

Springer Handbook of Auditory Research

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

The Auditory System at the Cocktail Party



ASA Press



Springer

Springer Handbook of Auditory Research

Volume 60

Series Editors

Richard R. Fay, Ph.D., Loyola University of Chicago
Arthur N. Popper, Ph.D., University of Maryland

Editorial Board

Karen Avraham, Ph.D., University of TelAviv, Israel
Andrew Bass, Ph.D., Cornell University
Lisa Cunningham, Ph.D., National Institutes of Health
Bernd Fritzsche, Ph.D., University of Iowa
Andrew Groves, Ph.D., Baylor University
Ronna Hertzano, M.D., Ph.D., School of Medicine, University of Maryland
Colleen Le Prell, Ph.D., University of Texas, Dallas
Ruth Litovsky, Ph.D., University of Wisconsin
Paul Manis, Ph.D., University of North Carolina
Geoffrey Manley, Ph.D., University of Oldenburg, Germany
Brian Moore, Ph.D., Cambridge University, UK
Andrea Simmons, Ph.D., Brown University
William Yost, Ph.D., Arizona State University

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

The ASA Press

The ASA Press imprint represents a collaboration between the Acoustical Society of America and Springer dedicated to encouraging the publication of important new books in acoustics. Published titles are intended to reflect the full range of research in acoustics. ASA Press books can include all types of books published by Springer and may appear in any appropriate Springer book series.

Editorial Board

Mark F. Hamilton (Chair), University of Texas at Austin
James Cottingham, Coe College
Diana Deutsch, University of California, San Diego
Timothy F. Duda, Woods Hole Oceanographic Institution
Robin Glosemeyer Petrone, Threshold Acoustics
William M. Hartmann, Michigan State University
James F. Lynch, Woods Hole Oceanographic Institution
Philip L. Marston, Washington State University
Arthur N. Popper, University of Maryland
Martin Siderius, Portland State University
Andrea M. Simmons, Brown University
Ning Xiang, Rensselaer Polytechnic Institute
William Yost, Arizona State University



ASA Press

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

The Auditory System at the Cocktail Party

With 41 Illustrations



Editors

John 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
USA

Arthur N. Popper
Department of Biology
University of Maryland
College Park, MD
USA

Richard R. Fay
Loyola University of Chicago
Chicago, IL
USA

Jonathan Z. Simon
Department of Electrical & Computer
Engineering, Department of Biology,
Institute for Systems Research
University of Maryland
College Park, MD
USA

ISSN 0947-2657 ISSN 2197-1897 (electronic)
Springer Handbook of Auditory Research
ISBN 978-3-319-51660-8 ISBN 978-3-319-51662-2 (eBook)
DOI 10.1007/978-3-319-51662-2

Library of Congress Control Number: 2017930799

© Springer International Publishing AG 2017

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

The Acoustical Society of America

On 27 December 1928 a group of scientists and engineers met at Bell Telephone Laboratories in New York City to discuss organizing a society dedicated to the field of acoustics. Plans developed rapidly and the Acoustical Society of America (ASA) held its first meeting 10–11 May 1929 with a charter membership of about 450. Today ASA has a world-wide membership of 7,000.

The scope of this new society incorporated a broad range of technical areas that continues to be reflected in ASA's present day endeavors. Today, ASA serves the interests of its members and the acoustics community in all branches of acoustics, both theoretical and applied. To achieve this goal, ASA has established technical committees charged with keeping abreast of the developments and needs of membership in specialized fields as well as identifying new ones as they develop.

The Technical Committees include: acoustical oceanography, animal bioacoustics, architectural acoustics, biomedical acoustics, engineering acoustics, musical acoustics, noise, physical acoustics, psychological and physiological acoustics, signal processing in acoustics, speech communication, structural acoustics and vibration, and underwater acoustics. This diversity is one of the Society's unique and strongest assets since it so strongly fosters and encourages cross-disciplinary learning, collaboration, and interactions.

ASA publications and meetings incorporate the diversity of these Technical Committees. In particular, publications play a major role in the Society. *The Journal of the Acoustical Society of America* (JASA) includes contributed papers and patent reviews. *JASA Express Letters* (JASA-EL) and *Proceedings of Meetings on Acoustics* (POMA) are online, open-access publications, offering rapid publication. *Acoustics Today*, published quarterly, is a popular open-access magazine. Other key features of ASA's publishing program include books, reprints of classic acoustics texts, and videos.

ASA's biannual meetings offer opportunities for attendees to share information, with strong support throughout the career continuum, from students to retirees. Meetings incorporate many opportunities for professional and social interactions and attendees find the personal contacts a rewarding experience. These experiences result in building a robust network of fellow scientists and engineers, many of whom become lifelong friends and colleagues.

From the Society's inception, members recognized the importance of developing acoustical standards with a focus on terminology, measurement procedures, and criteria for determining the effects of noise and vibration. The ASA Standard Program serves as the Secretariat for four American National Standards Institute Committees and provides administrative support for several international standards committees.

Throughout its history to present day ASA's strength resides in attracting the interest and commitment of scholars devoted to promoting the knowledge and practical applications of acoustics. The unselfish activity of these individuals in the development of the Society is largely responsible for ASA's growth and present stature.

Series Preface



The following preface is the one that we published in Volume 1 of the Springer Handbook of Auditory Research back in 1992. As anyone reading the original preface, or the many users of the series, will note, we have far exceeded our original expectation of eight volumes. Indeed, with books published to date and those in the pipeline, we are now set for over 60 volumes in SHAR, and we are still open to new and exciting ideas for additional books.

We are very proud that there seems to be consensus, at least among our friends and colleagues, that SHAR has become an important and influential part of the auditory literature. While we have worked hard to develop and maintain the quality and value of SHAR, the real value of the books is very much because of the numerous authors who have given their time to write outstanding chapters and to our many coeditors who have provided the intellectual leadership to the individual volumes. We have worked with a remarkable and wonderful group of people, many of whom have become great personal friends of both of us. We also continue to work with a spectacular group of editors at Springer. Indeed, several of our past editors have moved on in the publishing world to become senior executives. To our delight, this includes the current president of Springer US, Dr. William Curtis.

But the truth is that the series would and could not be possible without the support of our families, and we want to take this opportunity to dedicate all of the SHAR books, past and future, to them. Our wives, Catherine Fay and Helen Popper, and our children, Michelle Popper Levit, Melissa Popper Levinsohn, Christian Fay, and Amanda Fay Seirra, have been immensely patient as we developed and worked on this series. We thank them and state, without doubt, that this series could not have happened without them. We also dedicate the future of SHAR to our next generation of (potential) auditory researchers—our grandchildren—Ethan and Sophie Levinsohn, Emma Levit, and Nathaniel, Evan, and Stella Fay.

Preface 1992

The Springer Handbook of Auditory Research presents a series of comprehensive and synthetic reviews of the fundamental topics in modern auditory research. The volumes are aimed at all individuals with interests in hearing research including advanced graduate students, postdoctoral researchers, and clinical investigators. The volumes are intended to introduce new investigators to important aspects of hearing science and to help established investigators to better understand the fundamental theories and data in fields of hearing that they may not normally follow closely.

Each volume presents a particular topic comprehensively, and each serves as a synthetic overview and guide to the literature. As such, the chapters present neither exhaustive data reviews nor original research that has not yet appeared in peer-reviewed journals. The volumes focus on topics that have developed a solid data and conceptual foundation rather than on those for which a literature is only beginning to develop. New research areas will be covered on a timely basis in the series as they begin to mature.

Each volume in the series consists of a few substantial chapters on a particular topic. In some cases, the topics will be ones of traditional interest for which there is a substantial body of data and theory, such as auditory neuroanatomy (Vol. 1) and neurophysiology (Vol. 2). Other volumes in the series deal with topics that have begun to mature more recently, such as development, plasticity, and computational models of neural processing. In many cases, the series editors are joined by a co-editor having special expertise in the topic of the volume.

Arthur N. Popper, College Park, MD, USA

Richard R. Fay, Chicago, IL, USA

Volume Preface

The cocktail party is the archetype of a complex auditory scene: multiple voices compete for attention; glasses clink; background music plays. Other situations of daily life, including busy offices, crowded restaurants, noisy classrooms, and congested city streets, are no less acoustically complex. The normal auditory system exhibits a remarkable ability to parse these complex scenes. Even relatively minor hearing impairment, however, can disrupt this auditory scene analysis.

This volume grew out of the Presidential Symposium, “Ears and Brains at the Cocktail Party,” at the Midwinter Meeting of the Association for Research in Otolaryngology, held in 2013 in Baltimore, Maryland. In this volume, the authors describe both the conditions in which the auditory system excels at segregating signals of interest from distractors and the conditions in which the problem is insoluble, all the time attempting to understand the neural mechanisms that underlie both the successes and the failures. In Chap. 1, Middlebrooks and Simon introduce the volume and provide an overview of the cocktail party problem, putting it into the perspective of broader issues in auditory neuroscience. In Chap. 2, Shinn-Cunningham, Best, and Lee further set the stage by elaborating on the key concept of an *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. In Chap. 3, Culling and Stone address the challenges of low-level separation of signal from noise and consider the mechanisms by which those challenges may be overcome. They introduce the distinction between *energetic* and *informational* masking. Next, in Chap. 4, Kidd and Colburn develop the concept of informational masking by focusing on speech-on-speech masking.

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 useful computational models of the cocktail party problem. Then, in Chap. 6, Middlebrooks considers the importance of spatial separation of sound sources for stream segregation and reviews the psychophysics and physiological substrates of spatial stream segregation. Next, in Chap. 7, Simon reviews new developments in the field of experimental human auditory neuroscience.

A cocktail party is no place for infants and children. The auditory scene, however, is easily as acoustically complex on a noisy playground or in a crowded classroom. Young people apprehend these scenes with immature auditory systems and not-yet-crystallized language recognition. Werner, in Chap. 8, considers multiple stages and levels of development. Next, in Chap. 9, Pichora-Fuller, Alain, and Schneider consider older adults in whom maturity of language skills and stores of knowledge can to some degree compensate for senescence of the peripheral and central auditory systems. Finally, in Chap. 10, Litovsky, Goupell, Misurelli, and Kan consider the consequences of hearing impairment and the ways in which hearing can at least partially be restored.

Successful communication at the eponymous cocktail party as well as in other, everyday, complex auditory scenes 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 normal mature hearing, in early development, in aging, and in pathology.

John C. Middlebrooks, Irvine, CA, USA
Jonathan Z. Simon, College Park, MD, USA
Arthur N. Popper, College Park, MD, USA
Richard R. Fay, Chicago, IL, USA

Contents

1	Ear and Brain Mechanisms for Parsing the Auditory Scene	1
	John C. Middlebrooks and Jonathan Z. Simon	
2	Auditory Object Formation and Selection	7
	Barbara Shinn-Cunningham, Virginia Best, and Adrian K.C. Lee	
3	Energetic Masking and Masking Release	41
	John F. Culling and Michael A. Stone	
4	Informational Masking in Speech Recognition	75
	Gerald Kidd Jr. and H. Steven Colburn	
5	Modeling the Cocktail Party Problem	111
	Mounya Elhilali	
6	Spatial Stream Segregation	137
	John C. Middlebrooks	
7	Human Auditory Neuroscience and the Cocktail Party Problem	169
	Jonathan Z. Simon	
8	Infants and Children at the Cocktail Party	199
	Lynne Werner	
9	Older Adults at the Cocktail Party	227
	M. Kathleen Pichora-Fuller, Claude Alain, and Bruce A. Schneider	
10	Hearing with Cochlear Implants and Hearing Aids in Complex Auditory Scenes	261
	Ruth Y. Litovsky, Matthew J. Goupell, Sara M. Misurelli, and Alan Kan	

Contributors

Claude Alain Department of Psychology, The Rotman Research Institute, University of Toronto, Toronto, ON, Canada

Virginia Best Department of Speech, Language and Hearing Sciences, Boston University, Boston, MA, USA

H. Steven Colburn Department of Biomedical Engineering, Hearing Research Center, Boston University, Boston, MA, USA

John F. Culling School of Psychology, Cardiff University, Cardiff, UK

Mounya Elhilali Laboratory for Computational Audio Perception, Center for Speech and Language Processing, Department of Electrical and Computer Engineering, The Johns Hopkins University, Baltimore, MD, USA

Matthew J. Goupell Department of Hearing and Speech Sciences, University of Maryland, College Park, MD, USA

Alan Kan Waisman Center, University of Wisconsin–Madison, Madison, WI, USA

Gerald Kidd Jr. Department of Speech, Language and Hearing Sciences, Hearing Research Center, Boston University, Boston, MA, USA

Adrian K.C. Lee Department of Speech and Hearing Sciences, Institute for Learning and Brain Sciences (I-LABS), University of Washington, Seattle, WA, USA

Ruth Y. Litovsky Waisman Center, University of Wisconsin–Madison, Madison, WI, USA

John 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, USA

Sara M. Misurelli Department of Communication Sciences and Disorders, University of Wisconsin–Madison, Madison, WI, USA

M. Kathleen Pichora-Fuller Department of Psychology, University of Toronto, Mississauga, ON, Canada

Bruce A. Schneider Department of Psychology, University of Toronto, Mississauga, ON, Canada

Barbara Shinn-Cunningham Center for Research in Sensory Communication and Emerging Neural Technology, Boston University, Boston, MA, USA

Jonathan Z. Simon Department of Electrical & Computer Engineering, Department of Biology, Institute of Systems Research, University of Maryland, College Park, MD, USA

Michael A. Stone Manchester Centre for Audiology and Deafness, School of Health Sciences, University of Manchester, Manchester, UK

Lynne Werner Department of Speech and Hearing Sciences, University of Washington, Washington, USA

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

© Springer International Publishing AG 2017

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.

Chapter 2

Auditory Object Formation and Selection

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

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

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

B. Shinn-Cunningham (✉)

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

V. Best

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

A.K.C. Lee

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

2.1 Introduction

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

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

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

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

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

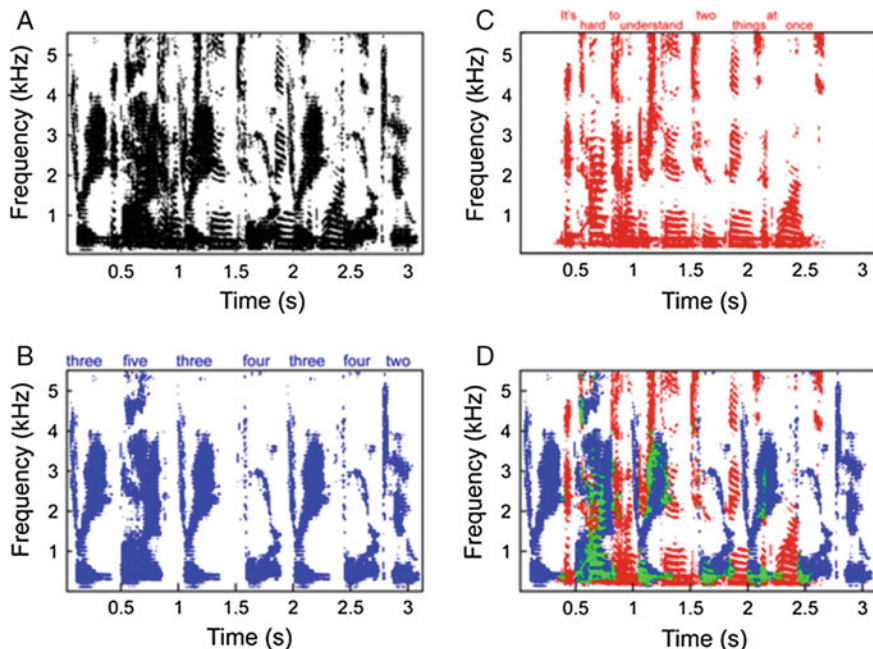


Fig. 2.1 Demonstration of how time–frequency sparseness leads to sound mixtures where clean “glimpses” of the component sounds are preserved, using two independent speech streams. **(A)** A thresholded spectrogram showing all time–frequency tiles with significant energy in a mixture of two sentences,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

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

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

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

2.1.2 *Object-Based Attention*

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

particular task, but that “comes along for the ride” when an observer focuses on one aspect of a perceptual object, is a hallmark of object-based attention.

2.1.3 Heterarchical Rather Than Hierarchical Processing

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

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

2.1.4 *A Historical Note*

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

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

A key realization that helped bridge this gap was that object formation and attention are best studied jointly (e.g., Shinn-Cunningham 2008). Interestingly, although the idea of object-based attention came from vision, there is relatively little discussion of the relationship between object formation and attention in that literature. It is not entirely clear why this is the case; historically, however, most visual attention studies use scenes consisting of very distinct, discrete objects (e.g., individual triangles and squares or individual letters), so that there is little ambiguity as to how to parse the inputs. In the auditory domain, failures of selective attention often arise because of failures to properly parse the acoustic scene into appropriate objects. Moreover, because auditory information (e.g., in speech) often

unfolds over relatively long time scales (seconds), auditory selective attention depends on properly tracking auditory objects through time, a concept commonly referred to as “streaming.” Given this, it may be that forming and streaming auditory objects is often inherently more challenging than forming visual objects.

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

2.2 Parsing the Acoustic Scene: Auditory Object Formation

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

“connected” (discussed in Sect. 2.2.1), and a yet longer time scale that causes locally grouped energy bursts to connect into auditory objects that extend through time, forming what Bregman referred to as “streams” (discussed in Sect. 2.2.2).

2.2.1 Local Spectrotemporal Cues Support “Syllable-Level” Object Formation

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

Intuitively, it seems as if the spatial cues of concurrent sounds should impact auditory grouping strongly. However, instantaneous spatial cues actually are relatively weak cues for grouping at the syllabic level (Culling and Stone, Chap. 3). For instance, sound elements that turn on and off together tend to fuse together even if they have spatial cues that are inconsistent with one another (Darwin and Hukin 1997); conversely, spatial cues influence local grouping only weakly, with effects that may be observable only when other spectrotemporal cues are ambiguous (e.g., Shinn-Cunningham et al. 2007; Schwartz et al. 2012). This counterintuitive result may reflect the fact that spatial cues are derived, requiring a comparison of the inputs to the two ears, whereas amplitude and harmonic cues are inherent in the peripheral representation of sounds. The modest influence of spatial cues on object formation may also reflect the fact that in the real world, spatial cues are quite unreliable owing to effects of reverberation as well as interference from other sound sources (Palomaki et al. 2004; Ihlefeld and Shinn-Cunningham 2011). While such effects can distort interaural time and level differences quite significantly, their

effects on amplitude modulation or harmonic structure are less pronounced; in line with this, moderate reverberant energy often degrades spatial cues significantly without interfering with perception of other sound properties, such as speech meaning (Culling et al. 1994; Ruggles et al. 2011). Although spatial cues have relatively weak effects on grouping at the syllabic level, when target and masker sources are at distinct locations, spatial cues can provide a strong basis for grouping of sequences of syllables into perceptual streams and for disentangling multiple interleaved sequences of sounds (Maddox and Shinn-Cunningham, 2012; Middlebrooks, Chap. 6).

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

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

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

individual syllables often are heard; the real challenge is tracking the stream of such syllables from a particular talker over time.

2.2.2 Higher-Order Features Link Syllables into “Streams”

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

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

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

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

2.2.3 *Open Questions*

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

A related question is whether the attentional “background,” comprising those parts of an acoustic scene that are not the focus of attention, is organized into objects or whether it remains undifferentiated. This question, although of great theoretical interest, is difficult to test, given that the only direct way to probe listeners' percepts of “the background” is to ask them what they perceive; however, the very act of asking this question is likely to cause them to focus attention on the

background, so that it flips to become the foreground. Studies of neural, rather than behavioral, responses may help shed light on this important question (e.g., Lepisto et al. 2009).

2.3 Focusing Attention: Selecting What to Process

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

2.3.1 *Top-Down Control Guides Selection*

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

2.3.2 *Bottom-up Salience Influences Attention*

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

2.3.3 *Extracting Meaning from Imperfect Objects*

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

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

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

2.4 Perceptual Consequences of Object-Based Auditory Selective Attention

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

2.4.1 *Failure to Divide Attention*

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

exists, it suggests that the appearance of a new event draws attention exogenously, whereas the disappearance of an unattended object does not.

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

2.4.2 Obligatory Interactions Between Formation and Selection

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

The influence of feature continuity on perception also supports the idea that objects are the focus of attention. As mentioned in Sect. 2.2.2, even when listeners try to ignore some task-irrelevant feature, the perceptual continuity of that feature influences the ability to extract information from a sound mixture. In particular, once a listener attends to one word, a subsequent word that shares some perceptual

feature with the attended word is automatically more likely to be the focus of attention than a word that does not match the preceding word (Bressler et al. 2014). This result supports the idea that auditory objects extend through time, and that the resulting stream is the unit of attention.

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

2.4.3 *Costs of Switching Attention*

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

2.5 **Neural Mechanisms Supporting Object Formation**

There are a multitude of hypotheses and models concerning the neural underpinnings of auditory object formation. One hypothesis postulates that sound elements segregate into separate streams whenever they activate well-separated populations of auditory neurons, such as when the streams do not overlap in frequency (Micheyl et al. 2005). However, sounds can bind together into one perceptual group even if they excite distinct neural populations (Elhilali et al., 2009b). The temporal coherence theory (TCT) of object formation accounts for these results by assuming that when neurons encoding various sound features have responses that modulate coherently through time, the features are bound together, perceptually (Shamma et al. 2011; O’Sullivan et al. 2015). A multifeature representation such as that proposed in TCT provides a general and flexible framework for explaining how

perceptual objects can emerge from a distributed neural code. The proposal that temporal coherence between different feature-selective neurons drives perceptual binding leverages two statistical aspects of a natural auditory scene: (1) In general, the strength of the response to a feature of a particular sound source will be proportional to the intensity of the source at a given moment, (2) The intensity of distinct sound sources, and thus the response to any associated features of the two sources, will be statistically independent over time. Attention has been hypothesized to influence object formation by modulating the temporal coherence of neural populations (O’Sullivan et al. 2015; see Gregoriou et al., 2009, for an example from the vision literature). When a listener selectively attends to a feature, this attentional focus is thought to up-regulate activity, which strengthens the binding of features that are temporally coherent with the attended feature.

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

2.6 Neural Mechanisms Supporting Object Selection

In the past two decades, the field of cognitive neuroscience has witnessed a growing interest in understanding the mechanisms controlling attentional selection. This may be partly due to the rapid advancement of recording techniques that have enabled scientists to study the brain while an observer is engaged in attentionally demanding tasks. Both noninvasive techniques, such as fMRI, EEG, and magnetoencephalography (MEG), and invasive electrocorticography (intracranial recording from the exposed surface of the brain, typically done in conjunction with

presurgery testing of epileptic patients) provide important, complementary information about how the human cortical response is modulated by attention. To a large degree, vision scientists have led the search for neural mechanisms underpinning attention. Given that the networks controlling attention seem at least partially to be shared across the senses (e.g., see Tark and Curtis 2009), understanding the attentional networks found by vision scientists is helpful for understanding the control of auditory attention. Thus, evidence about networks defined from visual studies is reviewed before returning to audition.

2.6.1 Visual Cognitive Networks Controlling Attention

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

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

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

attention. As discussed further in Sect. 2.6.2, there is clear support for the idea that this orienting network is engaged during auditory spatial processing (Tark and Curtis 2009; Michalka et al. 2015).

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

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

2.6.2 Auditory Spatial Attention Engages Visual Orienting and Reorienting Networks

In audition research, more effort has been devoted to understanding how we direct attention (i.e., select what sound source to attend) than to the alerting or executive function. This perhaps reflects the fundamental question at the heart of the cocktail party problem: How does one recognize what another person is saying when there

are multiple people speaking at the same time? As discussed in Sect. 2.3, many psychophysical studies have addressed how people orient attention or selectively attend to a particular sound object in a mixture.

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

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

2.6.3 Nonspatial Auditory Attention Differentially Engages Auditory-Specific Networks

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

2.6.4 Both Sensory Modality and Task Demands Affect Network Activity

A few studies underscore the emerging idea that which control networks are engaged by attention depends jointly on both the sensory modality of the input stimulus and the attributes to which attention is focused. In one fMRI study, directly contrasting activity during processing of auditory versus visual targets reveals interdigitated regions in lateral frontal cortex (LFC) that either favor visual attentional processing (superior precentral sulcus and inferior precentral sulcus) or auditory attentional processing (transverse gyrus intersecting precentral sulcus and caudal inferior frontal sulcus; see Michalka et al. 2015). These modality biases both are consistent with resting state analysis in individual subjects (i.e., the visual-biased LFC regions show intrinsic connectivity with visual sensory regions, whereas the auditory-biased LFC regions show intrinsic connectivity with auditory sensory regions; Michalka et al. 2015), and are also supported by analysis of

anatomical connectivity using data taken from the Human Connectome Project (Osher et al. 2015). These new findings can be resolved with previous reports that suggest broad, cross-modal control regions in LFC (e.g., see the review by Duncan 2010), in part by understanding that averaging brain regions across subjects (the approach normally taken) blurs away important distinctions in these regions because of the challenge of co-registration of activity in frontal cortex, where individual variations in anatomical and function patterns can be significant.

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

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

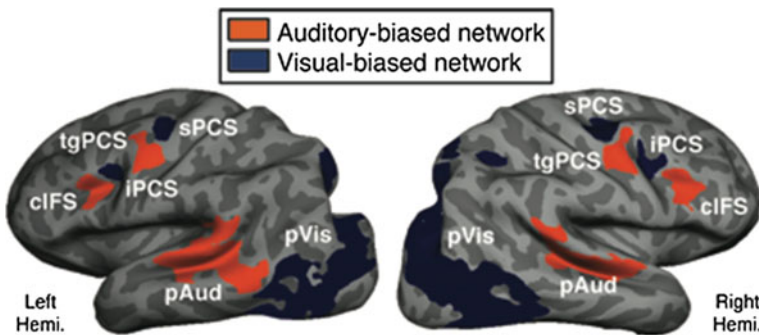


Fig. 2.2 Illustration of the brain regions making up auditory-biased (red) and vision-biased (blue) attentional control networks (derived from data reported in Michalka et al. 2015; figure provided by S. Michalka), shown on a “semi-inflated” map of the cortical surface (gyri shown in light gray; sulci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

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

2.6.5 Entrainment of Neural Responses to Attended Speech

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

Patients undergoing medical procedures that require implantation of electrodes into the brain (for instance, to discover the focal source of epileptic seizures for surgical planning) now often agree to participate in studies of brain function (producing what is known as electrocorticography [ECoG], measured from penetrating or surface electrodes on the brain). A number of such patients have participated in studies of auditory attention. Signals from these studies have provided further insight into the neural encoding of attended and unattended auditory signals. Whereas the cortical coverage of ECoG is driven exclusively by clinical needs, and thus provides only a limited window on cortical activity, ECoG yields exquisite

temporal and spatial resolution. In particular, the signal-to-noise ratio for high-frequency neural signals (especially in the high-gamma range of 80–150 Hz, which correlates with spiking activity in the underlying neural populations) is much greater in ECoG than with EEG or MEG.

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

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

2.6.6 Other Neural Signatures of Focused Auditory Attention

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

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

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

2.7 Summary Comments

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

selection bring one perceived sound source into attentional focus, allowing the listener to analyze that object in detail.

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

Compliance with Ethics Requirements

Barbara Shinn-Cunningham has no conflicts of interest.

Virginia Best has no conflicts of interest.

Adrian K. C. Lee has no conflicts of interest.

References

- Alain, C., Arnott, S. R., & Picton, T. W. (2001). Bottom-up and top-down influences on auditory scene analysis: Evidence from event-related brain potentials. *Journal of Experimental Psychology: Human Perception and Performance*, 27(5), 1072–1089.
- Alain, C., & Woods, D. L. (1997). Attention modulates auditory pattern memory as indexed by event-related brain potentials. *Psychophysiology*, 34(5), 534–546.
- Alho, K., Salmi, J., Koistinen, S., Salonen, O., & Rinne, T. (2015). Top-down controlled and bottom-up triggered orienting of auditory attention to pitch activate overlapping brain networks. *Brain Research*, 1626, 136–145.
- Arbogast, T. L., & Kidd, G., Jr. (2000). Evidence for spatial tuning in informational masking using the probe-signal method. *The Journal of the Acoustical Society of America*, 108(4), 1803–1810.
- Benard, M. R., Mensink, J. S., & Başkent, D. (2014). Individual differences in top-down restoration of interrupted speech: Links to linguistic and cognitive abilities. *The Journal of the Acoustical Society of America*, 135, EL88–94.

- Best, V., Gallun, F. J., Carlile, S., & Shinn-Cunningham, B. G. (2007a). Binaural interference and auditory grouping. *The Journal of the Acoustical Society of America*, *121*(2), 1070–1076.
- Best, V., Gallun, F. J., Ihlefeld, A., & Shinn-Cunningham, B. G. (2006). The influence of spatial separation on divided listening. *The Journal of the Acoustical Society of America*, *120*(3), 1506–1516.
- Best, V., Ozmeral, E. J., Kopco, N., & Shinn-Cunningham, B. G. (2008). Object continuity enhances selective auditory attention. *Proceedings of the National Academy of Sciences of the USA*, *105*(35), 13174–13178.
- Best, V., Ozmeral, E. J., & Shinn-Cunningham, B. G. (2007b). Visually-guided attention enhances target identification in a complex auditory scene. *Journal of the Association for Research in Otolaryngology*, *8*(2), 294–304.
- Bharadwaj, H. M., Lee, A. K. C., & Shinn-Cunningham, B. G. (2014). Measuring auditory selective attention using frequency tagging. *Frontiers in Integrative Neuroscience*, *8*, 6.
- Bharadwaj, H. M., Masud, S., Mehraei, G., Verhulst, S., & Shinn-Cunningham, B. G. (2015). Individual differences reveal correlates of hidden hearing deficits. *The Journal of Neuroscience*, *35*(5), 2161–2172.
- Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review*, *108*(3), 624–652.
- Bregman, A. S. (1990). *Auditory scene analysis: The perceptual organization of sound*. Cambridge, MA: MIT Press.
- Bressler, S., Masud, S., Bharadwaj, H., & Shinn-Cunningham, B. (2014). Bottom-up influences of voice continuity in focusing selective auditory attention. *Psychological Research*, *78*(3), 349–360.
- Bressler, S. L., Tang, W., Sylvester, C. M., Shulman, G. L., & Corbetta, M. (2008). Top-down control of human visual cortex by frontal and parietal cortex in anticipatory visual spatial attention. *The Journal of Neuroscience*, *28*(40), 10056–10061.
- Broadbent, D. E. (1954). The role of auditory localization in attention and memory span. *Journal of Experimental Psychology*, *47*(3), 191–196.
- Broadbent, D. E. (1956). Successive responses to simultaneous stimuli. *Quarterly Journal of Experimental Psychology*, 145–152.
- Broadbent, D. E. (1957). Immediate memory and simultaneous stimuli. *Quarterly Journal of Experimental Psychology*, *9*, 1–11.
- Broadbent, D. E. (1958). *Perception and communication*. New York: Pergamon Press.
- Brungart, D. S. (2001). Informational and energetic masking effects in the perception of two simultaneous talkers. *The Journal of the Acoustical Society of America*, *109*(3), 1101–1109.
- Bush, G., Luu, P., & Posner, M. I. (2000). Cognitive and emotional influences in anterior cingulate cortex. *Trends in Cognitive Sciences*, *4*(6), 215–222.
- Carlyon, R. P. (2004). How the brain separates sounds. *Trends in Cognitive Sciences*, *8*(10), 465–471.
- Carlyon, R. P., Plack, C. J., Fantini, D. A., & Cusack, R. (2003). Cross-modal and non-sensory influences on auditory streaming. *Perception*, *32*(11), 1393–1402.
- Chait, M., de Cheveigne, A., Poeppel, D., & Simon, J. Z. (2010). Neural dynamics of attending and ignoring in human auditory cortex. *Neuropsychologia*, *48*(11), 3262–3271.
- Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *The Journal of the Acoustical Society of America*, *25*, 975–979.
- Cherry, E. C., & Taylor, W. K. (1954). Some further experiments upon the recognition of speech, with one and with two ears. *The Journal of the Acoustical Society of America*, *26*, 554–559.
- Choi, I., Rajaram, S., Varghese, L. A., & Shinn-Cunningham, B. G. (2013). Quantifying attentional modulation of auditory-evoked cortical responses from single-trial electroencephalography. *Frontiers in Human Neuroscience*, *7*, 115.
- Conway, A. R., Cowan, N., & Bunting, M. F. (2001). The cocktail party phenomenon revisited: The importance of working memory capacity. *Psychonomic Bulletin Review*, *8*(2), 331–335.
- Cooke, M. (2006). A glimpsing model of speech perception in noise. *The Journal of the Acoustical Society of America*, *119*(3), 1562–1573.

