to a sound source's location. Binaural cues are primarily useful for deriving the location of a sound in the horizontal plane, because changes in elevation do not change interaural time or intensity differences much. To localize sounds in the vertical dimension, or to distinguish sounds coming from in front of the head from those from in back, listeners rely on a third source of information: the filtering of sounds by the body and ears. This filtering is direction specific, such that a spectral analysis can reveal peaks and valleys in the frequency spectrum that are signatures of location in the vertical dimension (Figure 2.12; discussed further below).

Interaural time differences (ITD) are typically a fraction of a millisecond, and just-noticeable differences in ITD (which determine spatial acuity) can be as low as 10 μs (Klump & Eady, 1956). This is striking, given that neural refractory periods (which determine the minimum interspike interval for a single neuron) are on the order of a millisecond, which one might think would put a limit on the temporal resolution of neural representations. Typical interaural

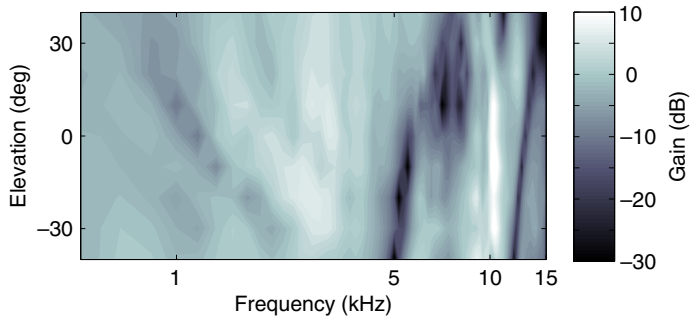

Figure 2.12 Head-related transfer functions. Sample HRTF for the left ear of one human listener. The gray level represents the amount by which a frequency originating at a particular elevation is attenuated or amplified by the torso, head, and ear of the listener. Sounds are filtered differently depending on their elevation, and the spectrum that is registered by the cochlea thus provides a localization cue. Note that most of the variation in elevation-dependent filtering occurs at high frequencies (above 4 kHz). SOURCE: From Zahorik, Bangayan, Sundareswaran, Wang, and Tam (2006). Reproduced with permission of AIP Publishing LLC.

level differences (ILD) for a single sound source in a quiet environment can be as large as 20 dB, with a just-noticeable difference of about 1 dB. ILDs result from the acoustic shadow cast by the head, and although the relationship between ILD and location is complex (Culling & Akeroyd, 2010), to first order, ILDs are more pronounced for high frequencies, as low frequencies are less affected by the acoustic shadow (because their wavelengths are comparable to the dimensions of the head). ITDs, in contrast, support localization most effectively at low frequencies, when the time difference between individual cycles of sinusoidal sound components can be detected via phase-locked spikes from the two ears (phase locking, as we discussed earlier, degrades at high frequencies). That said, ITDs between the envelopes of high-frequency sounds can also produce percepts of localization. The classical "duplex" view that localization is determined by either ILDs or ITDs, depending on the frequency (Rayleigh, 1907), is thus not fully appropriate for realistic natural sounds, which in general produce perceptible ITDs across the spectrum. It must also be noted that ITDs and ILDs recorded in natural conditions (i.e., with multiple sound sources and background noise) exhibit values and frequency dependence that are distinct from those expected from classical considerations of single sound sources in quiet (Mlynarski & Jost, 2014). More generally, localization in real-world conditions with multiple sources is understudied and remains poorly understood. See Middlebrooks and Green (1991) for a review of much of the classic behavioral work on sound localization.

The binaural cues to sound location are extracted in the superior olive, a subcortical region where inputs from the two ears are combined. There appears to be an elegant segregation of function, with ITDs being extracted in the medial superior olive (MSO) and ILDs being extracted in the lateral superior olive (LSO). In both cases, accurate coding of interaural differences is made possible by neural signaling with unusually high temporal precision. This precision is needed both to encode submillisecond ITDs as well as ILDs of brief transient events, for which the inputs from the ears must be aligned in time. Brain structures subsequent to the superior olive largely inherit its ILD and ITD sensitivity. See Yin and Kuwada (2010) for a recent review of the physiology of binaural localization.

Binaural cues are of little use in distinguishing sounds at different locations on the vertical dimension (relative to the head), or in distinguishing front from back, as interaural time and level differences are largely unaffected by changes across these locations. Instead, listeners rely on spectral cues provided by the filtering of a sound by the torso, head, and ears of a listener. The filtering results from the reflection and absorption of sound by the surfaces of a listener's body, with sound from different directions producing different patterns of reflection. The effect of these interactions on the sound that reaches the eardrum can be described by a linear filter known as the head-related transfer function (HRTF). The overall effect is that of amplifying some frequencies while attenuating others. A broadband sound entering the ear will thus be endowed with peaks and valleys in its frequency spectrum (Figure 2.12).

Compelling sound localization can be perceived when these peaks and valleys are artificially induced. The effect of the filtering is obviously confounded with the spectrum of the unfiltered sound source, and the listener must make some assumptions about the source spectrum. When these assumptions are violated, as with narrowband sounds whose spectral energy occurs at a peak in the HRTF of a listener, sounds are mislocalized (Middlebrooks, 1992). For broadband sounds, however, HRTF filtering produces signatures that are sufficiently distinct as to support localization in the vertical dimension to within 5 degrees or so in some cases, though some locations are more accurately perceived than others (Makous & Middlebrooks, 1990; Wightman & Kistler, 1989).

The bulk of the filtering occurs in the outer ear (the pinna), the folds of which produce distinctive pattern of reflections. Because pinna shapes vary across listeners, the HRTF is listener specific as well as location specific, with spectral peaks and valleys that are in different places for different listeners. Listeners appear to learn the HRTFs for their set of ears. When ears are artificially modified with plastic molds that change their shape, localization initially suffers considerably, but over a period of weeks, listeners regain the ability to localize with the modified ears (Hofman, Van Riswick, & van Opstal, 1998). Listeners thus learn at least some of the details of their particular HRTF through experience, although sounds can be localized even when the peaks and valleys of the pinna filtering are somewhat blurred (Kulkarni & Colburn, 1998). Moreover, compelling spatialization is often evident even if a generic HRTF is used.

The physiology of HRTF-related cues for localization is not as developed as it is for binaural cues, but there is evidence that midbrain regions may again be important. Many inferior colliculus neurons, for instance, show tuning to sound elevation (Delgutte, Joris, Litovsky, & Yin, 1999). The selectivity for elevation presumably derives from tuning to particular spectral patterns (peaks and valleys in the spectrum) that are diagnostic of particular locations (May, Anderson, & Roos, 2008).

Although the key cues for sound localization are extracted subcortically, lesion studies reveal that the cortex is essential for localizing sound behaviorally. Ablating the auditory cortex typically produces large deficits in localizing sounds (Heffner & Heffner, 1990), with unilateral lesions producing deficits specific to locations contralateral to the side of the lesion (Jenkins & Masterton, 1982). Consistent with these findings, tuning to sound location is widespread in auditory cortical neurons, with the preferred location generally positioned in the contralateral hemifield (Middlebrooks, 2000). Topographic representations of space have not been found to be evident within individual auditory cortical areas, though one recent report argues that such topography may be evident across multiple areas (Higgins, Storace,

Escabi, & Read, 2010). See Grothe, Pecka, and McAlpine (2010) for a recent review of the physiological basis of sound localization.

Pitch

Although the word "pitch" is often used colloquially to describe the perception of sound frequency, in hearing research it has a more specific meaning—pitch is defined as the perceptual correlate of periodicity. Vocalizations, instrument sounds, and some machine sounds are all often produced by periodic physical processes. Our vocal cords open and close at regular intervals, producing a series of clicks separated by regular temporal intervals. Instruments produce sounds via strings that oscillate at a fixed rate, or via tubes in which the air vibrates at particular resonant frequencies, to give two examples. Machines frequently feature rotating parts, which often produce sounds at every rotation. In all these cases, the resulting sounds are periodic—the sound pressure waveform consists of a single shape that repeats at a fixed rate (Figure 2.13A).

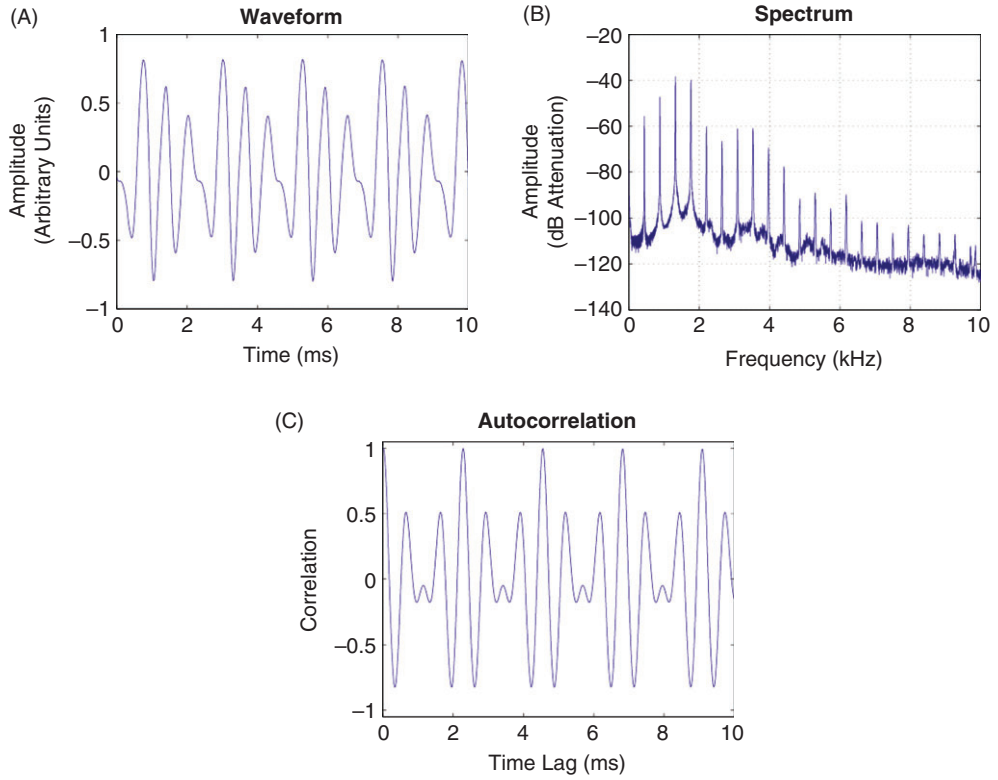

Figure 2.13 Periodicity and pitch. Periodicity and pitch. Waveform, spectrum, and autocorrelation function for a note (the A above middle C, with an F0 of 440 Hz) played on an oboe. (A) Excerpt of waveform. Note that the waveform repeats every 2.27 ms, which is the period. (B) Spectrum. Note the peaks at integer multiples of the F0, characteristic of a periodic sound. In this case the F0 is physically present, but the second, third, and fourth harmonics actually have higher amplitude. (C) Autocorrelation. The correlation coefficient is always 1 at a lag of 0 ms, but because the waveform is periodic, correlations close to 1 are also found at integer multiples of the period (2.27, 4.55, 6.82, and 9.09 ms, in this example).
SOURCE: From McDermott and Oxenham (2008a). Reprinted with permission of Elsevier.

Perceptually, such sounds are heard as having a pitch that can vary from low to high, proportional to the frequency at which the waveform repeats (the fundamental frequency, that is, the F0).

Pitch is important because periodicity is important—the period is often related to properties of the source that are useful to know, such as its size, or tension. Pitch is also used for communicative purposes, varying in speech prosody, for instance, to convey meaning or emotion. Pitch is a centerpiece of music, forming the basis of melody, harmony, and tonality. Listeners also use pitch to track sound sources of interest in auditory scenes.

Many physically different sounds—all those with a particular period—have the same pitch. The periodicity is unrelated to whether a sound's frequencies fall in high or low regions of the spectrum, for instance, though in practice periodicity and the center of mass of the spectrum are sometimes correlated. Historically, pitch has been a focal point of hearing research because it is an important perceptual property with a nontrivial relationship to the acoustic input. Debates on pitch and related phenomena date back at least to Helmholtz, and continue to occupy researchers today (Plack, Oxenham, Popper, & Ray, 2005).

One central debate concerns whether pitch is derived from an analysis of frequency or time. Periodic waveforms produce spectra whose frequencies are harmonically related—they form a harmonic series, being integer multiples of the fundamental frequency, whose period is the period of the waveform (Figure 2.13B). Pitch could thus conceivably be detected with harmonic templates applied to an estimate of a sound's spectrum obtained from the cochlea (Goldstein, 1973; Shamma & Klein, 2000; Terhardt, 1974; Wightman, 1973). Alternatively, periodicity could be assessed in the time domain, for instance via the autocorrelation function (Cariani & Delgutte, 1996; de Cheveigne and Kawahara, 2002; Meddis and Hewitt, 1991). The autocorrelation measures the correlation of a signal with a delayed copy of itself. For a periodic signal that repeats with some period, the autocorrelation exhibits peaks at multiples of the period (Figure 2.13C).

Such analyses are in principle functionally equivalent: The power spectrum is related to the autocorrelation via the Fourier transform, and detecting periodicity in one domain versus the other might simply seem a question of implementation. In the context of the auditory system, however, the two concepts diverge, because information is limited by distinct factors in the two domains. Time-domain models are typically assumed to utilize fine-grained spike timing (i.e., phase locking) with concomitant temporal resolution limits (because phase locking is absent for high frequencies). In contrast, frequency-based models (often known as "place models," in reference to the frequency-place mapping that occurs on the basilar membrane) rely on the pattern of excitation along the cochlea, which is limited in resolution by the frequency tuning of the cochlea (Cedolin & Delgutte, 2005). Cochlear frequency selectivity is present in time-domain models of pitch as well, but its role is typically not to estimate the spectrum but simply to restrict an autocorrelation analysis to a narrow frequency band (Bernstein & Oxenham, 2005), one consequence of which might be to improve its robustness in the presence of multiple sound sources. Reviews of the current debates and their historical origins are available elsewhere (de Cheveigne, 2004; Plack & Oxenham, 2005), and we will not discuss them exhaustively here.

Research on pitch has provided many important insights about hearing even though a conclusive account of pitch remains elusive. One contribution of pitch research has been to reveal the importance of the resolvability

of individual frequency components by the cochlea, a principle that has importance in other aspects of hearing as well. Because the frequency resolution of the cochlea is approximately constant on a logarithmic scale, whereas the components of a harmonic tone are equally spaced on a linear scale (separated by a fixed number of Hz, equal to the fundamental frequency of the tone; Figure 2.14A), multiple high-numbered harmonics fall within a single cochlear filter (Figure 2.14B). Because of the nature of the log scale, this is true regardless of whether the fundamental is low or high. As a result, the excitation pattern induced by a tone on the cochlea (of a human with normal hearing) is believed to contain resolvable peaks for only the first 10 or so harmonics (Figure 2.14).

