Chapter 1
Ear and Brain Mechanisms for Parsing the Auditory Scene

John C. Middlebrooks and Jonathan Z. Simon

Abstract The cocktail party is a popular metaphor for the complex auditory scene that is everyday life. In busy offices, crowded restaurants, and noisy streets, a listener is challenged to hear out signals of interest—most often speech from a particular talker—amid a cacophony of competing talkers, broadband machine noise, room reflections, and so forth. This chapter defines the problems that the auditory system must solve and introduces the ensuing chapters, which explore the relevant perception and physiology at all levels: in normal mature hearing, in early development, in aging, and in pathology.

Keywords Auditory object · Auditory scene analysis · Cocktail party problem · Energetic masking · Grouping · Informational masking · Stream segregation · Streaming

J.C. Middlebrooks (✉)
Department of Otolaryngology, Department of Neurobiology & Behavior, Department of Cognitive Sciences, Department of Biomedical Engineering, Center for Hearing Research, University of California, Irvine, CA 92697-5310, USA
e-mail: j.midd@uci.edu

J.Z. Simon
Department of Electrical & Computer Engineering, Department of Biology, Institute for Systems Research, University of Maryland, College Park, MD 20742, USA
e-mail: jzsimon@umd.edu

J.C. Middlebrooks et al. (eds.), *The Auditory System at the Cocktail Party*, Springer Handbook of Auditory Research 60,
DOI 10.1007/978-3-319-51662-2_1

1.1 Introduction

The cocktail party is the archetype of a complex auditory scene: multiple voices vie for attention; glasses clink; background music plays; all of which are shaken, not stirred, by room reflections. Colin Cherry (1953) brought hearing science to the cocktail party when he introduced the term "cocktail party problem." Cherry's cocktail party was rather dry: just two talkers reading narratives at the same time, either with one talker in each of earphones or with the two talkers mixed and played to both earphones. Real-life cocktail parties are far more acoustically complex, as are other auditory situations of daily life, such as busy offices, crowded restaurants, noisy classrooms, and congested city streets. Albert Bregman (1990) has referred to people's efforts to solve these everyday cocktail party problems as "auditory scene analysis."

The normal auditory system exhibits a remarkable ability to parse these complex scenes. As pointed out by Shinn-Cunningham, Best, and Lee (Chap. 2), the best efforts of present-day technology pale compared to the ability of even a toddler to hear out a special voice amid a crowd of distractors. Conversely, even a relatively minor hearing impairment can disrupt auditory scene analysis. People with mild to moderate hearing loss report that their inability to segregate multiple talkers or to understand speech in a noisy background is one of their greatest disabilities (Gatehouse and Nobel 2004).

1.2 Some Central Concepts

In attempting to make sense of the auditory scene, a listener must form distinct perceptual images—*auditory objects*—of one or more sound sources, where the sound sources might be individual talkers, musical lines, mechanical objects, and so forth. Formation of an auditory object requires *grouping* of the multiple sound components that belong to a particular source and *segregation* of those components from those of other sources. Grouping can happen instantaneously across frequencies, such as grouping of all the harmonics of a vowel sound or of all the sounds resulting from the release of a stop consonant. Grouping must also happen across time, such as in the formation of perceptual *streams* from the sequences of sounds from a particular source. In the cocktail party example, the relevant streams might be the sentences formed by the successions of phonemes originating from the various competing talkers. To a large degree, segregation of auditory objects takes place on the basis of low-level differences in sounds, such as fundamental frequencies, timbres, onset times, or source locations. Other, higher-level, factors for segregation include linguistic cues, accents, and recognition of familiar voices.

Failure to segregate the components of sound sources can impair formation of auditory objects: this is *masking*. When a competing sound coincides in frequency and time with a signal of interest, the resulting masking is referred to as *energetic*.

Energetic masking is largely a phenomenon of the auditory periphery, where signal and masker elicit overlapping patterns of activity on the basilar membrane of the cochlea and compete for overlapping auditory nerve populations. There is an extensive literature on the characteristics of energetic masking and on brain mechanisms that can provide some *release from energetic masking*.