- Corbetta, M., Patel, G., & Shulman, G. L. (2008). The reorienting system of the human brain: From environment to theory of mind. *Neuron*, 58(3), 306–324.
- Corbetta, M., & Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience*, 3(3), 201–215.
- Culling, J. F., & Darwin, C. J. (1993a). Perceptual separation of simultaneous vowels: Within and across-formant grouping by F0. *The Journal of the Acoustical Society of America*, 93(6), 3454–3467.
- Culling, J. F., & Darwin, C. J. (1993b). The role of timbre in the segregation of simultaneous voices with intersecting F0 contours. *Perception and Psychophysics*, 54(3), 303–309.
- Culling, J. F., Hodder, K. I., & Toh, C. Y. (2003). Effects of reverberation on perceptual segregation of competing voices. *The Journal of the Acoustical Society of America*, 114(5), 2871–2876.
- Culling, J. F., Summerfield, Q., & Marshall, D. H. (1994). Effects of simulated reverberation on the use of binaural cues and fundamental-frequency differences for separating concurrent vowels. *Speech Communication*, 14, 71–95.
- Cusack, R., Deeks, J., Aikman, G., & Carlyon, R. P. (2004). Effects of location, frequency region, and time course of selective attention on auditory scene analysis. *Journal of Experimental Psychology: Human Perception and Performance*, 30(4), 643–656.
- Cusack, R., & Roberts, B. (2000). Effects of differences in timbre on sequential grouping. *Perception and Psychophysics*, 62(5), 1112–1120.
- Dalton, P., & Fraenkel, N. (2012). Gorillas we have missed: Sustained inattentive deafness for dynamic events. *Cognition*, 124(3), 367–372.
- Dannenbring, G. L. (1976). Perceived auditory continuity with alternately rising and falling frequency transitions. *Canadian Journal of Psychology*, 30(2), 99–114.
- Darwin, C. J. (2005). Simultaneous grouping and auditory continuity. *Perception and Psychophysics*, 67(8), 1384–1390.
- Darwin, C. J. (2006). Contributions of binaural information to the separation of different sound sources. *International Journal of Audiology*, 45(Supplement 1), S20–S24.
- Darwin, C. J., Brungart, D. S., & Simpson, B. D. (2003). Effects of fundamental frequency and vocal-tract length changes on attention to one of two simultaneous talkers. *The Journal of the Acoustical Society of America*, 114(5), 2913–2922.
- Darwin, C. J., & Carlyon, R. P. (1995). Auditory grouping. In B. C. J. Moore (Ed.), *Hearing* (pp. 387–424). San Diego: Academic Press.
- Darwin, C. J., & Ciocca, V. (1992). Grouping in pitch perception: Effects of onset asynchrony and ear of presentation of a mistuned component. *The Journal of the Acoustical Society of America*, 91(6), 3381–3390.
- Darwin, C. J., & Hukin, R. W. (1997). Perceptual segregation of a harmonic from a vowel by interaural time difference and frequency proximity. *The Journal of the Acoustical Society of America*, 102(4), 2316–2324.
- Darwin, C. J., & Hukin, R. W. (2000). Effects of reverberation on spatial, prosodic, and vocal-tract size cues to selective attention. *The Journal of the Acoustical Society of America*, 108(1), 335–342.
- Darwin, C. J., Hukin, R. W., & al-Khatib, B. Y. (1995). Grouping in pitch perception: Evidence for sequential constraints. *The Journal of the Acoustical Society of America*, 98(2 Pt 1), 880–885.
- Darwin, C. J., & Sutherland, N. S. (1984). Grouping frequency components of vowels: When is a harmonic not a harmonic? *Quarterly Journal of Experimental Psychology*, 36A, 193–208.
- de Cheveigne, A., McAdams, S., & Marin, C. M. H. (1997). Concurrent vowel identification. II. Effects of phase, harmonicity, and task. *The Journal of the Acoustical Society of America*, 101, 2848–2856.
- De Sanctis, P., Ritter, W., Molholm, S., Kelly, S. P., & Foxe, J. J. (2008). Auditory scene analysis: The interaction of stimulation rate and frequency separation on pre-attentive grouping. *European Journal of Neuroscience*, 27(5), 1271–1276.

- Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review Neuroscience*, 18, 193–222.
- Devergie, A., Grimault, N., Tillmann, B., & Berthommier, F. (2010). Effect of rhythmic attention on the segregation of interleaved melodies. *The Journal of the Acoustical Society of America*, 128(1), EL1–7.
- Ding, N., & Simon, J. Z. (2012a). Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. *Journal of Neurophysiology*, 107(1), 78–89.
- Ding, N., & Simon, J. Z. (2012b). Emergence of neural encoding of auditory objects while listening to competing speakers. *Proceedings of the National Academy of Sciences of the USA*, 109(29), 11854–11859.
- Duncan, J. (2010). The multiple-demand (MD) system of the primate brain: Mental programs for intelligent behaviour. *Trends in Cognitive Sciences*, 14(4), 172–179.
- Elhilali, M., Ma, L., Micheyl, C., Oxenham, A. J., & Shamma, S. A. (2009a). Temporal coherence in the perceptual organization and cortical representation of auditory scenes. *Neuron*, 61(2), 317–329.
- Elhilali, M., Xiang, J., Shamma, S. A., & Simon, J. Z. (2009b). Interaction between attention and bottom-up saliency mediates the representation of foreground and background in an auditory scene. *PLoS Biology*, 7(6), e1000129.
- Engel, A. K., Fries, P., & Singer, W. (2001). Dynamic predictions: oscillations and synchrony in top-down processing. *Nature Reviews Neuroscience*, 2(10), 704–716.
- Engel, A. K., & Singer, W. (2001). Temporal binding and the neural correlates of sensory awareness. *Trends in Cognitive Sciences*, 5(1), 16–25.
- Eramudugolla, R., Irvine, D. R., McAnally, K. I., Martin, R. L., & Mattingley, J. B. (2005). Directed attention eliminates ‘change deafness’ in complex auditory scenes. *Current Biology*, 15(12), 1108–1113.
- Feldman, J. (2003). What is a visual object? *Trends in Cognitive Sciences*, 7(6), 252–256.
- Foxe, J. J., & Snyder, A. C. (2011). The role of alpha-band brain oscillations as a sensory suppression mechanism during selective attention. *Frontiers of Psychology*, 2, 154.
- Fritz, J. B., Elhilali, M., David, S. V., & Shamma, S. A. (2007). Auditory attention: Focusing the searchlight on sound. *Current Opinion in Neurobiology*, 17(4), 437–455.
- Fujisaki, W., & Nishida, S. (2005). Temporal frequency characteristics of synchrony-asynchrony discrimination of audio-visual signals. *Experimental Brain Research*, 166(3–4), 455–464.
- Gallun, F. J., Mason, C. R., & Kidd, G., Jr. (2007). The ability to listen with independent ears. *The Journal of the Acoustical Society of America*, 122(5), 2814–2825.
- Giraud, A. L., & Poeppel, D. (2012). Cortical oscillations and speech processing: Emerging computational principles and operations. *Nature Neuroscience*, 15(4), 511–517.
- Goldberg, M. E., & Bruce, C. J. (1985). Cerebral cortical activity associated with the orientation of visual attention in the rhesus monkey. *Vision Research*, 25(3), 471–481.
- Goodale, M. A., & Milner, A. D. (1992). Separate visual pathways for perception and action. *Trends in Neurosciences*, 15(1), 20–25.
- Greenberg, S., Carvey, H., Hitchcock, L., & Chang, S. (2003). Temporal properties of spontaneous speech—A syllable-centric perspective. *Journal of Phonetics*, 31(3–4), 465–485.
- Greenberg, G. Z., & Larkin, W. D. (1968). Frequency-response characteristic of auditory observers detecting signals of a single frequency in noise: The probe-signal method. *The Journal of the Acoustical Society of America*, 44(6), 1513–1523.
- Gregoriou, G. G., Gotts, S. J., Zhou, H., & Desimone, R. (2009). High-frequency, long-range coupling between prefrontal and visual cortex during attention. *Science*, 324(5931), 1207–1210.
- Griffiths, T. D., & Warren, J. D. (2004). What is an auditory object? *Nature Reviews Neuroscience*, 5(11), 887–892.
- Grimault, N., Bacon, S. P., & Micheyl, C. (2002). Auditory stream segregation on the basis of amplitude-modulation rate. *The Journal of the Acoustical Society of America*, 111(3), 1340–1348.

- Hall, J. W., 3rd, & Grose, J. H. (1990). Comodulation masking release and auditory grouping. *The Journal of the Acoustical Society of America*, 88(1), 119–125.
- Heller, L. M., & Richards, V. M. (2010). Binaural interference in lateralization thresholds for interaural time and level differences. *The Journal of the Acoustical Society of America*, 128(1), 310–319.
- Heller, L. M., & Trahiotis, C. (1996). Extents of laterality and binaural interference effects. *The Journal of the Acoustical Society of America*, 99(6), 3632–3637.
- Hill, K. T., & Miller, L. M. (2010). Auditory attentional control and selection during cocktail party listening. *Cerebral Cortex*, 20(3), 583–590.
- Hukin, R. W., & Darwin, C. J. (1995). Comparison of the effect of onset asynchrony on auditory grouping in pitch matching and vowel identification. *Perception and Psychophysics*, 57(2), 191–196.
- Hupe, J. M., Joffo, L. M., & Pressnitzer, D. (2008). Bistability for audiovisual stimuli: Perceptual decision is modality specific. *Journal of Vision*, 8(7), 11–15.
- Ihlefeld, A., & Shinn-Cunningham, B. G. (2011). Effect of source spectrum on sound localization in an everyday reverberant room. *The Journal of the Acoustical Society of America*, 130(1), 324–333.
- Jones, M. R., Kidd, G., & Wetzel, R. (1981). Evidence for rhythmic attention. *Journal of Experimental Psychology: Human Perception and Performance*, 7(5), 1059–1073.
- Kastner, S., & Ungerleider, L. G. (2001). The neural basis of biased competition in human visual cortex. *Neuropsychologia*, 39(12), 1263–1276.
- Kaya, E. M., & Elhilali, M. (2014). Investigating bottom-up auditory attention. *Frontiers in Human Neuroscience*, 8(327), 1–12.
- Kayser, C., Montemurro, M. A., Logothetis, N. K., & Panzeri, S. (2009). Spike-phase coding boosts and stabilizes information carried by spatial and temporal spike patterns. *Neuron*, 61(4), 597–608.
- Kerlin, J. R., Shahin, A. J., & Miller, L. M. (2010). Attentional gain control of ongoing cortical speech representations in a “cocktail party”. *The Journal of Neuroscience*, 30(2), 620–628.
- Kidd, G., Jr., Arbogast, T. L., Mason, C. R., & Gallun, F. J. (2005a). The advantage of knowing where to listen. *The Journal of the Acoustical Society of America*, 118(6), 3804–3815.
- Kidd, G., Mason, C. R., Brughera, A., & Hartmann, W. M. (2005b). The role of reverberation in release from masking due to spatial separation of sources for speech identification. *Acta Acustica united with Acustica*, 91(3), 526–536.
- Kidd, G., Jr., Mason, C. R., Richards, V. M., Gallun, F. J., & Durlach, N. I. (2008). Informational Masking. In W. Yost, A. Popper, & R. Fay (Eds.), *Auditory perception of sound sources* (pp. 143–189). New York: Springer Science+Business Media.
- Kitterick, P. T., Clarke, E., O’Shea, C., Seymour, J., & Summerfield, A. Q. (2013). Target identification using relative level in multi-talker listening. *The Journal of the Acoustical Society of America*, 133(5), 2899–2909.
- Kong, L., Michalka, S. W., Rosen, M. L., Sheremata, S. L., et al. (2014). Auditory spatial attention representations in the human cerebral cortex. *Cerebral Cortex*, 24(3), 773–784.
- Koreimann, S., Gula, B., & Vitouch, O. (2014). Inattentional deafness in music. *Psychological Research*, 78, 304–312.
- Lachter, J., Forster, K. I., & Ruthruff, E. (2004). Forty-five years after Broadbent (1958): Still no identification without attention. *Psychological Review*, 111(4), 880–913.
- Lakatos, P., Musacchia, G., O’Connel, M. N., Falchier, A. Y., Javitt, D. C., & Schroeder, C. E. (2013). The spectrotemporal filter mechanism of auditory selective attention. *Neuron*, 77(4), 750–761.
- Lalor, E. C., & Foxe, J. J. (2010). Neural responses to uninterrupted natural speech can be extracted with precise temporal resolution. *European Journal of Neuroscience*, 31(1), 189–193.
- Lalor, E. C., Power, A. J., Reilly, R. B., & Foxe, J. J. (2009). Resolving precise temporal processing properties of the auditory system using continuous stimuli. *Journal of Neurophysiology*, 102(1), 349–359.

- Larson, E., & Lee, A. K. C. (2013). Influence of preparation time and pitch separation in switching of auditory attention between streams. *The Journal of the Acoustical Society of America*, *134*(2), EL165–171.
- Larson, E., & Lee, A. K. C. (2014). Switching auditory attention using spatial and non-spatial features recruits different cortical networks. *NeuroImage*, *84*, 681–687.
- Lawo, V., & Koch, I. (2014). Dissociable effects of auditory attention switching and stimulus–response compatibility. *Psychological Research*, *78*, 379–386.
- Lee, A. K. C., Rajaram, S., Xia, J., Bharadwaj, H., et al. (2013). Auditory selective attention reveals preparatory activity in different cortical regions for selection based on source location and source pitch. *Frontiers in Neuroscience*, *6*, 190.
- Lepisto, T., Kuitunen, A., Sussman, E., Saalasti, S., et al. (2009). Auditory stream segregation in children with Asperger syndrome. *Biological Psychology*, *82*(3), 301–307.
- Loizou, P. C., Hu, Y., Litovsky, R., Yu, G., et al. (2009). Speech recognition by bilateral cochlear implant users in a cocktail-party setting. *The Journal of the Acoustical Society of America*, *125*(1), 372–383.
- Macken, W. J., Tremblay, S., Houghton, R. J., Nicholls, A. P., & Jones, D. M. (2003). Does auditory streaming require attention? Evidence from attentional selectivity in short-term memory. *Journal of Experimental Psychology: Human Perception and Performance*, *29*(1), 43–51.
- Maddox, R. K., Atilgan, H., Bizley, J. K., & Lee, A. K. (2015). Auditory selective attention is enhanced by a task-irrelevant temporally coherent visual stimulus in human listeners. *Elife*, *4*. doi:[10.7554/eLife.04995](https://doi.org/10.7554/eLife.04995)
- Maddox, R. K., & Shinn-Cunningham, B. G. (2012). Influence of task-relevant and task-irrelevant feature continuity on selective auditory attention. *Journal of the Association for Research in Otolaryngology*, *13*(1), 119–129.
- Marocco, R. T., & Davidson, M. C. (1998). Neurochemistry of attention. In R. Parasuraman (Ed.), *The attentive brain* (Vol. xii, pp. 35–50). Cambridge, MA: MIT Press.
- McCloy, D. R., & Lee, A. K. (2015). Auditory attention strategy depends on target linguistic properties and spatial configuration. *The Journal of the Acoustical Society of America*, *138*(1), 97–114.
- McDermott, J. H., Wroblewski, D., & Oxenham, A. J. (2011). Recovering sound sources from embedded repetition. *Proceedings of the National Academy of Sciences of the USA*, *108*, 1188–1193.
- Mesgarani, N., & Chang, E. F. (2012). Selective cortical representation of attended speaker in multi-talker speech perception. *Nature*, *485*(7397), 233–236.
- Michalka, S. W., Kong, L., Rosen, M. L., Shinn-Cunningham, B. G., & Somers, D. C. (2015). Short-term memory for space and time flexibly recruit complementary sensory-biased frontal lobe attention networks. *Neuron*, *87*(4), 882–892.
- Michalka, S. W., Rosen, M. L., Kong, L., Shinn-Cunningham, B. G., & Somers, D. C. (2016). Auditory spatial coding flexibly recruits anterior, but not posterior, visuotopic parietal cortex. *Cerebral Cortex*, *26*(3), 1302–1308.
- Micheyl, C., Tian, B., Carlyon, R. P., & Rauschecker, J. P. (2005). Perceptual organization of tone sequences in the auditory cortex of awake macaques. *Neuron*, *48*(1), 139–148.
- Molholm, S., Martinez, A., Shpaner, M., & Foxe, J. J. (2007). Object-based attention is multisensory: Co-activation of an object’s representations in ignored sensory modalities. *European Journal of Neuroscience*, *26*(2), 499–509.
- Moray, N. (1959). Attention in dichotic listening: Affective cues and the influence of instructions. *Quarterly Journal of Experimental Psychology*, *11*, 56–60.
- Naatanen, R., Teder, W., Alho, K., & Lavikainen, J. (1992). Auditory attention and selective input modulation: A topographical ERP study. *NeuroReport*, *3*(6), 493–496.
- Noyce, A. L., Cestero, N., Shinn-Cunningham, B. G., & Somers, D. C. (2016). Short-term memory stores organized by information domain. *Attention, Perception, & Psychophysics*, *78*(30), 960–970.

- O'Sullivan, J. A., Power, A. J., Mesgarani, N., Rajaram, S., et al. (2014). Attentional selection in a cocktail party environment can be decoded from single-trial EEG. *Cerebral Cortex*, *25*(7), 1697–1706.
- O'Sullivan, J. A., Shamma, S. A., & Lalor, E. C. (2015). Evidence for neural computations of temporal coherence in an auditory scene and their enhancement during active listening. *The Journal of Neuroscience*, *35*(18), 7256–7263.
- Osher, D., Tobyne, S., Congden, K., Michalka, S., & Somers, D. (2015). Structural and functional connectivity of visual and auditory attentional networks: Insights from the Human Connectome Project. *Journal of Vision*, *15*(12), 223.
- Oxenham, A. J. (2008). Pitch perception and auditory stream segregation: Implications for hearing loss and cochlear implants. *Trends in Amplification*, *12*(4), 316–331.
- Oxenham, A. J. (2012). Pitch perception. *The Journal of Neuroscience*, *32*(39), 13335–13338.
- Oxenham, A. J., & Dau, T. (2001). Modulation detection interference: Effects of concurrent and sequential streaming. *The Journal of the Acoustical Society of America*, *110*(1), 402–408.
- Palomaki, K. J., Brown, G. J., & Wang, D. L. (2004). A binaural processor for missing data speech recognition in the presence of noise and small-room reverberation. *Speech Communication*, *43*(4), 361–378.
- Pasley, B. N., David, S. V., Mesgarani, N., Flinker, A., et al. (2012). Reconstructing speech from human auditory cortex. *PLoS Biology*, *10*(1), e1001251.
- Pavani, F., & Turatto, M. (2008). Change perception in complex auditory scenes. *Perception and Psychophysics*, *70*(4), 619–629.
- Petersen, S. E., & Posner, M. I. (2012). The attention system of the human brain: 20 years after. *Annual Review of Neuroscience*, *35*, 73–89.
- Plack, C. J., Barker, D., & Prendergast, G. (2014). Perceptual consequences of “hidden” hearing loss. *Trends in Hearing*, *18*. doi:[10.1177/2331216514550621](https://doi.org/10.1177/2331216514550621)
- Posner, M. I., & Petersen, S. E. (1990). The attention system of the human brain. *Annual Review Neuroscience*, *13*, 25–42.
- Power, A. J., Foxe, J. J., Forde, E. J., Reilly, R. B., & Lalor, E. C. (2012). At what time is the cocktail party? A late locus of selective attention to natural speech. *European Journal of Neuroscience*, *35*(9), 1497–1503.
- Ptak, R. (2012). The frontoparietal attention network of the human brain: Action, saliency, and a priority map of the environment. *Neuroscientist*, *18*(5), 502–515.
- Pugh, K. R., Offywitz, B. A., Shaywitz, S. E., Fulbright, R. K., et al. (1996). Auditory selective attention: An fMRI investigation. *NeuroImage*, *4*(3 Pt 1), 159–173.
- Ruggles, D., Bharadwaj, H., & Shinn-Cunningham, B. G. (2011). Normal hearing is not enough to guarantee robust encoding of suprathreshold features important in everyday communication. *Proceedings of the National Academy of Sciences of the USA*, *108*(37), 15516–15521.
- Samuel, A. G. (1981). The role of bottom-up confirmation in the phonemic restoration illusion. *Journal of Experimental Psychology: Human Perception and Performance*, *7*(5), 1124–1131.
- Schaal, N. K., Williamson, V. J., & Banissy, M. J. (2013). Anodal transcranial direct current stimulation over the supramarginal gyrus facilitates pitch memory. *European Journal of Neuroscience*, *38*(10), 3513–3518.
- Scharf, B., Quigley, S., Aoki, C., Peachey, N., & Reeves, A. (1987). Focused auditory attention and frequency selectivity. *Perception and Psychophysics*, *42*(3), 215–223.
- Schwartz, A., McDermott, J. H., & Shinn-Cunningham, B. (2012). Spatial cues alone produce inaccurate sound segregation: The effect of interaural time differences. *The Journal of the Acoustical Society of America*, *132*(1), 357–368.
- Serences, J. T., & Yantis, S. (2006a). Selective visual attention and perceptual coherence. *Trends in Cognitive Sciences*, *10*(1), 38–45.
- Serences, J. T., & Yantis, S. (2006b). Spatially selective representations of voluntary and stimulus-driven attentional priority in human occipital, parietal, and frontal cortex. *Cerebral Cortex*, *17*(2), 284–293.
- Shamma, S. A., Elhilali, M., & Micheyl, C. (2011). Temporal coherence and attention in auditory scene analysis. *Trends in Neurosciences*, *34*(3), 114–123.

- Shinn-Cunningham, B. G. (2008). Object-based auditory and visual attention. *Trends in Cognitive Sciences*, *12*(5), 182–186.
- Shinn-Cunningham, B. G., & Best, V. (2008). Selective attention in normal and impaired hearing. *Trends in Amplification*, *12*(4), 283–299.
- Shinn-Cunningham, B. G., Lee, A. K. C., & Oxenham, A. J. (2007). A sound element gets lost in perceptual competition. *Proceedings of the National Academy of Sciences of the USA*, *104*(29), 12223–12227.
- Shuai, L., & Elhilali, M. (2014). Task-dependent neural representations of salient events in dynamic auditory scenes. *Frontiers in Neuroscience*, *8*(203), 1–11.
- Strauss, A., Wostmann, M., & Obleser, J. (2014). Cortical alpha oscillations as a tool for auditory selective inhibition. *Frontiers in Human Neuroscience*, *8*, 350.
- Sussman, E. S., Horvath, J., Winkler, I., & Orr, M. (2007). The role of attention in the formation of auditory streams. *Perception and Psychophysics*, *69*(1), 136–152.
- Tark, K. J., & Curtis, C. E. (2009). Persistent neural activity in the human frontal cortex when maintaining space that is off the map. *Nature Neuroscience*, *12*(11), 1463–1468.
- Teki, S., Chait, M., Kumar, S., Shamma, S., & Griffiths, T. D. (2013). Segregation of complex acoustic scenes based on temporal coherence. *Elife*, *2*, e00699.
- Terhardt, E. (1974). Pitch, consonance, and harmony. *The Journal of the Acoustical Society of America*, *55*(5), 1061–1069.
- Toscani, M., Marzi, T., Righi, S., Viggiano, M. P., & Baldassi, S. (2010). Alpha waves: A neural signature of visual suppression. *Experimental Brain Research*, *207*(3–4), 213–219.
- Treisman, A. M. (1960). Contextual cues in selective listening. *Quarterly Journal of Experimental Psychology*, *12*, 157–167.
- Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*(1), 97–136.
- Ungerleider, L. G., & Mishkin, M. (1982). Two cortical visual systems. In D. J. Ingle, M. A. Goodale, & R. J. W. Mansfield (Eds.), *Analysis of visual behaviour* (pp. 549–586). Cambridge, MA: MIT Press.
- Varghese, L., Bharadwaj, H. M., & Shinn-Cunningham, B. G. (2015). Evidence against attentional state modulating scalp-recorded auditory brainstem steady-state responses. *Brain Research*, *1626*, 146–164.
- Varghese, L. A., Ozmeral, E. J., Best, V., & Shinn-Cunningham, B. G. (2012). How visual cues for when to listen aid selective auditory attention. *Journal of the Association for Research in Otolaryngology*, *13*(3), 359–368.
- Vines, B. W., Schnider, N. M., & Schlaug, G. (2006). Testing for causality with transcranial direct current stimulation: Pitch memory and the left supramarginal gyrus. *NeuroReport*, *17*(10), 1047–1050.
- Vliegen, J., Moore, B. C., & Oxenham, A. J. (1999). The role of spectral and periodicity cues in auditory stream segregation, measured using a temporal discrimination task. *The Journal of the Acoustical Society of America*, *106*(2), 938–945.
- von Békésy, G. (1960). *Experiments in hearing* (1989th ed.). New York: Acoustical Society of America Press.
- Vossel, S., Geng, J. J., & Fink, G. R. (2014). Dorsal and ventral attention systems: Distinct neural circuits but collaborative roles. *Neuroscientist*, *20*(2), 150–159.
- Wardak, C., Ibos, G., Duhamel, J. R., & Olivier, E. (2006). Contribution of the monkey frontal eye field to covert visual attention. *The Journal of Neuroscience*, *26*(16), 4228–4235.
- Warren, R. M. (1970). Perceptual restoration of missing speech sounds. *Science*, *167*(917), 392–393.
- Warren, R. M., Wrightson, J. M., & Puretz, J. (1988). Illusory continuity of tonal and infratonal periodic sounds. *The Journal of the Acoustical Society of America*, *84*(4), 1338–1342.
- Welch, R. B., & Warren, D. H. (1980). Immediate perceptual response to intersensory discrepancy. *Psychological Bulletin*, *88*, 638–667.
- Whittingstall, K., & Logothetis, N. K. (2009). Frequency-band coupling in surface EEG reflects spiking activity in monkey visual cortex. *Neuron*, *64*(2), 281–289.

- Woldorff, M. G., Gallen, C. C., Hampson, S. A., Hillyard, S. A., et al. (1993). Modulation of early sensory processing in human auditory-cortex during auditory selective attention. *Proceedings of the National Academy of Sciences of the USA*, *90*(18), 8722–8726.
- Wood, N., & Cowan, N. (1995). The cocktail party phenomenon revisited: How frequent are attention shifts to one's name in an irrelevant auditory channel? *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *21*(1), 255–260.
- Woodruff, P. W., Benson, R. R., Bandettini, P. A., Kwong, K. K., et al. (1996). Modulation of auditory and visual cortex by selective attention is modality-dependent. *NeuroReport*, *7*(12), 1909–1913.
- Wright, B. A., & Fitzgerald, M. B. (2004). The time course of attention in a simple auditory detection task. *Perception and Psychophysics*, *66*(3), 508–516.
- Xiang, J., Simon, J., & Elhilali, M. (2010). Competing streams at the cocktail party: Exploring the mechanisms of attention and temporal integration. *The Journal of Neuroscience*, *30*(36), 12084–12093.
- Zion-Golombic, E. M., Ding, N., Bickel, S., Lakatos, P., et al. (2013). Mechanisms underlying selective neuronal tracking of attended speech at a “cocktail party”. *Neuron*, *77*(5), 980–991.
- Zion-Golombic, E., & Schroeder, C. E. (2012). Attention modulates ‘speech-tracking’ at a cocktail party. *Trends in Cognitive Sciences*, *16*(7), 363–364.

Chapter 3

Energetic Masking and Masking Release

John F. Culling and Michael A. Stone

Abstract Masking is of central interest in the cocktail party problem, because interfering voices may be sufficiently intense or numerous to mask the voice to which the listener is attending, rendering its discourse unintelligible. The definition of energetic masking is problematic, but it may be considered to consist of effects by which an interfering sound disrupts the processing of the speech signal in the lower levels of the auditory system. Maskers can affect speech intelligibility by overwhelming its representation on the auditory nerve and by obscuring its amplitude modulations. A release from energetic masking is obtained by using mechanisms at these lower levels that can recover a useful representation of the speech. These mechanisms can exploit differences between the target and masking speech such as in harmonic structure or in interaural time delay. They can also exploit short-term dips in masker strength or improvements in speech-to-masker ratio at one or other ear.

Keywords Better-ear listening · Binaural unmasking · Dip listening · Equalization—cancelation · Fundamental frequency difference · Modulation masking · Onset-time differences · Spatial release from masking

3.1 Introduction

Masking is of critical interest in the cocktail party problem, because interfering voices may be sufficiently intense or numerous to mask the voice to which the listener is attending, rendering its discourse unintelligible. Masking is defined as

J.F. Culling (✉)

School of Psychology, Cardiff University, Park Place, Cardiff CF10 3AT, UK
e-mail: CullingJ@cf.ac.uk

M.A. Stone

Manchester Centre for Audiology and Deafness, School of Health Sciences,
University of Manchester, Manchester M13 9PL, UK
e-mail: michael.stone@manchester.ac.uk

“the process by which the threshold of hearing for one sound is raised by the presence of another” (ANSI 2013, p. 61). For instance, in silence, a pure tone will be detected at a very low sound level, but if noise is present, the threshold sound level for detecting the tone (the masked detection threshold [MDT]) will be higher. For speech, this shift would usually be measured as an increase in the speech reception threshold (SRT), the sound level of speech at which a criterion proportion of that speech is understood (typically 50%). Usually, the sound level required for detecting a tone or understanding a sentence in a broadband noise increases in direct proportion to the masker sound level (Hawkins and Stevens 1950). Consequently, the MDT or the SRT remains at a constant signal-to-noise ratio (SNR) across a wide range of stimulus levels. However, if a mechanism of masking release can be exploited, the amount of masking can be reduced and the MDT or SRT will be lowered. The human auditory system is adept at exploiting a variety of mechanisms to enable release from masking. Consequently, humans can understand speech even at negative signal-to-noise ratios (SNRs), and are able, up to a point, to enjoy a conversation in a busy room. This capacity is not unique to humans and has been documented in many other species, so that they can identify mates, competitors, or offspring among a noisy background of competing calls (Bee and Micheyl 2008).