There is now abundant evidence that resolvability places strong constraints on pitch perception. For instance, human pitch perception is determined predominantly by low-numbered harmonics (harmonics 1–10 or so in the harmonic series), presumably owing to the peripheral resolvability of these harmonics. Moreover, pitch discrimination is much worse for tones synthesized with only high-numbered harmonics than for tones containing only low-numbered harmonics, an effect not accounted for simply by the frequency range in which the harmonics occur (Houtsma & Smurzynski, 1990; Shackleton & Carlyon, 1994). This might be taken as evidence that the spatial pattern of excitation, rather than the periodicity that could be derived from the autocorrelation, underlies pitch perception, but variants of autocorrelation-based models have also been proposed to account for the effect of resolvability (Bernstein & Oxenham, 2005). Resolvability has since been demonstrated to constrain sound segregation as well as pitch (Micheyl & Oxenham, 2010b); see below.

The past decade has seen considerable interest in the neural mechanisms of pitch perception in both humans and nonhuman animals. One question is whether pitch is analyzed in a particular part of the brain. If so, one might expect the region to respond more to stimuli with pitch than to those lacking it, other things being equal. Although initially controversial (Hall & Plack, 2009), it is now reasonably well established that a region of the human auditory cortex exhibits this response signature, responding more to harmonic tones than to spectrally matched noise when measured with fMRI (Norman-Haignere, Kanwisher, & McDermott, 2013; Patterson, Uppenkamp, Johnsrude, & Griffiths, 2002; Penagos, Melcher, & Oxenham, 2004; Schonwiesner & Zatorre, 2008). The region appears to be present in every normal hearing listener, and exhibits a stereotypical location, overlapping the low-frequency portion of primary auditory cortex and extending anterior into nonprimary cortex (Figures 2.15A and B). Moreover, the region is driven by resolved harmonics, mirroring their importance to pitch perception (Figures 2.15C and D) (Norman-Haignere et al., 2013; Penagos et al., 2004). It is also noteworthy that something similar to pitch selectivity emerges from decomposing responses to natural sounds into their underlying components (see Component 4 in Figure 2.11), indicating that it is one of the main selectivities present in the auditory cortex (Norman-Haignere et al., 2015).

Although it remains unclear whether a region with similar characteristics is present in nonhuman animals, periodicity-tuned neurons are present in a similarly located region of the marmoset auditory cortex (Bendor & Wang, 2005). It is thus conceivable that homologous regions exist for pitch processing in the two species (Bendor & Wang, 2006). Comparable neurophysiology results have yet to be reported in other species (Fishman, Reser, Arezzo, & Steinschneider, 1998), however, and some have argued that

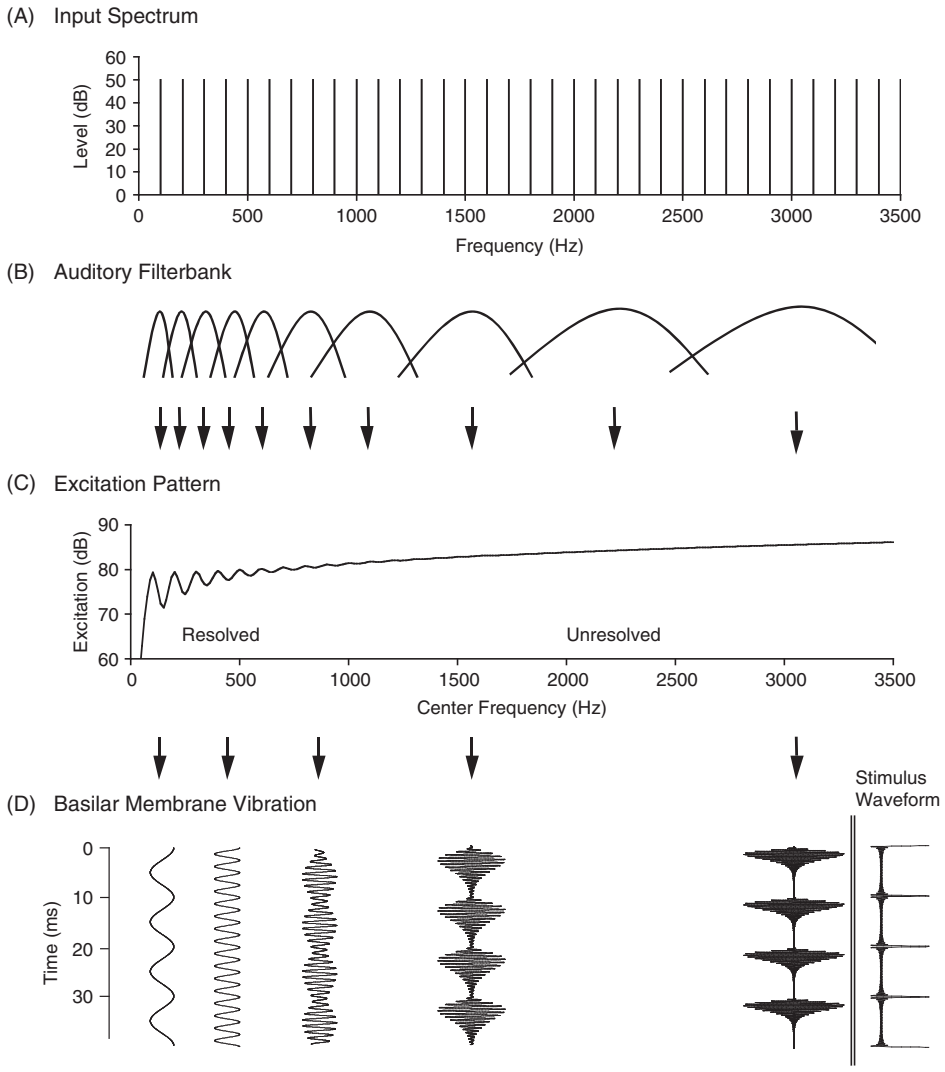

Figure 2.14 Resolvability. (A) Spectrum of a harmonic complex tone composed of 35 harmonics of equal amplitude. The fundamental frequency is 100 Hz—the lowest frequency in the spectrum and the amount by which adjacent harmonics are separated. (B) Frequency responses of auditory filters, each of which represents a particular point on the cochlea. Note that because a linear frequency scale is used, the filters increase in bandwidth with center frequency (compare to Figure 2.3a), such that many harmonics fall within the passband of the high-frequency filters. (C) The resulting pattern of excitation along the cochlea in response to the tone in (A). The excitation is the amplitude of vibration of the basilar membrane as a function of characteristic frequency (the frequency to which a particular point on the cochlea responds best, that is, the center frequency of the auditory filter representing the response properties of the cochlea at that point). Note that the first 10 or so harmonics produce resolvable peaks in the pattern of excitation, but that higher-numbered harmonics do not. The latter are thus said to be "unresolved." (D) The pattern of vibration that would be observed on the basilar membrane at several points along its length. When harmonics are resolved, the vibration is dominated by the harmonic close to the characteristic frequency, and is thus sinusoidal. When harmonics are unresolved, the vibration pattern is more complex, reflecting the multiple harmonics that stimulate the cochlea at those points.
SOURCE: Reprinted from Plack (2005). © Chris Plack.

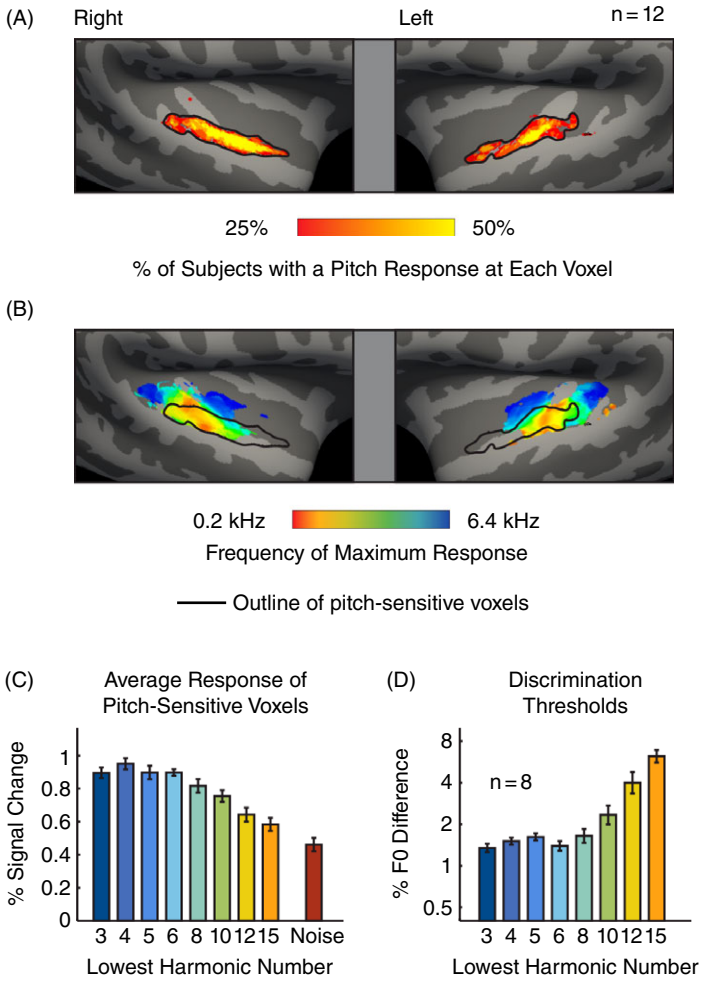

Figure 2.15 Pitch-responsive regions in humans. (A) Anatomical distribution of pitch-sensitive voxels, defined as responding significantly more to harmonic tones than to spectrally matched noise. (B) Tonotopic map obtained by plotting the frequency yielding the maximum response in each voxel (averaged across 12 listeners). Black outline replotted from (A) indicates the location of pitch responsive voxels, which overlap the low-frequency lobe of the tonotopic map and extend anteriorly. (C) Average response of pitch-sensitive voxels to tones varying in harmonic resolvability and to spectrally matched noise. Responses are lower for tones containing only high-numbered (unresolved) harmonics. (D) Pitch discrimination thresholds for the tones used in the fMRI experiment. Thresholds are higher for tones containing only high-numbered (unresolved) harmonics.
SOURCE: From Norman-Haignere, Kanwisher, and McDermott (2013). Reproduced with permission of the Society for Neuroscience.

pitch is encoded by ensembles of neurons with broad tuning rather than single neurons selective for particular fundamental frequencies (Bizley, Walker, King, & Schnupp, 2010). In general, pitch-related responses can be difficult to disentangle from artifactual responses to distortions introduced by the nonlinearities of the cochlea (de Cheveigne, 2010; McAlpine, 2004), though such distortions cannot account for the pitch

responses evident with fMRI in humans (Norman-Haignere & McDermott, 2016). See Walker, Bizley, King, and Schnupp (2011) and Winter (2005) for recent reviews of the brain basis of pitch.

It is important to note that hearing research has tended to equate pitch perception with the problem of estimating the F0 of a sound. However, in many real-world contexts (e.g., the perception of music or speech intonation), the changes in pitch over time are arguably more important than the absolute value of the F0. Pitch increases or decreases are what capture the identity of a melody, or the intention of a speaker. Less is known about how this *relative* pitch information is represented in the brain, but the right temporal lobe has been argued to be important, in part on the basis of brain-damaged patients with apparently selective deficits in relative pitch (Johnsrude et al., 2000). See McDermott and Oxenham (2008a) for a review of the perceptual and neural basis of relative pitch.

Loudness

Loudness is the perhaps the most immediate perceptual property of sound. To first order, loudness is the perceptual correlate of sound intensity. In real-world listening scenarios, loudness exhibits additional influences that suggest it serves to estimate the intensity of a sound source, as opposed to the intensity of the sound entering the ear (which changes with distance and the listening environment). However, loudness models that capture exclusively peripheral processing nonetheless have considerable predictive power. The ability to predict perceived loudness is important in many practical situations, and is a central issue in the fitting of hearing aids. The altered compression in hearing-impaired listeners affects the perceived loudness of sounds, and amplification runs the risk of making sounds uncomfortably loud unless compression is introduced artificially. There has thus been longstanding interest in quantitative models of loudness.

For a sound with a fixed spectral profile, such as a pure tone or a broadband noise, the relationship between loudness and intensity can be approximated via the classic Stevens power law (Stevens, 1955). However, the relation between loudness and intensity is not as simple as one might imagine. For instance, loudness increases with increasing bandwidth—a sound whose frequencies lie in a broad range will seem louder than a sound whose frequencies lie in a narrow range, even when their physical intensities are equal.

Standard models of loudness thus posit something somewhat more complex than a simple power law of intensity: that loudness is linearly related to the total amount of neural activity elicited by a stimulus at the level of the auditory nerve (ANSI, 2007; Moore & Glasberg, 1996). The effect of bandwidth on loudness is explained via the compression that occurs in the cochlea: Loudness is determined by the neural activity summed across nerve fibers, the spikes of which are generated after the output of a particular place on the cochlea is nonlinearly compressed. Because compression boosts low responses relative to high responses, the sum of several compressed responses to low amplitudes (produced by the several frequency channels stimulated by a broadband sound) is greater than a single compressed response to a large amplitude (produced by a single frequency channel responding to a narrowband sound of equal intensity). Loudness also increases with duration for durations up to half a second or so (Buus, Florentine, & Poulsen, 1997), suggesting that it is computed from neural activity integrated over some short window.

Loudness is also influenced in interesting ways by the apparent distance of a sound source. Because intensity attenuates with distance from a sound source, the intensity of

a sound at the ear is determined conjointly by the intensity and distance of the source. The auditory system appears to use loudness as a perceptual estimate of a source's intensity (i.e., the intensity at the point of origin): Sounds that appear more distant seem louder than those that appear closer but have the same overall intensity. Visual cues to distance have some influence on perceived loudness (Mershon, Desaulniers, Kiefer, Amerson, & Mills, 1981), but the cue provided by the amount of reverberation also seems to be important. The more distant a source, the weaker the direct sound from the source to the listener, relative to the reverberant sound that reaches the listener after reflection off of surfaces in the environment (Figure 2.14). This ratio of direct to reverberant sound appears to be used both to judge distance and to calibrate loudness perception (Zahorik & Wightman, 2001), though how the listener estimates this ratio from the sound signal remains unclear at present. Loudness thus appears to function somewhat like size or brightness perception in vision, in which perception is not based exclusively on retinal size or light intensity (Adelson, 2000).

AUDITORY SCENE ANALYSIS

Thus far we have discussed how the auditory system represents single sounds in isolation, as might be produced by a note played on an instrument, or a word uttered by someone talking. The simplicity of such isolated sounds renders them convenient objects of study, yet in many auditory environments isolated sounds are not the norm. It is often the case that lots of things make sound at the same time, causing the ear to receive a mixture of multiple sources as its input. Consider Figure 2.16, which displays cochleagrams of a single "target" speaker along with that of the mixture that results from adding to it the utterances of one, three, and seven additional speakers, as might occur in a social setting. The brain's task in this case is to take such a mixture as input and recover enough of the content of a target sound source to allow speech comprehension or otherwise support behavior. This is a nontrivial task. In the example of Figure 2.16, for instance, it is apparent that the structure of the target utterance is progressively obscured as more speakers are added to the mixture. The presence of competing sounds greatly complicates the computational extraction of just about any sound source property, from pitch (de Cheveigne, 2006) to location (Mandel, Weiss, & Ellis, 2010). Human listeners, however, parse auditory scenes with a remarkable degree of success. In the example of Figure 2.16, the target remains largely audible to most listeners even in the mixture of eight speakers. This is the classic "cocktail party problem" (Bee & Micheyl, 2008; Bregman, 1990; Bronkhorst, 2000; Cherry, 1953; Carlyon, 2004; Darwin, 1997; McDermott, 2009).