Another form of masking can occur in situations in which there is no spectrotemporal overlap of signal and masker: this is referred to as *informational masking*. In cases of informational masking, listeners fail to identify the signal amid the confusion of masking sounds. The magnitude of informational masking, tens of decibels in some cases, is surprising inasmuch as the spectral analysis by the cochlea presumably is doing its normal job of segregating activity from signal and masker components that differ in frequency. Given the presumed absence of interference in the cochlea, one assumes that informational masking somehow arises in the central auditory pathway. Chapters of this volume review the phenomena of informational masking and the possible central mechanisms for *release from informational masking*.

1.3 Overview of the Volume

The present volume addresses conditions in which the auditory system succeeds at segregating signals from distractors and conditions in which the cocktail party problem cannot be solved. Shinn-Cunningham, Best, and Lee (Chap. 2) set the stage by introducing the notion of the auditory object, which can be thought of as the perceptual correlate of an external auditory source and the unit on which target selection and attention operate. Sequences of auditory objects that are extended in time form auditory streams. Parsing of the auditory scene, then, consists of selection of particular auditory objects through some combination of bottom-up object salience and top-down attention, filtered by experience and expectation.

Culling and Stone (Chap. 3) address the challenges of low-level formation of auditory objects and consider some mechanisms by which those challenges can be overcome. They introduce the notion of energetic masking, in which interfering sounds disrupt the representation of speech signals at the level of the auditory nerve. Release from energetic masking can be achieved by exploiting differences between target and masker, such as differences in their harmonic structure or interaural time differences. In some conditions a listener can circumvent energetic masking by "listening in the dips," where "the dips" are moments at which masker amplitude is minimal. In addition, a listener might exploit the acoustic shadow of the head by attending to the ear at which the target-to-masker ratio is higher.

Understanding of a speech target can be impaired by the presence of a competing speech source even in the absence of energetic masking, that is, when there is no spectral or temporal overlap of target and masker. That residual *informational* masking is the topic of Chap. 4, by Kidd and Colburn. Focusing on speech-on-speech masking, the authors contrast and compare energetic and

informational masking, with historical views and with present-day understanding. The authors consider aspects of attention, memory, and linguistic processing that can support release from masking. Finally, they mine the extensive body of work on binaural mechanisms of energetic masking release as a resource for models of binaural solutions to the cocktail party problem.

Computational models can aid in formalizing the basic science understanding of a problem as well as in generating algorithms that exploit biological principles for use in solution of practical engineering problems. In Chap. 5, Elhilali considers the challenges of creating computational models of the cocktail party problem. These include the difficulty of even defining the theoretical foundations of the problem as well as the need to reconcile computational models with empirical data. The author samples the broad range of approaches that have been employed, from low-level biologically inspired to highly extracted engineering systems, including common automatic speech recognition systems that must perform their task well even when a user is not alone in a quiet room.

A cocktail party guest must segregate brief sounds (e.g., syllables) from multiple competing talkers, and then must piece together sequences of such sounds (e.g., sentences) into perceptual streams for interpretation. In Chap. 6, Middlebrooks considers the importance of spatial separation of sound sources for stream segregation. The psychophysics of spatial stream segregation is reviewed. Possible neural substrates, then, are evaluated at the level of single cortical neurons in animals and far-field recordings in humans. Available results suggest that the perception of two or more segregated streams might reflect the activity of a corresponding number of distinct populations of neurons.

New developments in the study of the neural mechanisms allowing the human brain to solve the cocktail party problem are reviewed by Simon in Chap. 7. The field of experimental human auditory neuroscience has shown some success in investigations of the foundations of auditory stream segregation, in general, and the neural processing of speech in the presence of maskers, in particular. Such investigations address the neural mechanisms by which acoustically faithful representations of an entire sound scene are somehow transformed into new stream-centric neural representations. It is these new representations that underlie the remarkably ordinary percept that the world is made of individual auditory objects that contribute separately and independently to the larger auditory scene.