Scientists interested in the cocktail party problem often subdivide the definition of masking into two components, originally calling them “masking” and “confusion” (Egan et al. 1954), but more recently preferring the adjectives “energetic” and “informational” (Brungart 2001). This and the next chapter of this book (Kidd and Colburn, Chap. 4) will examine the phenomena associated with each of these forms of masking. Although the distinction between energetic and informational masking is today quite common, its definition is elusive and deeply problematic (Durlach 2006).

Energetic masking can be narrowly defined to occur where target and interferer energy are present at similar times and frequencies (Brungart 2001), such that they directly compete with each other on the auditory nerve. This perspective has much in common with the neurophysiological concept of “line-busy” masking, in which the presence of a signal does not increase the average response rate of the auditory nerve above the response rate elicited by the masker. Broadly speaking, informational masking might then be described as any failure to successfully process a target signal that is not energetically masked. For speech signals that are extended in both time and frequency, energetic masking requires a masker that keeps all auditory nerves active, such as continuous broadband noise, leading many to describe continuous noise as an “energetic masker.”

Miller (1947) claimed that speech is best masked by continuous noise that shares the same frequency spectrum. The effect of spectral overlap has been successfully modeled by the speech intelligibility index (SII, ANSI 1997), a single-figure measure of the relative transmission of speech information through a communication channel. The SII is used to predict the intelligibility of target speech in a background noise based on the differences between their long-term power spectra (French and Steinberg 1947). However, defining energetic masking as a function of the difference between two power spectra would make it a very simple topic.

Moreover, it has become apparent that the effect of random noise on speech intelligibility is determined not only by the level of auditory nerve excitation it generates, but also by the energy fluctuations it contains (see Sect. 3.3). In this sense, even continuous random noise cannot be considered as an example of a purely energetic masker.

Durlach (2006) suggested that one might define some intermediate category of masking phenomena that was neither energetic nor informational. He pointed out that although many researchers tend to assume that energetic masking is a process that occurs at the auditory periphery, they also consider phenomena that occur at higher sites. It is interesting to consider binaural unmasking (Hirsh 1948) in this respect. Binaural unmasking (see Sect. 3.4.2) is often considered a release from energetic masking, but the processing involved must occur in the central nervous system at a level above the confluence of information from the two ears, because both ears are needed for the effect to occur. One might define energetic masking as a process that also occurs at these higher levels, but it should be noted that it is impossible to make distinctions in terms of level of processing with any great assurance.

Indeed, the processes involved in all masking phenomena are inadequately understood, and, as will become clear in the text that follows, accounts in terms of low-level and high-level processes often compete to explain the same phenomena. For the current purpose, however, a distinction must be drawn between energetic and informational masking, and this distinction will be framed in terms of the level of neural processing at which masking and masking release might operate. Under this tentative framework, energetic masking can be released by simple low-level processes such as masker filtering or cancellation, while informational masking is released by high-level processes of grouping and auditory scene analysis (Bregman 1990).

This chapter thus addresses the roles of low-level processes of masking and masking release in circumstances that are common at cocktail parties. Major factors that enable this release are (1) segregation by fundamental frequency that can occur when the interfering sound is periodic, (2) “dip listening” that can occur when the interfering sound is strongly modulated, and (3) spatial release from masking that can occur when speech and interfering sound come from different directions.

3.2 Segregation by Fundamental Frequency

When the human voice produces vowel sounds and sonorant consonants (such as nasals and liquids) the glottis releases regular pulses of air into the vocal tract. This results in periodicity in the acoustic waveform that repeats itself in a cycle linked to the rate of the glottal pulses. This rate, known as the fundamental frequency (F0), is the principal determinant of perceived pitch. The frequency spectrum of such a sound is composed of a series of harmonics, regularly spaced in frequency at intervals of the F0. Thus, the waveform is described as “periodic” and the spectrum

as “harmonic.” The male speaking voice, for instance, has a mean F0 of around 100 Hz, although the F0 varies continuously during speech over a range of about an octave (say 70–140 Hz). The female speaking voice has a register about an octave higher (i.e., 140–280 Hz). When people speak simultaneously, even with the same mean F0, the continuous independent variation means that the F0s of their voices are rarely the same at any given moment. It also means that neither waveform repeats itself exactly, so sounds can vary in their periodicity.

Differences in F0 ($\Delta F0$ s) can be used to perceptually segregate concurrent sounds. It has long been known that musical instruments playing the same note simultaneously (or notes an octave apart) can blend together to form a single combined timbre, but when they play different notes the distinct instrumental sounds are clearly heard (Helmholz 1895). One may think of this effect as occurring because the auditory system makes use of a $\Delta F0$ to perceptually segregate sound sources and so perceive each one individually. When the $\Delta F0$ is zero, this process cannot occur and sounds merge into one.

3.2.1 *The Effect of an F0 Difference*

Whereas most musical instruments generate steady tones of fixed F0, the human voice changes continuously in F0; it produces steady F0s only during singing and then often with some vibrato (cyclic variation of F0). To perform controlled experiments with spoken sentences, the F0 must be manipulated. Brokx and Nootboom (1982) used linear predictive coding (LPC) resynthesis to monotonize digital recordings of speech. The resulting speech sounds highly artificial (robot-like), but only by getting control of the F0 in this way was it possible to show the effect of having a nonzero $\Delta F0$ between target and interfering speech. In listening tests, Brokx and Nootboom found that errors in speech understanding were progressively less frequent as $\Delta F0$ increased from 0 to 3, 6, 9, or 20 Hz (above a baseline of 100 Hz), but increased again at the octave, when the target speech F0 was double that of the interfering speech (i.e., $\Delta F0 = 100$ Hz). These results mirror the effects that occur with musical tones, suggesting that segregation of concurrent speech by $\Delta F0$ is functionally very similar to segregation of musical tones, and is probably the product of the same perceptual mechanism.

At around the same time, Scheffers (1983) developed a simpler form of experiment that became known as the “double-vowel” paradigm. He synthesized eight different Dutch vowels with various fixed F0s using a formant synthesizer. These vowels were added together in pairs, but with six different $\Delta F0$ s ranging from zero to four semitones (each semitone is a 5.95% change in F0; 12 semitones = 1 octave). Scheffers asked listeners to identify both vowels in the pair and found that accuracy increased rapidly with small $\Delta F0$ s, reaching a plateau in performance at 1 semitone $\Delta F0$. An advantage of this paradigm was that while the stimuli were very good approximations of speech sounds, they could be tightly controlled for experimental purposes. For some years, this “double-vowel” paradigm became a

standard experimental approach for interrogating the mechanisms underlying the $\Delta F0$ effect, generally using the cascade formant synthesizer described by Klatt (1980).

Several potential mechanisms for the effect of $\Delta F0$ s have been proposed and explored using both modeling and experiments. Each remains controversial. These mechanisms operate in widely differing ways, and so the remainder of this section is devoted to explaining the differences and examining the strength of the evidence for their role. The discussion is limited to cases where $\Delta F0$ s exist between concurrent sounds, rather than sequences that may be “streamed” over time (Bregman 1990), but even for these cases, some of the proposed mechanisms could be quite central while others are very peripheral.

3.2.2 *Selecting Harmonic Components of a Common F0*

An obvious possibility is that listeners detect the $F0$ s (i.e., identify the pitches) of two competing voices, and use these $F0$ s to select harmonic components from the two voices, and so build a picture of the separated vowel timbres. The harmonic components of each voice are thus collected together into two separate groups. Such a mechanism would recover the spectral content of the two voices, and, because it parses the acoustic waveform into different sources, could be viewed as a form of auditory scene analysis.

Scheffers (1983) attempted to model such a grouping mechanism by first creating a cochlear excitation pattern for the stimulus (Fletcher 1930) that simulates the frequency analyzing power of the ear. He then used contemporary models of pitch perception to identify the two $F0$ s present, before sampling the excitation pattern at intervals of those two $F0$ s to recover the vowel spectra and identify the vowels. This model was not successful at producing any improvement in vowel identification with $\Delta F0$. Assmann and Summerfield (1990) also failed to model an increase in vowel identification with $\Delta F0$ by sampling the excitation pattern.

As analysis of the cochlear excitation pattern seemed to be insufficient to separate the two vowels, researchers began to examine spectrotemporal models. These models were based, once again, on those of pitch perception (e.g., Meddis and Hewitt 1991). Rather than relying on the total energy at each place in the cochlea, these models applied autocorrelation to the precise waveform extracted at each place in the cochlea. Autocorrelation is a process in which a waveform is correlated (point-by-point multiplication and summation) with a copy of itself at a range of time delays (lags) between the two copies. The potential benefit of using autocorrelation is that it can, in principle, separate the energy of the two sound sources at each cochlear place. When a mixture of two vowels is autocorrelated, the resulting function displays two sets of peaks at lags that are at multiples of the periods of the individual sounds (Fig. 3.1). The sizes of these peaks will be related to the relative intensities of the two sounds (although also to their relative periodicity). If this process is applied in different frequency channels then the relative

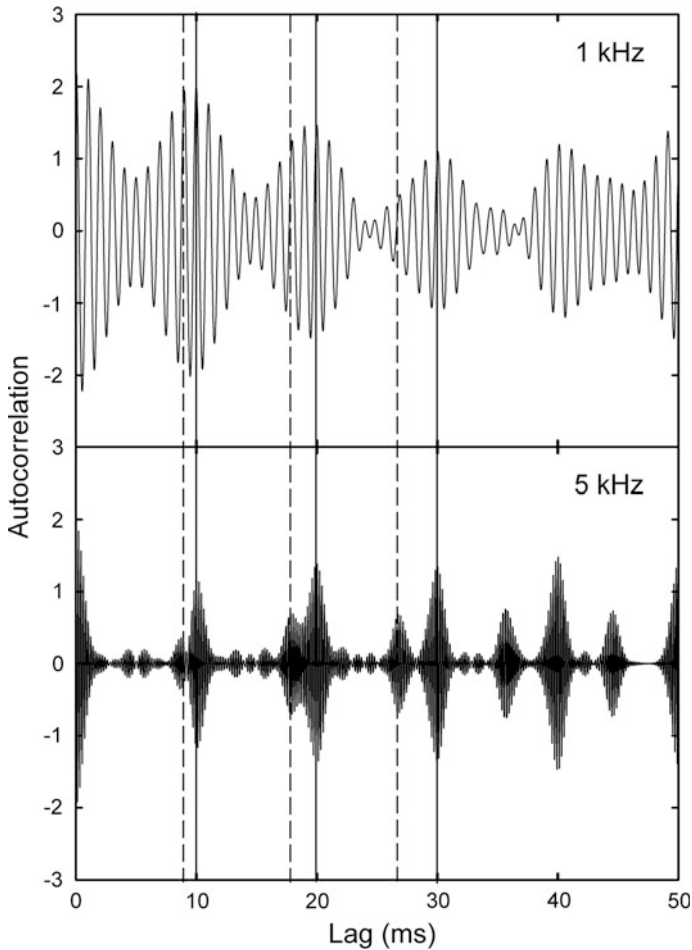


Fig. 3.1 Autocorrelation functions (ACFs) for mixtures of two harmonic complex tones within different frequency channels. The broader 5-kHz channel displays stronger modulation of the ACF and two clear series of peaks are visible. The dashed and solid vertical lines mark ACF lags that correspond to integer multiples of the two stimulus periods. At 5 kHz, it is evident that one tone has an F0 of 100 Hz (ACF lags of 10, 20, 30 ms) and the other, weaker tone has an F0 of 112.5 Hz (ACF lags of 8.9, 17.8, 26.7 ms). The F0s and relative levels of the two tones are less clear from the ACF at 1 kHz

intensities of the two sources in each channel can be recovered to some extent, and thus their respective power spectra. Assmann and Summerfield (1990) found some improvement in vowel identification using a model that worked in this way, although the improvement did not match that produced by human listeners. Furthermore, it seems doubtful that a segregation mechanism that relies on identification of two F0s could underlie human performance at this task, because it has subsequently been shown that humans are very poor at identifying both F0s from

these stimuli at ΔF_0 s smaller than four semitones (Assmann and Paschall 1998), whereas identification of the vowels reaches its maximum at only one semitone ΔF_0 .

A much better match to the data was achieved by Meddis and Hewitt (1992). Only one “dominant” F_0 was identified from the sum of the channel autocorrelation functions, and rather than parsing the energy in each frequency channel into two, they allocated whole frequency channels to one vowel or the other. If the most prominent autocorrelation peak in a given channel matched the period of the dominant F_0 , that channel was allocated to the first vowel, while the rest were allocated to the second vowel. This modeling method produced a satisfactory fit to the empirical data. Moreover, the use of one dominant F_0 seems consistent with the later finding that it is unnecessary for both vowels to have a regular harmonic structure (de Cheveigné et al. 1995). Nonetheless, the fact that the model reproduced the data does not necessarily imply that the model captured the underlying perceptual mechanism.

A troubling aspect of all autocorrelation-based mechanisms is that autocorrelation is better at identifying periodicity in high-frequency channels. Because these channels are broader, they admit a large number of frequency components that produce a well-modulated autocorrelation function with strong peaks at the F_0 (Fig. 3.1). In contrast, human performance seems more driven by low-frequency channels. Culling and Darwin (1993), using double vowels, and Bird and Darwin (1998), using competing speech, showed that ΔF_0 s need exist only in the first formant region for the greater part of the ΔF_0 effect to be observed. Several other possible mechanisms exist, but, as will be argued in the text that follows, most of them either conflict with the pattern of perceptual data or occur only in limited laboratory conditions.

3.2.3 *Temporal Analysis*

One alternative possibility was that listeners could separate the two vowels in the time domain. Because the repetition in a vocal waveform is generated by the release of glottal pulses, the energy of the voice tends to be concentrated into quite brief but regular intervals. When the F_0 differs between concurrent voices, the relative timing of the glottal pulses from each voice would be constantly shifting, so they would never consistently mask each other as they might if they stayed in synchrony. To look at this, Summerfield and Assmann (1991) used a double-vowel experiment in which the F_0 s of both vowels were the same, but the relative timing of the glottal pulses between the two vowels was varied. They found that when the glottal pulses of the two vowels were timed to alternate, a benefit in vowel recognition could occur, but only when the F_0 was 50 Hz and not when it was at a more realistic value of 100 Hz. It appeared, therefore, that although this mechanism worked in principle, the auditory system had insufficient temporal resolution to solve the task in this way at ecologically relevant F_0 s.

A second alternative was that waveform interactions between the two vowels could be exploited by the brain to help perform the task. The idea was that, for small ΔF_0 s, harmonic components of the two F_0 s would be very close in frequency and would beat together (i.e., produce amplitude modulation at the difference frequency) to some extent. This beating would create a constantly changing short-term spectrum. The fluctuating spectrum might either make the timbre of one vowel or the other more dominant at one moment or another, or might make the overall spectrum more characteristic of the particular vowel pair at some moment.

Culling and Darwin (1994) demonstrated that improvement in identification with ΔF_0 could occur even if harmonic frequencies were allocated in alternation to the two different vowels. This manipulation largely preserved the beating pattern, but would have disrupted any mechanism based on segregating harmonic components of one F_0 or the other. Moreover, they showed that computer models based on selecting moments of timbral clarity were able to predict the improvement on that basis. Meanwhile, Assmann and Summerfield (1994) found evidence that identification performance for each double-vowel stimulus could be based on one brief time frame within the stimulus. They showed that when double vowels were divided into 50-ms segments, the individual segment yielding the highest vowel identification produced equivalent identification to the complete stimulus. These studies together suggested that the steep increase in identification at very small ΔF_0 s in double vowel stimuli was probably based on this beating cue.

After publication of these studies, it became clear that the conventional double-vowel paradigm was too unrealistic and interest in it began to wane. New experiments were either based on connected speech or used adaptive threshold techniques to measure ΔF_0 effects. With connected speech, the intrinsically changing spectrum of the speech should mask any beating produced by harmonic components, while, using an adaptive technique, the difference in level that (usually) existed at threshold between the target and masking vowels should preclude the existence of any substantial beating between them.

3.2.4 Effects of Peripheral Nonlinearity

A third alternative was that nonlinearities in the auditory periphery were responsible for the effect of ΔF_0 . This mechanism again relied on the concentration of energy in the glottal pulse, but here the idea was that the compressive response of the basilar membrane would reduce the encoding of a more intense periodic masking sound such as an interfering voice.

Summers and Leek (1998) found that speech masked by a flat-spectrum harmonic masker produces an SRT up to 10 dB lower (better) when that masker had harmonic components in positive Schroeder phase than in negative Schroeder phase (Schroeder 1970). These maskers feature a particular phase for each component that results in a rapid frequency modulation of a fixed amplitude tone within a specified band. The significance of these phase structures is that although both of them have

relatively unmodulated acoustic waveforms, positive Schroeder phase results in a very pulsatile waveform on the basilar membrane whereas the resulting waveform from negative Schroeder phase remains relatively unmodulated (Kohler and Sander 1995). The effect on SRT was level dependent, which is consistent with the idea that it is related to cochlear nonlinearity. However, masker-phase effects occurred only at relatively high masker levels (Summers and Leek 1998) and relatively low masker F0s (Deroche et al. 2013), whereas effects of $\Delta F0$ have been measured across a wide range of these parameters. Consequently, it seems that other mechanisms of segregation by $\Delta F0$ must still exist. It is perhaps not surprising that the auditory system does not rely on the concentration of energy in the glottal pulse, because even small amounts of reverberation would produce temporal smearing of the previously discrete glottal pulses, which are only 5–10 ms apart, rendering any resulting cue unusable.

3.2.5 Cancellation Mechanisms

A final, alternative, mechanism focuses on the importance of masker harmonicity. De Cheveigné et al. (1995) found that identification of double vowels was dependent on the harmonicity of the masking vowel, rather than on that of the target vowel. This observation encouraged the idea that the main mechanism underlying the effect of $\Delta F0$ s is one of cancellation. Such a mechanism would naturally display a dependence on masker harmonicity alone, as only the masker need be cancelled. De Cheveigné (1998) proposed a model of such a process, in which the waveform in each frequency channel is subjected to *auto-cancellation* (point-by-point self-subtraction and summation) at a range of different delays. The residue from this cancellation at the delay of the masker F0 will reflect only the energy of the target, albeit with any energy at frequencies shared with the masker removed. This mechanism is a form of “comb” filtering, where the frequency response is low at multiples of F0, and high at frequencies in between.

The key evidence in favor of cancellation mechanisms is the effect of disturbing the harmonicity of a masking complex, typically by introducing frequency offsets for each frequency component. An interesting caveat to this technique was pointed out by Deroche et al. (2014). They noted that when the frequency components of a harmonic complex are disturbed from a regular sequence in this way, the resulting excitation pattern is reduced in mean level in the region of resolved frequency components. As the complex becomes inharmonic, some components become clustered together while others are more widely separated. In the excitation pattern, this clustering results in higher peaks and deeper troughs, but, on average, the troughs deepen more than the peaks climb (Fig. 3.2), resulting in lower overall excitation and lower overall masking power for an inharmonic complex compared to a harmonic complex, even though they have the same number of equal-amplitude components. Hence, for the benefit for a harmonic masker to be observed, this intrinsic advantage for inharmonicity must first be overcome.

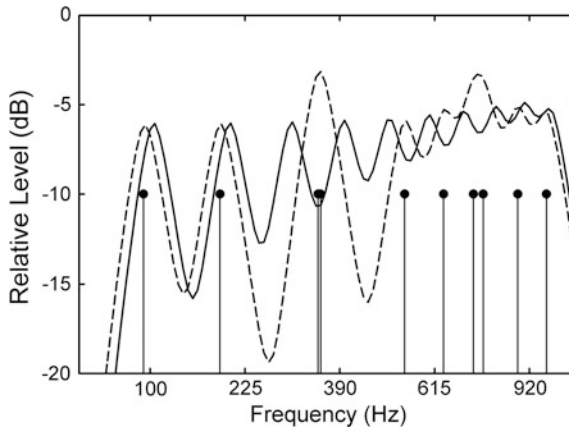


Fig. 3.2 Cochlear excitation patterns for harmonic and inharmonic complex sounds. The solid line is an excitation pattern for a harmonic complex and the dashed line the pattern for an example inharmonic complex with the same number of equal-amplitude components. The vertical lines show the line spectrum of the inharmonic complex. Because the third and fourth components of the inharmonic complex are close together, there is an enhanced peak in the excitation pattern at that frequency, but deep troughs are also formed on either side of this cluster

Deroche and Culling (2011a) demonstrated a similar dependency on masker harmonicity using connected speech and a more realistic form of inharmonicity. They independently varied the level of reverberation applied to target speech and a masking complex tone. Such variations could occur naturally if these sources were at different distances from the listener in a reverberant space. The F0s of the speech and the tone were either monotonized or modulated sinusoidally, such that the combination of modulation and reverberation would cause the resulting sound to be inharmonic. The inharmonicity occurs because the myriad reflected sound paths in a reverberant room are all of different lengths and so introduce differently delayed copies of the source waveform. Since the source F0 is modulated, the variously delayed copies all have different F0s by the time they reach the receiver. SRTs were elevated by about 6–8 dB where the combination of modulation and reverberation was applied to the masking complex, but there was no such modulation \times reverberation interaction for the target speech.

3.2.6 *Level of Processing*

It seems unlikely that the very low-level mechanism of peripheral compression is solely responsible for the effect of $\Delta F0$. The mechanisms for selecting energy at a given F0 described in Sects. 3.2.2 and 3.2.3 all presume that the F0 is first identified. A parsimonious assumption would be that this identification process is the

same as the process by which a pitch percept is derived from that F0. The cancellation mechanism (Sect. 3.2.5), on the other hand, does not necessarily have to identify an F0. For instance, it might work to suppress the dominant periodicity in each frequency channel regardless of the strongest period detected elsewhere.

This issue, as well as the question of how effective the mechanism was at different frequencies, was explored by Deroche and Culling (2011b). They measured MDTs for narrow bands of noise centered on fixed-frequency components of a masking complex. The SNR in the target frequency channel was thereby kept approximately constant, while the harmonicity of the rest of the harmonic complex was manipulated. A difference in MDT of 3–5 dB was found between cases where the masker’s overall structure was harmonic or inharmonic, so the process was not entirely determined by a single-frequency channel centered on the target. These differences occurred for target frequencies up to 2.5 kHz, but were negligible at higher frequencies. By varying the band of masker frequencies that were harmonic rather than inharmonic, it was found that the effect increased progressively the more masker components were harmonic, although the harmonicity of components close to the target frequency appeared to exert greater influence.

Further work will be required to ascertain whether these basic psychoacoustic phenomena can account for the effects of $\Delta F0$ in speech stimuli. Nonetheless, these results suggest that the mechanism at work does extract information about the dominant F0 using a wide range of frequency channels consistent with a pitch identification mechanism. Unfortunately, the level of processing at which pitch is determined from F0 and, with it, the level of processing at which segregation by F0 may occur, remains uncertain.

3.2.7 *Conclusions*

In many ways, the literature on the effect of $\Delta F0$ has been stalled over the question of mechanism. A number of plausible mechanisms have been discussed: each of them is supported to at least some extent by empirical data. It is probable that each of them plays some role in certain circumstances, but, for most of them, it seems likely that those circumstances are quite limited. The cancellation mechanism is the only one that appears to combine consistency with the data and the potential for robust application to real-world scenarios. Our knowledge of this mechanism is very limited, however. Its existence is largely confirmed by eliminating the other possibilities. Only recently have data been collected that can characterize its behavior. More data are needed to guide the design of models, because without a model that can predict a substantial body of empirical data, the contribution that $\Delta F0$ s make to unmasking in real-life listening situations is difficult to determine.

3.3 Masking and Masking Release by Envelope Fluctuations

It is usually easier to understand speech in background noise if modulation is applied to the background noise. This effect was first observed by Miller (1947) and Miller and Licklider (1950) using square-wave modulation of the interfering noise, which produced what they termed “interrupted” noise. Miller and Licklider (1950) found that speech intelligibility at negative SNRs was optimal at noise interruption rates of 10 Hz (see their Fig. 8). This effect is thought to occur through some form of selective processing of the target sound during the silent phases of the interfering noise when the SNR is momentarily more advantageous; this is also referred to as “dip listening.” The process of modulating the noise does not change its long-term spectrum (provided certain trivial constraints are met). For that reason, a simple comparison of signal and masker spectra, as in the SII model (ANSI 1997), is not a sufficient explanation for the improved performance.

The effect of interrupting the noise can be very large. De Laat and Plomp (1983) found a difference in SRT of 21 dB between continuous and 10-Hz interrupted noise with a 50% duty cycle (i.e., 50 ms on, 50 ms off). However, at a real cocktail party, interfering speech will vary in intensity in a more erratic and less profound way, with a peak modulation frequency of around 4 Hz, which corresponds to the syllable rate (Plomp 1983). In addition, the dips are less correlated across frequency. Festen and Plomp (1990) measured SRTs in speech-modulated noise to evaluate the potential benefit of the intrinsic modulation of a speech interferer. Speech-modulated noise was created by extracting the temporal envelope of wideband speech and using it to modulate a speech-spectrum-shaped noise carrier. The resulting noise contained variable levels of dips but co-timed across a frequency band (either one, full-spectrum band, or two, high- and low-pass bands). SRTs in this noise were only about 4 dB better than in continuous speech-shaped noise.

To examine the effect of co-timing of the dips in modulated noise, Howard-Jones and Rosen (1993) generated a masker in which adjacent frequency bands were modulated in alternation rather than together. They applied square wave modulation in multiple bands to a continuous pink noise, where the square wave was in anti-phase in adjacent bands, to produce what they termed “checkerboard” noise (so called because of the patterning of the resulting spectrogram). Using a modulation frequency of 10 Hz, they found that the benefit of the dips decreased as the number of bands increased. There was negligible benefit for checkerboard modulation in eight bands or more. The data for multiple bands of speech modulated noise are not yet so clear. Festen and Plomp (1990) found no difference in SRTs for sentences presented against one- or two-band speech-envelope modulated noises. Extending their work to a larger number of bands carries the problem that the noise starts to carry intelligible phonetic information when multiple channels are processed in this way.

3.3.1 *Listening in the Dips*

To exploit dips in the interfering noise, one would think that it would be necessary to determine in some way when the masking noise has dipped, so that the moments of improved SNR can be exploited. However, the benefit of dip listening has been modeled quite successfully without taking any account of the timing of masker peaks or dips. Rhebergen and Versfeld (2005) applied the SII measure to individual time frames of the stimulus. Deriving the SII normally involves measuring the long-term SNR in a series of audio frequency bands, limiting the range of measured values to ± 15 dB (which is intended to reflect the range of speech that is both potentially audible and contributes to intelligibility), and weighting the resulting values with a weighting function that reflects the relative contribution of different frequency bands to speech understanding. The use of short time frames permitted Rhebergen and Versfeld to produce a time series of short-term SNRs in each frequency band. The durations of the time frames, approximately 9–35 ms, varied across frequency to mimic the variation of the ear’s temporal resolution across frequency and were sufficiently short for this frame-based measure of SII to fluctuate with the dips. Their “extended SII” was formed by taking the average SII of a time sequence of such values. The relative success of this revised model therefore did not need to assume any selective processing of the stimulus during masker dips to capture the basic effect.

Some aspects of the data were not so well captured by Rhebergen and Versfeld’s model, though. Sinusoidal intensity modulation at various rates produced SRTs whose variation was anticorrelated with the model’s predictions; high rates of modulation (up to 32 Hz) gave lower SRTs than lower rates of modulation (down to 4 Hz), while the model predicted the reverse. In addition, it seems likely that a form of selective processing has to be involved, because familiar masker modulation functions that are based on the same speech envelope, offering the listener an opportunity to learn the moments when the SNR will be high, are less disruptive than novel functions that are based on a new speech envelope for each trial and consequently produce less predictable variation of the instantaneous SNR (Collin and Lavandier 2013).

3.3.2 *Effects of Peripheral Nonlinearity*

It is possible that peripheral compression may also play a role. As with the effect of ΔF_0 , peripheral compression can lead to a reduced response at points in time when the masker is intense compared to when it is weak. However, evidence for such an effect is mixed.

Listeners with hearing loss have less compression, so one would expect them to display a reduced benefit from dip listening. Although this is indeed found (de Laat and Plomp 1983), such observations may arise from an experimental confound

(see Sect. 3.3.6), and more analytic experiments fail to show a direct link. For instance, Stone et al. (2012) tested the relative importance to intelligibility of peaks and dips in the masker. They applied a form of distortion to a speech-plus-interfering-talker signal to different degrees in the dips of the signal relative to the peaks of the signal. When the dips were selectively distorted, intelligibility dropped. By measuring intelligibility as a function of the signal level at which the distortion was applied, they could map out the range of levels that contributed to intelligibility. They found that listeners with moderate hearing loss were able to benefit from a relatively undistorted signal over a similar magnitude of dynamic range to that found in listeners with normal hearing (Stone et al. 2010), but shifted 3 dB higher in level, possibly indicating a form of internal distortion. The similarity of range observed in these two listener groups indicates that peripheral nonlinearity is not an important contributor to masking release.

3.3.3 Modulation Masking

Although dips in interfering noise power have been found to be beneficial, it has become clear that the modulations of a masking noise can also be detrimental. Modulation masking has not usually been observed in psychoacoustic experiments, due to the pervasive use of noise maskers. Noise maskers contain intrinsic modulations that themselves produce modulation masking so, in these experiments, no baseline existed against which its effects could be observed. Modulation masking works by obscuring modulations of the signal. Indeed, when masker modulation is less prominent, a modulated signal is quite easy to detect. Buus (1985) reported that two narrowly spaced sinusoids presented in a narrowband noise masker were easier to detect than a single sinusoid of the same energy. The close spacing of his two sinusoids relative to the width of the auditory filter would imply little change in the excitation pattern compared to that of the single sinusoid, and therefore spectral equality for the two targets. Buus obtained a release from masking of up to 25 dB when detecting the two sinusoids, compared to the single sinusoid. The primary difference between the two targets was in their modulation spectrum; the two sinusoids produced a strong beating sensation that was not easily masked by a narrowband masker with its relatively smooth temporal envelope.

Kwon and Turner (2001, expt. 2) showed that, in some circumstances, applying modulation to a masking noise could *increase* masking of speech information rather than reduce it. They found that when bands of noise and speech were in the same frequency region, modulation of the noise released masking, but when the noise was in a different frequency region to the speech, modulation increased masking. Modulation in these remote frequency regions was interfering with the detection of modulations within the speech. The interplay between these two conflicting processes of masking release and modulation masking is still being explored.