Historically, the "cocktail party problem" has referred to two conceptually distinct problems that in practice are closely related. The first, known as sound segregation, is the problem of deriving representations of individual sound sources from a mixture of sounds. The second, selective attention, entails the task of directing attention to one source among many, as when listening to a particular speaker at a party. These two problems are related because the ability to segregate sounds is probably dependent on attention (Carlyon, Cusack, Foxton, & Robertson, 2001; Shinn-Cunningham, 2008; Woods & McDermott, 2015), though the extent and nature of this dependence remains an active area of study (Macken, Tremblay, Houghton, Nicholls, & Jones, 2003; Masutomi, Barascud, Kashino, McDermott, & Chait, 2016). Here we will focus on the

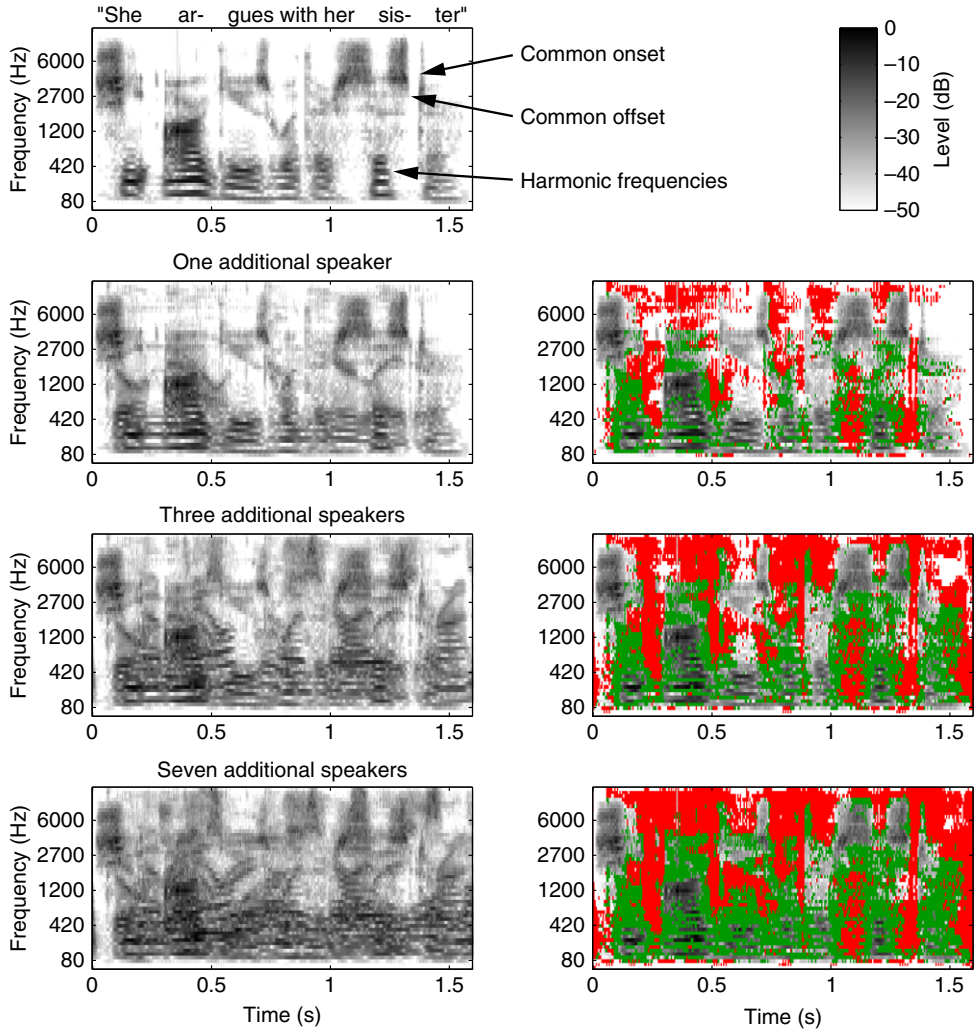

Figure 2.16 The cocktail party problem. Cochleagrams of a single "target" utterance (top row), and the same utterance mixed with one, three, and seven additional speech signals from different speakers. The mixtures approximate the signal that would enter the ear if the additional speakers were talking as loudly as the target speaker, but were standing twice as far away from the listener (to simulate cocktail party conditions). The grayscale denotes attenuation from the maximum energy level across all of the signals (in dB), such that gray levels can be compared across cochleagrams. Cochleagrams in the right column are identical to those on the left except for the superimposed color masks. Pixels labeled green are those where the original target speech signal is more than –50 dB but the mixture level is at least 5 dB higher, thus masking the target speech. Pixels labeled red are those where the target had less and the mixture had more than –50 dB energy. Cochleagrams were computed from a filter bank with bandwidths and frequency spacing similar to those in the ear. Each pixel is the rms amplitude of the signal within a frequency band and time window.

SOURCE: From McDermott (2009). Reprinted with permission of Elsevier.

first problem, of sound segregation, which is typically studied under conditions where listeners pay full attention to a target sound. Al Bregman, a Canadian psychologist, is typically credited with drawing interest to this problem and pioneering its study (Bregman, 1990).

Sound Segregation and Acoustic Grouping Cues

Sound segregation is a classic example of an ill-posed problem in perception. Many different sets of sounds are physically consistent with the mixture that enters the ear (in that their sum is equal to the mixture). The auditory system must infer the set of sounds that actually occurred. As in other ill-posed problems, this inference is only possible with the aid of assumptions that constrain the solution. In this case the assumptions concern the nature of sounds in the world, and are presumably learned from experience with natural sounds (or perhaps hardwired into the auditory system via evolution).

Grouping cues (i.e., sound properties that dictate whether sound elements are heard as part of the same sound) are examples of these assumptions. For instance, natural sounds that have pitch, such as vocalizations, contain frequencies that are harmonically related, evident as banded structures in the cochleagram of the target speaker in Figure 2.16. Harmonically related frequencies are unlikely to occur from the chance alignment of multiple different sounds, and thus when they are present in a mixture, they are likely to be due to the same sound, and are generally heard as such (de Cheveigne, McAdams, Laroche, & Rosenberg, 1995). Moreover, a component that is mistuned (in a tone containing otherwise harmonic frequencies) segregates from the rest of the tone (Hartmann, McAdams, & Smith, 1990; Moore, Glasberg, & Peters, 1986; Roberts & Brunstrom, 1998). Understanding sound segregation requires understanding the acoustic regularities, such as harmonicity, that characterize natural sound sources, and that are used by the auditory system. It is my view that these regularities should be revealed by analysis of natural sounds, but for now research in this area is mostly being driven by intuitions about sound properties that might be important for segregation.

Perhaps the most important generic acoustic grouping cue is common onset: Frequency components that begin and end at the same time are likely to belong to the same sound. Onset differences, when manipulated experimentally, cause frequency components to perceptually segregate from each other (Cutting, 1975; Darwin, 1981). Interestingly, a component that has an earlier or later onset than the rest of a set of harmonics has reduced influence over the perceived pitch of the entire tone (Darwin & Ciocca, 1992), suggesting that pitch computations operate on frequency components that are deemed likely to belong together, rather than on the raw acoustic input.

Onset may be viewed as a special case of co-modulation—amplitude modulation that is common to different spectral regions. In some cases relatively slow co-modulation promotes grouping of different spectral components (Hall, Haggard, & Fernandes, 1984), though abrupt onsets seem to be most effective. Common offset also promotes grouping, though is less effective than common onset (Darwin, 1984), perhaps because abrupt offsets are less common in natural sounds (Cusack & Carlyon, 2004).

Not every intuitively plausible grouping cue produces a robust effect when assessed psychophysically. For instance, frequency modulation (FM) that is shared ("coherent") across multiple frequency components, as in voiced speech, has been proposed to promote their grouping (Bregman, 1990; McAdams,

1989). However, listeners are poor at discriminating coherent from incoherent FM if the component tones are not harmonically related, indicating that sensitivity to FM coherence may simply be mediated by the deviations from harmonicity that occur when harmonic tones are incoherently modulated (Carlyon, 1991).

Failures of segregation are often referred to as "informational masking," so-called because they often manifest as masking-like effects on the detectability of a target tone, but cannot be explained in terms of classical "energetic masking" (in which the response to the target is swamped by a masker that falls within the same peripheral channel). Demonstrations of informational masking typically present a target tone along with other tones that lie outside a "protected region" of the spectrum, such that they do not stimulate the same filters as the target tone. These "masking" tones nonetheless often elevate the detection threshold for the target, sometimes quite dramatically (Durlach et al., 2003; Lutfi, 1992; Neff, 1995; Watson, 1987). The effect is presumably due to impairments in the ability to segregate the target tone from the masker tones.

One might also think that the task of segregating sounds would be greatly aided by the tendency of distinct sound sources in the world to originate from distinct locations. In practice, spatial cues are indeed of some benefit, for instance in hearing a target sentence from one direction amid distracting utterances from other directions (Bronkhorst, 2000; Hawley, Litovsky, & Culling, 2004; Ihlefeld & Shinn-Cunningham, 2008; Kidd, Arbogast, Masson, & Gallun, 2005). However, spatial cues are surprisingly ineffective at segregating one frequency component from a group of others (Culling & Summerfield, 1995), especially when pitted against other grouping cues such as onset or harmonicity (Darwin & Hukin, 1997). The benefit of listening to a target with a distinct location (Bronkhorst, 2000; Hawley et al., 2004; Ihlefeld & Shinn-Cunningham, 2008; Kidd et al., 2005) may thus be due to the ease with which the target can be attentively tracked over time amid competing sound sources, rather than to a facilitation of auditory grouping per se (Darwin & Hukin, 1999). Moreover, humans are usually able to segregate monaural mixtures of sounds without difficulty, demonstrating that spatial separation is often not necessary for high performance. For instance, much popular music of the 20th century was released in mono, and yet listeners have no trouble distinguishing many different instruments and voices in any given recording. Spatial cues thus contribute to sound segregation, but their presence or absence does not seem to fundamentally alter the problem.

The weak effect of spatial cues on segregation may reflect their fallibility in complex auditory scenes. Binaural cues can be contaminated when sounds are combined or degraded by reverberation (Brown & Palomaki, 2006), and can even be deceptive, as when caused by echoes (whose direction is generally different from the original sound source). It is possible that the efficacy of different grouping cues in general reflects their reliability in natural conditions. Evaluating this hypothesis will require statistical analysis of natural auditory scenes, an important direction for future research.

Sequential Grouping

Because the cochleagram approximates the input that the cochlea provides to the rest of the auditory system, it is common to view the problem of sound segregation as one of deciding how to group the various parts of the cochleagram (Bregman, 1990). However, the brain does not receive an entire spectrogam at once—sound arrives gradually

over time. Many researchers thus distinguish between the problem of simultaneous grouping (determining how the spectral content of a short segment of the auditory input should be segregated) and sequential grouping (determining how the groups from each segment should be linked over time, for instance to form a speech utterance, or a melody) (Bregman, 1990).

Although most of the classic grouping cues (onset/comodulation, harmonicity, ITD, etc.) are quantities that could be measured over short timescales, the boundary between what is simultaneous or sequential is unclear for most real-world signals, and it may be more appropriate to view grouping as being influenced by processes operating at multiple timescales rather than two cleanly divided stages of processing. There are, however, contexts in which the bifurcation into simultaneous and sequential grouping stages is natural, as when the auditory input consists of discrete sound elements that do not overlap in time. In such situations interesting differences are sometimes evident between the grouping of simultaneous and sequential elements. For instance, spatial cues, which are relatively weak as a simultaneous cue, have a stronger influence on sequential grouping of tones (Darwin & Hukin, 1997).

Another clear case of sequential processing can be found in the effects of sound repetition. Sounds that occur repeatedly in the acoustic input are detected by the auditory system as repeating, and are inferred to be a single source. Perhaps surprisingly, this is true even when the repeating source is embedded in mixtures with other sounds, and is never presented in isolation (McDermott et al., 2011). In such cases the acoustic input itself does not repeat, but the source repetition induces correlations in the input that the auditory system detects and uses to extract the repeating sound. The informativeness of repetition presumably results from the fact that mixtures of multiple sounds tend not to occur repeatedly, such that when a structure does repeat, it is likely to be a single source.

Effects of repetition are also evident in classic results on "informational masking," in that the detectability of a target tone amid masking tones can be increased when the target is repeatedly presented (Kidd, Mason, Deliwala, & Woods, 1994; Kidd, Mason, & Richards, 2003). Similarly, when repeating patterns of tones appear in a random background of tones, they reliably "pop out" and are detected (Teki, Chait, Kumar, Shamma, & Griffiths, 2013). Moreover, segregation via repetition seems to be robust to inattention—listeners are able to make judgments about target sources embedded in mixtures even when the repetitions that enable target detection and discrimination occur while the listeners perform a difficult second concurrent task (Masutomi et al., 2016). Although repeating structure is rarely present in speech, it is common in music, and in many animal vocalizations, which often consist of a short call repeated several times in quick succession, perhaps facilitating their segregation from background sounds.