A cocktail party is no place for infants and children. The auditory scenes encountered on a noisy playground or in a crowded classroom, however, are easily as acoustically complex. Young people apprehend these scenes with immature auditory systems and with not-yet-crystallized language recognition. Werner (Chap. 8) considers multiple stages and levels of development. These include early phases of central representations of sound during infancy; maturation of spatial hearing and auditory–visual correspondence during early childhood; improving ability to group components of complex sounds; and development of selective attention.

At the other end of the lifespan, in older adults, multiple factors can have opposing effects on auditory scene analysis (Pichora-Fuller, Alain, and Schneider,

Chap. 9). Some, but far from all, of the decline in performance can be blamed on age-related senescence of the auditory periphery. That decline is mitigated to a degree, however, by contextual factors that include an older adult's command of language and stores of knowledge. In the most highly demanding conditions, the auditory rigors of age plus greater likelihood of mild-to-moderate acquired hearing loss are likely to produce some social isolation of older people. That, in turn, places them at greater risk for development of cognitive impairment.

In the final chapter, Litovsky, Goupell, Misurelli, and Kan consider the consequences of hearing impairment and the repercussions of attempts to (at least partially) restore hearing. Motivated by the body of research on binaural and spatial cues for auditory scene analysis, present-day clinical practice strives to provide hearing in both ears. The chapter reviews the auditory cues that are available through hearing aids and/or cochlear implants and considers cues that are faithfully transmitted, cues that are degraded, and cues that are not transmitted at all by various forms of auditory prosthesis. Also considered are the consequences of patients' unique hearing histories.

1.4 Ears and Brains

The phenomena of hearing in complex auditory scenes highlight the notion that humans (and other animals) hear with their brains, not just with their ears. Humans clearly rely on the active mechanics of the cochlea to provide the initial analysis of sound spectra. Nevertheless, information from just one cochlea is not enough. Any normal-hearing listener can demonstrate this to him- or herself simply by plugging one ear at a cocktail party or other multitalker setting, thereby disrupting the critical binaural cues for scene analysis. The demonstration of a need for binaural input implicates the binaural nuclei of the brainstem. Central, presumably brainstem, mechanisms also are required for any analysis that spans a wide frequency range, such as analysis of multicomponent harmonic structure. The phenomenon of stream segregation occurs on a time scale of hundreds of milliseconds. That time scale points to an involvement of early auditory–cortical mechanisms; that implication is supported by animal studies that show stream segregation by single neurons in the primary auditory cortex. Involvement of even higher-level cortical areas is implied by the ability of listeners to perform stream segregation on the basis of linguistic cues; again, there are human neurophysiological results demonstrating extra-primary cortical substrates of such behavior.

The success of the auditory system in parsing the auditory scene is a marvel of auditory processing. Future investigation of this topic surely will provide new insights into the basic science of hearing. Psychophysical studies continue to define the relevant perceptual algorithms; animal models yield insights into peripheral and central mechanisms at the levels of single neurons and networks of neurons; and human neurophysiology and functional imaging give us increasingly sophisticated understanding of the links between brain function and cognition.

The fascination of exploring the beautiful ear and brain mechanisms that support hearing in complex auditory scenes provides no end of motivation for the scientists who do this work. Nevertheless, all are motivated by a desire to exploit new understanding of the auditory system for the benefit of those many people who suffer from limitations in their ability to hear in complex auditory scenes. Some key questions for ongoing research are: How can sound processing by hearing aids and cochlear implants be improved to preserve and, possibly, exaggerate monaural and binaural cues for auditory scene analysis? Can auditory training programs overcome maladaptive effects of abnormal auditory experience, or of ordinary aging? What are the critical central auditory structures that should be the targets for diagnosis and therapy?

Successful communication at the eponymous cocktail party, as well as in the complex auditory scenes of everyday life, demands all the resources of the auditory system, from basic coding mechanisms in the periphery to high-order integrative processes. The chapters of this volume are intended to be a resource for exploration of these resources at all levels: in normally functioning mature hearing, in early development, in aging, and in pathology.

Compliance with Ethics Requirements
John Middlebrooks has no conflicts of interest.
Jonathan Simon has no conflicts of interest.

References

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

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

Gatehouse, S., & Nobel, W. (2004). The speech, spatial, and qualities of hearing scale (SSQ). *International Journal of Audiology, 43*, 85–99.