3.3.4 *Intrinsic Modulation in Noises*

Nearly any effort to quantify modulation masking is confounded by a baseline modulation intrinsic to the waveform of the masking noise. In many measures of speech intelligibility it has been common to use a random noise, usually with a Gaussian amplitude distribution, as the masker. Even though that noise itself contains appreciable envelope fluctuations, masking release is usually quantified by comparing speech intelligibility in such a noise with intelligibility in the same noise but with a deterministic modulation applied, such as a low-rate sinusoid. One approach to analyzing the effects of modulation masking has been to try to remove the amplitude modulation intrinsic to the Gaussian noise so that any observed modulation masking or masking release is due primarily to the applied modulation in the test condition. The desired characteristic of the reference noise is that it should have a relatively flat temporal envelope, having a low “crest factor,” defined as a low ratio of the peak sample value to the root-mean-square value.

As mentioned in Sect. 3.2.4, Schroeder-phase complexes have a low crest factor, although only negative Schroeder phase is thought to retain its low crest factor when it reaches the basilar membrane. However, stimuli generated with this technique produce several effects in the peripheral auditory system that make their use unsuitable for examining the effects of modulation masking. A notable consequence of these effects is that positive Schroeder phase, which is more modulated on the basilar membrane, gives *lower* tone detection thresholds (i.e., less masking) than the less modulated negative Schroeder phase. Several other methods have been developed for producing masker waveforms with minimal intrinsic modulation, generally involving gradient descent or iterative methods to adjust component timings, while preserving a random structure (Pumplin 1985; Hartmann and Pumplin 1988; Kohlrausch et al. 1997).

Hartmann and Pumplin (1988) generated maskers using the “low-noise noise” (LNN) technique of Pumplin (1985). For relatively narrowband maskers, they observed that MDTs for pure tones were about 5 dB lower in their LNN than in Gaussian noise with the same spectrum. Analysis of the results from a similar experiment by Kohlrausch et al. (1997) showed how the addition of the target tone modified the modulation spectrum of the masking noise. This is shown in Fig. 3.3.

The dark green line shows the modulation spectrum for a 100-Hz-wide band of Gaussian noise (GN). GN has a modulation spectrum that decreases with increasing modulation frequency up to the bandwidth of the noise. The dark purple line shows the modulation spectrum for a similar band of LNN, whose modulation spectrum increases with increasing modulation frequency up to the bandwidth of the noise (the higher levels at low modulation rates are a consequence of the signal duration used in the experiments, here 250 ms). Counterintuitively, the addition of the *unmodulated* target tone at the discrimination threshold measured by Kohlrausch et al. (1997) increases the modulation spectrum for rates up to half the noise bandwidth: the modulation spectrum of the GN increases slightly in level (light green line), but the modulation spectrum for the LNN increases markedly (light purple line). This

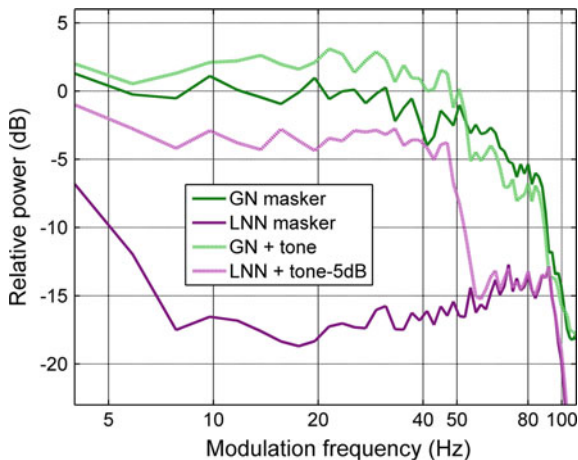


Fig. 3.3 Modulation spectra plots for the signals used in Kohlrausch et al. (1997). The dark-colored lines show the spectra for noise with Gaussian statistics (GN, dark green) and with low-noise statistics (LNN, dark purple). The light green and light purple lines show the spectra for the same noise signals but with a tone added at the center frequency of the noise, with the same relative level as measured in the discrimination thresholds reported in Kohlrausch et al. Discrimination of the tone in LNN from LNN alone occurred with a tone level 5 dB less than the level of the tone required to achieve discrimination of a tone in GN from the GN alone

occurs because the LNN has frequency modulation but little amplitude modulation; when the tone is added, the phase relationships that suppress the amplitude modulation are disrupted and amplitude modulation is reintroduced. Hence the low tone thresholds in LNN observed by Kohlrausch et al. (1997) were interpreted as detection of increased amplitude modulation from a very low baseline. The greater intrinsic modulation in the GN may have masked changes in modulation caused by the addition of the tone, resulting in relatively elevated thresholds for the GN condition.

For their broader bandwidth maskers, these effects disappeared (Hartmann and Pumplin, 1988), because their narrow bandwidth masker was within the span of the auditory filter centered on the masker. The broader bandwidth maskers spanned more than one auditory filter, thereby destroying the amplitude and phase relationships of components that gave rise to the low-noise property of the masker. Consequently, while the selection of appropriate phases can produce a noise-like stimulus that has a low crest factor, this low crest factor is not preserved in sub-bands of the stimulus, which retain Gaussian characteristics. As a result, it is very difficult to create a noise that has a low crest factor in every auditory frequency channel, a property necessary to examine the influence of modulation masking for a broadband signal like speech. However, Hilkuysen and Machery (2014) developed a technique that does just that, with a crest factor intermediate between those of Gaussian and narrowband LNN within all individual frequency channels. This

technique offers promise of possibly more accurate simulation of cochlear implant processing, and the potential for reduced development time of prototype algorithms.

Stone and Moore (2014) used an alternative strategy to produce a broadband masker with a low crest factor. Instead of taking a noise masker and processing it to reduce modulations, they generated a complex tone with sparsely and inharmonically distributed frequency components. The distribution was performed according to the frequency analyzing power of the cochlea (Glasberg and Moore 1990), with one component for each consecutive auditory frequency channel. An auditory filter centered on one of these components would also transmit some of the adjacent components but at a lower level, because these adjacent components would pass through the edges of the auditory filter. The presence of multiple components in the filter output would give rise to a fluctuation in the envelope of the signal, due to the beating between the components, which is a form of modulation. To minimize further the interactions between adjacent components and reduce this modulation, odd-numbered components were presented to one ear and even-numbered to the opposite ear. By this process, they constructed a masker that primarily produced “pure” energetic masking, without the confound of introducing intrinsic modulations, which would otherwise produce modulation masking, as occurs with the conventional use of random noises. The speech signal was filtered into narrow bands centered on each masker component and similarly divided between the ears, directing odd channels to one ear and even channels to the other. Speech intelligibility was measured with the maskers either unmodulated or sinusoidally modulated at different rates. This modulation of their energetic masker could transform it into a modulation masker or create opportunities for dip listening, but primarily at only one modulation frequency. They found that masker modulations at rates of 2 Hz and greater elevated SRTs rather than reducing them, indicating modulation masking; only modulation at rates of 1 Hz produced an improvement in SRT similar to that from interrupted-noise experiments. Consequently, only in a very restricted condition were they able to demonstrate a classic dip-listening effect. Since they showed that masking by random noise was dominated by modulation masking, one can infer that some benefits previously described as “masking release” were release from the modulation masking, although the mechanism by which this release occurs is unclear. The data of Stone and Moore (2014) also permitted an estimate of the relative effectiveness of modulation versus “pure” energetic masking of speech: for equal-energy maskers of similar spectral shape, the modulations inherent in a Gaussian noise were about 7–8 dB more effective at masking than their pure energetic masker.

3.3.5 Models Based on Modulation Filter Banks

To account for the effects of modulation masking, Jørgensen and Dau (2011) presented a model to predict speech reception in noise based on a modulation filter bank. Like the SII, this model calculates the long-term difference in the power

spectrum between the speech and speech+noise signal, but this time in the modulation frequency domain. The model separately analyzes the modulation power produced by either the noise or by the speech+noise within each of 22 audio frequency channels. The modulation power is calculated in each of seven modulation bands for each frequency channel, to form a two-dimensional array of speech-to-noise power ratios. A “modulation SNR” is subsequently calculated by taking the square root of the sum of all 154 elements in this array. Predictions of resulting intelligibility, in percent correct, are obtained by transforming this modulation SNR via a nonlinear function that incorporates measures of the difficulty of the speech material used in the experiment. They showed that this model could accurately predict the effects of interfering noise, a noise-reduction algorithm, and reverberation.

Jørgensen and Dau’s (2011) model was thus rather successful at predicting modulation masking, but real-life listening is likely to involve a mixture of modulation masking and dip listening. To address this problem, Jørgensen et al. (2013) developed a version of the model that operated in short time windows in a similar way to Rhebergen and Versfeld’s (2005) extended SII model. The newer model was quite successful in predicting effects of both modulation masking and dip listening within the same framework.

3.3.6 Dip Listening in the Hearing Impaired

Although dip listening has been regularly observed when using listeners with normal hearing, observation of the same effect with listeners with impaired hearing has been elusive. Festen and Plomp (1990) reported that, for listeners with a moderate degree of sensorineural hearing loss, there was no appreciable dip-listening effect when speech was presented in noise maskers modulated with either one- or two-band speech envelopes compared to the unmodulated noise alone. Nelson et al. (2003) also reported a failure to observe dip listening in users of cochlear implants, even when using square-wave-modulated noise, the modulation pattern with the easiest of dips to glimpse. Many other researchers have similarly replicated these findings using a variety of modulated interferers.

One explanation of this problem was proposed by Oxenham and Simonson (2009). They noted that even when using listeners with normal hearing, a dip-listening effect was only observed for SNRs of 0 dB or less, and larger dip-listening effects were observed for more negative SNRs. Experimenters comparing performance between listeners with normal hearing to those with impaired hearing usually compared measured SRTs from both groups, but those with hearing impairment required higher SNRs, around 0 dB, to reach the usual criterion of 50% for SRT. Consequently, the higher SNR at which their SRTs were measured might explain the lack of an observable dip-listening effect in hearing impaired listeners.

To explore this possibility more thoroughly, Bernstein and Grant (2009) enabled their impaired listeners to understand speech at negative SNRs by presenting the

target speech audiovisually to facilitate lip reading. Once an SRT could be measured at negative SNRs, they observed dip listening with their hearing impaired listeners. However, the magnitude of the dip-listening effect was still 1–5 dB less for these listeners than for the normal-hearing listeners, when referenced to performance at the same SNR in the steady noise. This residual difference suggests that other deficits associated with hearing impairment may also be affecting their performance. The previous failures to observe any dip-listening effect in hearing impaired listeners were therefore largely due to a confound in the experimental design rather than a deficit of impaired hearing.

Bernstein and Grant offered a model to explain their results, in terms of the Intensity Importance Function (IIF) (Studebaker and Sherbecoe 2002). The IIF is based on the observation that speech is a signal that fluctuates about its mean level. Although the fluctuations contribute to intelligibility, fluctuations at different levels do not all contribute equally. The IIF describes the relative contribution of each level to the overall intelligibility. As the low-level parts of speech commonly comprise the decay from voicing or frication or reverberation, they do not carry much information. Similarly, a peak, by definition, is prominent from the remainder of the speech, and so is very unlikely to be masked. For levels in between valley and peak, the IIF rises to a peak. The explanation by Bernstein and Grant is illustrated in Fig. 3.4. In each of the two panels, the bell-shaped curve denotes the IIF for speech presented in a speech masker (as measured by Stone et al. 2010). Levels near 0 dB relative to the signal RMS provide the most information, while levels further away, above or below RMS, contribute less to intelligibility. In both panels the thick dash-dotted line, labeled N_{RMS} , denotes the RMS of an added noise. If the noise is continuous, with only minor fluctuations, speech levels below this line are masked and contribute no information to intelligibility. The contribution from information in speech levels exceeding this line produces the measured intelligibility. For speech presented in noise at a negative SNR, top panel, only a small part of the total distribution is unmasked. At an SNR of 0 dB, lower panel, more than half of the distribution is unmasked. Now consider the situation when the noise contains major fluctuations, but still has the same RMS as the continuous noise.

Occasionally there will be noise peaks that prevent access to the information in the light gray areas, while the noise dips will unmask some speech permitting access to information in the medium gray areas. The benefit, or otherwise, of these fluctuations relative to the continuous noise can be seen to be a trade-off between the losses (the area under the curve in the light gray areas) and gains (area under the curve in the medium gray areas) in information. At negative SNRs (Fig. 3.4, top) the losses are much less than the gains, so intelligibility improves in a fluctuating noise relative to that in the continuous noise. As the SNR increases toward 0 dB, lower panel, the losses become more similar in area to the gains, so there is decreasing benefit from the fluctuations. As mentioned previously, IIFs measured for hearing-impaired listeners (Stone et al. 2012) place the peak at a slightly higher level than found for normal-hearing listeners (Stone et al. 2010). Consequently, in the lower panel, the IIF could even be displaced to the right for hearing impaired

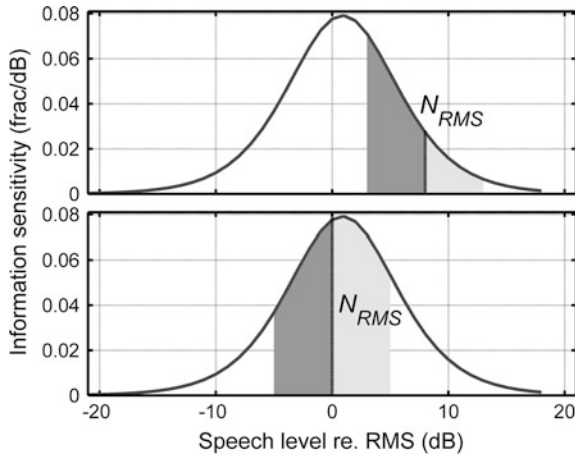


Fig. 3.4 An explanation for the observation of reduced masking release for SNRs approaching 0 dB. In both panels, an example IIF for speech is shown as a curved solid line, while continuous noise is denoted by the vertical line, labeled N_{RMS} . For a modulated noise with the same RMS, peaks will mask speech information in the light gray areas while dips will unmask speech information in the darker gray areas. In the top panel, with a negative SNR, more information is unmasked than masked so positive masking release is observed. In the lower panel, with an RMS of 0 dB, the converse holds. (Adapted from Bernstein and Grant 2009.)

listeners. It should be noted that the 0-dB SNR dip-listening observation by Oxenham and Simonson (2009) appears to hold true for full-bandwidth unmanipulated speech. However, this may not be so for manipulated signals. For example, Christiansen and Dau (2012) observed masking release in high-pass filtered speech +noise where the SRT in continuous noise was at a positive SNR.

3.3.7 Conclusions

For speech presented in realistic noises, such as the babble of a cocktail party, the masking release gained by the listener is primarily a trade-off between the modulation masking produced by the noise and the benefit of access to the target speech at better SNRs produced by the spectrotemporal dips in the noise. Intelligibility is thus driven not just by SNR in the audio frequency domain but also by SNR in the modulation domain. Models of masking that rely only on spectral differences to predict intelligibility are thus of limited applicability owing to the variety of possible real-world interference. It has become increasingly apparent that consideration of the modulations within and across audio frequency bands is crucially important. Although a dip temporarily permits the target speech to become audible at a more advantageous SNR in the audio frequency domain, a dip implies modulation, which implies a less advantageous SNR at some rates in the modulation domain. When the

masker contains minimal fluctuations, introduction of dips does not improve intelligibility, except at very low modulation rates.

The frequently reported failure to observe dip listening in hearing impairment is largely the result of a confound in experimental technique, rather than the hoped-for doorway to understanding hearing impairment and possible remediation.

Finally, the experiments with maskers generated with low intrinsic modulations show that the previously assumed energetic masking produced by Gaussian noise maskers is not pure, but additionally contains a powerful component of modulation masking.

3.4 Spatial Release from Masking

Speech is easier to understand in background noise when speech and noise are spatially separated than when they are in the same location. The difference in SRT between these two cases is known as the spatial release from masking (SRM). SRM itself cannot be categorized as either energetic or informational, but two of the component processes, better-ear listening and binaural unmasking, can reasonably be categorized as energetic masking release. Other components of SRM include spatial release of informational masking (Kidd and Colburn, Chap. 4), and spatial segregation of competing sequences of sounds (Middlebrooks, Chap. 6). Our understanding and characterization of better-ear listening and binaural unmasking are relatively well developed, so this section will focus on an exposition of these mechanisms, as well as identifying areas in which their characterization could be improved.

Better ear listening and binaural unmasking are linked to the interaural differences that occur when a sound comes from one side: interaural level differences (ILDs) and interaural time differences (ITDs). Although ILD and ITD cues are used to localize sound sources, localization does not seem necessary for unmasking to occur (Edmonds and Culling 2005). Instead, localization and unmasking seem to operate independently through quite different mechanisms.

It is possible to separate the effects of better-ear listening and binaural unmasking experimentally. In one approach, the stimuli are processed to remove either ILDs or ITDs and isolate the remaining one (Bronkhorst and Plomp 1988). In a second approach, it is assumed that the effect of better-ear listening can be isolated by presenting the signal+noise stimulus to one ear only, while the effect of binaural unmasking can be isolated by looking at the difference in SRT between this condition and a fully binaural condition (Hawley et al. 2004). This latter technique makes a strong theoretical assumption that the effects of better-ear listening and binaural unmasking are additive. The studies based on the ILD/ITD removal technique both suggest that the effects are not quite fully additive, but models that assume additivity have been quite successful in reproducing the overall patterns of data (e.g., Jelfs et al. 2011).

3.4.1 *Better-Ear Listening*

Better-ear listening occurs when the spatial separation of target and interfering sounds results in different SNRs at each ear. The differences in SNR at the ears arise from the effect of the head on the sound levels of both the target and the interferer. Consider a target sound on one side of the head and an interfering sound on the other. At the ear on the target side, the target level will be enhanced somewhat by the reflection of sound from head, while the interfering sound will be attenuated due to the head shadow, both effects combining to increase the SNR. Conversely, SNR is worsened by these mechanisms at the ear on the side of the interfering sound. If the listener is able to take advantage of an improved SNR at the better ear, without being affected by the worsened SNR at the other ear, a release of energetic masking occurs. In a sense, better-ear listening occurs before the signal enters the auditory system, because the improved SNR is already present at the ear. While better-ear listening may appear a trivial case of SRM, there are still some questions about how higher-level processes exploit differences in SNR between the ears.

When speech and noise are separated in space, the SNR will always improve at one ear, but it will usually worsen at the other ear. It seems that the ear with the poorer SNR does not impair speech understanding. Rather, binaural presentation invariably improves signal detection and identification over monaural presentation. The question arises, therefore, of how the brain selects the appropriate ear. Several possibilities exist. The brain may select the ear with the better SNR, in which case one must answer the question of how it identifies which ear that is. Alternatively, it may employ perceived sound lateralization to determine on which side of the head the target sound source is located. There are some interesting spatial configurations in which these two strategies would lead the listener to employ different ears, and these configurations could be used to differentiate between these possibilities. The brain might also select the ear at which the target sound source is most intense, regardless of the sound level of the masker. Finally, there may be no selection process as such, but the information from the two ears may always be gathered independently and then integrated; the ear with the poorer SNR provides a “second look” that adds to the total information provided by the two ears.

There is some evidence that listeners do integrate information gathered independently at the two ears. When the two ears receive identical stimuli, there is sometimes a measurable advantage, termed “summation,” in detection or identification of signals in noise (e.g., Bronkhorst and Plomp 1988), but it is not always observed (e.g., Hawley et al. 2004). Over and above this effect, when signals are masked by independent noise at each ear, there is always a binaural benefit. The latter phenomenon has also been attributed to the auditory system getting a “second look,” but other interpretations in terms of binaural unmasking mechanisms are also possible. These observations imply that different information can be gathered at the two ears, but that some sort of selection process is also involved.

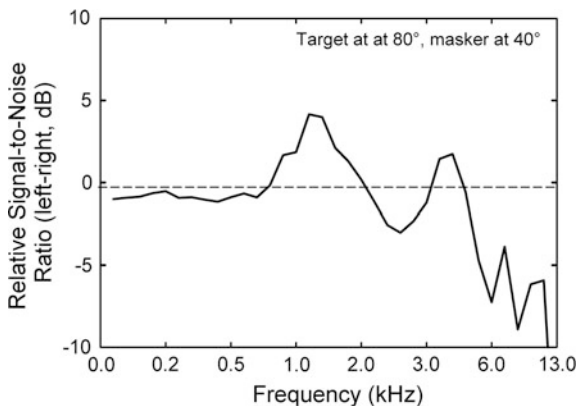
Culling and Mansell (2013) tested how quickly any selection process might adapt to changing circumstances by making the better ear switch back and forth

from one side to the other in a repeating cycle. Such a situation can occur naturally in circumstances where there is more than one independently fluctuating interferer in different spatial locations. They found that SRTs were strongly dependent on the switching rate. Because the information available was not changing as a function of time, but only the ear at which the best information was supplied, these results suggest that there is a selection mechanism, but that it is quite “sluggish.” It is able to switch between ears only up to a certain rate.

A second question concerning better-ear listening relates to integration of information across frequency. When sound diffracts around the head and its appended structures, such as the pinnae, there is a complex, frequency-dependent pattern of constructive and destructive interference over its surface. This pattern varies with both azimuth and elevation of the source relative to the head. As a result, although the effect of head-shadow tends to increase with both frequency and with sound source azimuth, these trends vary erratically. These frequency-and-location-dependent variations raise the possibility that the better ear at one frequency may not be the better ear at another. Figure 3.5 shows that, for a scenario in which a speech target is at 80° and a speech-shaped masker is at 40° azimuth, the relative SNR at the two ears switches several times as a function of frequency between favoring either left or right. This scenario is not uncommon. It tends to occur whenever target and masker are in different locations on the same side of the head. Reverberation can also give rise to effects of this sort. One can therefore ask whether the process of selecting the better ear selects that ear at all frequencies, or whether a “within-channel” process is deployed for which processing is independent in each frequency channel and different ears can be selected at different frequencies.

Edmonds and Culling (2006) examined whether the binaural system was capable of using a different better ear within different frequency bands. Using speech from two male talkers as target and masker, they split the frequency spectrum of both the target and the masking speech at each of three different frequencies and distributed the resulting bands to different ears. In three separate conditions, the target and

Fig. 3.5 Long-term average signal-to-noise ratio at the left ear compared to the right as a function of frequency when a speech target is at $+80^\circ$ azimuth and a speech-shaped noise masker is at $+40^\circ$ azimuth



masking speech bands were either directed to the same ear, consistently to different ears, or “swapped,” such that the low frequencies of the target speech were at one ear and the high frequencies at the other ear and vice versa for the masking speech. SRTs in the swapped condition were intermediate between those for the other two conditions, indicating that the auditory system was not able to gather all the information available from the swapped condition. A second experiment showed that if the two speech sources were directed to different ears only in one of the two frequency regions, and mixed at the same ear in the other, SRTs from the swapped condition were always matched by one of these conditions. This result suggested that listeners were selecting information from the same ear at all frequencies. They seemed to be able to identify the optimum ear to select, but not to select different ears at different frequencies. Further experiments with noise interferers should clarify whether this result holds for interfering noise rather than speech.

3.4.2 *Binaural Unmasking*

Binaural unmasking occurs when the interaural phases of the target and the masker differ from each other. The interaural phases of each source are introduced by the ITDs that occur when there is a difference in path length from that sound source location to each ear. If two sound sources are in the same location, their interaural phase differences will be identical. However, if they are spatially separated, the differences in path length to each ear will be different for each source location. The processing of these interaural differences to improve signal detection was first observed in pure-tone detection by Hirsh (1948), who used the simple expedient of inverting the waveform (a 180° phase shift) of either a tonal signal or a masking noise at one ear. This observation was rapidly extended to speech using a similar technique (Licklider 1948). For tone detection in broadband noise, MDTs are up to 15 dB lower when the two sources differ in interaural phase. The improvement in MDT is known as the binaural masking level difference (BMLD). The largest BMLD at any frequency occurs when the noise is identical at the two ears (diotic) and the signal has a 180° (π radians) phase difference. This condition is often termed N_0S_π in the literature. Importantly, N_0S_π is also the optimal configuration when measuring the intelligibility of a speech signal (e.g., Schubert 1956), despite the fact that such a stimulus is entirely artificial and can be created only using headphone presentation. It is thus the relative interaural phase that determines the magnitude of these effects, rather than the relative ITD.

Binaural unmasking clearly involves some processing by the brain. There is, of course, the question of what this process is, but this question has been extensively reviewed elsewhere (Colburn and Durlach 1978; Colburn 1996). For the present purpose, it is sufficient to note that the equalization–cancellation (E–C) theory (Durlach 1963, 1972; Culling 2007) seems to provide a sufficient framework to account for the principal effects of binaural unmasking and has been a popular choice among those who have sought to model the unmasking of speech (e.g.,

Beutelmann et al. 2010). According to E–C theory, the stimulus at the two ears is compared centrally within each frequency channel, delayed and scaled internally, such that the masker waveforms at the two ears are optimally aligned in time and amplitude (equalization), and then the two waveforms are subtracted one from the other (cancellation). If a signal with a different interaural phase/delay is present in the same channel, then some of its energy will survive this process, whereas that of the noise will be largely removed. According to E–C theory, a π -radians difference in interaural phases (0 vs. π) produces the greatest unmasking because it maximizes the signal residue.

Because the equalization process involves the determination of an internal delay, one can ask whether the E–C process requires consistent interaural differences across frequency, or whether it operates within channels. In contrast to the findings with better-ear listening, which seemed to require the SNR to be better at the same ear across frequency, it appears that binaural unmasking does not require any consistency across frequency in the ITD of the noise. The earliest evidence for this comes from the very first study of binaural unmasking for speech. Licklider (1948) found that the $N_{\pi}S_0$ condition, in which the noise is inverted at one ear relative to the other, produced a strong masking release for speech, despite the fact that the phase inversion of the noise requires a different equalization delay in every frequency channel. Moreover, Edmonds and Culling (2005) measured SRT for target speech masked by either noise or by a competing voice. In either case, applying different ITDs to low- and high-frequency bands of the masker and the target speech had no detrimental effect on SRTs. In particular, they found that SRTs were the same whether the two bands of speech had the same ITD and the masker a different ITD, or whether the low frequencies of the masker shared the same ITD as the high frequencies of the target speech and vice versa. These data suggest that the factor that determines the degree of unmasking is the difference in interaural phase within each frequency channel, and not the relationships between the phases/ITDs in different frequency channels.

3.4.3 *The Problem of “Sluggishness”*

A number studies have shown that the binaural system is, in various contexts, slow to adapt to changes in the configuration of the stimulus. This slowness to react to change is in marked contrast to the very high temporal precision (tens of microseconds) which the binaural system uses to, for instance, detect ITDs (Klumpp and Eady 1956). In the context of binaural unmasking, Grantham and Wightman (1979) found that the MDT of a pure tone against a masker with sinusoidally varying interaural correlation increased steeply with modulation frequency, and the effects of binaural unmasking were all but eliminated at a modulation rate of just 2 Hz. One way of thinking about this sluggishness is to suppose that the very high temporal precision displayed by the binaural system demands a long integration time to gather sufficient information. This long integration time, of

the order of 100 ms, has been termed the “binaural temporal window” (Culling and Summerfield 1998). This coarse temporal resolution raises questions about how the binaural system is able to recover speech from noise.

If the shortest time interval within which binaural unmasking can report information about a signal in noise is 100 ms, then any modulations in the intensity of the signal during that time will be temporally smeared at the output of the process. Because the temporal modulation of speech is crucial to intelligibility (Houtgast and Steeneken 1985), such smearing would potentially undermine the ability of binaural processes to contribute to intelligibility. Culling and Colburn (2000) examined this problem. They first used non-speech stimuli (repeated pure-tone arpeggios) in noise to test the idea that binaural sluggishness can interfere with the discrimination of complex spectrotemporal patterns. Listeners’ thresholds for discriminating ascending from descending arpeggios were lower when these signals were presented in the N_0S_π configuration than when both were presented diotically (referred to as N_0S_0), but the difference between these conditions shrank as the repetition rate was increased, suggesting that the sluggishness does smear the spectrotemporal representation of the target. They then performed a similar experiment in which the modulation rate of speech was controlled by using a digital-signal-processing technique to increase the speech rate. They found a robust benefit of binaural unmasking at the original speech rate, but the unmasking benefit again shrank as the speech rate was increased up to double the original speaking rate. It therefore appears that although sluggishness can limit SRM, the binaural system is responsive enough to markedly improve the intelligibility of speech at normal articulation rates.

3.4.4 *Models of SRM*

Spatial release from energetic masking is sufficiently well understood that some very effective predictive models have been developed. Models from Beutelmann et al. (2010), Jelfs et al. (2011), and Wan et al. (2014) each employ E–C theory and give good predictions of SRM in both anechoic and reverberant conditions.

3.4.5 *Conclusions*

SRM is the best-understood process of energetic masking release. The two component processes, better-ear listening and binaural unmasking, have been explored sufficiently well to permit the development of accurate predictive models. However, there are still some open questions about the workings of these mechanisms that could improve these models’ predictions in some circumstances. The models need to be able to predict the effects of variations across frequency and time in the

interaural differences on which they rely. They also need to reflect listeners' ability to combine information presented identically to both ears ("summation").

3.5 Other Mechanisms

There are some other potential mechanisms at work in the release from energetic masking. The phenomena attributed to these mechanisms are somewhat ambiguous in their interpretation. They could be interpreted as reflecting processes of auditory grouping and scene analysis, but they could also be interpreted as resulting from the operation of simpler mechanisms.

3.5.1 *Effect of Frequency Modulation on Prominence*

McAdams (1989) demonstrated the effects of vocal vibrato on the prominence of one voice among several. When three synthesized vowels were presented simultaneously to participants, their ratings of the "prominence" of the different vowels depended on whether a sinusoidal modulation of F0 had been applied to that vowel. When vibrato is applied, all the individual frequency components of that vowel move up and down in frequency together. One interpretation of this effect is that the modulation in F0 allowed the different frequency components of that vowel to be grouped together more strongly than those of a vowel with a static F0. Consequently, modulated vowels stand out from a background of static ones. This interpretation draws analogies with the visual system, which is clearly able to group together visual elements that share a common trajectory of movement, allowing moving forms to be detected and identified.

Culling and Summerfield (1995) tested this interpretation by creating vowels whose frequency components moved independently; all frequency components shared a common rate of frequency modulation, but their phases of modulation were randomized, to produce incoherent modulation. An undesired consequence of this manipulation is that these vowels quickly become inharmonic. To avoid a difference in harmonicity between these vowels and coherently modulated or static vowels, all types of vowel were made inharmonic with randomly offset frequency components. Listeners' ability to identify different vowels when masked by interfering vowels was measured. It was confirmed that modulation of a target vowel made it easier to identify vowels (rather than just make them more prominent) when the interfering vowel was unmodulated. However, the improvement in vowel identification was observed regardless of whether the modulation of frequency components was coherent or incoherent, indicating that the common movement of the frequency components was not a factor in the effect.