Streaming

One type of sequential segregation effect has particularly captured the imagination of the hearing community, and merits special mention. When two pure tones of different frequency are repeatedly presented in alternation (Figure 2.17A), one of two perceptual states is commonly reported by listeners: one in which the two repeated tones are heard as a single "stream" whose pitch varies over time, and one in which two streams are heard, one with the high tones and one with the low tones (Bregman & Campbell, 1971). If the frequency separation between the two tones is small, and if the rate of alternation is slow, one stream is generally heard. When the

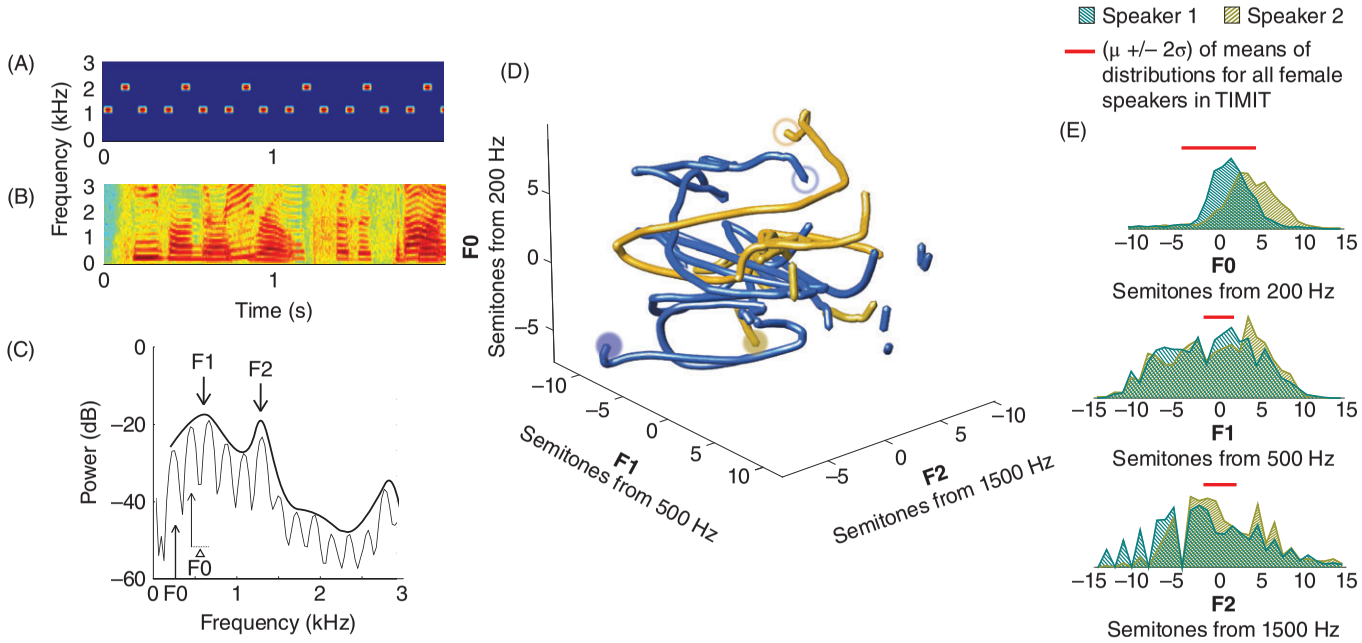

Figure 2.17 Streaming. (A) Spectrogram of "A-B-A" alternating tone stimulus commonly used to investigate streaming. Over time the two repeating tones segregate into separate streams. (B) Spectrogram of two concurrent utterances by two female speakers. (C) Power spectrum of a short excerpt from one of the utterances in (B). The spectrum exhibits peaks corresponding to harmonics of the fundamental frequency (i.e., F0) that determines the pitch, as well as peaks in the spectral envelope, known as formants, that determine the vowel being produced. (D) Pitch and formant trajectories for the two utterances in (B). The yellow line plots the trajectory for the utterance in (C). Open and closed circles denote the beginning and end of the trajectories, respectively. (E) Marginal distributions of F0, F1, and F2 measured across 10 sentences for the two speakers in (B). Red bars show the range of mean values of F0, F1, and F2 for the 53 female speakers in the TIMIT speech database. Note that differences between the average features of speakers are small relative to the variability produced by a single speaker, such that most pairs of speakers of the same gender will have overlapping trajectories.

SOURCE: Created by Kevin Woods; panels D–E from Woods and McDermott (2015). Reprinted with permission of Elsevier.

frequency separation is larger or the rate is faster, two streams tend to be heard, in which case "streaming" is said to occur (van Noorden, 1975).

An interesting hallmark of this phenomenon is that when two streams are perceived, judgments of the temporal order of elements in different streams are impaired (Bregman & Campbell, 1971; Micheyl & Oxenham, 2010a). This latter finding provides compelling evidence for a substantive change in the representation underlying the two percepts. Subsequent research has demonstrated that separation along most dimensions of sound can elicit streaming (Moore & Gockel, 2002). The streaming effects in these simple stimuli may be viewed as a variant of grouping by similarity—elements are grouped together when they are similar along some dimension, and segregated when they are sufficiently different, presumably because this similarity reflects the likelihood of having been produced by the same source.

Although two-tone streaming continues to be widely studied, its relevance to real-world streaming is unclear. The canonical (and arguably most important) real-world streaming problem is that of following one voice amid others, and mixtures of speech utterances are different in almost every respect from the A-B-A streaming stimulus. As shown in Figure 2.17B, mixtures of speakers physically overlap in both time and frequency. Moreover, even when represented in terms of perceptually relevant features such as pitch and the first two formants (which help define vowels; Figure 2.17C), two speakers of the same gender follow highly intertwined trajectories (Figure 2.17D). It is thus not obvious that insights from A-B-A streaming will translate straightforwardly to the streaming of speech and other natural sounds, though there are occasions in classical music in which alternating pitches stream (Bregman, 1990). One alternative approach is to synthesize stimuli that bear more similarity to the mixtures of sources we encounter in the world. For instance, when presented with mixtures of synthetic stimuli that evolve over time like spoken utterances, human listeners can track one of two "voices" with the aid of selective attention (Woods & McDermott, 2015).

Sound Texture

Although most work on scene analysis has focused on the perception of individual sound sources occurring concurrently with a few others, many natural scenes feature large numbers of similar sound elements, as produced by rain, fire, or groups of insects or animals (Figure 2.18A). The superposition of such similar acoustic events collectively gives rise to aggregate statistical properties, and the resulting sounds are referred to as "textures" (Saint-Arnaud & Popat, 1995; McDermott & Simoncelli, 2011). Sound textures are ubiquitous in the world, and commonly form the background for "foreground" sounds we want to recognize, such as someone talking. Textures also convey information themselves about the surrounding environment. Textures have only recently begun to be studied in audition, but are an appealing starting point for understanding auditory representation—they have rich, behaviorally relevant structure, and yet their properties do not change over time, simplifying the representational requirements.

Human listeners can reliably identify many different textures (McDermott & Simoncelli, 2011), raising the question of how they do so. Motivated in part by the observation that textures tend to be temporally homogeneous, we have proposed that they are represented with statistics: time averages of acoustic measurements made in the early auditory system. One piece of evidence for this idea comes from sound texture

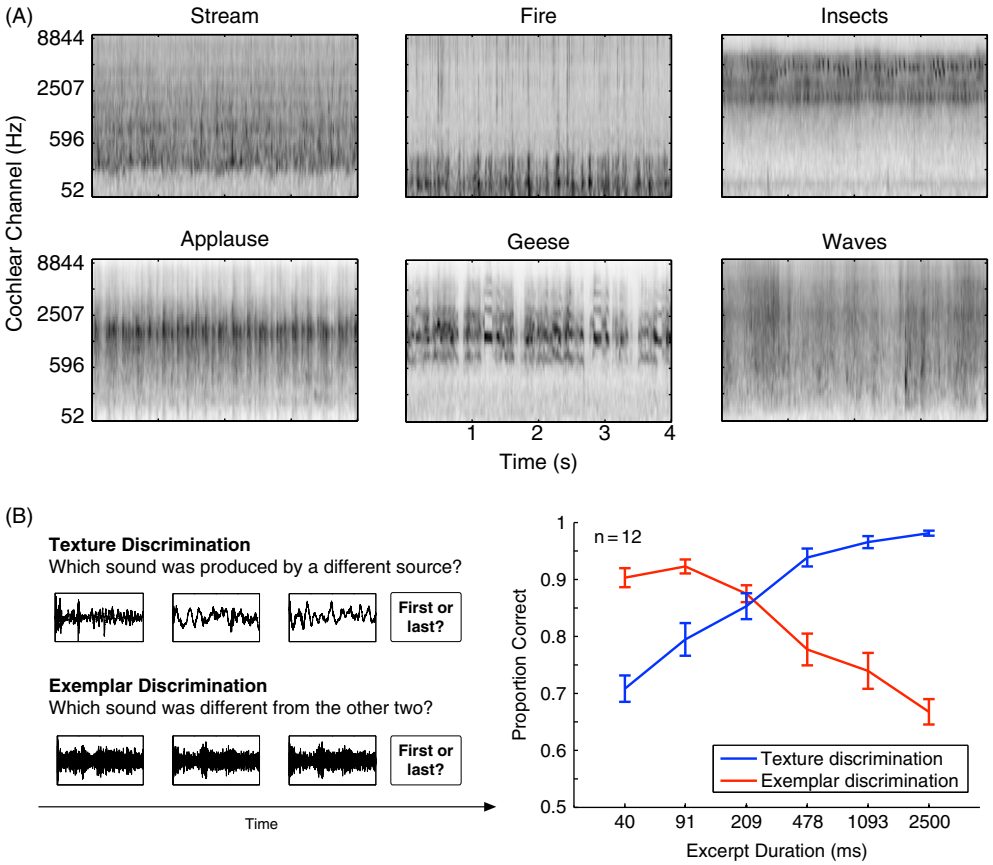

Figure 2.18 Sound texture. (A) Cochleagrams of six sample sound textures. Note the high degree of temporal homogeneity. (B) Texture and exemplar discrimination. Listeners were presented with three sound signals. In the texture discrimination task, two of the excerpts were from one texture, and one was from a different texture, and listeners identified the excerpt that was produced by a distinct source. In the exemplar discrimination task, two of the excerpts were identical, and the third was a distinct excerpt that was the same texture. Listeners identified the excerpt that was different from the other two.
SOURCE: (A) From McDermott and Simoncelli (2011). (B) From McDermott, Schemitsch, and Simoncelli (2013). Reproduced with permission of Macmillan.

synthesis, in which signals are synthesized to have statistics matched to those of particular real-world sounds (McDermott, Oxenham, & Simoncelli, 2009; McDermott & Simoncelli, 2011). The logic of this procedure is that if such statistics underlie our perception, then synthetic signals that share the statistical properties of real-world textures should sound like them (Heeger & Bergen, 1995; Portilla & Simoncelli, 2000). We found that realistic-sounding examples of many textures (water, fire, insects, etc.) can be synthesized from relatively simple statistics (moments and correlations) computed from the auditory model of Figure 2.7, suggesting that such statistics could underlie our perception.

Further evidence for statistical representations of texture comes from psychophysical experiments (Figure 2.18B). When human listeners are asked to discriminate excerpts from different textures, their performance improves with the excerpt duration, as

expected (longer excerpts provide more information on which to base discrimination). However, when asked to discriminate different excerpts of the same texture, performance declines with duration, even though longer excerpts again provide more information with which to tell the excerpts apart (McDermott et al., 2013). This result is what one would expect if the texture excerpts were represented with "summary" statistics that are averaged over time: As duration increases, the summary statistics of different excerpts of the same texture converge to the same value, rendering the excerpts difficult to tell apart if only their statistics are retained. The results suggest that the acoustic details that compose a texture are accumulated into statistical summaries but then become inaccessible. These statistical summaries permit distinct textures to be differentiated, but limit the ability to distinguish temporal detail.

Because texture is a relatively new topic of study, many interesting questions remain open. Texture statistics are time averages, and the mechanisms for computing the averages remain uncharacterized, as does their neural locus. It also remains unclear how textures and the other sound sources that are often superimposed on them are segregated, and whether the averaging process for computing texture statistics is selective, averaging only those acoustic details that are likely to belong to the texture rather than to other sources.

Filling In

Although it is common to view sound segregation as the problem of grouping the spectrogram-like output of the cochlea across frequency and time, this cannot be the whole story. That is because large swathes of a sound's time-frequency representation are often physically obscured (masked) by other sources, and are thus not physically available to be grouped. Masking is evident in the green pixels of Figure 2.16, which represent points where the target source has substantial energy, but where the mixture exceeds it in level. If these points are simply assigned to the target, or omitted from its representation, its level at those points will be misconstrued, and the sound potentially misidentified. To recover an accurate estimate of the target source, it is necessary to infer not just the grouping of the energy in the cochleagram, but also the structure of the target source in the places where it is masked.

There is considerable evidence that the auditory system in many cases infers "missing" portions of sound sources when evidence suggests that they are likely to have been masked. For instance, tones that are interrupted by noise bursts are "filled in" by the auditory system, such that they are heard as continuous in conditions in which physical continuity is likely given the stimulus (Warren, Obusek, & Ackroff, 1972). Known as the continuity effect, it occurs only when the interrupting noise bursts are sufficiently intense in the appropriate part of the spectrum to have masked the tone should it have been present continuously. Continuity is also heard for frequency glides (Ciocca & Bregman, 1987; Kluender & Jenison, 1992), as well as oscillating frequency- or amplitude-modulated tones (Carlyon, Micheyl, Deeks, & Moore, 2004; Lyzenga, Carlyon, & Moore, 2005). The perception of continuity across intermittent maskers was actually first reported for speech signals interrupted by noise bursts (Warren, 1970). For speech the effect is often termed "phonemic restoration," and likely indicates that knowledge of speech acoustics (and perhaps of other types of sounds as well) influences the inference of the masked portion of sounds. Similar effects occur for spectral gaps in sounds—they are perceptually filled in when evidence indicates they are

likely to have been masked (McDermott & Oxenham, 2008b). Filling in effects in hearing are conceptually similar to completion under and over occluding surfaces in vision, though the ecological constraints provided by masking (involving the relative intensity of two sounds) are distinct from those provided by occlusion (involving the relative depth of two surfaces). Neurophysiological evidence indicates that the representation of tones in primary auditory cortex reflects the perceived continuity, responding as though the tone were continuously present despite being interrupted by noise (Petkov, O'Connor, & Sutter, 2007; Riecke, van Opstal, Goebel, & Formisano, 2007).

Separating Sound Sources From the Environment

Thus far we have mainly discussed how the auditory system segregates the signals from multiple sound sources, but listeners face a second important scene-analysis problem. The sound that reaches the ear from a source is almost always altered to some extent by the surrounding environment, and these environmental influences complicate the task of recognizing the source. Typically the sound produced by a source reflects off multiple surfaces prior to reaching the ears, such that the ears receive some sound directly from the source, but also many reflected copies (Figure 2.19). These reflected copies (echoes) are delayed, as their path to the ear is lengthened, but generally also have altered frequency spectra, as reflective surfaces absorb some frequencies more than others. Because each reflection can be well described with a linear filter applied to the source signal, the signal reaching the ear, which is the sum of the direct sound along with all the reflections, can be described simply as the result of applying a single composite linear filter to the source (Gardner, 1998). Significant filtering of this sort occurs in almost every natural listening situation, such that sound produced in anechoic conditions (in which all surfaces are minimally reflective) sounds noticeably strange and unnatural.

Listeners are often interested in the properties of sound sources, and one might think of the environmental effects as a nuisance that should simply be discounted. However, environmental filtering imbues the acoustic input with useful information; for instance, about the size of a room where sound is produced and the distance of the source from the listener (Bronkhorst & Houtgast, 1999). It may thus be more appropriate to think of separating source and environment, at least to some extent, rather than simply recovering the source (Traer & McDermott, 2016). Reverberation is commonly used in music production, for instance, to create a sense of space, or to give a different feel to particular instruments or voices.

The loudness constancy phenomena discussed earlier (Zahorik & Wightman, 2001) are one example where the brain appears to infer the properties of the sound source separately from that of the environment, but there are several others. One of the most interesting involves the treatment of echoes in sound localization. The echoes that are common in most natural environments pose a problem for localization, as they generally come from directions other than that of the source (Figure 2.19B). The auditory system appears to solve this problem by perceptually fusing similar impulsive sounds that occur within a short duration of each other (on the order of 10 ms or so), and using the sound that occurs first to determine the perceived location. This "precedence effect," so-called because of the dominance of the sound that occurs first, was described and named by Hans Wallach (Wallach, Newman, & Rosenzweig, 1949), one of the great Gestalt psychologists, and has since been the subject of an interesting literature.

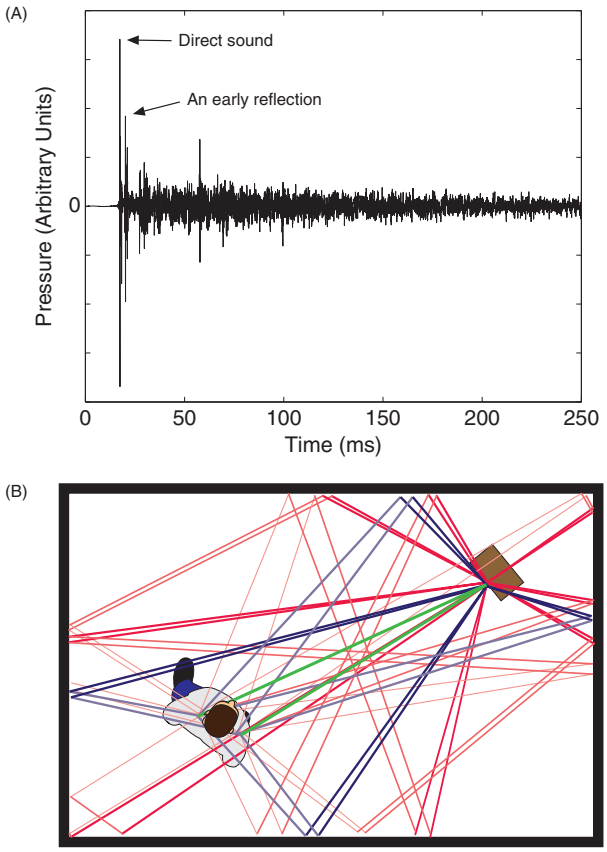

Figure 2.19 Reverberation. (A) Impulse response for a classroom. This is the sound waveform recorded in this room in response to a click (impulse) produced at a particular location in the room. The top arrow indicates the impulse that reaches the microphone directly from the source (that thus arrives first). The lower arrow indicates one of the subsequent reflections (i.e., echoes). After the early reflections, a gradually decaying reverberation tail is evident (cut off at 250 ms for clarity). The sound signal resulting from a different kind of source could be produced by convolving the sound from the source with this impulse response. (B) Schematic diagram of the sound reflections that contribute to the signal that reaches a listener's ears in a typical room. The brown box in the upper right corner depicts the speaker producing sound. The green lines depict the path taken by the direct sound to the listener's ears. Blue and green lines depict sound reaching the ears after one and two reflections, respectively. Sound reaching the ear after more than two reflections is not shown.
SOURCE: Part B from Culling and Akeroyd (2010). Reproduced with permission of Oxford University Press.

For instance, the maximum delay at which echoes are perceptually suppressed increases as two pairs of sounds are repeatedly presented (Freyman, Clifton, & Litovsky, 1991), presumably because the repetition provides evidence that the second sound is indeed an echo of the first, rather than being due to a distinct source (in which case it would not occur at a consistent delay following the first sound). Moreover, reversing the order of presentation can cause an abrupt breakdown of the effect, such that two sounds are heard rather than one, each with a different location. See Brown, Stecker, and Tollin (2015)

and Litovsky, Colburn, Yost, and Guzman (1999) for reviews of the precedence-effect literature.

Reverberation also poses a challenge for sound recognition, because different environments alter the sound from a source in different ways. Large amounts of reverberation (with prominent echoes at very long delays), as are present in some large auditoriums, can in fact greatly reduce the intelligibility of speech. Moderate amounts of reverberation, however, as are present in most spaces, typically have minimal effect on our ability to recognize speech and other sounds. Our robustness to everyday reverberation appears to be due in part to implicit knowledge of the regularities of real-world environmental acoustics. A recent large-scale study of impulse responses of everyday spaces found that they typically exhibit stereotyped properties (Traer & McDermott, 2016), presumably due to regularities in the way that common materials and environmental geometries reflect sound. Reverberant sound almost always decays exponentially, for instance, and at rates that depend on frequency in a consistent way (mid frequencies decay slowest and high frequencies the fastest). When these regularities are violated in synthetic reverberation, the resulting sound does not resemble reverberation to human listeners, and the properties of the underlying sound source are more difficult to separate from the effects of the environment (Traer & McDermott, 2016).

Part of our robustness to reverberation also likely derives from a process that adapts to the history of echo stimulation. In reverberant conditions, the intelligibility of a speech utterance has been found to be higher when preceded by another utterance than when not, an effect that does not occur in anechoic conditions (Brandewie & Zahorik, 2010; Watkins, 2005). Such results, like those of the precedence effect, are consistent with the idea that listeners construct a model of the environment's contribution to the acoustic input and use it to partially discount it when judging properties of a source. Analogous effects have been found with nonspeech sounds. When listeners hear instrument sounds preceded by speech or music that has been passed through a filter that "colors" the spectrum, the instrument sound is identified differently, as though listeners internalized the filter, assume it to be an environmental effect, and discount it to some extent when identifying the sound (Stilp, Alexander, Kiefte, & Kluender, 2010).

THE FUTURE OF HEARING RESEARCH

Hearing science is one of the oldest areas of psychology and neuroscience, with a strong research tradition dating back over 100 years, yet there remain many important open questions. Historically, hearing science tended to focus on the periphery—sound transduction and "early" sensory processing. This focus can be explained in part by the challenge of understanding the cochlea, the considerable complexity of the early auditory system, and the clinical importance of peripheral audition. However, the focus on the periphery has left many central aspects of audition underexplored, and recent trends in hearing research reflect a shift toward the study of these neglected mid- and high-level questions.

One important set of questions concerns the interface of audition with the rest of cognition, via attention and memory. Attention research ironically flourished in hearing early on (with the classic dichotic listening studies of Cherry [1953]), but then largely moved to the visual domain. Recent years have seen renewed interest, but much is still unclear about the role of attention in perceptual organization, about the representation of sound

outside of focused attention, and about the mechanisms of attentional selection in the auditory system.

Another promising research area involves working memory. Auditory short-term memory may have some striking differences with its visual counterpart (Demany, Trost, Serman, & Semal, 2008), and appears closely linked to attention and perhaps auditory scene analysis (Conway, Cowan, & Bunting, 2001). Studies of these topics in audition also hold promise for informing us more generally about the structure of cognition, as the similarities and differences with respect to visual cognition will reveal much about whether attention and memory mechanisms are domain general (perhaps exploiting central resources) or specific to particular sensory systems.

Interactions between audition and the other senses are also attracting increased interest. Information from other sensory systems likely plays a crucial role in hearing, given that sound on its own often provides ambiguous information. The sounds produced by rain and applause, for instance, can in some cases be quite similar, such that multisensory integration (using visual, somatosensory, or olfactory input) may help to correctly recognize the sound source. Cross-modal interactions in localization (Alais & Burr, 2004) are similarly powerful. Understanding cross-modal effects within the auditory system (Bizley, Nodal, Bajo, Nelken, & King, 2007; Ghazanfar, 2009; Kayser, Petkov, & Logothetis, 2008) and their role in behavior will be a significant direction of research going forward.

In addition to the uncharted territory in perception and cognition, there remain important open questions about peripheral processing. Some of these unresolved issues, such as the mechanisms of outer hair cell function, have great importance for understanding hearing impairment. Others may dovetail with higher-level function. For instance, the role of efferent connections to the cochlea is still uncertain, with some hypothesizing a role in attention or segregation (Guinan, 2006). The role of phase locking in frequency encoding and pitch perception is another basic issue that remains controversial and debated, and that has widespread relevance to mid-level audition.

As audition continues to evolve as a field, I believe useful guidance will come from a computational analysis of the inference problems the auditory system must solve (Marr, 1982). This necessitates thinking about the behavioral demands of real-world listening situations, as well as the constraints imposed by the way that information about the world is encoded in a sound signal. Many of these issues are becoming newly accessible with recent advances in computational power and signal-processing techniques.

For instance, one of the most important tasks a listener must perform with sound is surely that of recognition—determining what it was in the world that caused a sound, be it a particular type of object, or of a type of event, such as something falling on the floor (Gaver, 1993; Lutfi, 2008). Recognition is computationally challenging because the same type of occurrence in the world typically produces a different sound waveform each time it occurs. A recognition system must generalize across the variation that occurs within categories, but not the variation that occurs across categories (DiCarlo & Cox, 2007). Realizing this computational problem allows us to ask how the auditory system solves it. One place where these issues have been explored to some extent is speech perception (Holt & Lotto, 2010). The ideas explored there—about how listeners achieve invariance across different speakers and infer the state of the vocal apparatus along with the accompanying intentions of the speaker—could perhaps be extended to audition more generally (Rosenblum, 2004).

The inference problems of audition can also be better appreciated by examining real-world sound signals, and formal analysis of these signals seems likely to yield valuable clues. As discussed in previous sections, statistical analysis of natural sounds has been a staple of recent computational auditory neuroscience (Carlson, Ming, & DeWeese, 2012; Harper & McAlpine, 2004; Mlynarski, 2015; Rodriguez et al., 2010; Smith & Lewicki, 2006), where natural sound statistics have been used to explain the mechanisms observed in the peripheral auditory system. However, sound analysis seems likely to provide insight into mid- and high-level auditory problems as well. For instance, the acoustic grouping cues used in sound segregation are almost surely rooted to some extent in natural sound statistics, and examining such statistics could reveal unexpected cues. Similarly, because sound recognition must generalize across the variability that occurs within sounds produced by a particular type of source, examining this variability in natural sounds may provide clues to how the auditory system achieves the appropriate invariance in this domain.

The study of real-world auditory competence will also necessitate measuring auditory abilities and physiological responses with more realistic sound signals. The tones and noises that have been the staple of classical psychoacoustics and auditory physiology have many uses, but also have little in common with many everyday sounds. One challenge of working with realistic signals is that actual recordings of real-world sounds are often uncontrolled, and may introduce confounds associated with their familiarity. Methods of synthesizing novel sounds with naturalistic properties (Cavaco & Lewicki, 2007; McDermott & Simoncelli, 2011) are thus likely to be useful experimental tools. Considering more realistic sound signals will in turn necessitate more sophisticated models, particularly of cortical neural responses. The modulation filterbank models of Figures 2.7C and 2.8B have served hearing researchers well, but are clearly inadequate as models of complex auditory behavior and of cortical neural responses to natural sounds (Norman-Haignere et al., 2015).

We must also consider more realistic auditory behaviors. Hearing does not normally occur while we are seated in a quiet room, listening over headphones, and paying full attention to the acoustic stimulus, but rather in the context of everyday activities in which sound is a means to some other goal. The need to respect this complexity while maintaining sufficient control over experimental conditions presents a challenge, but not one that is insurmountable. For instance, neurophysiology experiments involving naturalistic behavior are becoming more common, with preparations being developed that will permit recordings from freely moving animals engaged in vocalization (Eliades & Wang, 2008) or locomotion—ultimately, perhaps a real-world cocktail party.

REFERENCES

Adelson, E. H. (2000). Lightness perception and lightness illusions. In M. S. Gazzaniga (Ed.), *The new cognitive neurosciences* (2nd ed., pp. 339–351). Cambridge, MA: MIT Press.

Ahrens, M. B., Linden, J. F., & Sahani, M. (2008). Influences in auditory cortical responses modeled with multilinear spectrotemporal methods. *Journal of Neuroscience*, *28*(8), 1929–1942.

Ahveninen, J., Jääskeläinen, I. P., Raij, T., Bonmassar, G., Devore, S., Hämäläinen, M., … Belliveau, J. W. (2006). Task-modulated "what" and "where" pathways in human auditory cortex. *Proceedings of the National Academy of Sciences, USA*, *103*(39), 14608–14613.

Alain, C., Arnott, S. R., Hevenor, S., Graham, S., & Grady, C. L. (2001). "What" and "where" in the human auditory system. *Proceedings of*

the National Academy of Sciences, USA, *98*, 12301–12306.

Alais, D., & Burr, D. E. (2004). The ventriloquist effect results from near-optimal bimodal integration. *Current Biology*, *14*, 257–262.

ANSI. (2007). *American National Standard Procedure for the Computation of Loudness of Steady Sounds (ANSI S3.4)*. Retrieved from https://global.ihs.com/doc_detail.cfm?&csf=ASA&document_name=ANSI%2FASA%20S3%2E4&item_s_key=00009561

Ashmore, J. (2008). Cochlear outer hair cell motility. *Physiological Review*, *88*, 173–210.

Attias, H., & Schreiner, C. E. (1997). Temporal low-order statistics of natural sounds. In M. Mozer, M. Jordan, & T. Petsche (Eds.), *Advances in neural information processing* (pp. 27–33). Cambridge, MA: MIT Press.

Attneave, F., & Olson, R. K. (1971). Pitch as a medium: A new approach to psychophysical scaling. *American Journal of Psychology*, *84*(2), 147–166.

Bacon, S. P., & Grantham, D. W. (1989). Modulation masking: Effects of modulation frequency, depth, and phase. *Journal of the Acoustical Society of America*, *85*, 2575–2580.

Bandyopadhyay, S., Shamma, S. A., & Kanold, P. O. (2010). Dichotomy of functional organization in the mouse auditory cortex. *Nature Neuroscience*, *13*(3), 361–368.

Barbour, D. L., & Wang, X. (2003). Contrast tuning in auditory cortex. *Science*, *299*, 1073–1075.

Barton, B., Venezia, J. H., Saberi, K., Hickok, G., & Brewer, A. A. (2012). Orthogonal acoustic dimensions define auditory field maps in human cortex. *Proceedings of the National Academy of Sciences, USA*, *109*(50), 20738–20743.

Baumann, S., Griffiths, T. D., Sun, L., Petkov, C. I., Thiele, A., & Rees, A. (2011). Orthogonal representation of sound dimensions in the primate midbrain. *Nature Neuroscience*, *14*(4), 423–425.

Baumann, S., Petkov, C. I., & Griffiths, T. D. (2013). A unified framework for the organization of the primate auditory cortex. *Frontiers in Systems Neuroscience*, *7*, 1–8.

Bee, M. A., & Micheyl, C. (2008). The cocktail party problem: What is it? How can it be solved? And why should animal behaviorists study it? *Journal of Comparative Psychology*, *122*(3), 235–251.

Belin, P., Zatorre, R. J., Lafaille, P., Ahad, P., & Pike, B. (2000). Voice-selective areas in human auditory cortex. *Nature*, *403*, 309–312.

Bendor, D., & Wang, X. (2005). The neuronal representation of pitch in primate auditory cortex. *Nature*, *426*, 1161–1165.

Bendor, D., & Wang, X. (2006). Cortical representations of pitch in monkeys and humans. *Current Opinion in Neurobiology*, *16*, 391–399.

Bendor, D., & Wang, X., (2008). Neural response properties of primary, rostral, and rostrotemporal core fields in the auditory cortex of marmoset monkeys. *Journal of Neurophysiology*, *100*(2), 888–906.

Bernstein, J. G. W., & Oxenham, A. J. (2005). An autocorrelation model with place dependence to account for the effect of harmonic number on fundamental frequency discrimination. *Journal of the Acoustical Society of America*, *117*(6), 3816–3831.

Binder, J. R., Frost, J. A., Hammeke, T. A., Bellgowan, P. S. F., Springer, J. A., Kaufman, J. N., & Possing, E. T. (2000). Human temporal lobe activation by speech and nonspeech sounds. *Cerebral Cortex*, *10*(5), 512–528.