Culling and Summerfield concluded that the brain must possess some low-level mechanism that detects movement in frequency. For instance, a constantly moving

frequency component may be less susceptible to adaptation than a steady one. Alternatively, there may be a central mechanism that detects movement in frequency regardless of its coherence. In favor of the movement-detector interpretation, modulation of a masking vowel did not reduce listeners' ability to identify an unmodulated target vowel. If adaptation were at work, a modulated masker would have taken over the representation on the auditory nerve, making it a more effective masker. The fact that this did not occur suggests that the modulated vowel entered a separate processing channel after the auditory nerve.

3.5.2 Onset-Time Differences and the Potential Role of Adaptation

If one sound begins before another, the two are perceived as being separate, with individual characteristics, but if they begin at the same time they are likely to be integrated into a single percept. At a cocktail party, this would mean that maskers that have different temporal envelopes from the target speech, such as competing voices, will interfere less with the identification of sounds from the target speech. Such phenomena are often attributed to a perceptual grouping/segregation process by which concurrent sounds that have begun at different times will be parsed into two, perhaps through some spectral subtraction operation in which the frequency content of the first sound is subtracted from the frequency content of the combined sound. However, simple adaptation (either peripherally or more centrally) can have similar effects because it reduces the neural response at frequencies that have recently been stimulated, thereby emphasizing the representation of newly added sounds.

In a classic demonstration, Darwin (1984) manipulated the onset time of a tone added to a synthetic vowel sound at the same frequency as one of the vowel's harmonic components. If the onset time of the tone preceded the onset of the vowel, this individual component would be heard throughout the duration of the vowel as a separate sound, but if the tone began at the same time as the vowel, no separate sound was heard. Rather than rely on these perceptual impressions, however, Darwin looked for objective evidence that the tone was not integrated into the vowel when its onset was asynchronous. He demonstrated that asynchronous onset resulted in a change in the perceived identity of the vowel and that the change was consistent with exclusion of the tone from the vowel percept.

Darwin considered whether this effect might have been a simple effect of peripheral adaptation, whereby the longer duration of the tone meant that the neural response to that frequency was reduced by the time the vowel started, creating an effect somewhat similar to spectral subtraction. Although a contribution from adaptation could not be excluded, he noted that asynchronous offsets had a similar, but less powerful effect on vowel identity, and that a "captor tone" played synchronously with the leading portion of the tone, reduced its effect on the vowel

(Darwin and Sutherland 1984). The idea of this manipulation is that the captor tone must have provided an alternative perceptual group for the leading portion of the tone, forming it into a separate perceptual object from the vowel, whereas it could not have had any effect on adaptation at the tone frequency.

Roberts and Holmes (2006) and Holmes and Roberts (2011) reexamined this capture effect. They found that the effect of the captor tone depended neither on temporal synchrony with the leading portion of the added tone, nor on any harmonic relationship between the two. According to ideas about grouping, the strength of capture should depend on both of these properties. Instead, they found that the effects seemed more consistent with a combination of adaptation to the added tone reducing the representation of that frequency in the vowel, and an inhibitory effect of the “captor” on the added tone. Although these results provide a simpler explanation of the effect of an onset asynchrony, the effect of an offset asynchrony (in which the tone ends after the vowel) cannot be explained with adaptation. It therefore appears that effects mediated by widely different levels of neural processing can contribute to the same phenomenon.

3.6 Summary

This chapter has discussed relatively low-level processes that can act on a masker to reduce the degree to which it interferes with speech perception. It appears that a periodic masker can be suppressed, probably by some form of harmonic cancellation mechanism. A strongly modulated masker can be evaded by listening in the dips in its energy level, although its modulations may also interfere with detection of the modulations intrinsic to speech. A masker that lies in a different direction from the target speech can be both evaded by listening to the ear with the better signal-to-noise ratio and apparently reduced by an interaural cancellation mechanism. In addition to these effects, we have seen that low-level processes also contribute to some phenomena more often associated with perceptual grouping/segregation; the prominence of vibrato, and the segregating effect of differences in onset time.

At the same time, it is clear that, even for mechanisms as simple as better-ear listening, higher-level processes must be involved to select and combine signals that have been separated out by these mechanisms. The high-level versus low-level distinction is thus rather unworkable. Processes of audio frequency analysis, modulation frequency analysis, compression, adaptation, suppression, cancellation, interference, segregation, grouping, and streaming may all contribute at once to a single auditory event. Arguably, therefore, masking would be better discussed in terms of those individual auditory processes and their complex interaction rather than in the rather-too-broad classifications of energetic and informational.

Compliance with Ethics Requirements

John Culling has no conflicts of interest.

Michael Stone has no conflicts of interest.

References

- ANSI. (1997). ANSI S3.5-1997. *Methods for the calculation of the speech intelligibility index*. Washington, DC: American National Standards Institute.
- ANSI. (2013). ANSI S1.1-2013. *Acoustical terminology*. Washington, DC: American National Standard Institute.
- Assmann, P. F., & Paschall, D. D. (1998). Pitches of concurrent vowels. *The Journal of the Acoustical Society of America*, *103*, 1150–1160.
- Assmann, P. F., & Summerfield, Q. (1990). Modeling the perception of concurrent vowels: Vowels with different fundamental frequencies. *The Journal of the Acoustical Society of America*, *88*, 680–697.
- Assmann, P. F., & Summerfield, Q. (1994). The contribution of waveform interactions to the perception of concurrent vowels. *The Journal of the Acoustical Society of America*, *95*, 471–484.
- 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*, 235–251.
- Bernstein, J. G. W., & Grant, K. W. (2009). Auditory and auditory-visual speech intelligibility in fluctuating maskers for normal-hearing and hearing-impaired listeners. *The Journal of the Acoustical Society of America*, *125*, 3358–3372.
- Beutelmann, R., Brand, T., & Kollmeier, B. (2010). Revision, extension, and evaluation of a binaural speech intelligibility model. *The Journal of the Acoustical Society of America*, *127*, 2479–2497.
- Bird, J., & Darwin, C. J. (1998). Effects of a difference in fundamental frequency in separating two sources. In A. R. Palmer, A. Rees, A. Q. Summerfield, & R. Meddis (Eds.), *Psychophysical and physiological advances in hearing*. London: Whurr.
- Bregman, A. S. (1990). *Auditory scene analysis*. Cambridge, MA: MIT Press.
- Brox, J. P., & Nootboom, S. G. (1982). Intonation and the perceptual separation of simultaneous voices. *Journal of Phonetics*, *10*, 23–36.
- Bronkhorst, A. W., & Plomp, R. (1988). The effect of head-induced interaural time and level differences on speech intelligibility in noise. *The Journal of the Acoustical Society of America*, *83*, 1508–1516.
- Brungart, D. S. (2001). Informational and energetic masking effects in the perception of two simultaneous talkers. *The Journal of the Acoustical Society of America*, *109*, 1101–1109.
- Buus, S. (1985). Release from masking caused by envelope fluctuations. *The Journal of the Acoustical Society of America*, *78*, 1958–1965.
- Christiansen, C., & Dau, T. (2012). Relationship between masking release in fluctuating maskers and speech reception thresholds in stationary noise. *The Journal of the Acoustical Society of America*, *132*, 1655–1666.
- Colburn, H. S. (1996). Computational models of binaural processing. In H. L. Hawkins, T. A. McMullen, A. N. Popper, & R. R. Fay (Eds.), *Auditory computation* (pp. 332–400). New York: Springer.
- Colburn, H. S., & Durlach, N. I. (1978). Models of binaural interaction. In E. C. Carterette (Ed.), *Handbook of perception* (Vol. IV, pp. 467–518). New York: Academic Press.

- Collin, B., & Lavandier, M. (2013). Binaural speech intelligibility in rooms with variations in spatial location of sources and modulation depth of noise interferers. *The Journal of the Acoustical Society of America*, *134*, 1146–1159.
- Culling, J. F. (2007). Evidence specifically favoring the equalization-cancellation theory of binaural unmasking. *The Journal of the Acoustical Society of America*, *122*(5), 2803–2813.
- Culling, J. F., & Colburn, H. S. (2000). Binaural sluggishness in the perception of tone sequences. *The Journal of the Acoustical Society of America*, *107*, 517–527.
- Culling, J. F., & Darwin, C. J. (1993). Perceptual separation of simultaneous vowels: Within and across-formant grouping by F0. *The Journal of the Acoustical Society of America*, *93*, 3454–3467.
- Culling, J. F., & Darwin, C. J. (1994). Perceptual and computational separation of simultaneous vowels: Cues arising from low-frequency beating. *The Journal of the Acoustical Society of America*, *95*, 1559–1569.
- Culling, J. F., & Mansell, E. R. (2013). Speech intelligibility among modulated and spatially distributed noise sources. *The Journal of the Acoustical Society of America*, *133*, 2254–2261.
- Culling, J. F., & Summerfield, Q. (1995). The role of frequency modulation in the perceptual segregation of concurrent vowels. *The Journal of the Acoustical Society of America*, *98*, 837–846.
- Culling, J. F., & Summerfield, Q. (1998). Measurements of the binaural temporal window. *The Journal of the Acoustical Society of America*, *103*, 3540–3553.
- Darwin, C. J. (1984). Perceiving vowels in the presence of another sound: Constraints on formant perception. *The Journal of the Acoustical Society of America*, *76*, 1636–1647.
- Darwin, C. J., & Sutherland, N. S. (1984). Grouping frequency components of vowels: When is a harmonic not a harmonic? *Quarterly Journal of Experimental Psychology*, *36A*, 193–208.
- de Cheveigné, A. (1998). Cancellation model of pitch perception. *The Journal of the Acoustical Society of America*, *103*, 1261–1271.
- de Cheveigné, A., McAdams, S., Laroche, J., & Rosenberg, M. (1995). Identification of concurrent harmonic and inharmonic vowels: A test of Theory of harmonic cancellation and enhancement. *The Journal of the Acoustical Society of America*, *97*, 3736–3748.
- de Laat, J. A. P. M., & Plomp, R. (1983). The reception threshold of interrupted speech for hearing-impaired listeners. In R. Klinke & R. Hartmann (Eds.), *Hearing—Physiological bases and psychophysics* (pp. 359–363). Berlin, Heidelberg: Springer.
- Deroche, M. L. D., & Culling, J. F. (2011a). Voice segregation by difference in fundamental frequency: Evidence for harmonic cancellation. *The Journal of the Acoustical Society of America*, *130*, 2855–2865.
- Deroche, M. L. D., & Culling, J. F. (2011b). Narrow noise band detection in a complex masker: Masking level difference due to harmonicity. *Hearing Research*, *282*, 225–235.
- Deroche, M. L. D., Culling, J. F., & Chatterjee, M. (2013). Phase effects in masking by harmonic complexes: Speech recognition. *Hearing Research*, *306*, 54–62.
- Deroche, M. L. D., Culling, J. F., Chatterjee, M., & Limb, C. J. (2014). Speech recognition against harmonic and inharmonic complexes: Spectral dips and periodicity. *The Journal of the Acoustical Society of America*, *135*, 2873–2884.
- Durlach, N. I. (1963). Equalization and cancellation theory of binaural masking-level differences. *The Journal of the Acoustical Society of America*, *35*, 416–426.
- Durlach, N. I. (1972). Binaural signal detection: Equalization and cancellation theory. In J. V. Tobias (Ed.), *Foundations of modern auditory theory* (Vol. II, p. 365462). New York: Academic Press.
- Durlach, N. (2006). Auditory masking: Need for improved conceptual structure. *The Journal of the Acoustical Society of America*, *120*, 1787–1790.
- Edmonds, B. A., & Culling, J. F. (2005). The spatial unmasking of speech: Evidence for within-channel processing of interaural time delay. *The Journal of the Acoustical Society of America*, *117*, 3069–3078.
- Edmonds, B. A., & Culling, J. F. (2006). The spatial unmasking of speech: Evidence for better-ear listening. *The Journal of the Acoustical Society of America*, *120*, 1539–1545.

- Egan, J., Carterette, E., & Thwing, E. (1954). Factors affecting multichannel listening. *The Journal of the Acoustical Society of America*, 26, 774–782.
- Festen, J., & Plomp, R. (1990). Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing. *The Journal of the Acoustical Society of America*, 88, 1725–1736.
- Fletcher, H. (1930). A space-time pattern theory of hearing. *The Journal of the Acoustical Society of America*, 1, 311–343.
- French, N. R., & Steinberg, J. C. (1947). Factors governing the intelligibility of speech sounds. *The Journal of the Acoustical Society of America*, 19, 90–119.
- Glasberg, B. R., & Moore, B. C. (1990). Derivation of auditory filter shapes from notched-noise data. *Hearing Research*, 47, 103–138.
- Grantham, D. W., & Wightman, F. L. (1979). Detectability of a pulsed tone in the presence of a masker with time-varying interaural correlation. *The Journal of the Acoustical Society of America*, 65, 1509–1517.
- Hartmann, W. M., & Pumphlin, J. (1988). Noise power fluctuations and the masking of sine signals. *The Journal of the Acoustical Society of America*, 83, 2277–2289.
- Hawkins, J. E., & Stevens, S. S. (1950). The masking of pure tones and of speech by white noise. *The Journal of the Acoustical Society of America*, 22, 6–13.
- 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. *The Journal of the Acoustical Society of America*, 115, 833–843.
- Hilkhuysen, G., & Machery, O. (2014). Optimizing pulse-spreading harmonic complexes to minimize intrinsic modulations after cochlear filtering. *The Journal of the Acoustical Society of America*, 136, 1281–1294.
- Hirsh, I. J. (1948). The influence of interaural phase on interaural summation and inhibition. *The Journal of the Acoustical Society of America*, 20, 536–544.
- Holmes, S. D., & Roberts, B. (2011). The influence of adaptation and inhibition on the effects of onset asynchrony on auditory grouping. *Journal of Experimental Psychology. Human Perception and Performance*, 37, 1988–2000.
- Houtgast, T., & Steeneken, H. J. M. (1985). A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria. *The Journal of the Acoustical Society of America*, 77, 1069–1077.
- Howard-Jones, P. A., & Rosen, S. (1993). Unmodulated glimpsing in ‘checkerboard’ noise. *The Journal of the Acoustical Society of America*, 93, 2915–2922.
- Jelfs, S., Culling, J. F., & Lavandier, M. (2011). Revision and validation of a binaural model for speech intelligibility in noise. *Hearing Research*, 275, 96–104.
- Jørgensen, S., & Dau, T. (2011). Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing. *The Journal of the Acoustical Society of America*, 130, 1475–1487.
- Jørgensen, S., Ewert, S. D., & Dau, T. (2013). A multi-resolution envelope-power based model for speech intelligibility. *The Journal of the Acoustical Society of America*, 134, 436–446.
- Klatt, H. (1980). Software for a cascade/parallel formant synthesizer. *The Journal of the Acoustical Society of America*, 67, 971–995.
- Klumpp, R. G., & Eady, H. R. (1956). Some measurements of interaural time difference thresholds. *The Journal of the Acoustical Society of America*, 28, 859–860.
- Kohlrausch, A., Fassel, R., van der Heijden, M., Kortekaas, R., et al. (1997). Detection of tones in low-noise noise: Further evidence for the role of envelope fluctuations. *Acta Acustica united with Acustica*, 83, 659–669.
- Kohlrausch, A., & Sander, A. (1995). Phase effects in masking related to dispersion in the inner ear. II. Masking period patterns of short targets. *The Journal of the Acoustical Society of America*, 97, 1817–1829.
- Kwon, B. J., & Turner, C. W. (2001). Consonant identification under maskers with sinusoidal modulation: Masking release or modulation interference? *The Journal of the Acoustical Society of America*, 110, 1130–1140.

- Licklider, J. C. R. (1948). The influence of interaural phase relations upon the masking of speech by white noise. *The Journal of the Acoustical Society of America*, 20, 150–159.
- McAdams, S. (1989). Segregation of concurrent sounds. I: Effects of frequency modulation coherence. *The Journal of the Acoustical Society of America*, 86, 2148–2159.
- Meddis, R., & Hewitt, M. J. (1991). Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: Pitch identification. *The Journal of the Acoustical Society of America*, 89, 2866–2882.
- Meddis, R., & Hewitt, M. J. (1992). Modeling the identification of concurrent vowels with different fundamental frequencies. *The Journal of the Acoustical Society of America*, 91, 233–245.
- Miller, G. A. (1947). The masking of speech. *Psychological Bulletin*, 44, 105–129.
- Miller, G. A., & Licklider, J. C. R. (1950). The intelligibility of interrupted speech. *The Journal of the Acoustical Society of America*, 22, 167–173.
- Nelson, P., Jin, S.-H., Carney, A. E., & Nelson, D. A. (2003). Understanding speech in modulated interference: Cochlear implant users and normal-hearing listeners. *The Journal of the Acoustical Society of America*, 113, 961–968.
- Oxenham, A., & Simonson, A. M. (2009). Masking release for low- and high-pass-filtered speech in the presence of noise and single-talker interference. *The Journal of the Acoustical Society of America*, 125, 457–468.
- Plomp, R. (1983). The role of modulation in hearing. In R. Klinke & R. Hartmann (Eds.), *Hearing—Physiological bases and psychophysics* (pp. 270–276). Heidelberg: Springer.
- Pumplin, J. (1985). Low-noise noise. *The Journal of the Acoustical Society of America*, 78, 100–104.
- Rhebergen, K. S., & Versfeld, N. J. (2005). A Speech Intelligibility Index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners. *The Journal of the Acoustical Society of America*, 117, 2181–2192.
- Roberts, B., & Holmes, S. D. (2006). Asynchrony and the grouping of vowel components: Captor tones revisited. *The Journal of the Acoustical Society of America*, 119, 2905–2918.
- Scheffers, T. M. (1983). *Sifting vowels: Auditory pitch analysis and sound segregation*. Doctoral thesis, University of Groningen.
- Schroeder, M. R. (1970). Synthesis of low-peak-factor signals and binary sequences with low autocorrelation. *IEEE Transactions on Information Theory*, 16, 85–89.
- Schubert, E. D. (1956). Some preliminary experiments on binaural time delay and intelligibility. *The Journal of the Acoustical Society of America*, 28, 895–901.
- Stone, M. A., Anton, K., & Moore, B. C. J. (2012). Use of high-rate envelope speech cues and their perceptually relevant dynamic range for the hearing impaired. *The Journal of the Acoustical Society of America*, 132, 1141–1151.
- Stone, M. A., Füllgrabe, C., & Moore, B. C. J. (2010). Relative contribution to speech intelligibility of different envelope modulation rates within the speech dynamic range. *The Journal of the Acoustical Society of America*, 128, 2127–2137.
- Stone, M. A., & Moore, B. C. J. (2014). On the near non-existence of “pure” energetic masking release for speech. *The Journal of the Acoustical Society of America*, 135, 1967–1977.
- Studebaker, G. A., & Sherbecoe, R. L. (2002). Intensity-importance functions for bandlimited monosyllabic words. *The Journal of the Acoustical Society of America*, 111, 1422–1436.
- Summerfield, Q., & Assmann, P. F. (1990). Perception of concurrent vowels: Effects of harmonic misalignment and pitch-period asynchrony. *The Journal of the Acoustical Society of America*, 89, 1364–1377.
- Summerfield, Q., & Assmann, P. F. (1991). Perception of concurrent vowels: Effects of harmonic misalignment and pitch-period asynchrony. *The Journal of the Acoustical Society of America*, 89, 1364–1377.
- Summers, V., & Leek, M. R. (1998). Masking of tones and speech by Schroeder-phase harmonic complexes in normally hearing and hearing-impaired listeners. *Hearing Research*, 118, 139–150.
- von Helmholtz, H. (1895). *On the sensations of tone as a physiological basis for Theory of music*. London: Longmans.
- Wan, R., Durlach, N. I., & Colburn, H. S. (2014). Application of a short-time version of the equalization–cancellation model to speech intelligibility experiments with speech maskers. *The Journal of the Acoustical Society of America*, 136, 768–776.

Chapter 4

Informational Masking in Speech Recognition

Gerald Kidd Jr. and H. Steven Colburn

Abstract Solving the “cocktail party problem” depends on segregating, selecting, and comprehending the message of one specific talker among competing talkers. This chapter reviews the history of study of speech-on-speech (SOS) masking, highlighting the major ideas influencing the development of theories that have been proposed to account for SOS masking. Much of the early work focused on the role of spectrotemporal overlap of sounds, and the concomitant competition for representation in the auditory nervous system, as the primary cause of masking (termed energetic masking). However, there were some early indications—confirmed and extended in later studies—of the critical role played by central factors such as attention, memory, and linguistic processing. The difficulties related to these factors are grouped together and referred to as informational masking. The influence of methodological issues—in particular the need for a means of designating the target source in SOS masking experiments—is emphasized as contributing to the discrepancies in the findings and conclusions that frequent the history of study of this topic. Although the modeling of informational masking for the case of SOS masking has yet to be developed to any great extent, a long history of modeling binaural release from energetic masking has led to the application/adaptation of binaural models to the cocktail party problem. These models can predict some, but not all, of the factors that contribute to solving this problem. Some of these models, and their inherent limitations, are reviewed briefly here.

G. Kidd Jr. (✉)

Department of Speech, Language and Hearing Sciences, Hearing Research Center,
Boston University, 635 Commonwealth Avenue, Boston, MA 02215, USA
e-mail: gkidd@bu.edu

H.S. Colburn

Department of Biomedical Engineering, Hearing Research Center, Boston University,
44 Cummington Street, Boston, MA 02215, USA
e-mail: colburn@bu.edu

Keywords Adverse listening conditions · Auditory masking · Auditory scene analysis · Binaural models · Cocktail party problem · Energetic masking · Informational masking · Speech comprehension · Speech in noise · Speech perception

4.1 Introduction

Of all of the important uses for the sense of hearing, human listeners are perhaps most dependent in their everyday lives on selectively attending to one talker among concurrent talkers and following the flow of communication between participants in conversation. This ability is fundamental to a wide range of typical social interactions and, for listeners with normal hearing at least, usually is accomplished successfully and fairly effortlessly (see Mattys et al. 2012; Carlile 2014; and Bronkhorst 2015 for recent reviews). It has long been recognized, though, that these are highly complex tasks that must be solved by the concerted actions of the ears and the brain (and, in many cases, the eyes as well). Extracting a stream of speech from one talker among a mixture of talkers or other sounds depends on perceptually segregating the different sound sources, selecting one to focus attention on, and then recognizing and comprehending the flow of information emanating from the chosen source. These tasks usually are performed while the listener remains attuned—to some degree—to sources outside of the primary focus of attention in the event that attention needs to be redirected. The sounds a listener may wish to receive (“targets”) often overlap in time and frequency with competing sounds (“maskers”), resulting in what is known as “energetic masking” (EM). Even in the absence of spectral or temporal overlap, however, a variety of other factors may act to limit target speech recognition. These factors are broadly categorized as “informational masking” (IM).

The present chapter compares and contrasts EM and IM for the case of speech-on-speech (SOS) masking. The chapter is divided into three sections. First, the early work on the masking of speech by speech and other sounds is reviewed in an attempt to explain how the major ideas developed and the evidence on which they were based. Second, the issues involved in measuring SOS masking are discussed, focusing on how the distinction between EM and IM is made. Finally, some models of masking are considered—in particular those binaural models addressing the benefit of spatial separation of sources—with respect to how they may be applied to the masking of speech by other speech.

4.2 The History of Study of the Special Case of SOS Masking

In his seminal article describing the masking of speech, George A. Miller (1947) writes, “It is said that the best place to hide a leaf is in the forest, and presumably the best place to hide a voice is among other voices” (p. 118). Although he concluded in that article that the masking of speech by other speech was largely a consequence of overlapping energy in time and frequency, this analogy serves to illustrate a fundamental problem in the design of speech masking experiments: when the sound field contains many distinct but similar sources, how do we ask the question of whether one specific source is present or what information is being conveyed by that particular source? In a typical communication situation comprising multiple concurrent talkers, a listener normally may use a variety of cues—often relying heavily on context—to segregate the sounds and determine which source should be the focus of attention.

Cherry (1953) suggested several factors facilitating the process of separating one talker from others, including differences in source direction, lip-reading and gestures, differences in vocal properties and accents between talkers, and various transition probabilities. In designing experiments in the laboratory to measure aspects of this formidable ability, such as determining the strength of source segregation cues or measuring the ability to shift attention from one source to another, the means by which one source is designated as the target and so distinguished from those sources that are maskers may exert a profound influence on the outcome of the experiment. Thus, assessing the potential benefit that might result from another variable under test is strongly influenced by the way that the target is designated as the target, and a different answer about the role of such factors may be obtained with a different means for designating the source. This issue pervades the literature on SOS masking and has become increasingly relevant as a finer distinction is drawn between the sources of interference from competing talkers (i.e., whether they produce primarily EM or IM).

The issue of source designation in SOS masking was raised early on by Broadbent (1952a), who demonstrated that the manner of target designation could affect the amount of masking produced by a concurrent talker. In summarizing a study of the factors that underlie the recognition of the speech of one talker in competition with another, he observes: “From the practical point of view, these experiments show that there is a possibility, when two messages arrive simultaneously, of identification of the message to be answered becoming a more serious problem than the understanding of it once identified” (p. 126). However, because the majority of early work on the topic of masking relied on noise maskers—regardless of whether the target was speech or other sounds such as pure tones—the issues of source designation and listener uncertainty (e.g., possibility for source confusions) were not given extensive consideration (a notable exception is the topic of signal frequency uncertainty; cf. Kidd et al. 2008a). Likewise, in Cherry’s (1953)

study, designating one ear as containing the target with the other ear containing the masker provided a simple, unambiguous means of source designation.

The findings from much of the early work on SOS masking were, in fact, largely consistent with the more general view of masking that was prevalent at the time: that is, that one sound interferes with the reception and processing of another sound primarily by obscuring or covering up the energy of the target sound within the frequency channels (“critical bands”; Fletcher 1940) containing the target. This perspective, which is based on EM, led to the original methods proposed for predicting speech recognition in noise (e.g., Egan and Weiner 1946; French and Steinberg 1947) as well as later refinements of those methods such as the speech intelligibility index (SII; cf. ANSI 1997). The connection between detecting a tone in noise and understanding speech in noise seemed obvious. For example, Beranek (1947) states, “Of great importance in understanding the ability of the ear to interpret transmitted speech is the way in which various noises mask desired sounds. Extensive tests have shown that for noises with a continuous spectrum, it is the noise in the immediate frequency region of the masked tone which contributes to the masking.... The bandwidth at which the masking just reaches its stable value is known as a “critical band”... Bands of speech appear to be masked by continuous-spectra noises in much the same way as pure tones are masked by them. For this reason, it is possible to divide the speech spectrum into narrow bands and study each band independently of the others” (p. 882).

Using noise as a masker has many advantages: it is easy to specify based on its underlying statistical properties, and it produces masking that tends to be more repeatable across trials and subjects than that produced by speech maskers (e.g., Freyman et al. 1999; Brungart 2001; Arbogast et al. 2002). Also, importantly, one need not worry about the listener confusing the target with the masker so that attention is unlikely to be misdirected, nor does noise typically carry any special information that commands our interest (however, the effect of Gaussian noise is not confined to EM although it often is used as a high-EM control condition for comparison; cf. Culling and Stone, Chap. 3; Schubotz et al., 2016).

Some of the early findings that supported EM as the basis for SOS masking include Miller’s (1947) report that the masking produced by unintelligible speech from a language other than that of the listener was about the same as for intelligible speech in the primary language. Similarly, Miller noted that uncertainty about the content or production of speech also had little effect on masking: “The content of the masking speech is a more difficult factor to evaluate [than masking by noise or other non-speech sounds]. Conversational voices were compared with loud, excited voices liberally interspersed with laughter, cheering and improbable vocal effects. The two sounds could be likened to the chatter at a friendly dinner-party versus the din of a particularly riotous New Year’s Eve celebration” (p. 119). These findings led Miller to state: “Once again, it is necessary to conclude that the crucial factor is the masking spectrum. The particular way in which the spectrum is produced is of secondary importance” (p. 120). Although this work was limited by the methods available at the time, and later work produced findings inconsistent with this broad

conclusion, Miller's comments presaged both the "cocktail party problem" and, importantly, the role that uncertainty could play in SOS masking.¹

The proposition that central factors—and not just peripheral overlap—may contribute to speech signals masking other speech signals was given strong empirical support by Broadbent (1952b). In a clever paradigm, he interleaved target and masker words in sequence finding that, despite the fact that the words had no spectrotemporal overlap and therefore ostensibly no EM, performance in target speech recognition nonetheless was degraded by the presence of the intervening masker words. Furthermore, certain nonacoustic aspects of the stimuli (e.g., familiar target voice; cf. Johnsrude et al. 2013; Samson and Johnsrude 2016) also influenced performance. Broadbent considered that his results revealed a "failure of attention in selective listening" because a perfect selection mechanism could simply gate "on" only the target words and gate "off" the masker words so that they would have no masking effect. Later, Broadbent (Broadbent 1958; pp. 11–29) concluded that these findings provided strong evidence for "central factors" in masking.

In an article that identified and evaluated several factors contributing to SOS masking that involved both peripheral and central mechanisms, Schubert and Schultz (1962) measured the benefit of imposing differences in interaural timing between the target talker and masker talkers. This study exemplified some of the difficulties inherent to the study of SOS masking because multiple variables influenced the results, but it also identified several ways that SOS masking could be released by central factors. The binaural differences they imposed were phase inversion (i.e., the target was π radians out of phase at the two ears while the masker was in-phase at the two ears; $S_{\pi}M_0$) or broadband time delays. Those manipulations were logical extensions of earlier work demonstrating masking level differences (MLDs) for detecting tones in noise (e.g., Hirsh 1948) and intelligibility gains for speech in noise (Licklider 1948), and therefore aimed to reduce EM (see Sect. 4.4). Other manipulations tried by Schubert and Schultz (1962), however, appear to have stemmed from intuitions about the perceptual basis upon which sources are segregated. This is apparent in their Table 1, in which they proposed a hierarchical arrangement of the effects of the masking stimuli according to a rough, qualitative estimate of similarity to the target. In that hierarchy, the most similar masker was the target talker's own voice, followed by single same-sex talker, single different-sex talker, multiple talkers, and ultimately multiple talkers reversed in time. It is clear from that hierarchy that their choice of masking stimuli reflected an expectation about an interaction between the binaural manipulations and these similarity-based masker properties.

In a study that has been widely cited because it identified both the masking of speech that could not be attributed to peripheral processes and the release from

¹Irwin Pollack (2002; personal communication) attributed his use of the term "informational masking" to influential comments by George A. Miller at a seminar presented by Pollack describing the masking of speech by bands of filtered noise. According to Pollack, Miller objected to (Pollack's) use of noise as a masker considering its effects to be "secondary" to the "informational content of the messages" contained in speech maskers.

masking of speech beyond that predicted by traditional models of binaural unmasking, Carhart et al. (1969a) reported several instances of “excess masking.” As with the Schubert and Schultz (1962) study, Carhart et al. (1969a) were interested primarily in understanding binaural release from masking for speech. However, that interest inevitably led to consideration of the cause of masking to begin with. It became clear that explanations were required for this excess masking effect—which they termed “perceptual masking”—that extended beyond traditional EM-based theories and models (see also Carhart et al. 1969b).