Bitterman, Y., Mukamel, R., Malach, R., Fried, I., & Nelken, I. (2008). Ultra-fine frequency tuning revealed in single neurons of human auditory cortex. *Nature*, *451*(7175), 197–201.

Bizley, J. K., Nodal, F. R., Bajo, V. M., Nelken, I., & King, A. J. (2007). Physiological and anatomical evidence for multisensory interactions in auditory cortex. *Cerebral Cortex*, *17*, 2172–2189.

Bizley, J. K., Walker, K. M. M., King, A. J., & Schnupp, J. W. (2010). Neural ensemble codes for stimulus periodicity in auditory cortex. *Journal of Neuroscience*, *30*(14), 5078–5091.

Bizley, J. K., Walker, K. M. M., Silverman, B. W., King, A. J., & Schnupp, J. W. (2009). Interdependent encoding of pitch, timbre, and spatial location in auditory cortex. *Journal of Neuroscience*, *29*(7), 2064–2075.

Boemio, A., Fromm, S., Braun, A., & Poeppel, D. (2005). Hierarchical and asymmetric temporal

sensitivity in human auditory cortices. *Nature Neuroscience*, *8*, 389–395.

Brandewie, E., & Zahorik, P. (2010). Prior listening in rooms improves speech intelligibility. *Journal of the Acoustical Society of America*, *128*, 291–299.

Bregman, A. S. (1990). Auditory scene analysis: The perceptual organization of sound. Cambridge, MA: MIT Press.

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

Bronkhorst, A. W. (2000). The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions. *Acta Acustica united with Acustica*, *86*, 117–128.

Bronkhorst, A. W., & Houtgast, T. (1999). Auditory distance perception in rooms. *Nature*, *397*, 517–520.

Brown, A. D., Stecker, G. C., & Tollin, D. J. (2015). The precedence effect in sound localization. *Journal of the Association for Research in Otolaryngology*, *16*, 1–28.

Brown, G. J., & Palomaki, K. J. (2006). Reverberation. In D. Wang & G. J. Brown (Eds.), *Computational auditory scene analysis: Principles, algorithms, and applications* (pp. 209–250). Hoboken, NJ: Wiley.

Buus, S., Florentine, M., & Poulsen, T. (1997). Temporal integration of loudness, loudness discrimination, and the form of the loudness function. *Journal of the Acoustical Society of America*, *101*, 669–680.

Cariani, P. A., & Delgutte, B. (1996). Neural correlates of the pitch of complex tones. I. Pitch and pitch salience. *Journal of Neurophysiology*, *76*, 1698–1716.

Carlson, N. L., Ming, V. L., & DeWeese, M. R. (2012). Sparse codes for speech predict spectrotemporal receptive fields in the inferior colliculus. *PLoS Computational Biology*, *8*(7), e1002594.

Carlyon, R. P. (1991). Discriminating between coherent and incoherent frequency modulation of complex tones. *Journal of the Acoustical Society of America*, *89*, 329–340.

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

Carlyon, R. P., Cusack, R., Foxton, J. M., & Robertson, I. H. (2001). Effects of attention and unilateral neglect on auditory stream segregation. *Journal of Experimental Psychology: Human Perception and Performance*, *27*(1), 115–127.

Carlyon, R. P., Micheyl, C., Deeks, J. M., & Moore, B. C. (2004). Auditory processing of real and illusory changes in frequency modulation (FM) phase. *Journal of the Acoustical Society of America*, *116*(6), 3629–3639.

Cavaco, S., & Lewicki, M. S. (2007). Statistical modeling of intrinsic structures in impact sounds. *Journal of the Acoustical Society of America*, *121*(6), 3558–3568.

Cedolin, L., & Delgutte, B. (2005). Pitch of complex tones: Rate-place and interspike interval representations in the auditory nerve. *Journal of Neurophysiology*, *94*, 347–362.

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and two ears. *Journal of the Acoustical Society of America*, *25*(5), 975–979.

Chi, T., Ru, P., & Shamma, S. A. (2005). Multiresolution spectrotemporal analysis of complex sounds. *Journal of the Acoustical Society of America*, *118*(2), 887–906.

Christianson, G. B., Sahani, M., & Linden, J. F. (2008). The consequences of response nonlinearities for interpretation of spectrotemporal receptive fields. *Journal of Neuroscience*, *28*(2), 446–455.

Ciocca, V., & Bregman, A. S. (1987). Perceived continuity of gliding and steady-state tones through interrupting noise. *Perception & Psychophysics*, *42*, 476–484.

Cohen, Y. E., Russ, B. E., Davis, S. J., Baker, A. E., Ackelson, A. L., & Nitecki, R. (2009). A functional role for the ventrolateral prefrontal cortex in non-spatial auditory cognition. *Proceedings of the National Academy of Sciences, USA*, *106*, 20045–20050.

Conway, A. R. A., Cowan, N., & Bunting, M. F. (2001). The cocktail party phenomenon

revisited: The importance of working memory capacity. *Psychonomic Bulletin & Review*, *8*, 331–335.

Culling, J. F., & Akeroyd, M. A. (2010). Spatial hearing. In C. J. Plack (Ed.), *The Oxford handbook of auditory science: Hearing* (pp. 123–144). New York, NY: Oxford University Press.

Culling, J. F., & Summerfield, Q. (1995). Perceptual separation of concurrent speech sounds: Absence of across-frequency grouping by common interaural delay. *Journal of the Acoustical Society of America*, *98*(2), 785–797.

Cusack, R., & Carlyon, R. P. (2004). Auditory perceptual organization inside and outside the laboratory. In J. G. Neuhoff (Ed.), *Ecological psychoacoustics* (pp. 15–48). San Diego, CA: Elsevier/Academic Press.

Cutting, J. E. (1975). Aspects of phonological fusion. *Journal of Experimental Psychology: Human Perception and Performance*, *104*: 105–120.

Dallos, P. (2008). Cochlear amplification, outer hair cells and prestin. *Current Opinion in Neurobiology*, *18*, 370–376.

Darwin, C. (1984). Perceiving vowels in the presence of another sound: Constraints on formant perception. *Journal of the Acoustical Society of America*, *76*(6), 1636–1647.

Darwin, C. J. (1981). Perceptual grouping of speech components differing in fundamental frequency and onset-time. *Quarterly Journal of Experimental Psychology Section A*, *33*(2), 185–207.

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

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

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

Darwin, C. J., & Hukin, R. W. (1999). Auditory objects of attention: The role of interaural time differences. *Journal of Experimental Psychology: Human Perception and Performance*, *25*(3), 617–629.

Dau, T., Kollmeier, B., & Kohlrausch, A. (1997). Modeling auditory processing of amplitude modulation. I. Detection and masking with narrow-band carriers. *Journal of the Acoustical Society of America*, *102*(5), 2892–2905.

David, S. V., Mesgarani, N., Fritz, J. B., & Shamma, S. A. (2009). Rapid synaptic depression explains nonlinear modulation of spectro-temporal tuning in primary auditory cortex by natural stimuli. *Journal of Neuroscience*, *29*(11), 3374–3386.

de Cheveigne, A. (2004). Pitch perception models. In C. J. Plack & A. J. Oxenham (Eds.), *Pitch* (pp. 169–233). New York, NY: Springer Verlag.

de Cheveigne, A. (2006). Multiple F0 estimation. In D. Wang & G. J. Brown (Eds.), *Computational auditory scene analysis: Principles, algorithms, and applications* (pp. 45–80). Hoboken, NJ: Wiley.

de Cheveigne, A. (2010). Pitch perception. In C. J. Plack (Ed.), *The Oxford handbook of auditory science: Hearing*. New York, NY: Oxford University Press.

de Cheveigne, A., & Kawahara, H. (2002). YIN, a fundamental frequency estimator for speech and music. *Journal of the Acoustical Society of America*, *111*, 1917–1930.

de Cheveigne, A., McAdams, S., Laroche, J., & Rosenberg, M. (1995). Identification of concurrent harmonic and inharmonic vowels: A test of the theory of harmonic cancellation and enhancement. *Journal of the Acoustical Society of America*, *97*(6), 3736–3748.

Delgutte, B., Joris, P. X., Litovsky, R. Y., & Yin, T. C. T. (1999). Receptive fields and binaural interactions for virtual-space stimuli in the cat inferior colliculus. *Journal of Neurophysiology*, *81*, 2833–2851.

Demany, L., & Semal, C. (1990). The upper limit of "musical" pitch. *Music Perception*, *8*, 165–176.

Demany, L., Trost, W., Serman, M., & Semal, C. (2008). Auditory change detection: Simple sounds are not memorized better than complex sounds. *Psychological Science*, *19*, 85–91.

Depireux, D. A., Simon, J. Z., Klein, D. J., & Shamma, S. A. (2001). Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex. *Journal of Neurophysiology*, *85*(3), 1220–1234.

DiCarlo, J. J., & Cox, D. D. (2007). Untangling invariant object recognition. *Trends in Cognitive Sciences*, *11*, 333–341.

Durlach, N. I., Mason, C. R., Kidd, G., Arbogast, T. L., Colburn, H. S., & Shinn-Cunningham, B. G. (2003). Note on informational masking. *Journal of the Acoustical Society of America*, *113*(6), 2984–2987.

Elgoyhen, A. B., & Fuchs, P. A. (2010). Efferent innervation and function. In P. A. Fuchs (Ed.), *The Oxford handbook of auditory science: The ear* (pp. 283–306). New York, NY: Oxford University Press.

Eliades, S. J., & Wang, X. (2008). Neural substrates of vocalization feedback monitoring in primate auditory cortex. *Nature*, *453*, 1102–1106.

Escabi, M. A., Miller, L. M., Read, H. L., & Schreiner, C. E. (2003). Naturalistic auditory contrast improves spectrotemporal coding in the cat inferior colliculus. *Journal of Neuroscience*, *23*, 11489–11504.

Field, D. J. (1987). Relations between the statistics of natural images and the response profiles of cortical cells. *Journal of the Optical Society of America A*, *4*(12), 2379–2394.

Fishman, Y. I., Reser, D. H., Arezzo, J. C., & Steinschneider, M. (1998). Pitch vs. spectral encoding of harmonic complex tones in primary auditory cortex of the awake monkey. *Brain Research*, *786*, 18–30.

Formisano, E., Kim, D., Di Salle, F., van de Moortele, P., Ugurbil, K., & Goebel, R. (2003). Mirror-symmetric tonotopic maps in human primary auditory cortex. *Neuron*, *40*(4), 859–869.

Freyman, R. L., Clifton, R. K., & Litovsky, R. Y. (1991). Dynamic processes in the precedence effect. *Journal of the Acoustical Society of America*, *90*, 874–884.

Gardner, W. G. (1998). Reverberation algorithms. In M. Kahrs & K. Brandenburg (Eds.), *Applications of digital signal processing to audio and acoustics*. Norwell, MA: Kluwer Academic.

Gaver, W. W. (1993). What in the world do we hear? An ecological approach to auditory source perception. *Ecological Psychology*, *5*(1), 1–29.

Ghazanfar, A. A. (2009). The multisensory roles for auditory cortex in primate vocal communication. *Hearing Research*, *258*, 113–120.

Ghitza, O. (2001). On the upper cutoff frequency of the auditory critical-band envelope detectors in the context of speech perception. *Journal of the Acoustical Society of America*, *110*(3), 1628–1640.

Giraud, A., Lorenzi, C., Ashburner, J., Wable, J., Johnsrude, I. S., Frackowiak, R., & Kleinschmidt, A. (2000). Representation of the temporal envelope of sounds in the human brain. *Journal of Neurophysiology*, *84*(3), 1588–1598.

Glasberg, B. R., & Moore, B. C. J. (1990). Derivation of auditory filter shapes from notched-noise data. *Hearing Research*, *47*, 103–138.

Goldstein, J. L. (1973). An optimum processor theory for the central formation of the pitch of complex tones. *Journal of the Acoustical Society of America*, *54*, 1496–1516.

Grothe, B., Pecka, M., & McAlpine, D. (2010). Mechanisms of sound localization in mammals. *Physiological Review*, *90*, 983–1012.

Guinan, J. J. (2006). Olivocochlear efferents: Anatomy, physiology, function, and the measurement of efferent effects in humans. *Ear and Hearing*, *27*(6), 589–607.

Gygi, B., Kidd, G. R., & Watson, C. S. (2004). Spectral-temporal factors in the identification of environmental sounds. *Journal of the Acoustical Society of America*, *115*(3), 1252–1265.

Hall, D. A., & Plack, C. J. (2009). Pitch processing sites in the human auditory brain. *Cerebral Cortex*, *19*(3), 576–585.

Hall, J. W., Haggard, M. P., & Fernandes, M. A. (1984). Detection in noise by spectro-temporal pattern analysis. *Journal of the Acoustical Society of America*, *76*, 50–56.

Harper, N. S., & McAlpine, D. (2004). Optimal neural population coding of an auditory spatial cue. *Nature*, *430*, 682–686.

Hartmann, W. M., McAdams, S., & Smith, B. K. (1990). Hearing a mistuned harmonic in an otherwise periodic complex tone. *Journal of the Acoustical Society of America*, *88*, 1712–1724.

Hawley, M. L., Litovsky, R. Y., & Culling, J. F. (2004). The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer. *Journal of the Acoustical Society of America*, *115*(2), 833–843.

Heeger, D. J., & Bergen, J. (1995). Pyramid-based texture analysis/synthesis. In S. G. Mair & R. Cook (Eds.), *SIGGRAPH 95 Visual Proceedings: 22nd International ACM Conference on Computer Graphics and Interactive Techniques* (*Computer Graphics: Annual Conference Series*) (pp. 229–238). New York, NY: ACM.

Heffner, H. E., & Heffner, R. S. (1990). Effect of bilateral auditory cortex lesions on sound localization in Japanese macaques. *Journal of Neurophysiology*, *64*(3), 915–931.

Heinz, M. G., Colburn, H. S., & Carney, L. H. (2001). Evaluating auditory performance limits: I. One-parameter discrimination using a computational model for the auditory nerve. *Neural Computation*, *13*, 2273–2316.

Herdener, M., Esposito, F., Scheffler, K., Schneider, P., Logothetis, N. K., Uludag, K., & Kayser, C. (2013). Spatial representations of temporal and spectral sound cues in human auditory cortex. *Cortex*, *49*(10), 2822–2833.

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience*, *8*(5), 393–402.

Higgins, N. C., Storace, D. A., Escabi, M. A., & Read, H. L. (2010). Specialization of binaural responses in ventral auditory cortices. *Journal of Neuroscience*, *30*(43), 14522–14532.

Hofman, P. M., Van Riswick, J. G. A., & van Opstal, A. J. (1998). Relearning sound localization with new ears. *Nature Neuroscience*, *1*(5), 417–421.

Holt, L. L., & Lotto, A. J. (2010). Speech perception as categorization. *Attention, Perception, and Psychophysics*, *72*(5), 1218–1227.

Hood, L. J., Berlin, C. I., & Allen, P. (1994). Cortical deafness: A longitudinal study. *Journal of the American Academy of Audiology*, *5*, 330–342.

Houtgast, T. (1989). Frequency selectivity in amplitude-modulation detection. *Journal of the Acoustical Society of America*, *85*, 1676–1680.

Houtsma, A. J. M., & Smurzynski, J. (1990). Pitch identification and discrimination for complex tones with many harmonics. *Journal of the Acoustical Society of America*, *87*(1), 304–310.

Hsu, A., Woolley, S. M., Fremouw, T. E., & Theunissen, F. E. (2004). Modulation power and phase spectrum of natural sounds enhance neural encoding performed by single auditory neurons. *Journal of Neuroscience*, *24*, 9201–9211.

Hudspeth, A. J. (2008). Making an effort to listen: Mechanical amplification in the ear. *Neuron*, *59*(4), 530–545.

Humphries, C., Liebenthal, E., & Binder, J. R. (2010). Tonotopic organization of human auditory cortex. *NeuroImage*, *50*(3), 1202–1211.

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

Javel, E., & Mott, J. B. (1988). Physiological and psychophysical correlates of temporal processes in hearing. *Hearing Research*, *34*, 275–294.

Jenkins, W. M., & Masterton, R. B. (1982). Sound localization: Effects of unilateral lesions in central auditory system. *Journal of Neurophysiology*, *47*, 987–1016.

Johnson, D. H. (1980). The relationship between spike rate and synchrony in responses of auditory-nerve fibers to single tones. *Journal of the Acoustical Society of America*, *68*, 1115–1122.

Johnsrude, I. S., Penhune, V. B., & Zatorre, R. J. (2000). Functional specificity in the right human auditory cortex for perceiving pitch direction. *Brain*, *123*(1), 155–163.

Joris, P. X., Bergevin, C., Kalluri, R., McLaughlin, M., Michelet, P., van der Heijden, M., & Shera, C. A. (2011). Frequency selectivity in Old-World monkeys corroborates sharp cochlear tuning in humans. *Proceedings of the National Academy of Sciences, USA*, *108*(42), 17516–17520.

Joris, P. X., Schreiner, C. E., & Rees, A. (2004). Neural processing of amplitude-modulated sounds. *Physiological Review*, *84*, 541–577.

Kaas, J. H., & Hackett, T. A. (2000). Subdivisions of auditory cortex and processing streams in

primates. *Proceedings of the National Academy of Sciences, USA*, *97*, 11793–11799.

Kanwisher, N. (2010). Functional specificity in the human brain: A window into the functional architecture of the mind. *Proceedings of the National Academy of Sciences, USA*, *107*, 11163–11170.

Kawase, T., Delgutte, B., & Liberman, M. C. (1993). Anti-masking effects of the olivocochlear reflex, II: Enhancement of auditory-nerve response to masked tones. *Journal of Neurophysiology*, *70*, 2533–2549.

Kayser, C., Petkov, C. I., & Logothetis, N. K. (2008). Visual modulation of neurons in auditory cortex. *Cerebral Cortex*, *18*(7), 1560–1574.

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

Kidd, G., Mason, C. R., Deliwala, P. S., & Woods, W. S. (1994). Reducing informational masking by sound segregation. *Journal of the Acoustical Society of America*, *95*(6), 3475–3480.

Kidd, G., Mason, C. R., & Richards, V. M. (2003). Multiple bursts, multiple looks, and stream coherence in the release from informational masking. *Journal of the Acoustical Society of America 114*(5), 2835–2845.

Kikuchi, Y., Horwitz, B., & Mishkin, M. (2010). Hierarchical auditory processing directed rostrally along the monkey's supratemporal plane. *Journal of Neuroscience*, *30*(39), 13021–13030.

Kluender, K. R., & R. L. Jenison (1992). Effects of glide slope, noise intensity, and noise duration on the extrapolation of FM glides through noise. *Perception & Psychophysics*, *51*, 231–238.

Klump, R. G., & Eady, H. R. (1956). Some measurements of interural time difference thresholds. *Journal of the Acoustical Society of America*, *28*, 859–860.

Kulkarni, A., & Colburn, H. S. (1998). Role of spectral detail in sound-source localization. *Nature*, *396*, 747–749.

Langner, G., Sams, M., Heil, P., & Schulze, H. (1997). Frequency and periodicity are represented in orthogonal maps in the human auditory cortex: Evidence from magnetoencephalography. *Journal of Comparative Physiology*, *181*, 665–676.

Lerner, Y., Honey, C. J., Silbert, L. J., & Hasson, U. (2011). Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. *Journal of Neuroscience*, *31*(8), 2906–2915.

Lewicki, M. S. (2002). Efficient coding of natural sounds. *Nature Neuroscience*, *5*(4), 356–363.

Liberman, M. C. (1982). The cochlear frequency map for the cat: labeling auditory-nerve fibers of known characteristic frequency. *Journal of the Acoustical Society of America*, *72*, 1441–1449.

Liebenthal, E., Binder, J. R., Spitzer, S. M., Possing, E. T., & Medler, D. A. (2005). Neural substrates of phonemic perception. *Cerebral Cortex*, *15*(10), 1621–1631.

Litovsky, R. Y., Colburn, H. S., Yost, W. A., & Guzman, S. J. (1999). The precedence effect. *Journal of the Acoustical Society of America*, *106*, 1633–1654.

Lomber, S. G., & Malhotra, S. (2008). Double dissociation of "what" and "where" processing in auditory cortex. *Nature Neuroscience*, *11*(5), 609–616.

Lorenzi, C., Gilbert, G., Carn, H., Garnier, S., & Moore, B. C. J. (2006). Speech perception problems of the hearing impaired reflect inability to use temporal fine structure. *Proceedings of the National Academy of Sciences, USA*, *103*, 18866–18869.

Lutfi, R. A. (1992). Informational processing of complex sounds. III. Interference. *Journal of the Acoustical Society of America*, *91*, 3391–3400.

Lutfi, R. A. (2008). Sound source identification. In W. A. Yost & A. N. Popper (Eds.), *Springer handbook of auditory research: Auditory perception of sound sources*. New York, NY: Springer-Verlag.

Lyzenga, J., Carlyon, R. P., & Moore, B. C. J. (2005). Dynamic aspects of the continuity illusion: Perception of level and of the depth, rate, and phase of modulation. *Hearing Research*, *210*, 30–41.

Machens, C. K., Wehr, M. S., & Zador, A. M. (2004). Linearity of cortical receptive fields

measured with natural sounds. *Journal of Neuroscience*, *24*, 1089–1100.

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*, 43–51.

Makous, J. C., & Middlebrooks, J. C. (1990). Two-dimensional sound localization by human listeners. *Journal of the Acoustical Society of America*, *87*, 2188–2200.

Mandel, M. I., Weiss, R. J., & Ellis, D. P. W. (2010). Model-based expectation maximization source separation and localization. *IEEE Transactions on Audio, Speech, and Language Processing*, *18*(2), 382–394.

Marr, D. C. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. New York, NY: Freeman.

Masutomi, K., Barascud, N., Kashino, M., McDermott, J. H., & Chait, M. (2016). Sound segregation via embedded repetition is robust to inattention. *Journal of Experimental Psychology: Human Perception and Performance*, *42*(3), 386–400.

May, B. J., Anderson, M., & Roos, M. (2008). The role of broadband inhibition in the rate representation of spectral cues for sound localization in the inferior colliculus. *Hearing Research*, *238*, 77–93.

May, B. J., & McQuone, S. J. (1995). Effects of bilateral olivocochlear lesions on pure-tone discrimination in cats. *Auditory Neuroscience*, *1*, 385–400.

McAdams, S. (1989). Segregation of concurrent sounds: I. Effects of frequency modulation coherence. *Journal of the Acoustical Society of America*, *86*, 2148–2159.

McAlpine, D. (2004). Neural sensitivity to periodicity in the inferior colliculus: Evidence for the role of cochlear distortions. *Journal of Neurophysiology*, *92*, 1295–1311.

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

McDermott, J. H. (2013). Audition. In K. Ochsner & S. Kosslyn (Eds.), *The Oxford handbook of cognitive neuroscience*. Oxford, United Kingdom: Oxford University Press.

McDermott, J. H., & Oxenham, A. J. (2008a). Music perception, pitch, and the auditory system. *Current Opinion in Neurobiology*, *18*, 452–463.

McDermott, J. H., & Oxenham, A. J. (2008b). Spectral completion of partially masked sounds. *Proceedings of the National Academy of Sciences, USA*, *105*(15), 5939–5944.

McDermott, J. H., Oxenham, A. J., & Simoncelli, E. P. (2009, October). Sound texture synthesis via filter statistics. In *2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (pp. 297–300), New Paltz, NY.

McDermott, J. H., Schemitsch, M., & Simoncelli, E. P. (2013). Summary statistics in auditory perception. *Nature Neuroscience*, *16*(4), 493–498.

McDermott, J. H., & Simoncelli, E. P. (2011). Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis. *Neuron*, *71*, 926–940.

McDermott, J. H., Wrobleski, D., & Oxenham, A. J. (2011). Recovering sound sources from embedded repetition. *Proceedings of the National Academy of Sciences, USA*, *108*(3), 1188–1193.

Meddis, R., & Hewitt, M. J. (1991). Virtual pitch and phase sensitivity of a computer model of the auditory periphery: Pitch identification. *Journal of the Acoustical Society of America*, *89*, 2866–2882.

Mershon, D. H., Desaulniers, D. H., Kiefer, S. A., Amerson, T. L. J., & Mills, J. T. (1981). Perceived loudness and visually–determined auditory distance. *Perception*, *10*, 531–543.

Mesgarani, N., Cheung, C., Johnson, K., & Chang, E. F. (2014). Phonetic feature encoding in human superior temporal gyrus. *Science*, *343*(6174), 1006–1010.

Mesgarani, N., David, S. V., Fritz, J. B., & Shamma, S. A. (2008). Phoneme representation and classification in primary auditory cortex. *Journal of the Acoustical Society of America*, *123*(2), 899–909.

Mesgarani, N., & Shamma, S. A. (2011, May). Speech processing with a cortical representation of audio. In *IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP)* (pp. 5872–5875). Prague, Czech Republic.

Micheyl, C., & Oxenham, A. J. (2010a). Objective and subjective psychophysical measures of auditory stream integration and segregation. *Journal of the Association for Research in Otolaryngology*, *11*(4), 709–724.

Micheyl, C., & Oxenham, A. J. (2010b). Pitch, harmonicity and concurrent sound segregation: Psychoacoustical and neurophysiological findings. *Hearing Research*, *266*, 36–51.

Middlebrooks, J. C. (1992). Narrow-band sound localization related to external ear acoustics. *Journal of the Acoustical Society of America*, *92*(5), 2607–2624.

Middlebrooks, J. C. (2000). Cortical representations of auditory space. In M. S. Gazzaniga (Ed.), *The new cognitive neurosciences* (2nd ed., pp. 425–436). Cambridge, MA: MIT Press.

Middlebrooks, J. C., & Green, D. M. (1991). Sound localization by human listeners. *Annual Review of Psychology*, *42*, 135–159.

Miller, L. M., Escabi, M. A., Read, H. L., & Schreiner, C. E. (2001). Spectrotemporal receptive fields in the lemniscal auditory thalamus and cortex. *Journal of Neurophysiology*, *87*, 516–527.

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

Miller, R. L., Schilling, J. R., Franck, K. R., & Young, E. D. (1997). Effects of acoustic trauma on the representation of the vowel /e/ in cat auditory nerve fibers. *Journal of the Acoustical Society of America*, *101*(6), 3602–3616.

Mlynarski, W. (2015). The opponent channel population code of sound location is an efficient representation of natural binaural sounds. *PLoS Computational Biology*, *11*(5), e1004294.

Mlynarski, W., & Jost, J. (2014). Statistics of natural binaural sounds. *PLOS ONE*, *9*(10), e108968.

Moore, B. C., & Glasberg, B. R. (1996). A revision of Zwicker's loudness model. *Acta Acustica united with Acustica*, *82*(2), 335–345.

Moore, B. C. J. (1973). Frequency differences limens for short-duration tones. *Journal of the Acoustical Society of America*, *54*, 610–619.

Moore, B. C. J. (2003). *An introduction to the psychology of hearing*. San Diego, CA: Academic Press.

Moore, B. C. J., Glasberg, B. R., & Peters, R. W. (1986). Thresholds for hearing mistuned partials as separate tones in harmonic complexes. *Journal of the Acoustical Society of America*, *80*, 479–483.

Moore, B. C. J., & Gockel, H. (2002). Factors influencing sequential stream segregation. *Acta Acustica united with Acustica*, *88*, 320–332.

Moore, B. C. J., & Oxenham, A. J. (1998). Psychoacoustic consequences of compression in the peripheral auditory system. *Psychological Review*, *105*(1), 108–124.

Morosan, P., Rademacher, J., Schleicher, A., Amunts, K., Schormann, T., & Zilles, K. (2001). Human primary auditory cortex: Cytoarchitectonic subdivisions and mapping into a spatial reference system. *NeuroImage*, *13*, 684–701.

Moshitch, D., Las, L., Ulanovsky, N., Bar Yosef, O., & Nelken, I. (2006). Responses of neurons in primary auditory cortex (A1) to pure tones in the halothane-anesthetized cat. *Journal of Neurophysiology*, *95*(6), 3756–3769.

Neff, D. L. (1995). Signal properties that reduce masking by simultaneous, random-frequency maskers. *Journal of the Acoustical Society of America*, *98*, 1909–1920.

Nelken, I., Bizley, J. K., Nodal, F. R., Ahmed, B., King, A. J., & Schnupp, J. W. (2008). Responses of auditory cortex to complex stimuli: Functional organization revealed using intrinsic optical signals. *Journal of Neurophysiology*, *99*(4), 1928–1941.

Norman-Haignere, S., Kanwisher, N., & McDermott, J. H. (2013). Cortical pitch regions in humans respond primarily to resolved harmonics and are located in specific tonotopic regions of anterior auditory cortex. *Journal of Neuroscience*, *33*(50), 19451–19469.

Norman-Haignere, S., Kanwisher, N., & McDermott, J. H. (2015). Distinct cortical pathways for music and speech revealed by hypothesis-free voxel decomposition. *Neuron*, *88*, 1281–1296.

Norman-Haignere, S., & McDermott, J. H. (2016). Distortion products in auditory fMRI research: Measurements and solutions. *NeuroImage*, *129*, 401–413.

Obleser, J., Zimmermann, J., Van Meter, J., & Rauschecker, J. P. (2007). Multiple stages of auditory speech perception reflected in event-related FMRI. *Cerebral Cortex*, *17*(10), 2251–2257.

Olshausen, B. A., & Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, *381*, 607–609.

Overath, T., McDermott, J. H., Zarate, J. M., & Poeppel, D. (2015). The cortical analysis of speech-specific temporal structure revealed by responses to sound quilts. *Nature Neuroscience*, *18*, 903–911.

Palmer, A. R., & Russell, I. J. (1986). Phase-locking in the cochlear nerve of the guinea-pig and its relation to the receptor potential of inner hair-cells. *Hearing Research*, *24*, 1–15.

Patterson, R. D., Uppenkamp, S., Johnsrude, I. S., & Griffiths, T. D. (2002). The processing of temporal pitch and melody information in auditory cortex. *Neuron*, *36*(4), 767–776.

Penagos, H., Melcher, J. R., & Oxenham, A. J. (2004). A neural representation of pitch salience in nonprimary human auditory cortex revealed with functional magnetic resonance imaging. *Journal of Neuroscience*, *24*(30), 6810–6815.