4.3 Determining Energetic and Informational Masking in SOS Masking

Although there are several methods that researchers have employed in an attempt to separate energetic and informational factors in masking experiments, the two most common are—broadly speaking—to vary the degree of target and/or masker uncertainty in the task and to control the amount of spectrotemporal overlap that is present between target and masker. In the former case, this is usually accomplished by manipulating the variability in the stimulus or the manner in which it is presented to the listener. In the latter case, an attempt is made to hold EM constant (or is taken into account by modeling) while factors that do not influence EM (e.g., linguistic aspects of speech) are varied, with the rationale being that any observed changes in performance may then be attributed to the influences of IM.

4.3.1 Uncertainty

Manipulating observer uncertainty by imposing stimulus variability is an empirical approach that was commonly employed in the early studies of IM using nonspeech stimuli (see Kidd et al. 2008a for a review). For example, in the series of studies by Watson and colleagues (summarized in Watson 2005), the task often was to detect an alteration in the frequency or intensity of a tone pulse embedded in a sequence of similar pulses or “context tones.” The way that the context tones were presented—specifically, whether they varied in composition from trial to trial within a block of trials or were held constant across trials within a block—was used to manipulate listener uncertainty and often produced large differences in performance. Although less common in the SOS masking literature, analogous manipulations are possible. Brungart and Simpson (2004) explicitly varied the degree of uncertainty in a SOS masking paradigm. They used a closed-set, forced-choice, speech identification task (the “Coordinate Response Measure,” CRM, test) in which the target voice is followed throughout the sentence after a specified “callsign” occurs until two test words—a color and a number—are presented (cf. Brungart 2001; Iyer et al. 2010).

Both the masker talkers and/or the semantic content could be fixed or randomized across trials. Somewhat surprisingly based on a logical extrapolation of the findings from the nonspeech IM literature, increasing masker uncertainty caused little decrement in performance, with variability in semantic content producing the only statistically significant difference. Similarly, Freyman et al. (2007) tested a condition in which masker sentences were held constant across trials or varied randomly across trials. Consistent with the small effects of masker uncertainty reported by Brungart and Simpson (2004), no significant effect on performance was found due to masker uncertainty for variation in talker, content, or target-to-masker ratio (T/M). The open-set target speech materials used by Freyman and colleagues were nonsense sentences while the maskers were similar nonsense sentences from a different corpus. It is possible that the time available to focus on these relatively long stimuli allowed the listener to overcome any initial uncertainty about the characteristics of the target source. With a clear cue to source designation (e.g., the callsign for the CRM test), the ability to select the target source was sufficient to overcome the relatively minor uncertainty caused by the stimulus variation that was present.

Uncertainty about some aspects of the stimulus or its presentation *can* affect the amount of IM in SOS masking. For example, Kidd et al. (2005) demonstrated that uncertainty about the spatial location of a target talker influenced speech identification performance in a multiple-talker sound field. By manipulating the a priori probability of target presentation (one of three concurrent talkers) from one of three locations separated in azimuth, Kidd and colleagues found large differences in performance depending on whether the listener was provided with the cue designating the target sentence (the “callsign”) before or after the stimulus. When the listener had no a priori knowledge about target location and did not receive the callsign designating the target until after the stimulus, performance was relatively poor—near the value expected simply from choosing to focus attention on only one of the three locations. When the target sentence was cued/designated before the trial, but location was uncertain, performance improved significantly relative to the uncued case. When the probabilities about source location were provided before the stimulus, performance improved significantly for both cued and uncued conditions. If the location of the target was certain, proportion correct identification performance was higher than 0.9 independent of whether the target was cued beforehand. These findings are shown in Fig. 4.1A. Similar effects of location uncertainty have been reported by Best and colleagues (2007) and by Kidd and colleagues (2014) using different paradigms. In those studies, as in the Kidd et al. (2005) study just described, the conclusion was that a priori knowledge about target source location can improve speech recognition under multiple-talker competition..

An example of the type of error analysis that reveals confusions among sources is found in Fig. 4.1B, reproduced from Kidd et al. (2005). This panel shows a breakdown of error types for each condition. For the condition with the greatest certainty about location, the most frequent error was to mix one target word (color or number) with one masker word. For the uncertain cases, the most common error was to report both color and number words from one of the two masker sources.

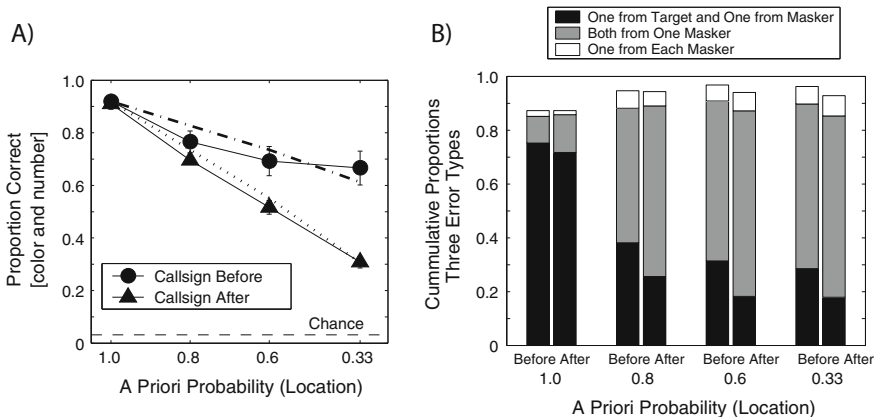


Fig. 4.1 (A) Proportion correct speech identification scores as a function of the a priori probability of occurrence at one of three locations. The data points are group means with standard errors. The straight lines are predictions of a simple probability-based model. The circles show performance when the callsign designating the target sentence was provided before the stimulus while the triangles show performance when the callsign was provided after the stimulus. Chance performance is indicated by the dashed line at the bottom. (B) The error analysis associated with the results shown in A. The bars are composite histograms indicating the proportions of error types that occurred. (A and B from Kidd et al. 2005, *The Journal of the Acoustical Society of America*, with permission.)

The difference between the height of each composite bar and 1.0 indicates the proportion of errors not attributable to confusions that could be due to EM. The authors concluded that in nearly all cases the three talkers likely were each audible but that errors occurred because of source confusions/misdirected attention.

It is clear from the preceding discussion that the structure of the SOS masking task can affect the outcome of the experiment. This observation may seem obvious but what is (or historically has been) less obvious is that it applies much more strongly for speech masked by other speech than for speech masked by noise and is at the heart of the IM–EM distinction. The conditions that produce the highest IM tend to be those in which confusions are possible such as happens when both target and masker share similar low-level features (e.g., same-sex talkers or even same talker as masker) and the masker words are allowable response alternatives in closed-set paradigms (see Webster 1983 for a review of early work on closed-set speech tests). Using very different types of materials for target and masker(s) can greatly reduce uncertainty and therefore reduce IM. Natural communication situations may of course vary widely in the degree to which source or message uncertainty is present and expectation based on context and a priori knowledge often determines success.

4.3.2 *Controlling/Estimating Energetic Masking*

When two or more independent talkers are speaking concurrently, the acoustic overlap between the sounds varies considerably from moment to moment. The spectrotemporal overlap of the speech from different sources depends on a variety of factors including inherent differences in source characteristics (e.g., size and shape of the vocal apparatus, acquired speaking patterns, etc.), the speech materials that are being uttered by the various sources, and the acoustic environment (e.g., reverberation), among others. Moreover, speech sources in real sound fields typically originate from different locations meaning that the waveforms arrive at the listener's ears with differing interaural time and intensity values. For this reason, perhaps, much of the work on the "cocktail party problem" has addressed multiple source segregation and selection cues that occur concurrently and include such explicit factors as binaural difference cues and fundamental frequency/formant resonance differences, etc., in addition to the source designation methods discussed in Sect. 4.2. Ultimately, determining the precise way that the sounds overlap in their representations in the auditory system can be a very complex problem involving models of how the ear codes the relevant sound parameters dynamically and the interaural differences in the sound inputs.

Because the early stages of the peripheral auditory system are tonotopically organized, one nearly universal way of thinking about EM is to divide the stimulus into physiologically inspired frequency channels and to consider how the representations of the competing speech sounds are preserved within these channels over time. To test hypotheses about how these representations interact under different assumptions, a variety of experimental approaches have been devised that reduce the acoustic stimulus to limited frequency regions so as to manipulate the overlap that occurs within auditory channels.

Among the first studies to attempt to separate EM from IM in SOS masking by taking advantage of the tonotopic organization of sounds in the auditory system was Arbogast et al. (2002). They used a tone-vocoding procedure to process two independent speech sources into acoustically mutually exclusive frequency channels (within the limits of the procedure). This is illustrated in Fig. 4.2.

The upper panels show the magnitude spectra of the processed target plus masker while the lower panels show the waveforms. The two types of masker shown are "different-band speech" (DBS), which consists of intelligible speech in narrow frequency bands that do not contain target speech and "different-band noise" (DBN), which consists of equally narrow (unintelligible) bands of noise in the bands that do not contain target speech. Pilot tests showed that sufficient speech information was present in the envelopes of the small number of spectral bands for the target and masker speech sources each to be intelligible separately. To solve the task the listener had to distinguish the target speech from another similar CRM sentence (DBS condition) spoken by a different talker. The key to determining the amount of IM present was to compare performance obtained using the speech masker (DBS) with the performance obtained using the noise masker (DBN).

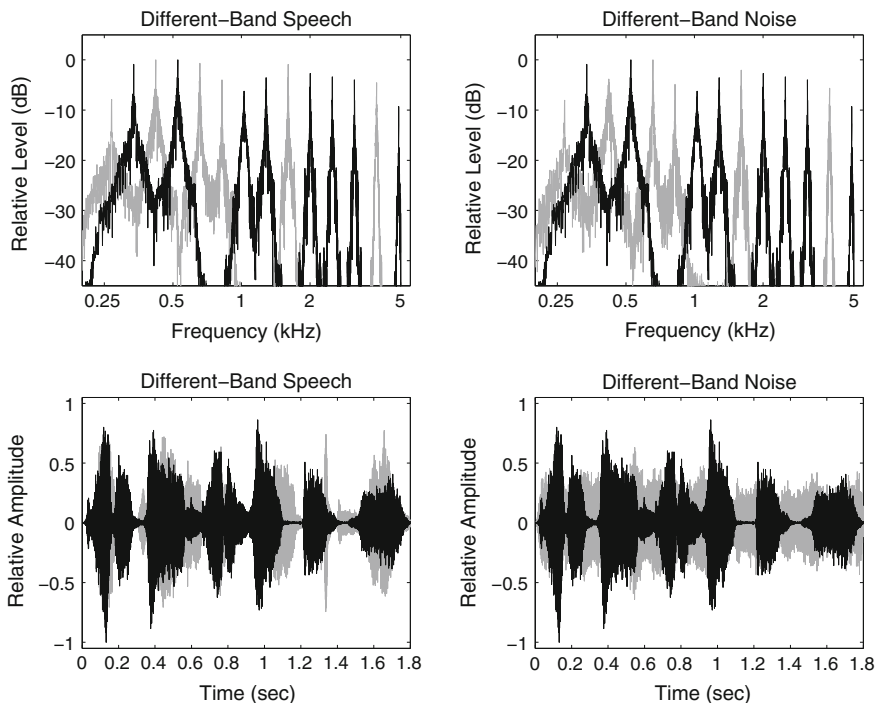


Fig. 4.2 The *upper two panels* show the magnitude spectra for the “different-band speech” and “different-band noise” maskers (light gray) plus target (dark gray) while the *lower two panels* show the associated waveforms (same shading). As may be seen from the upper panels, the target and maskers are processed into mutually exclusive frequency channels that are chosen randomly on every presentation. (Adapted from Arbogast et al. 2002, *The Journal of the Acoustical Society of America*, with permission.)

Because the amount of EM for the DBS and DBN maskers was expected to be about the same, the greater masking caused by the speech (about 18 dB) was attributed to IM. The large amount of IM found in this experiment depended in part on the specific way that the stimuli were processed which was designed to minimize EM while preserving enough of the speech for high intelligibility. The important finding from Arbogast et al. (2002) for the current discussion is that maskers that were equated for EM were shown to produce significantly different amounts of IM depending on whether the masker was intelligible.

Brungart et al. (2006) proposed a method of processing speech into highly quantized elements so that the EM and IM present in SOS masking could be estimated/controlled. Not only did they analyze the speech stimulus into narrow frequency channels but they also then subdivided each channel into brief time intervals. Essentially, the result was a matrix of values representing energy contained in fine time–frequency (T–F) units. Based on a priori knowledge of the stimuli, the T/M in each bin was computed and a criterion for sorting the bins based

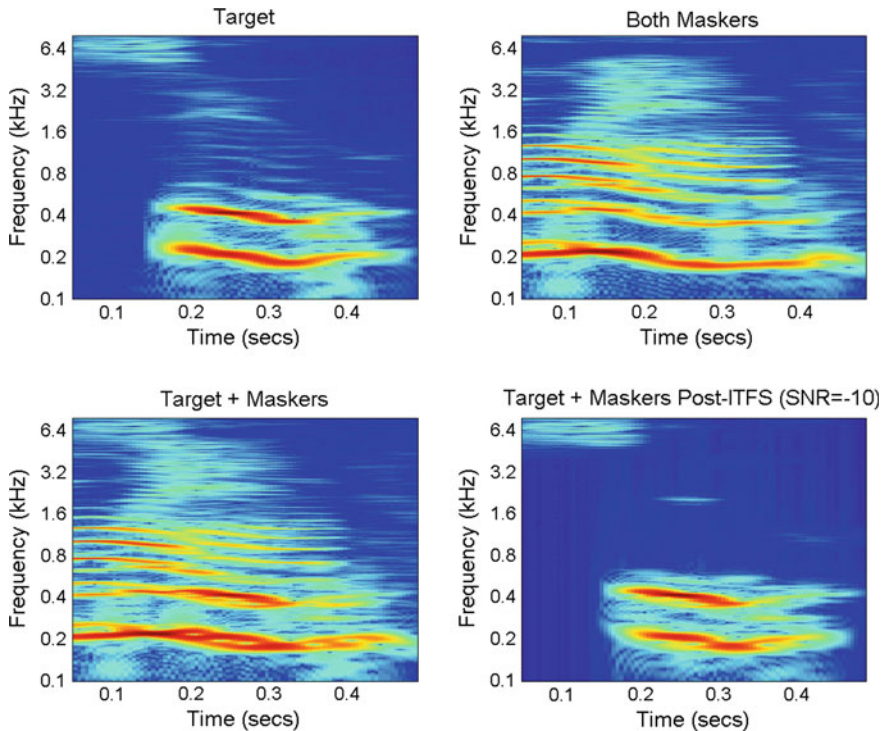


Fig. 4.3 Results of the processing of target and masker stimuli into time–frequency bins following the procedure used by Brungart et al. (2006). The abscissa is time while the ordinate is frequency on a log scale. Red/blue shading represents high/low intensity. The *upper left panel* shows the spectrogram of the target; the *upper right panel* shows the spectrogram of the two-talker masker; the *lower left panel* shows the combination of the target and maskers; and the *lower right panel* shows the T–F units of the combined stimulus for which $T > M$ (stimuli and analysis)

on T/M was applied. The criterion could be used to exclude bins based on T/M —discarding the bins below the criterion—with the remaining bins reassembled into a speech stimulus. The results of this procedure applied to multiple speech sources are shown in Fig. 4.3.

The top left panel is a spectrogram of the original target speech; the top right panel shows the masker speech (two independent maskers); the lower left panel shows the mixture of target and masker signals, while the lower right panel shows only the T–F units that remain after discarding those in which the masker energy is greater than the target energy (an “ideal binary mask”). In the procedure used by Brungart et al. (2006) the difference in intelligibility between the two sets of stimuli shown in the lower panels is taken as an estimate of IM. The finding of a significant improvement in speech identification performance by removal of the low T/M bins argued for a strong role of IM. This is a crucial finding on a theoretical level because the usual assumption about combining the information from different T–F

units based on speech in noise tasks is that each unit containing target energy contributes some increment—even if infinitesimal—to overall intelligibility. The worst a T–F unit could do is to produce no appreciable gain. However, the presence of IM means that the presence of units with little or no target information *reduces* overall intelligibility. In fact, not only will including these units “garble” the target, they also may yield an alternate, intelligible source that is confused with the target source. In contrast, a parallel manipulation using noise as a masker revealed minor detrimental effects of presentation of the unprocessed versus processed stimulus thereby eliminating differences in EM as the cause of the effect. The findings of Brungart et al. (2006) were significant not only because they provided a quantitative means for separating EM from IM in SOS mixtures but also because their results revealed a dominant role of IM in SOS masking for the stimuli and conditions tested. In a later study using the procedure described above, Brungart et al. (2009) found that increasing the number of independent masker talkers to the point where the individual voices are lost in an incomprehensible—but obviously speech—babble increased EM while decreasing IM. The idea that increasing the number of similar individual elements in the sound field (like increasing the number of leaves in the forest envisioned by Miller 1947), increases EM while it (ultimately) decreases IM, is a common theme in contemporary auditory masking studies (cf. Kidd et al. 2008a). The use of unintelligible babble as a speech masker, coupled with strong target segregation/designation cues, likely contributed to the conclusion from some early studies that SOS masking was predictable solely on the basis of spectrotemporal overlap of the competing sources.

4.3.3 *Linguistic Variables*

A persistent question in the SOS literature is whether the degree of meaningfulness of competing maskers affects the masking that is observed. For example, randomly selected words with no syntactic structure and little semantic value are less meaningful than coherent discourse, but do they mask target speech any less? If so, does this imply that the greater the meaning, or perceived potential to carry meaning, a masker possesses the more it invokes some degree of obligatory processing? If linguistic variables affect SOS masking, then an explanation based purely on peripheral overlap of excitation falls short of providing a satisfactory account of the underlying processes governing performance. Although this point has been recognized for decades, the evidence often has been inconclusive and sometimes contradictory—partly for reasons discussed in Sect. 4.2 concerning differences in methodology. Here we review work intended to determine the role that linguistic variables play in masking target speech.

4.3.3.1 Time Reversal

Among the more obvious ways of evaluating the influence of lexical factors in SOS masking is to degrade the meaning of speech by reversing it in time. Historically, reversed speech has been an intriguing stimulus because it largely maintains its frequency and envelope spectra while losing intelligibility (cf. Kellogg 1939; Cherry 1953; Schubert and Schultz 1962). Differences in the amount of masking produced by time-forward speech and the same speech time-reversed therefore could be due to the difference in “meaningfulness.” Based on the premise that “speech perception cannot be explained by principles that apply to perception of sounds in general” (p. 208), Hygge et al. (1992) reasoned that “...it can be expected that a normal background speech condition should interfere more with a speech comprehension task than a noise control that does not carry any phonological information (and)...normal (i.e., forward) speech should interfere more than the same speech played in reverse...” With respect to early work examining this issue, an article by Dirks and Bower (1969) was particularly influential. In their careful and systematic study, short “synthetic” sentences (Speaks and Jerger 1965) spoken by a male talker were masked by unrelated, continuous discourse spoken by the same talker played forward or backward. The observed performance-level functions indicated nearly identical results in all cases. Likewise, in the Hygge et al. (1992) study, in which the target talker was female and the masker was a single male talker, no significant difference in the amount of masking (using a subjective “just understandable” criterion and method of adjustment) was found when the masker talker was presented normally versus time reversed. In this case the speech materials (both target and maskers) were relatively long (3 min) passages of connected speech. The conclusion drawn from these studies, supported by the original findings from Miller (1947) noted in Sect. 4.2, was that the main determinant of SOS masking is the spectrotemporal overlap of the sounds and that linguistic factors per se were of little import. These studies suggest that the outcomes of SOS masking experiments are very sensitive to the specific methods that are used. When the masker differs in fundamental ways from the target—on a semantic level, as is the case with very different types of speech materials, or on a more basic acoustic level as with the differences in source characteristics for male versus female talkers—uncertainty may be minimal and subsequent manipulations intended to examine other factors (e.g., masker time reversal) may produce negligible effects.

In a pivotal article in the IM literature concerning speech, Freyman et al. (2001) reported a large difference (4–8 dB) between the masking effectiveness of forward and time-reversed masker speech. The speech corpus for both target and masker consisted of simple sentences spoken by female talkers that were semantically implausible but syntactically correct. Importantly for the discussion that follows regarding spatial release from IM, the additional release from IM (beyond that obtained by time reversal) due to a perceived difference in location between target and masker was relatively small when compared to the same perceived location difference for forward speech. These findings suggested that the high IM produced by the SOS masking conditions tested could be released by *either* time reversing the

masker—causing it to be unintelligible—*or* by perceptually segregating the apparent locations of the sources.

The large benefit due to time-reversing the masker obtainable in some SOS conditions subsequently has been confirmed in several other studies. Marrone et al. (2008; see also Best et al. 2012) used the closed-set CRM test spoken by a female target talker masked by two female masker talkers with the specific voices randomized from trial to trial. Marrone and colleagues varied the locations from which the maskers were presented using the co-located case as a reference for determining spatial benefit. When target and masker talkers were co-located, time-reversing the maskers yielded a large advantage over natural presentation with the T/Ms at threshold lower by about 12 dB—nearly the same release from masking as was obtained from spatial separation of sources. Even larger reductions in T/M due to masker time reversal—about 17 dB, on average—in co-located conditions have been reported by Kidd et al. (2010). They used a different closed-set speech identification test with female target and two female masker talkers uttering five-word sentences with all of the syntactically correct sentences drawn from the same corpus. As with Marrone et al. (2008), the specific talkers were selected randomly from a small closed set of talkers on every trial. The large “reversed masking release” (RMR) reported by Marrone et al. and Kidd et al. in the co-located condition likely reflects a reduction in IM based on the assumption that the amount of EM remains the same when the masker is time reversed. However, the extent to which time reversal preserves the EM of a speech masker is a matter of some conjecture. It is possible, for example, that time reversal affects the temporal masking that one phoneme can exert on another. Moreover, closed-set tests that use the same syntactic structure for target and masker speech, with some degree of synchrony, could result in more EM if the envelopes were highly correlated reducing “glimpses” of the target in masker envelope minima.

Rhebergen et al. (2005) proposed that time reversal of masking speech may not produce EM that is equivalent to natural speech. They noted that the envelopes of speech produced naturally often tend to exhibit an asymmetric shape with quick onsets (attributed to plosive sounds) followed by slower decays. Time reversal alters this shape so that the rise is more gradual and the offset more abrupt. The consequence of this reversal is that some soft sounds would be masked (via forward masking) in one case but not in the other so that EM could effectively differ. In the key finding from their study, the masking produced by a different-sex masking talker uttering a language that was not known to the listeners was greater when the masker was time reversed than when it was played forward. The greater amount of masking from the reversed speech was small, about 2 dB, but was judged to be significant. The greater EM for reversed speech means that release from IM due to time reversal may be *underestimated* by an amount that depends on the increase in EM due to greater forward masking from the reversed envelope.

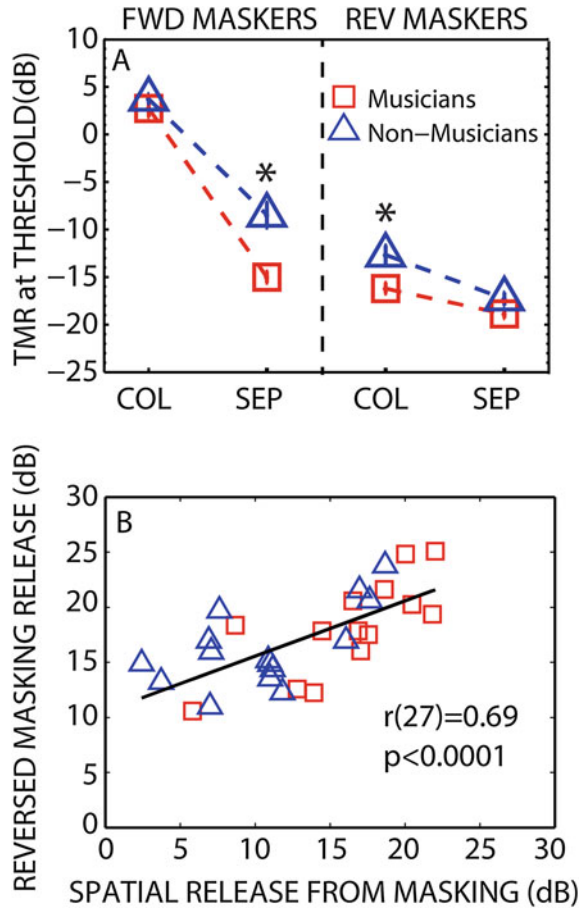
Concerns about potential differences in EM due to time reversal, and the possibility that these differences are exacerbated when the target and masker sentences are similar in structure and spoken nearly in cadence, led Marrone et al. (2008) to test a “control” condition explicitly examining whether time-reversed speech

generated greater EM than the same speech played forwards. In their experiment, the target speech was masked by two independent speech-spectrum-shaped speech-envelope-modulated noises that were co-located with the target. The speech envelopes that modulated the maskers were presented time-forward versus time-reversed. No significant difference was observed in threshold T/Ms between these two noise masker conditions, suggesting that EM was the same for both because the small amount of IM expected from modulated noise maskers would be the same as well. They concluded that the large reduction in masking found in the actual SOS conditions (about 12 dB) therefore was due to a release from IM and not to a difference in EM. Recent work from Kidd et al. (2016) using the ideal T-F segregation technique (e.g., Fig. 4.3) applied to time-forward and time-reversed speech supports the conclusion by Marrone and colleagues that the amount of EM for the two cases is the same. It should be noted that both Marrone and colleagues and Kidd and colleagues used (different) closed-set speech tests that have been shown to produce high IM. It is not yet clear whether the conclusion above generalizes to other types of speech materials and testing procedures and perhaps accounts for the small difference with the findings by Rhebergen et al. (2005) noted earlier in this section.

Further evidence that the meaningfulness of the masker may exert a strong effect in SOS masking comes from Kidd et al. (2008b; see also Best et al. 2011), who employed a variation on the “every other word” paradigm devised by Broadbent (1952b). In that paradigm, as implemented by Kidd and colleagues, five-word sentences from a closed-set corpus consisting of one random selection from each of five word categories (name, verb, number, adjective, object) were used to generate syntactically correct sentences (e.g., “Sue bought four old toys”). On any given trial, the target words formed the odd-numbered elements in a sequence with the even-numbered elements being masker words, time-reversed masker words, or noise bursts. When the masker was bursts of noise, performance was the same as when no masker was present. A small decrement in performance was found for the time-reversed speech masker but much less than was found for the meaningful time-forward speech (however, as noted in Sect. 4.3.3, masker syntax did not affect performance). This is a clear case in which speech caused significant IM with little or no EM. It should be pointed out that the small difference between the effect of the noise masker and the time-reversed speech masker is consistent with the view that even unintelligible speech—or masking stimuli that mimic the properties of speech such as speech-shaped speech-envelope-modulated noise—produces some amount of IM.

Swaminathan et al. (2015) reported large reductions (16–18.5 dB) in T/M at threshold for a target talker masked by two independent, same-sex masker talkers when the masker talkers were time reversed relative to when they were presented naturally. These large threshold reductions were obtained using the same closed-set speech materials employed by Kidd et al. (2010) in the study noted earlier in this section. Swaminathan and colleagues examined one factor potentially related to the individual differences observed between subjects: musical training. The results of this study are shown in Fig. 4.4A.

Fig. 4.4 (A) Group mean thresholds (target-to-masker ratio, TMR, in decibels) and standard errors for co-located (COL) and spatially separated (SEP) conditions for natural (FWD) and time-reversed (REV) speech maskers. The squares show the results from the musician group while triangles are for nonmusicians. The asterisks indicate statistically significant group differences. (From Swaminathan et al. 2015, *Scientific Reports*, with permission.) (B) Results from individual listeners plotted as reversed masking release (RMR) as a function of spatial masking release (SRM)



Group mean T/Ms at threshold are plotted for musicians and nonmusicians for time-forward and -reversed speech maskers presented in co-located and spatially separated configurations. Thresholds in the co-located condition for the forward speech maskers were about the same for the different subject groups, with relatively small differences observed across subjects. Either spatial separation or time reversal produced large reductions in T/Ms at threshold. Musicians as a group showed greater masking release for both variables than did their nonmusician counterparts. Large individual differences were observed for both subject groups. This is illustrated in Fig. 4.4B, in which the spatial release from masking (SRM) is plotted against the reduction in threshold that occurred due to masker time reversal (RMR) for individual subjects. The two subject groups are indicated by different symbols. The significant correlation between these variables suggests that subjects tended to exhibit a similar proficiency in using either variable to overcome IM

(see also Kidd et al. 2016). It also is clear that, despite the overlap in the distributions, most individual musically trained listeners exhibited greater masking release than the nonmusicians. Supporting evidence for this finding was reported by Clayton et al. (2016; see also Başkent and Gaudrain 2016) who found that the best predictors of individual differences in SRM were musicianship and performance on a visual selective attention task. Swaminathan and colleagues argued that the differences between groups were more likely due to central factors related to training and/or innate ability than to differences in peripheral auditory mechanisms. They employed a physiologically inspired model of the responses of the auditory nerve (AN) to determine whether the large RMRs found experimentally could be accounted for by a decrease in EM. The performance predicted by the AN model, however, was roughly equivalent for the time-forward and -reversed conditions. Swaminathan and colleagues concluded that the large RMRs found in their study were not due to differences in EM but rather to differences in IM.

4.3.3.2 Familiar Versus Unfamiliar Languages as Maskers

As noted in Sect. 4.2, the attempt to determine whether the masking produced by a familiar language was greater than that produced by an unfamiliar language dates at least to the report by Miller (1947). Although early work did not find much evidence that SOS masking varied depending on whether the masker was understandable or not, more recent work clearly has shown that this can be the case. Freyman et al. (2001) reported small differences in masking between Dutch and English sentence-length maskers on the intelligibility of English target speech by native English listeners who did not understand Dutch. The differences they reported were as large as 10 percentage points at low T/Ms in a reference condition in which the target and masker were co-located (the study focused on the benefit of perceptual segregation of sources based on apparent location differences). In the Rhebergen et al. (2005) study discussed in Sect. 4.3.3.1, only a 2-dB difference in masked speech reception thresholds (SRTs) was found for maskers in familiar (Dutch) versus unfamiliar (Swedish) languages.

In an important study specifically designed to determine whether knowledge of the language spoken by the masker talker affects the amount of SOS masking, Van Engen and Bradlow (2007) tested the recognition of simple meaningful English sentences masked by speech in either a known (English) or unknown (Mandarin) language. The maskers were two or six concurrent talkers uttering semantically anomalous (implausible) sentences. The target speech was distinguished from the masking speech by the nature of the materials and by a temporal offset between the masker and the target. Van Engen and Bradlow found that speech recognition performance was poorer when the masker was English, particularly at low T/Ms, and comprised two masker talkers rather than six. The broad conclusion was that greater masking occurs when the masker is intelligible to the listener. Thus, English is a more effective masker than Mandarin for English-speaking listeners, especially

when the maskers comprise distinctly individual, salient sources as opposed to multitalker babble.

A number of other studies have provided evidence that the amount of masking obtained in SOS masking experiments is greater when the masker language is familiar to the listener than when it is not, even after accounting for language-specific acoustic differences (e.g., Calandruccio et al. 2010, 2013). When the masker language is unfamiliar to the listener, there is little reason to expect that the masking it produces is substantially different from that produced by a familiar language that is unintelligible due to time reversal. The relatively small effects of maskers in familiar versus unfamiliar languages reported to date thus seems inconsistent with the large—and in some cases very large—masking release found for masker time reversal noted in Sect. 4.3.3 (e.g., 15–19 dB by Kidd et al. 2010 and Swaminathan et al. 2015). The reason for this discrepancy is not clear at present but may be due in part to differences in the procedures that have been used to study these issues.

The semantic content of speech may influence its effectiveness as a masker when the language in which it is spoken is native or otherwise well known to the listener. However, a much more complicated case arises when the target speech or the masker speech, or both, are spoken in a language known to the listener but are not the native or primary language (e.g., Cooke et al. 2008; Brouwer et al. 2012; Calandruccio et al. 2013). There are several possible combinations of talker–listener languages that may occur, and there are the further complications of the linguistic similarity between the target and masker speech together with the possibility that the unfamiliar language is actually partially comprehensible by the listener. If the target speech is in a language that is not well known/native to the listener, so that it requires greater effort and/or time for the listener to decipher, then it may be more susceptible to interference from other speech, especially if that speech is in the primary language. Conversely, if the target is in the primary language but the masker speech is not, the masker speech likely may be less distracting than if it is easily recognized (in the limit, as above, a completely unfamiliar language would cause relatively little IM). A general principle that appears to summarize many of the observations about primary and secondary language SOS masking, as well as other higher-level effects, was proposed by Brouwer and colleagues (2012) and is referred to as the “linguistic similarity” hypothesis.

In a study that emphasized the importance of linguistic factors, Ezzatian et al. (2010) measured SOS performance when the target and masker speech were in an acquired secondary language (English) as a function of the age of acquisition by the listener and compared performance to that of native English language listeners. They measured performance at different T/Ms for two spatial conditions, one where the target and masker were co-located and a second where target and masker were perceived at different spatial locations using the method of Freyman et al. (1999). The key findings are depicted in Fig. 4.5; open and filled symbols represent co-located and spatially/perceptually separated conditions, respectively.

The left column shows the results from a noise masker used as a high-EM control while the right panel shows results from a two-talker same-sex masker.

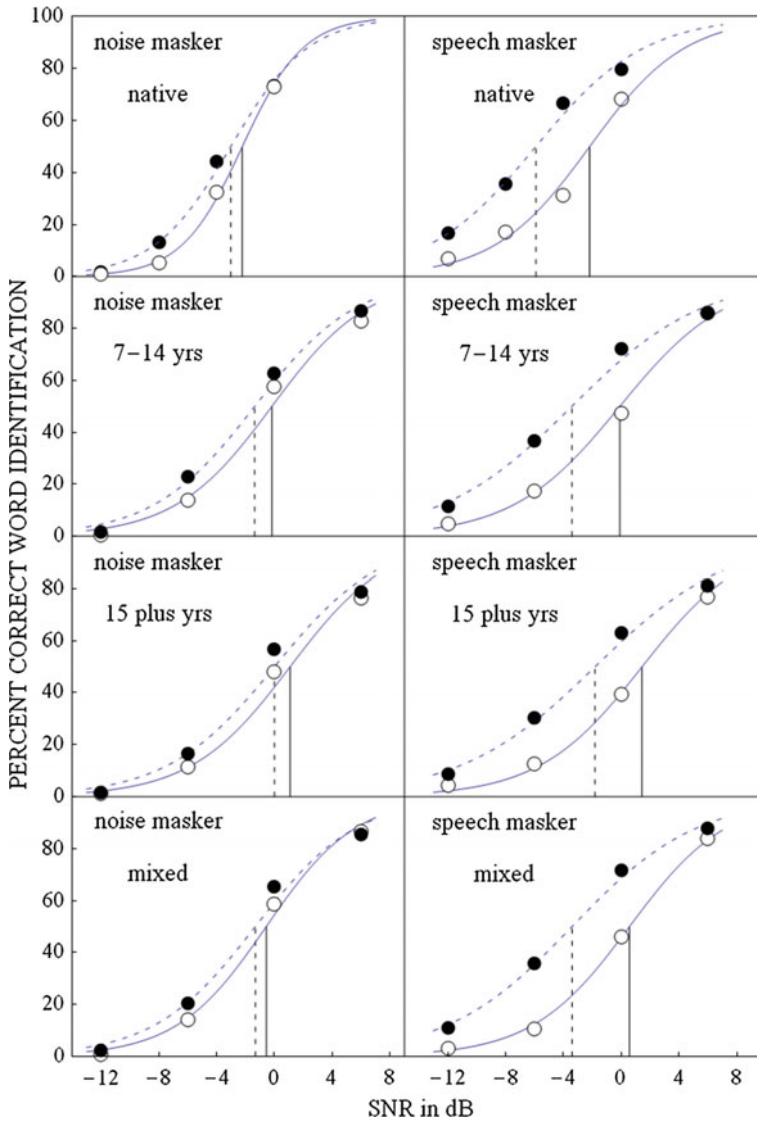


Fig. 4.5 Word identification performance as a function of signal-to-noise ratio (SNR) in decibels for four groups based on age of acquisition of English (native listeners; 7-14 years; 15-plus years; mixed: those who were raised in a non-English environment but learned to speak English at an early age). The left column is for a noise masker while the right column is for a speech masker. The open circles/solid lines represent spatially co-located target and masker. Solid circles/dashed lines indicate target and masker perceived from different locations. Thresholds (50% points on the psychometric functions) are indicated by the solid vertical lines for the co-located conditions and by the dashed vertical lines for the separated conditions. (From Ezzatian et al. 2010, *Speech Communication*, with permission.)

The rows are for different groups divided according to the age at which English was acquired. The important findings for this discussion are that performance was generally better (masking was less) when English was native (top row) or acquired early (7–14 years of age) as opposed to later (15 years or older) or in a “mixed” language environment where there was exposure to English from childhood but not as the primary language. Age of acquisition was less of a factor for the noise masker. A related finding concerning age of language acquisition was reported by Newman (2009). She tested infants’ ability to recognize their own names (respond preferentially re other names) against different types of backgrounds including the speech of a single talker presented naturally or reversed in time. She concluded that linguistic influences on IM develop as language is acquired and that infants have not yet acquired language to the point where meaningful speech interferes more than similar nonmeaningful speech. In a recent study, Newman et al. (2015) found that the greater masking effectiveness for meaningful speech, compared to the same speech rendered unintelligible by time reversal, was apparent for children by the age of 4–6 years. These findings suggest that susceptibility to IM in SOS masking is influenced by the degree of linguistic competence in the target language, at least as indicated by age/length of time of acquisition (see also Buss et al., 2016, and Calandruccio et al., 2016).

4.3.3.3 Syntactic and Semantic Content: Predictability and Obligatory Processing

Cherry’s (1953) seminal article exploring the factors governing communication performance in a “cocktail party” environment continues to be cited frequently for highlighting the importance of binaural processing of sounds and, less frequently, for identifying other relevant factors for separating competing talkers such as vocal characteristics and speech reading. However, what is often overlooked is that Cherry also emphasized the important role of predictability in natural communication and, indeed, the first experiment in his 1953 article was devoted to determining the effect of varying the predictability of speech by manipulating speaker transition probabilities. He states, “The logical principles involved in the recognition of speech seem to require that the brain have a vast “store” of probabilities, or at least of probability rankings. Such a store enables prediction to be made, noise or disturbances to be combatted, and maximum-likelihood estimates to be made” (p. 976). A number of speech corpora and tests have been developed subsequently that explicitly varied target speech predictability (e.g., Speaks and Jerger 1965; Kalikow et al. 1977; Uslar et al. 2013; Helfer and Jesse 2015).

Recently, Kidd et al. (2014) provided evidence suggesting that the predictability of sequences of words, as reflected by the conformance to a known syntax, can be beneficial in selectively attending to one of three spatially distributed speech sources. In their experiment, the intelligibility of target speech that comprised randomly selected words was compared to similar target speech arranged into brief,

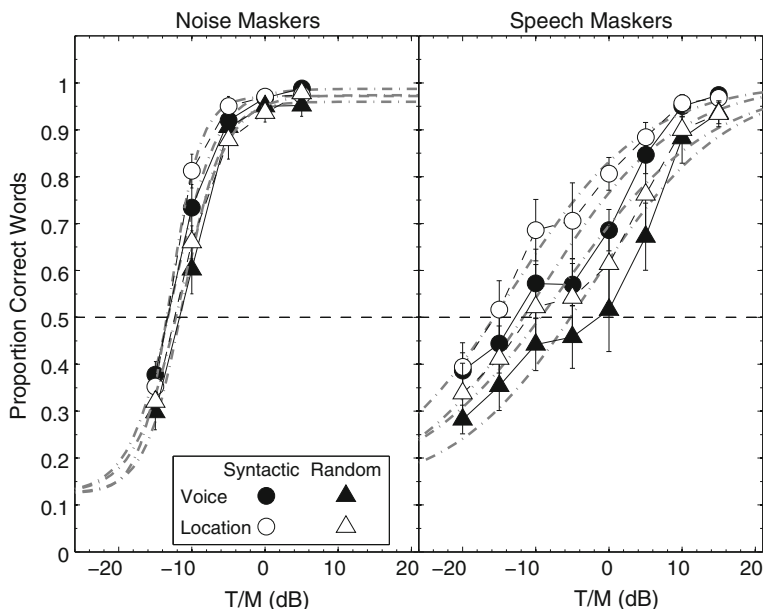


Fig. 4.6 Speech identification performance as a function of target-to-masker ratio (T/M) in decibels. The *left panel* contains the results for noise maskers while the *right panel* contains the results for speech maskers. The data points are group mean proportion correct scores and standard errors of the means. The fits are logistic functions (dashed-dotted lines) from which thresholds were obtained at the 0.5 proportion correct point (horizontal dashed line). The filled symbols are for conditions in which the target was indicated by constant voice while the open symbols are for conditions in which the target was indicated by constant location. Circles indicate that the target sentence was syntactically correct (*syntactic*) while triangles are for syntactically incorrect (*random*) target sentences. (From Kidd et al. 2014, *The Journal of the Acoustical Society of America*, with permission.)

syntactically correct, simple sentences masked by two competing talkers or by noise. Group mean results from that study are depicted in Fig. 4.6.

The left panel shows the intelligibility results obtained under two competing noise maskers while the right panel shows the results for two competing speech maskers. The primary cues to the target were constant talker voice or location, which were paired with correct or random target sentence syntax. In all cases, performance was better when the target conformed to correct syntax, but the differences—expressed as a reduction in T/M—were much larger when the maskers were speech. The authors concluded that the predictability of the target words conforming to a known syntax was particularly beneficial under conditions that were high in IM.

An earlier study by Freyman et al. (2004) demonstrated that priming a target sentence could improve performance under speech masking (but not noise masking) conditions relative to unprimed sentence presentation. They provided a prime by presenting a fragment of the target sentence spoken by the same talker that

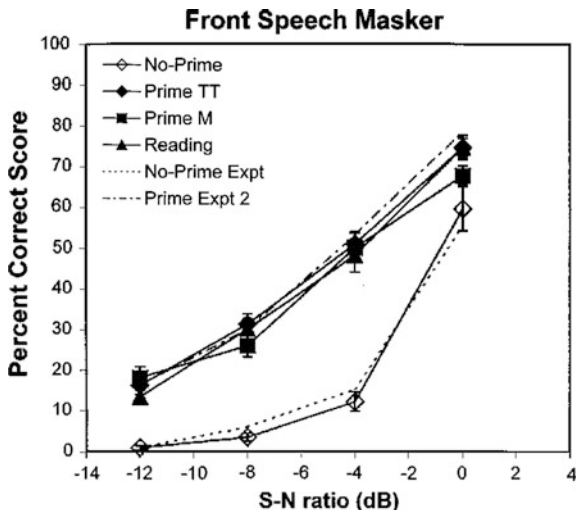


Fig. 4.7 Comparison of group mean percent correct scores and standard errors as a function of signal-to-noise ratio (S-N) for different priming conditions with target and masker co-located in the front. The control was the “no-prime” condition (open diamonds). “Prime TT” (filled diamonds) refers to the condition in which the target talker produced the priming utterance. “Prime M” (filled squares) is the condition in which the priming utterance was produced by a male (nontarget) talker. “Reading” (filled triangles) refers to the prime presented in print. Dashed/dotted lines without symbols show the primed and unprimed percent correct scores obtained in a separate experiment. (From Freyman et al. 2004, *The Journal of the Acoustical Society of America*, with permission.)

subsequently repeated the entire sentence as well as primes that were a different same-sex talker uttering the sentence fragment prime or the sentence fragment presented in written, rather than spoken, form. The results of this experiment are shown in Fig. 4.7. Rather remarkably, these three primes were equally effective in enhancing speech recognition performance. These effects were obtained using syntactically correct nonsense sentences masked by similar sentences from a different corpus for two co-located same-sex (as the target) masker talkers. Freyman and colleagues concluded that the benefit of the prime was to partially release IM by reducing the attentional resources devoted to the maskers.

Brouwer et al. (2012) proposed that the greater the degree of linguistic similarity between target and masker speech sources, the greater the IM that results. To test this “linguistic similarity hypothesis,” they varied the language of the target and masker talkers (i.e., target in one language, masker in the same or different language), measuring performance when the languages were primary, secondary, or the masker was not understood by the listener. They also varied the semantic value of the target and masker speech. For both manipulations—language and semantic content—the observed amount of masking increased when the target and masker speech were similar, as compared to dissimilar according to their criteria. Some of their findings are shown in Fig. 4.8.

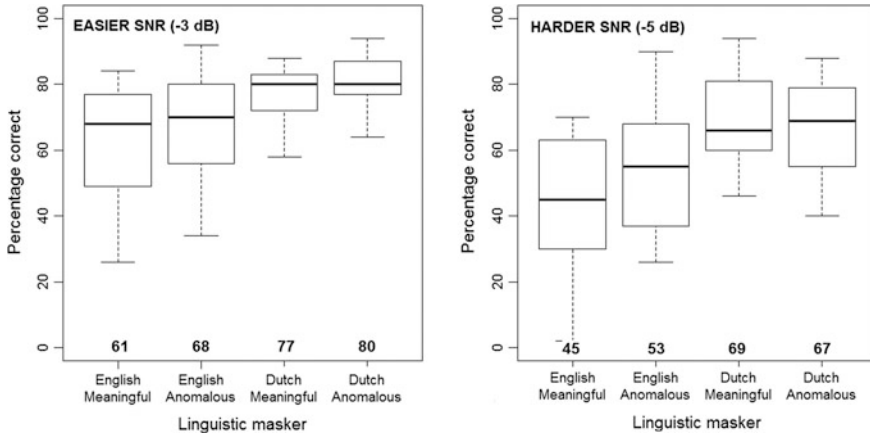


Fig. 4.8 Boxplots showing the interquartile ranges of intelligibility scores (in % correct) for English listeners on English target sentence recognition. The two panels show results at different signal-to-noise ratios (SNRs). The abscissa indicates masker type ordered according to decreasing linguistic similarity to the target. The mean percent correct score is given at the bottom of each plot. (From Brouwer et al. 2012, *The Journal of the Acoustical Society of America*, with permission.)

In general, the patterns of results were interpreted as being consistent with the linguistic similarity hypothesis. For these English-speaking listeners and meaningful English targets, performance was poorer when the masking speech was also in English than when the masking speech was in Dutch, a language that was not intelligible to these listeners. The differences due to language were more pronounced at the lower T/M. Furthermore, the “meaningful” English masker sentences produced more masking than did the semantically “anomalous” English sentences. These differences in performance due to linguistic factors occurred even in the absence of reliable differences in “general auditory distance” (low-level segregation cues) between stimuli. This idea of IM increasing in proportion to linguistic similarity was further supported by Calandruccio et al. (2013), who measured the masking of English target/masker speech for English-speaking listeners and compared it to that found for two maskers in languages unfamiliar to the subjects: Dutch and Mandarin. Furthermore, they attempted to control acoustically for differences in EM across languages so that the changes in performance that were found could then be attributed to IM. Their results indicated that comprehensible English was the most effective masker of English while Dutch maskers, which were judged to be more similar linguistically to English than were Mandarin maskers, produced more masking than Mandarin even though both the Dutch and Mandarin maskers were unintelligible. All three languages produced more masking than did a speech-spectrum-shaped noise-masker control.

Although qualitative differences between conditions can readily be specified, quantifying the degree of linguistic similarity may prove to be as challenging as quantifying the degree of IM in general. Furthermore, not all of the SOS masking

results support the linguistic similarity hypothesis. The Kidd et al. (2008b) study mentioned in Sect. 4.3.3 that used an adaptation of Broadbent's (1952b) "every other word" paradigm found no significant difference between masker speech that was syntactically correct versus the same speech that was not syntactically correct (presented in random word order). The target speech was also syntactically correct short sentences. A logical extrapolation of the linguistic similarity hypothesis discussed earlier in the preceding paragraph would seem to predict greater masking for the more similar masker; that is, the syntactically correct masker. However, the target sentences used by Kidd and colleagues, while syntactically correct, were low in semantic value and perhaps for that reason differences due to masker syntax were not apparent. Furthermore, although this method eliminates EM as a factor, the linguistic structure—as noted by Broadbent (1958)—may be so different than normal communication that it invokes a different form of processing than occurs in natural speech, perhaps reducing the effects of linguistic similarity that would otherwise occur.

To summarize, the available evidence suggests that predictability and linguistic similarity may exert a strong influence on the outcome of SOS masking experiments. However, disentangling linguistic effects from other factors, in particular low-level segregation cues or high-level selective attention, may be challenging and depends on the interactions of many variables such as the means of target source designation, the speech corpora used, and the specific methods that are employed. The extent to which linguistic factors govern performance in natural listening environments remains an intriguing question, with the answer likely to depend on obtaining a better understanding of the role of context and predictability in realistic sound fields.

4.4 Models of Binaural Analysis Applied to SOS Masking

As noted in Sect. 4.2, Cherry (1953) identified several factors that could affect human performance in solving the cocktail party problem. Of those factors, the spatial separation of sound sources subsequently received the greatest attention in the literature, and this attention helped to inspire the development and testing of models of the processes underlying spatial release from masking. The efforts to model binaural factors in SOS masking largely have been limited to extensions of the binaural models that have been developed to explain tone-in-noise and speech-in-noise stimulus configurations. Thus, although the speech masker produces a complex pattern of spectrotemporal overlap with a speech target, the underlying mechanism limiting performance is assumed to be energetic masking. The lack of explicit modeling applied to the issues specific to SOS masking (e.g., linguistic and cognitive factors influencing IM) likely is due, at least in part, to the multiplicity and complexity of the factors involved. Although it may be possible to construct experiments to isolate and control some of these factors, incorporating all

of these influences—and their interactions—into a comprehensive model of binaural analysis is a daunting task.

In the following paragraphs, the work to date is summarized, starting with the traditional waveform-based models of EM as developed originally for detecting tones in noise, followed by a discussion of the specializations that were incorporated to extend these models to predicting speech intelligibility in noise. A brief presentation of recent work is then provided that considers the importance of the significant spectrotemporal fluctuations found in speech masked by speech. None of these existing models explicitly account for the role of IM, but by comparing predictions of models that include as many of the factors as currently may be described, it is then possible to estimate the masking that is unaccounted for and to begin to develop new models that may be more comprehensive.

The earliest binaural models were based solely on differences in the interaural values of target and masker waveforms. Stimulated by the postulate from Jeffress (1948) of a network of coincidence detectors that were sensitive to interaural time delay/difference (ITD) in the binaural stimulus, Webster (1951) suggested that ITD might be the basis for binaural advantages in detection of tones in noise (i.e., MLDs). This concept received notable support from the findings of Jeffress and colleagues (1956), and it remains a viable hypothesis about the mechanism underlying binaural advantages for detection. Another early model devised to account for binaural detection advantages was proposed by Durlach (1963) and was termed the “equalization-cancellation (EC) model.” Put simply, the EC model postulated a binaural equalization of the masker using interaural time and level compensations followed by a (partial) cancellation of masker energy resulting in an improved target-to-masker ratio in the internal representation of the stimulus. Even today, these two models, or variations of these two models, form the bases for most explanations of binaural masking release and there continues to be active discussion and debate about the possible physiological mechanisms that might implement their processing.

These two models, and the modifications proposed to accommodate variations in model parameters, have evolved over the decades. Initially, work focused on tone-in-noise masking experiments with the goal of accounting for variations in parameters such as frequency and duration, and eventually the interaural parameters, of the target tone. Similar studies of how detection thresholds depended on the parameters of the Gaussian masking noise, including level, center frequency and bandwidth, and the interaural difference parameters (e.g., time delay, phase, level, and their interactions) contributed to the refinement of these models. A summary of much of this early work may be found in Colburn and Durlach (1978).

As was the case with SOS masking in general, the early models that attempted to account for the release from masking of speech resulting from interaural differences in target and masker focused on the case of speech masked by noise and assumed that the masking that occurred was predominantly EM. This view of binaural masking release for speech found considerable support from the work of Levitt and Rabiner (1967a, b), who combined the known frequency dependence of the MLD with Articulation Index (AI) theory (French and Steinberg 1947) to successfully

predict both the improvements in masked speech detection and recognition scores for different interaural parameters for target speech masked by noise. The empirical manipulations tested by Levitt and Rabiner involved reversing the interaural phase or delaying the waveform of target speech relative to the masking noise and doing so for various frequency regions. The binaural gain in intelligibility for the independently contributing bands of the AI was assumed to follow directly from the magnitude of the MLD for that frequency band. The success of this approach also was extended to the case of speech masked by a speech spectrum-shaped noise in a free-field environment by Zurek (1993), who accounted for the effects of head shadow in addition to the binaural analysis underlying the MLD measured under earphones. The maximum benefit predicted by Zurek's model was 8–10 dB, divided roughly equally between interaural differences in timing (MLD) and level (head shadow). Zurek's work provided a very good description of the spatial dependence of thresholds on the angle of the target speech and the angle of the masking noise. Performance with monaural listening alone was also considered. Overall, this work gave excellent support to the idea that, for these noise-masker cases, frequency bands were processed independently and combined to exploit the signal-to-noise advantages that were available in each band. In Levitt and Rabiner (1967a, b) and Zurek (1993) the underlying mechanism responsible for the binaural advantages found empirically was not specified but was assumed to be the same as that producing the MLDs on which the model predictions were based.

It is notable that all of the modeling discussed to this point was based on interference in speech reception caused by noise, which differs from the interference caused by speech in multiple ways. In terms of acoustic differences, speech-masker envelopes have greater fluctuations than steady-state noise maskers (even narrow-band filtered maskers), and there are times when the level of a speech masker may be negligible within one or more frequency bands (e.g., during gaps between the words comprising sentences or in lower-level phonemes such as voiceless consonants). One way to address this opportunity to “listen in the dips” of the masker envelope is to analyze binaural performance using a weighted combination of signal-to-noise ratios within individual time–frequency slices (T–F units; cf. Brungart and Iyer 2012; Best et al. 2015). This approach was used to model monaural speech intelligibility by Rhebergen and colleagues (2006) for both amplitude-modulated noise and speech maskers.

This time-dependent processing approach was extended to binaural models in a series of studies by a variety of investigators. Beutelmann et al. (2010) extended their binaural modeling of speech in wideband noise (see also Beutelmann et al. 2009) by allowing processing parameters to vary across time. Basically, they considered processing in separate T–F slices so that they could use appropriate parameters for maskers that were modulated in time. Their modeling was quite successful in comparing the different types of maskers. They concluded that listeners were able to process the stimulus binaurally according to separate T–F units, which supported the proposition that binaural model parameters could vary accordingly. This time-dependent EC processing was also suggested and used by Wan et al. (2010, 2014) to model the case of multiple speech maskers. They

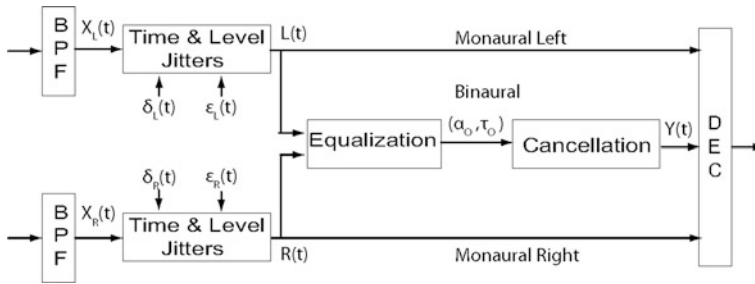


Fig. 4.9 Equalization–cancellation model of Durlach (1963) modified to include time-varying jitter. The *leftmost boxes* indicate the bandpass filtering stage (BPF) and the added time and level “jitter” for the left and right monaural channels. The binaural processing stages of equalization and cancellation are shown in *center boxes* followed by a decision mechanism (DEC). In the short-time EC (STEC) model used here, the equalization parameters α_0 and T_0 are adjusted to optimize cancellation within each time window. (From Wan et al. 2014, *The Journal of the Acoustical Society of America*, with permission.)

reasoned that independent speech-masker sources dominate in different time and frequency intervals, and so processing that was tuned to the dominant source would allow more efficient cancellation. Wan et al. (2014) demonstrated that many (but not all) of the spatial attributes of speech discrimination in the presence of multiple-speech maskers can be described with this type of binaural model. All of these model variations are based on extensions of the EC model. The basic processing implemented by these models is illustrated schematically in Fig. 4.9.

The inputs to the model are the acoustic waveforms arriving at the left and right ears. Each of these waveforms is passed through a bank of contiguous bandpass filters. The signals are represented in both the binaural pathway and the two monaural pathways and are corrupted by time-varying “jitter” in time and amplitude. These values are applied independently in each frequency channel and the equalization and cancellation process occurs in each time-frequency unit independently. A 20-ms sliding time window that is rectangular in shape is applied with an overlap between adjacent time windows of 10 ms.

These binaural models applied to multiple speech sources have not yet been modified to explicitly include IM. When target and masker speech sources are co-located there are no spatial cues to separate masker and target and, depending on the other source separation cues available, source confusions may occur resulting in a significant amount of IM. However, when speech interferers are spatially separated from the target, confusions about whether the target words come from the target source direction or from the masker source direction are greatly reduced, which in turn reduces source confusions and IM. This is illustrated for the case of two speech maskers in Fig. 4.10, which shows the results of applying the short-time EC (STEC) model to conditions with independent maskers on both sides as a function of separation of the maskers from the centered target.

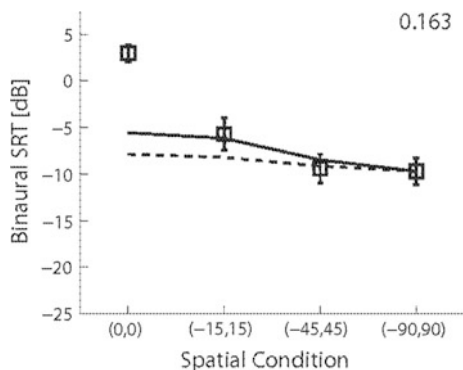


Fig. 4.10 Simulated and measured binaural speech reception thresholds (SRTs) as a function of spatial separation of two speech maskers from the target talker at 0° azimuth. Symbols are the measurements from Marrone and colleagues (2008), and the error bar is one standard error. Predicted values are connected by solid lines (short-term EC model) and dashed lines (steady-state EC model). The number in the upper right corner of the plot gives the value of the Speech Intelligibility Index criterion, which was chosen to match STEC prediction and data in the $(-90^\circ, +90^\circ)$ condition. (From Wan et al. 2014, *The Journal of the Acoustical Society of America*, with permission.)

More specifically, this figure shows the obtained (Marrone et al. 2008) and predicted speech-reception thresholds for conditions in which the speech target was presented from straight ahead of the listener (0° azimuth) and two independent speech maskers were presented from locations symmetrically separated from the target. The predictions of the STEC model are connected by the solid lines. The dashed lines connect predictions from the steady-state EC (SSEC) model without selectivity with respect to time (from Wan et al. 2010). Note that the predicted values were fit to the threshold for the widely separated $(-90^\circ, +90^\circ)$ masker condition (by adjusting a constant in the model). The thresholds for $\pm 15^\circ$ and $\pm 45^\circ$ angular separation are captured relatively well by the model, reflecting the ability of the model to describe the spatial aspects of the release from masking. The lack of IM in the model is illustrated by the poor fit for the co-located case where the amount of masking is almost ten decibels greater than in the $(-15^\circ, +15^\circ)$ separation case. This large increase in masking when the sources are co-located is consistent with significant confusions between the speech masker and the speech target. Because of the strong similarity between the targets and maskers (both were CRM sentences), performance in some cases was no better than would be expected simply from attending to the more intense (louder) talker. The resulting threshold of about 4 dB in the co-located condition is consistent with the idea of choosing the target on the basis of its higher level. This illustrates the need to incorporate IM in binaural models of SOS masking. Even when the T/Ms are sufficient to extract target information in a reasonable subset of T-F slices, the difficulty of perceiving/recognizing which samples contain information about the target itself leads to errors.

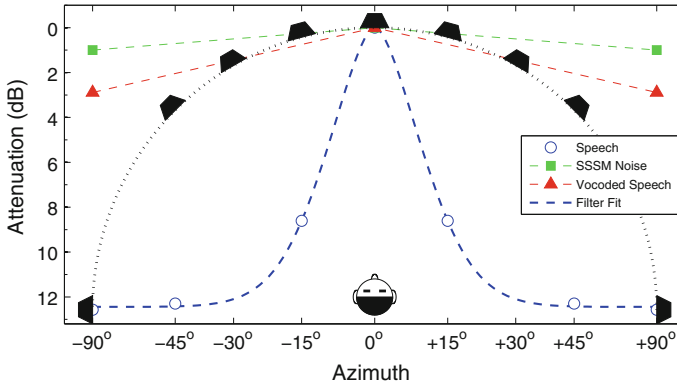


Fig. 4.11 Spatial tuning schematic showing attenuation of off-axis sources due to an attentional filter operating on interaural differences caused by different source locations (azimuth in degrees). The filter is oriented symmetrically around the point corresponding to 0° azimuth (directly in front of the simulated listener) and 0 dB attenuation. The amount of attenuation is assumed to be equal to the spatial release from masking (SRM) from human speech recognition experiments (Marrone et al. 2008) plotted in decibels and the roex filter function is a least-squares fit to the data. Overlaid on the spatial filter plot is a second schematic representing the location and arrangement of the listener and loudspeakers in a typical speech recognition experiment as used to measure SRM. The open circles on the filter function are group mean results for two independent speech maskers; the squares are data obtained using the same subjects and procedures but for two independent speech-shaped speech envelope–modulated noise maskers (also from Marrone et al. 2008) and the triangles are from eight-channel noise-vocoded speech maskers separated by $\pm 600 \mu\text{s}$ ITDs under earphones, one to the extreme left and the other to the extreme right

The net result of binaural processing may be viewed conceptually as implementing a “spatial filter” that attenuates sounds along the horizontal (azimuthal) dimension (or potentially other spatial dimensions). This perspective was proposed by Arbogast and Kidd (2000), who used a variation of the “probe-signal” method to obtain accuracy and response-time measures that exhibited “tuning” in azimuth in sound field conditions high in IM. The basic idea is illustrated schematically in Fig. 4.11.