Petkov, C. I., C. Kayser, M. Augath & N. K. Logothetis (2006). Functional imaging reveals numerous fields in the monkey auditory cortex. *PLoS Biology 4*(7), 1213–1226.

Petkov, C. I., Kayser, C., Steudel, T., Whittingstall, K., Augath, M., & Logothetis, N. K. (2008). A voice region in the monkey brain. *Nature Neuroscience*, *11*, 367–374.

Petkov, C. I., O'Connor, K. N., & Sutter, M. L. (2007). Encoding of illusory continuity in primary auditory cortex. *Neuron*, *54*, 153–165.

Plack, C. J. (2005). *The sense of hearing*. Mahwah, NJ: Lawrence Erlbaum.

Plack, C. J., & Oxenham, A. J. (2005). The psychophysics of pitch. In C. J. Plack, A. J. Oxenham, R. R. Fay, & A. J. Popper (Eds.), *Pitch: Neural coding and perception* (pp. 7–25). New York, NY: Springer Verlag.

Plack, C. J., Oxenham, A. J., Popper, A. J., & Fay, R. R. (Eds.) (2005). *Pitch: Neural coding and perception*. New York, NY: Springer.

Poeppel, D. (2003). The analysis of speech in different temporal integration windows: cerebral lateralization as "asymmetric sampling in time." *Speech Communication*, *41*, 245–255.

Poremba, A., Saunders, R. C., Crane, A. M., Cook, M., Sokoloff, L., & Mishkin, M. (2003). Functional mapping of the primate auditory system. *Science*, *299*, 568–572.

Portilla, J., & Simoncelli, E. P. (2000). A parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision*, *40*(1), 49–71.

Rajan, R. (2000). Centrifugal pathways protect hearing sensitivity at the cochlea in noisy environments that exacerbate the damage induced by loud sound. *Journal of Neuroscience*, *20*, 6684–6693.

Rauschecker, J. P., & Tian, B. (2004). Processing of band-passed noise in the lateral auditory belt cortex of the rhesus monkey. *Journal of Neurophysiology*, *91*, 2578–2589.

Rayleigh, L. (1907). On our perception of sound direction. *Philosophical Magazine*, *3*, 456–464.

Recanzone, G. H. (2008). Representation of conspecific vocalizations in the core and belt areas of the auditory cortex in the alert macaque monkey. *Journal of Neuroscience*, *28*(49), 13184–13193.

Rhode, W. S. (1971). Observations of the vibration of the basilar membrane in squirrel monkeys using the Mossbauer technique. *Journal of the Acoustical Society of America*, *49*, 1218–1231.

Rhode, W. S. (1978). Some observations on cochlear mechanics. *Journal of the Acoustical Society of America*, *64*, 158–176.

Riecke, L., van Opstal, J., Goebel, R., & Formisano, E. (2007). Hearing illusory sounds

in noise: Sensory-perceptual transformations in primary auditory cortex. *Journal of Neuroscience*, *27*(46), 12684–12689.

Roberts, B., & Brunstrom, J. M. (1998). Perceptual segregation and pitch shifts of mistuned components in harmonic complexes and in regular inharmonic complexes. *Journal of the Acoustical Society of America*, *104*(4), 2326–2338.

Rodriguez, F. A., Chen, C., Read, H. L., & Escabi, M. A. (2010). Neural modulation tuning characteristics scale to efficiently encode natural sound statistics. *Journal of Neuroscience*, *30*, 15969–15980.

Romanski, L. M., B. Tian, J. B. Fritz, M. Mishkin, P. S. Goldman-Rakic & J. P. Rauschecker (1999). Dual streams of auditory afferents target multiple domains in the primate prefrontal cortex. *Nature Neuroscience*, *2*(12), 1131–1136.

Rose, J. E., Brugge, J. F., Anderson, D. J., & Hind, J. E. (1967). Phase-locked response to low-frequency tones in single auditory nerve fibers of the squirrel monkey. *Journal of Neurophysiology*, *30*, 769–793.

Rosen, S. (1992). Temporal information in speech: Acoustic, auditory and linguistic aspects. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *336*, 367–373.

Rosenblum, L. D. (2004). Perceiving articulatory events: Lessons for an ecological psychoacoustics. In J. G. Neuhoff (Ed.), *Ecological Psychoacoustics* (pp. 219–248). San Diego, CA: Elsevier Academic Press.

Rothschild, G., Nelken, I., & Mizrahi, A. (2010). Functional organization and population dynamics in the mouse primary auditory cortex. *Nature Neuroscience*, *13*(3), 353–360.

Rotman, Y., Bar Yosef, O., & Nelken, I. (2001). Relating cluster and population responses to natural sounds and tonal stimuli in cat primary auditory cortex. *Hearing Research*, *152*, 110–127.

Ruggero, M. A. (1992). Responses to sound of the basilar membrane of the mammalian cochlea. *Current Opinion in Neurobiology*, *2*, 449–456.

Ruggero, M. A., & Rich, N. C. (1991). Furosemide alters organ of Corti mechanics: Evidence for feedback of outer hair cells upon the basilar membrane. *Journal of Neuroscience*, *11*, 1057–1067.

Ruggero, M. A., Rich, N. C., Recio, A., & Narayan, S. S. (1997). Basilar-membrane responses to tones at the base of the chinchilla cochlea. *Journal of the Acoustical Society of America*, *101*, 2151–2163.

Saint-Arnaud, N., & Popat, K. (1995). Analysis and synthesis of sound texture. In *AJCAI Workshop on Computational Auditory Scene Analysis* (pp. 293–308). Montreal, Canada.

Samson, F., Zeffiro, T. A., Toussaint, A., & Belin, P. (2011). Stimulus complexity and categorical effects in human auditory cortex: An Activation Likelihood Estimation meta-analysis. *Frontiers in Psychology*, *1*, 1–23.

Scharf, B., Magnan, J., & Chays, A. (1997). On the role of the olivocochlear bundle in hearing: 16 case studies. *Hearing Research*, *103*, 101–122.

Schonwiesner, M., & Zatorre, R. J. (2008). Depth electrode recordings show double dissociation between pitch processing in lateral Heschl's gyrus. *Experimental Brain, Research 187*(97–105).

Schonwiesner, M., & Zatorre, R. J. (2009). Spectro-temporal modulation transfer function of single voxels in the human auditory cortex measured with high-resolution fMRI. *Proceedings of the National Academy of Sciences, USA*, *106*(34), 14611–14616.

Schreiner, C. E., & Urbas, J. V. (1986). Representation of amplitude modulation in the auditory cortex of the cat. I. Anterior auditory field. *Hearing Research*, *21*, 227–241.

Schreiner, C. E., & Urbas, J. V. (1988). Representation of amplitude modulation in the auditory cortex of the cat. II. Comparison between cortical fields. *Hearing Research*, *32*, 49–64.

Scott, S. K., Blank, C. C., Rosen, S., & Wise, R. J. S. (2000). Identification of a pathway for intelligible speech in the left temporal lobe. *Brain*, *123*(12), 2400–2406.

Shackleton, T. M., & Carlyon, R. P. (1994). The role of resolved and unresolved harmonics in pitch perception and frequency modulation discrimination. *Journal of the Acoustical Society of America*, *95*(6), 3529–3540.

Shamma, S. A., & Klein, D. (2000). The case of the missing pitch templates: How harmonic templates emerge in the early auditory system. *Journal of the Acoustical Society of America*, *107*, 2631–2644.

Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science*, *270*(5234), 303–304.

Shera, C. A., Guinan, J. J., & Oxenham, A. J. (2002). Revised estimates of human cochlear tuning from otoacoustic and behavioral measurements. *Proceedings of the National Academy of Sciences, USA*, *99*(5), 3318–3323.

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

Singh, N. C., & Theunissen, F. E. (2003). Modulation spectra of natural sounds and ethological theories of auditory processing. *Journal of the Acoustical Society of America*, *114*(6), 3394–3411.

Smith, E. C., & Lewicki, M. S. (2006). Efficient auditory coding. *Nature*, *439*, 978–982.

Smith, Z. M., Delgutte, B., & Oxenham, A. J. (2002). Chimaeric sounds reveal dichotomies in auditory perception. *Nature*, *416*, 87–90.

Stevens, S. S. (1955). The measurement of loudness. *Journal of the Acoustical Society of America*, *27*(5), 815–829.

Stilp, C. E., Alexander, J. M., Kiefte, M., & Kluender, K. R. (2010). Auditory color constancy: Calibration to reliable spectral properties across nonspeech context and targets. *Attention, Perception, and Psychophysics*, *72*(2), 470–480.

Sumner, C. J., & Palmer, A. R. (2012). Auditory nerve fibre responses in the ferret. *European Journal of Neuroscience*, *36*, 2428–2439.

Sweet, R. A., Dorph-Petersen, K., & Lewis, D. A. (2005). Mapping auditory core, lateral belt, and parabelt cortices in the human superior temporal gyrus. *Journal of Comparative Neurology*, *491*, 270–289.

Talavage, T. M., Sereno, M. I., Melcher, J. R., Ledden, P. J., Rosen, B. R., & Dale, A. M. (2004). Tonotopic organization in human auditory cortex revealed by progressions of frequency sensitivity. *Journal of Neurophysiology*, *91*, 1282–1296.

Tansley, B. W., & Suffield, J. B. (1983). Time-course of adaptation and recovery of channels selectively sensitive to frequency and amplitude modulation. *Journal of the Acoustical Society of America*, *74*, 765–775.

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

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

Theunissen, F. E., David, S. V., Singh, N. C., Hsu, A., Vinje, W. E., & Gallant, J. L. (2001). Estimating spatio-temporal receptive fields of auditory and visual neurons from their responses to natural stimuli. *Network*, *12*(3), 289–316.

Theunissen, F. E., Sen, K., & Doupe, A. J. (2000). Spectral-temporal receptive fields of non-linear auditory neurons obtained using natural sounds. *Journal of Neuroscience*, *20*, 2315–2331.

Tian, B., & J. P. Rauschecker (2004). Processing of frequency-modulated sounds in the lateral auditory belt cortex of the rhesus monkey. *Journal of Neurophysiology*, *92*, 2993–3013.

Tian, B., Reser, D., Durham, A., Kustov, A., & Rauschecker, J. P. (2001). Functional specialization in rhesus monkey auditory cortex. *Science 292*, 290–293.

Traer, J., & McDermott, J. H. (2016). Statistics of natural reverberation enable perceptual separation of sound and space. *Proceedings of the National Academy of Sciences, USA*, *113*, E7856–E7865.

van Noorden, L. P. A. S. (1975). *Temporal coherence in the perception of tone sequences* (Doctoral dissertation). Eindhoven University of Technology, The Netherlands.

Walker, K. M. M., Bizley, J. K., King, A. J., & Schnupp, J. W. (2011). Cortical encoding of pitch: Recent results and open questions. *Hearing Research*, *271*(1–2), 74–87.

Wallace, M. N., Anderson, L. A., & Palmer, A. R. (2007). Phase-locked responses to pure tones in the auditory thalamus. *Journal of Neurophysiology*, *98*(4), 1941–1952.

Wallach, H., Newman, E. B., & Rosenzweig, M. R. (1949). The precedence effect in sound localization. *American Journal of Psychology*, *42*, 315–336.

Warren, J. D., Zielinski, B. A., Green, G. G. R., Rauschecker, J. P., & Griffiths, T. D. (2002). Perception of sound-source motion by the human brain. *Neuron*, *34*, 139–148.

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

Warren, R. M., Obusek, C. J., & Ackroff, J. M. (1972). Auditory induction: Perceptual synthesis of absent sounds. *Science*, *176*, 1149–1151.

Watkins, A. J. (2005). Perceptual compensation for effects of reverberation in speech identification. *Journal of the Acoustical Society of America*, *118*, 249–262.

Watson, C. S. (1987). Uncertainty, informational masking and the capacity of immediate auditory memory. In W. A. Yost & C. S. Watson (Eds.), *Auditory processing of complex sounds* (pp. 267–277). Hillsdale, NJ: Erlbaum.

Wightman, F. (1973). The pattern-transformation model of pitch. *Journal of the Acoustical Society of America*, *54*, 407–416.

Wightman, F., & Kistler, D. J. (1989). Headphone simulation of free-field listening. II: Psychophysical validation. *Journal of the Acoustical Society of America*, *85*(2), 868–878.

Willmore, B. D. B., & Smyth, D. (2003). Methods for first-order kernel estimation: simple-cell receptive fields from responses to natural scenes. *Network*, *14*, 553–577.

Winslow, R. L., & Sachs, M. B. (1987). Effect of electrical stimulation of the crossed olivocochlear bundle on auditory nerve response to tones in noise. *Journal of Neurophysiology*, *57*(4), 1002–1021.

Winter, I. M. (2005). The neurophysiology of pitch. In C. J. Plack, A. J. Oxenham, R. R. Fay & A. J. Popper (Eds.), *Pitch: Neural coding and perception*. New York, NY: Springer Verlag.

Woods, K. J. P., & McDermott, J. H. (2015). Attentive tracking of sound sources. *Current Biology*, *25*, 2238–2246.

Woods, T. M., Lopez, S. E., Long, J. H., Rahman, J. E., & Recanzone, G. H. (2006). Effects of stimulus azimuth and intensity on the single-neuron activity in the auditory cortex of the alert macaque monkey. *Journal of Neurophysiology*, *96*(6), 3323–3337.

Woolley, S. M., Fremouw, T. E., Hsu, A., & Theunissen, F. E. (2005). Tuning for spectro-temporal modulations as a mechanism for auditory discrimination of natural sounds. *Nature Neuroscience*, *8*(10), 1371–1379.

Yates, G. K. (1990). Basilar membrane nonlinearity and its influence on auditory nerve rate-intensity functions. *Hearing Research*, *50*, 145–162.

Yin, T. C. T., & Kuwada, S. (2010). Binaural localization cues. In A. Rees & A. R. Palmer (Eds.), *The Oxford handbook of auditory science: The auditory brain* (pp. 271–302). New York, NY: Oxford University Press.

Young, E. D. (2010). Level and spctrum. In A. Rees & A. R. Palmer (Eds.), *The Oxford handbook of auditory science: The auditory brain* (pp. 93–124). New York, NY: Oxford University Press.

Zahorik, P., Bangayan, P., Sundareswaran, V., Wang, K., & Tam, C. (2006). Perceptual recalibration in human sound localization: Learning to remediate front-back reversals. *Journal of the Acoustical Society of America*, *120*(1), 343–359.

Zahorik, P., & Wightman, F. L. (2001). Loudness constancy with varying sound source distance. *Nature Neuroscience*, *4*(1), 78–83.

Zatorre, R. J. (1985). Discrimination and recognition of tonal melodies after unilateral cerebral excisions. *Neuropsychologia 23*(1), 31–41.

Zatorre, R. J., & Belin, P. (2001). Spectral and temporal processing in human auditory cortex. *Cerebral Cortex*, *11*, 946–953.

Zatorre, R. J., Belin, P., & Penhune, V. B. (2002). Structure and function of auditory cortex: Music and speech. *Trends in Cognitive Sciences*, *6*(1), 37–46.