In this illustration, a listener is located in the center of a semicircle of loudspeakers from which target and masker sound sources may be presented. This physical layout is illustrated by the sketch of loudspeakers along the dotted-line semicircle; this sketch is not related to the labeling of the axes, which is used for the empirical data plotted as open circles, green squares, and red triangles. These data are all for the case with the target source at 0° azimuth and with interfering sources symmetrically located at the azimuths where the data are plotted (and so as to appear filter-like are mirrored in the two hemispheres). The ordinate denotes the attenuation by the hypothetical “spatial filter.” The filter is shown by the smoothed function that peaks at 0 dB/ 0° azimuth and attenuates sounds off-axis symmetrically around the target location. The arbitrarily chosen filter function has the “rounded exponential” shape often used to represent filtering in the auditory system

along the frequency dimension. The values for the filter parameters were obtained from least-squares fits to SOS masking data from Marrone et al. (2008) and those thresholds are plotted as open circles along the fitted function. In the Marrone and colleagues experiment, there were two independent speech maskers that, when separated, were located symmetrically around the target location (one to either side). Conceptually, the attenuation of the filter is proportional to the amount of SRM measured in speech identification experiments; in this case, the data from Marrone and colleagues were obtained using the CRM materials/procedures. The maximum attenuation—equal to the maximum SRM—is about 12 dB. Two other sets of thresholds are also plotted representing results obtained with maskers producing lower levels of IM: one set obtained using two independent speech-shaped speech-modulated noises (also from Marrone et al. 2008) and the other obtained using “distorted” but intelligible eight-channel noise-vocoded speech (Best et al. 2012) separated by ITDs ($\pm 600 \mu\text{s}$). These thresholds emphasize the point that the amount of “attenuation” of masking (i.e., masking release) that is possible by the attention-based spatial filter is limited by the amount of IM that is present.

4.5 Summary

Early in the history of study of SOS masking, the potential influence of nonperipheral mechanisms was considered by leading auditory and speech scientists. Although the empirical work available at the time often did not support drawing strong conclusions about peripheral versus central components of masking, it is clear from Miller’s (1947) work that factors such as the intelligibility of competing speech or the uncertainty of the listening situation (e.g., “improbable vocal effects”) motivated the design of his experiments. In his famous article that coined the term “cocktail party problem,” Cherry (1953) elaborated several factors that human observers could use to solve the SOS masking problem, some of which fundamentally involved significant processing beyond the auditory periphery. The evidence he presented indicating that listeners perceived only certain attributes of unattended sounds presented to one ear while engaged in the recognition of speech in the contralateral attended ear demonstrated the existence of central effects and encouraged decades of study of the binaural processing of sounds. Perhaps as importantly, though, sophisticated higher-level mechanisms were implicated in Cherry’s observations about the importance of the transition probabilities inherent to normal speech. The idea that aspects of natural speech communication—e.g., turn-taking in conversation, sources joining or leaving the auditory “scene,” the unpredictable mixing of speech and nonspeech competition—involve the exploitation of predictability (e.g., that “a vast store of probabilities allows...noise or disturbances to be combatted”) is an underappreciated observation that has found increasing relevance as tools for studying perception in natural sound fields have been developed. Unambiguous evidence for SOS masking that could not be accounted for by peripheral overlap of sounds was provided by Broadbent (1952a,

b), who later argued convincingly for the important role of central factors. The importance of these factors in solving SOS masking problems led Carhart et al. (1969a, b) to propose a separate category of masking, termed perceptual masking, to account for otherwise unexplained results.

Numerous examples of the influence of what is now termed IM may be found in the modern-day literature. That is, reports of large masking effects beyond those that can be attributed to EM are commonplace and variables that lead to perceptual segregation of sources—without accompanying reductions in EM—have been found to produce significant release from masking in SOS conditions. In many instances, clear demonstrations of the role of linguistic variables in producing, or releasing, SOS masking have been reported that cannot be attributed to changes in the peripheral overlap of sounds. Historically, Theories explaining the masking of speech paralleled those of masking in general. Although such theories provide a good account of conditions dominated by EM, they are less successful in accounting for conditions dominated by IM. With respect to the causes of IM, even early work (e.g., Broadbent, 1952b) implicated the important role of failures of selective attention. However, the complex interaction of attention and memory and, particularly, the complications inherent to the comprehension of multiple streams of speech, caution against assigning IM to simple categories or attributing its effects exclusively to any single mechanism or process (cf. Watson 2005; Kidd et al. 2008a; Mattys et al. 2012).

The benefits of interaural differences between target and masker have been the subject of considerable modeling efforts over the years. These models originally were intended to account for the empirical findings from experiments in which tones or speech were masked by noise. As these models developed over time they were adapted to account for some of the spectrotemporal fluctuations of speech maskers and thus allowed the model parameters to vary across frequency channels or even small T-F units. The underlying physiological mechanism that could achieve this fine-grained parameter variation—whether it would respond solely to low-level stimulus features common to T-F units from the same source or would require some higher-level influence—presently is unclear. However, the underlying assumptions of even these refinements of traditional models of binaural analysis do not adequately provide for IM, as discussed in Sect. 4.4 The assumption that only the channels (or T-F units) containing target energy govern performance—and all other channels/units may be disregarded—does not provide for the deleterious effects of those units that are dominated by masker energy. It is clear from studies of SOS masking, however, that humans cannot disregard the nontarget energy in such units that may exert a profound influence on overall performance. Thus, what often could matter the most is not improving the T/M in units with significant target energy as much as it is minimizing masker energy in units where it is dominant. Current modeling approaches may be adapted to account for such circumstances (e.g., the EC model could null locations containing high-IM sources) but the higher-level processes that come into play with such putative mechanisms are quite complex.

Acknowledgements The authors are indebted to Christine Mason for her comments on this chapter and for her assistance with its preparation. Thanks also to Elin Roverud and Jing Mi for providing comments on an earlier version and to the members of the Psychoacoustics Laboratory, Sargent College graduate seminar SLH 810, and Binaural Group for many insightful discussions of these topics. We are also grateful to those authors who generously allowed their figures to be reprinted here and acknowledge the support of the National Institutes of Health/National Institute on Deafness and Other Communication Disorders and Air Force Office of Scientific Research for portions of the research described here.

Compliance with Ethic Requirements

Gerald Kidd, Jr. declares that he has no conflict of interest.

H. Steven Colburn declares that he has no conflict of interest.

References

- ANSI (American National Standards Institute). (1997). *American National Standard: Methods for calculation of the speech intelligibility index*. Melville, NY: Acoustical Society of America.
- Arbogast, T. L., & Kidd, G., Jr. (2000). Evidence for spatial tuning in informational masking using the probe-signal method. *The Journal of the Acoustical Society of America*, *108*(4), 1803–1810.
- Arbogast, T. L., Mason, C. R., & Kidd, G., Jr. (2002). The effect of spatial separation on informational and energetic masking of speech. *The Journal of the Acoustical Society of America*, *112*(5), 2086–2098.
- Başkent, D. & Gaudrain, E. (2016). Musician advantage for speech-on-speech perception. *The Journal of the Acoustical Society of America*, *139*(3), EL51–EL56.
- Beranek, L. (1947). Design of speech communication systems. *Proceedings of the Institute of Radio Engineers*, *35*(9), 880–890.
- Best, V., Marrone, N., Mason, C. R., & Kidd, G., Jr. (2012). The influence of non-spatial factors on measures of spatial release from masking. *The Journal of the Acoustical Society of America*, *131*(4), 3103–3110.
- Best, V., Mason, C. R., Kidd, G. Jr., Iyer, N., & Brungart, D. S. (2015). Better ear glimpsing efficiency in hearing-impaired listeners. *The Journal of the Acoustical Society of America*, *137*(2), EL213–EL219.
- Best, V., Mason, C. R., & Kidd, G., Jr. (2011). Spatial release from masking as a function of the temporal overlap of competing maskers. *The Journal of the Acoustical Society of America*, *129*(3), 1616–1625.
- Best, V., Ozmeral, E. J., & Shinn-Cunningham, B. G. (2007). Visually-guided attention enhances target identification in a complex auditory scene. *The Journal of the Association for Research in Otolaryngology*, *8*, 294–304.
- Beutelmann, R., Brand, T., & Kollmeier, B. (2009). Prediction of binaural speech intelligibility with frequency-dependent interaural phase differences. *The Journal of the Acoustical Society of America*, *126*(3), 1359–1368.
- Beutelmann, R., Brand, T., & Kollmeier, B. (2010). Revision, extension, and evaluation of a binaural speech intelligibility model. *The Journal of the Acoustical Society of America*, *127*(4), 2479–2497.
- Broadbent, D. E. (1952a). Listening to one of two synchronous messages. *The Journal of Experimental Psychology*, *44*(1), 51–55.
- Broadbent, D. E. (1952b). Failures of attention in selective listening. *The Journal of Experimental Psychology*, *44*(6), 428–433.
- Broadbent, D. E. (1958). *Perception and communication*. Oxford: Pergamon Press.
- Bronkhorst, A. W. (2015). The cocktail-party problem revisited: Early processing and selection of multi-talker speech. *Attention, Perception, & Psychophysics*, *77*(5), 1465–1487.

- Brouwer, S., Van Engen, K., Calandruccio, L., & Bradlow, A. R. (2012). Linguistic contributions to speech-on-speech masking for native and non-native listeners: Language familiarity and semantic content. *The Journal of the Acoustical Society of America*, *131*(2), 1449–1464.
- Brungart, D. S. (2001). Informational and energetic masking effects in the perception of two simultaneous talkers. *The Journal of the Acoustical Society of America*, *109*(3), 1101–1109.
- Brungart, D. S., Chang, P. S., Simpson, B. D., & Wang, D. (2006). Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation. *The Journal of the Acoustical Society of America*, *120*(6), 4007–4018.
- Brungart, D. S., Chang, P. S., Simpson, B. D., & Wang, D. (2009). Multitalker speech perception with ideal time-frequency segregation: Effects of voice characteristics and number of talkers. *The Journal of the Acoustical Society of America*, *125*(6), 4006–4022.
- Brungart, D. S., & Iyer, N. (2012). Better-ear glimpsing efficiency with symmetrically-placed interfering talkers. *The Journal of the Acoustical Society of America*, *132*(4), 545–2556.
- Brungart, D. S., & Simpson, B. D. (2004). Within-ear and across-ear interference in a dichotic cocktail party listening task: Effects of masker uncertainty. *The Journal of the Acoustical Society of America*, *115*(1), 301–310.
- Buss, E., Grose, J., & Hall, J. W., III. (2016). Effect of response context and masker type on word recognition. *The Journal of the Acoustical Society of America*, *140*(2), 968–977.
- Calandruccio, L., Brouwer, S., Van Engen, K., Dhar, S., & Bradlow, A. (2013). Masking release due to linguistic and phonetic dissimilarity between the target and masker speech. *American Journal of Audiology*, *22*(1), 157–164.
- Calandruccio, L., Dhar, S., & Bradlow, A. R. (2010). Speech-on-speech masking with variable access to the linguistic content of the masker speech. *The Journal of the Acoustical Society of America*, *128*(2), 860–869.
- Calandruccio, L., Leibold, L. J., & Buss, E. (2016). Linguistic masking release in school-age children and adults. *American Journal of Audiology*, *25*, 34–40.
- Carhart, R., Tillman, T. W., & Greetis, E. S. (1969a). Release from multiple maskers: Effects of interaural time disparities. *The Journal of the Acoustical Society of America*, *45*(2), 411–418.
- Carhart, R., Tillman, T. W., & Greetis, E. S. (1969b). Perceptual masking in multiple sound backgrounds. *The Journal of the Acoustical Society of America*, *45*(3), 694–703.
- Carlile, S. (2014). Active listening: Speech intelligibility in noisy environments. *Acoustics Australia*, *42*, 98–104.
- Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and two ears. *The Journal of the Acoustical Society of America*, *25*(5), 975–979.
- Clayton, K. K., Swaminathan, J., Yazdanbakhsh, A., Patel, A. D., & Kidd, G., Jr. (2016). Executive function, visual attention and the cocktail party problem in musicians and non-musicians. *PLoS ONE*, *11*(7), e0157638.
- Colburn, H. S., & Durlach, N. I. (1978). Models of binaural interaction. In E. Carterette & M. Friedman (Eds.), *Handbook of perception: Hearing* (Vol. 4, pp. 467–518). New York: Academic Press.
- Cooke, M., Lecumberri, M. G., & Barker, J. (2008). The foreign language cocktail party problem: Energetic and informational masking effects in non-native speech perception. *The Journal of the Acoustical Society of America*, *123*(1), 414–427.
- Dirks, D. D., & Bower, D. R. (1969). Masking effects of speech competing messages. *Journal of Speech and Hearing Research*, *12*(2), 229–245.
- Durlach, N. I. (1963). Equalization and cancellation theory of binaural masking-level differences. *The Journal of the Acoustical Society of America*, *35*(8), 1206–1218.
- Egan, J. P., & Wiener, F. M. (1946). On the intelligibility of bands of speech in noise. *The Journal of the Acoustical Society of America*, *18*(2), 435–441.
- Ezzatian, P., Avivi, M., & Schneider, B. A. (2010). Do nonnative listeners benefit as much as native listeners from spatial cues that release speech from masking? *Speech Communication*, *52*(11), 919–929.
- Fletcher, H. (1940). Auditory patterns. *Review of Modern Physics*, *12*(1), 47–65.

- French, N. R., & Steinberg, J. C. (1947). Factors governing the intelligibility of speech sounds. *The Journal of the Acoustical Society of America*, 19(1), 90–119.
- Freyman, R. L., Balakrishnan, U., & Helfer, K. S. (2001). Spatial release from informational masking in speech recognition. *The Journal of the Acoustical Society of America*, 109(5), 2112–2122.
- Freyman, R. L., Balakrishnan, U., & Helfer, K. S. (2004). Effect of number of masker talkers and auditory priming on informational masking in speech recognition. *The Journal of the Acoustical Society of America*, 115(5), 2246–2256.
- Freyman, R. L., Helfer, K. S., & Balakrishnan, U. (2007). Variability and uncertainty in masking by competing speech. *The Journal of the Acoustical Society of America*, 121(2), 1040–1046.
- Freyman, R. L., Helfer, K. S., McCall, D. D., & Clifton, R. K. (1999). The role of perceived spatial separation in the unmasking of speech. *The Journal of the Acoustical Society of America*, 106(6), 3578–3588.
- Helfer, K. S., & Jesse, A. (2015). Lexical influences on competing speech perception in younger, middle-aged, and older adults. *The Journal of the Acoustical Society of America*, 138(1), 363–376.
- Hirsh, I. J. (1948). The influence of interaural phase on interaural summation and inhibition. *The Journal of the Acoustical Society of America*, 20(4), 536–544.
- Hygge, S., Ronnberg, J., Larsby, B., & Arlinger, S. (1992). ‘Normal hearing and hearing-impaired subjects’ ability to just follow conversation in competing speech, reversed speech, and noise backgrounds. *Journal of Speech and Hearing Research*, 35(1), 208–215.
- Iyer, N., Brungart, D. S., & Simpson, B. D. (2010). Effects of target-masker contextual similarity on the multimasker penalty in a three-talker diotic listening task. *The Journal of the Acoustical Society of America*, 128(5), 2998–3010.
- Jeffress, L. A. (1948). A place theory of sound localization. *Journal of Comparative and Physiological Psychology*, 41(1), 35–39.
- Jeffress, L. A., Blodgett, H. C., Sandel, T. T., & Wood, C. L. III. (1956). Masking of tonal signals. *The Journal of the Acoustical Society of America*, 28(3), 416–426.
- Johnsrude, I. S., Mackey, A., Hakymez, H., Alexander, E., et al. (2013). Swinging at a cocktail party: Voice familiarity aids speech perception in the presence of a competing voice. *Psychological Science*, 24, 1995–2004.
- Kalikow, D. N., Stevens, K. N., & Elliot, L. L. (1977). Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. *The Journal of the Acoustical Society of America*, 61(5), 1337–1351.
- Kellogg, E. W. (1939). Reversed speech. *The Journal of the Acoustical Society of America*, 10(4), 324–326.
- Kidd, G., Jr., Arbogast, T. L., Mason, C. R., & Gallun, F. J. (2005). The advantage of knowing where to listen. *The Journal of the Acoustical Society of America*, 118(6), 3804–3815.
- Kidd, G., Jr., Best, V., & Mason, C. R. (2008a). Listening to every other word: Examining the strength of linkage variables in forming streams of speech. *The Journal of the Acoustical Society of America*, 124(6), 3793–3802.
- Kidd, G., Jr., Mason, C. R., & Best, V. (2014). The role of syntax in maintaining the integrity of streams of speech. *The Journal of the Acoustical Society of America*, 135(2), 766–777.
- Kidd, G., Jr., Mason, C. R., Best, V., & Marrone, N. L. (2010). Stimulus factors influencing spatial release from speech on speech masking. *The Journal of the Acoustical Society of America*, 128(4), 1965–1978.
- Kidd, G., Jr., Mason, C. R., Richards, V. M., Gallun, F. J., & Durlach, N. I. (2008b). Informational masking. In W. A. Yost, A. N. Popper, & R. R. Fay (Eds.), *Auditory perception of sound sources* (pp. 143–190). New York: Springer Science + Business Media.
- Kidd, G., Jr., Mason, C. R., Swaminathan, J., Roverud, E., et al. (2016). Determining the energetic and informational components of speech-on-speech masking. *The Journal of the Acoustical Society of America*, 140(1), 132–144.
- Levitt, H., & Rabiner, L. R. (1967a). Binaural release from masking for speech and gain in intelligibility. *The Journal of the Acoustical Society of America*, 42(3), 601–608.

- Levitt, H., & Rabiner, L. R. (1967b). Predicting binaural gain in intelligibility and release from masking for speech. *The Journal of the Acoustical Society of America*, 42(4), 820–829.
- Licklider, J. C. R. (1948). The influence of interaural phase relations upon the masking of speech by white noise. *The Journal of the Acoustical Society of America*, 20(2), 150–159.
- Marrone, N. L., Mason, C. R., & Kidd, G., Jr. (2008). Tuning in the spatial dimension: Evidence from a masked speech identification task. *The Journal of the Acoustical Society of America*, 124(2), 1146–1158.
- Mattys, S. L., Davis, M. H., Bradlow, A. R., & Scott, S. K. (2012). Speech recognition in adverse conditions: A review. *Language and Cognitive Processes*, 27(7–8), 953–978.
- Miller, G. A. (1947). The masking of speech. *Psychological Bulletin*, 44(2), 105–129.
- Newman, R. (2009). Infants' listening in multitalker environments: Effect of the number of background talkers. *Attention, Perception, & Psychophysics*, 71(4), 822–836.
- Newman, R. S., Morini, G., Ahsan, F., & Kidd, G., Jr. (2015). Linguistically-based informational masking in preschool children. *The Journal of the Acoustical Society of America*, 138(1), EL93–EL98.
- Rhebergen, K. S., Versfeld, N. J., & Dreschler, W. A. (2005). Release from informational masking by time reversal of native and non-native interfering speech. *The Journal of the Acoustical Society of America*, 118(3), 1274–1277.
- Rhebergen, K. S., Versfeld, N. J., & Dreschler, W. A. (2006). Extended speech intelligibility index for the prediction of the speech reception threshold in fluctuating noise. *The Journal of the Acoustical Society of America*, 120(6), 3988–3997.
- Samson, F., & Johnsrude, I. S. (2016). Effects of a consistent target or masker voice on target speech intelligibility in two- and three-talker mixtures. *The Journal of the Acoustical Society of America*, 139(3), 1037–1046.
- Schubert, E. D., & Schultz, M. C. (1962). Some aspects of binaural signal selection. *The Journal of the Acoustical Society of America*, 34(6), 844–849.
- Schubotz, W., Brand, T., Kollmeier, B., & Ewert, S. D. (2016). Monaural speech intelligibility and detection in maskers with varying amounts of spectro-temporal speech features. *The Journal of the Acoustical Society of America*, 140(1), 524–540.
- Speaks, C., & Jerger, J. (1965). Method for measurement of speech identification. *Journal of Speech and Hearing Research*, 8(2), 185–194.
- Swaminathan, J., Mason, C. R., Streeter, T. M., Best, V. A., et al. (2015). Musical training and the cocktail party problem. *Scientific Reports*, 5, 1–10, No. 11628.
- Uslar, V. N., Carroll, R., Hanke, M., Hamann, C., et al. (2013). Development and evaluation of a linguistically and audiological controlled sentence intelligibility test. *The Journal of the Acoustical Society of America*, 134(4), 3039–3056.
- Van Engen, K. J., & Bradlow, A. R. (2007). Sentence recognition in native- and foreign-language multi-talker background noise. *The Journal of the Acoustical Society of America*, 121(1), 519–526.
- Wan, R., Durlach, N. I., & Colburn, H. S. (2010). Application of an extended equalization-cancellation model to speech intelligibility with spatially distributed maskers. *The Journal of the Acoustical Society of America*, 128(6), 3678–3690.
- Wan, R., Durlach, N. I., & Colburn, H. S. (2014). Application of a short-time version of the equalization-cancellation model to speech intelligibility experiments. *The Journal of the Acoustical Society of America*, 136(2), 768–776.
- Watson, C. S. (2005). Some comments on informational masking. *Acta Acustica united with Acustica*, 91(3), 502–512.
- Webster, F. A. (1951). The influence of interaural phase on masked thresholds. I: The role of interaural time-deviation. *The Journal of the Acoustical Society of America*, 23(4), 452–462.
- Webster, J. C. (1983). Applied research on competing messages. In J. V. Tobias & E. D. Schubert (Eds.), *Hearing research and theory* (Vol. 2, pp. 93–123). New York: Academic Press.
- Zurek, P. M. (1993). Binaural advantages and directional effects in speech intelligibility. In G. A. Studebaker & I. Hochberg (Eds.), *Acoustical factors affecting hearing aid performance* (pp. 255–276). Boston: Allyn and Bacon.

Chapter 5

Modeling the Cocktail Party Problem

Mounya Elhilali

Abstract Modeling the cocktail party problem entails developing a computational framework able to describe what the auditory system does when faced with a complex auditory scene. While completely intuitive and omnipresent in humans and animals alike, translating this remarkable ability into a quantitative model remains a challenge. This chapter touches on difficulties facing the field in terms of defining the theoretical principles that govern auditory scene analysis, as well as reconciling current knowledge about perceptual and physiological data with their formulation into computational models. The chapter reviews some of the computational theories, algorithmic strategies, and neural infrastructure proposed in the literature for developing information systems capable of processing multisource sound inputs. Because of divergent interests from various disciplines in the cocktail party problem, the body of literature modeling this effect is equally diverse and multifaceted. The chapter touches on the various approaches used in modeling auditory scene analysis from biomimetic models to strictly engineering systems.

Keywords Computational auditory scene analysis • Feature extraction • Inference model • Multichannel audio signal • Population separation • Receptive field • Source separation • Stereo mixture • Temporal coherence

5.1 Introduction

In everyday life, humans are constantly challenged to attend to specific sound sources or follow particular conversations in the midst of competing background chatter—a phenomenon referred to as the “cocktail party problem” (Cherry 1953). Whether at a real cocktail party, walking down a busy street, or having a conver-

M. Elhilali (✉)

Laboratory for Computational Audio Perception, Center for Speech and Language Processing, Department of Electrical and Computer Engineering, The Johns Hopkins University, 3400 North Charles Street, Baltimore, MD 21218, USA
e-mail: mounya@jhu.edu

sation in a crowded coffee shop, sounds reaching a listener's ears from a particular sound source almost never exist in isolation. They persistently occur in the presence of other competing sources and distractors that form a person's acoustic environment. This soundscape needs to be organized into meaningful percepts, a process formally called "auditory scene analysis" (ASA) (Cherry 1957; Bregman 1990).

The ASA challenge is not confined to humans. Animals too, including mammals, penguins, songbirds, and fishes, have to overcome similar difficulties to navigate their complex auditory scenes, avoid predators, mate, and locate their newborns (Izumi 2002; Aubin 2004). A similar challenge also faces engineering systems, from military communication and surveillance devices to smart phones. Much like biological systems, these technologies have to navigate their soundscapes to pick out relevant sound sources (e.g., speech) while ignoring interference from the surround (Loizou 2013).

It is important to note that auditory scene analysis is not a monolithic process that is easily defined within an exact framework. Despite its seemingly effortless and intuitive nature, it is a multifaceted challenge that encompasses various processes. It underlies the brain's ability to detect, identify, and classify sound objects; to robustly represent and maintain these representations amidst severe distortions; to guide actions and behaviors in line with complex goals and shifting acoustic soundscapes; to adapt to and learn from the environment; as well as to integrate potentially multimodal sensory cues with information in memory, prior knowledge, and expectations to provide a complete understanding of the scene.

Given its multilayered nature, modeling auditory scene analysis has often been faced with a lack of a unified vision or agreed-on benchmarks that clearly define the objectives to be achieved. These goals have varied from tracking only relevant targets in a scene to a complete scan of all elements in the scene. Despite this complexity, interest in addressing the problem computationally is driven by a number of aims: (1) The ability of the brain to parse informative sensory inputs and track targets of interests amidst severe, unknown, and dynamic interferers is ultimately what gives the biological system its lead over state-of-the-art engineering systems. Modern technologies strive to replicate this intelligent processing in computational systems. This goal remains one of the holy grails of audio and speech systems (Wang and Brown 2006). (2) Computational models of ASA can provide a strong perspective in guiding neural and perceptual investigations of the problem in both humans and animals (Cisek et al. 2007). (3) Defining theoretical principles that govern aspects of the cocktail party problem will guide the field to develop better benchmarks to compare performance across systems as well as match up different implementations against the biological system for well-defined subtasks. (4) Mathematical ASA models can also act as a platform to examine commonalities across different sensory modalities and shed light on questions of optimality and efficiency of performance of the biological or engineering system under different operating conditions and for various tasks and environments.

5.2 Defining the Problem in the Cocktail Party Problem

Exploring the computational principles of the cocktail party challenge requires articulating the exact nature of the problem itself as well as considering the architecture of models that could tackle this challenge. As is the case with the study of any complex system, it is important to define the system's input to the task at hand and the nature of the output. At the input level, the most biologically reasonable expectation of the input is the acoustic signal that impinges on the listener's ears either monaurally or binaurally. This corresponds to a single-microphone or two-microphone recording of the soundscape. Naturally, some engineering applications extend this notion to the possibility of multiple microphones, which expands the spatial resolution of the system, though taking it away from the realm of biological plausibility. This design takes full advantage of the role of spatial processing in analyzing complex soundscapes without limiting the engineering application to the same constraints of the biology. This view has indeed opened the door to many successful "solutions" to certain aspects of the cocktail party problem by using independent component analysis (ICA) (Hyvarinen et al. 2001) and other blind source separation (BSS) (Naik and Wang 2014) and beamforming techniques (van der Kouwe et al. 2001).

While choosing the number of input channels for a computational model is a relatively straightforward decision based on the desired fidelity of the model to biological processes, defining the actual goal for modeling the cocktail party problem is an ill-posed query (Haykin and Chen 2005; Lewicki et al. 2014). Brain mechanisms engaged in processing complex scenes can be interpreted at many levels. One level is as an *analysis* or *segmentation* goal that defines auditory scene analysis as a stream segregation problem, as envisioned by Bregman and Campbell (Bregman and Campbell 1971; Bregman 1981). In this view, the cocktail party problem describes the task whereby a listener is confronted with intertwined sound sequences from multiple sources and the brain must form separate perceptual streams (or "sound objects"). A computational implementation of this level focuses on segregating different sound sources based on their acoustic attributes, including their spatial location, and binding the appropriate elements together to represent the perceived streams in a multisource auditory scene. Although this definition identifies a goal for the computational algorithm, it maintains a significant degree of ambiguity when it comes to defining the exact relationship between the physical nature of the sound source and the perceived stream, which is not a one-to-one mapping.

Think, for instance, of a scenario in which the audience at a symphony hall is enjoying an orchestral concert. Although the sound sources can be discretely distinguished in acoustic space, the perceptual experience of this rich auditory scene is not trivial to segregate. Should the model distinguish woodwinds from the rest of the instruments or should it focus on flutes versus clarinets versus bassoons? Uniquely defining the granularity of this segregation task is simply impossible and ultimately depends either on the goals of the model/system, or—in the case of

modeling human behavior—on the specific task given to the listener along with any behavioral metrics. This subsequently raises questions as to the limits of incorporating information about the sources in the segregation process. Should the model have knowledge about what a flute or a clarinet sounds like?

More importantly, the segmentation of an auditory scene poses additional, larger, questions: should the segregation be confined to a two-stream problem consisting of segregating a foreground (or target) stream from the background that incorporates the entire remainder of the scene; or should the segregation truly represent “all” possible individual sound streams within the scene itself? When framed as a figure–ground segregation problem, the degree of complexity is greatly reduced. It is still incomplete, however, until additional processes (e.g., selective attention) are incorporated to help dictate what the target or foreground characteristics are. It also requires specifying the underlying priors as to “what” the target (or target class) is, what its attributes are, and whether there are descriptive or statistical models that define them.

Alternatively, one can take a different approach and cast the overall goal of the cocktail party model as arising from a *recognition* point of view. In this case, the objective is to provide a recognizable label of the soundscape. This view aligns with frameworks commonly employed in computer vision and traditions of visual scene perception (Riesenhuber and Poggio 2002; Xu and Chun 2009) and has found applications in many sound technologies and speech systems (Chen and Jokinen 2010). Such systems are developed to provide various informative descriptors about a given a scene; e.g. is human speech present in a recording? Which melody is playing right now? Can footsteps be tracked in a surveillance microphone? Is there an abnormal heart murmur in a stethoscope signal? Clearly, the range of information that can be potentially conveyed from an auditory scene can be limitless.

Existing technologies have successfully focused on particular aspects of this recognition task, especially recognizing a single target amidst interfering backgrounds such as human speech (Virtanen et al. 2012) or tune/melody recognition systems (Collins 2009). Alternatively, some systems focus on recognizing the environment that gave rise to the scene itself (Patil and Elhilali 2013; Barchiesi et al. 2015), while other systems target abnormal or unexpected events in a scene for surveillance and medical systems (Anemuller et al. 2008; Kaya and Elhilali 2013) or even attempt to learn from the surrounding soundscape (Buxton 2003).

Finally, another body of work interprets the cocktail party problem from a *synthesis* point of view, where the intent of the computational model is to synthesize individual streams following the segregation process (e.g., musical track separation [Collins 2009]), or extract cleaner or denoised versions of a target stream by suppressing undesired backgrounds, echoes, and reverberations, as is goal of speech enhancement (Loizou 2013). In these systems, the ultimate goal is to generate a simplified or cleaned version of the auditory scene that captures only one or a few signals of interest.

Overall, the lack of uniformity across the body of work addressing the computational bases of auditory scene analysis raises additional challenges when it comes to assessing the success of such systems: it becomes task dependent and

contingent on the perspective of the modeler. The lack of well-defined goals is one of the main hurdles that restricts progress in the field, constrains comparative studies of existing models, and limits incremental innovation that builds on the existing body of work. Ultimately, the cocktail party problem is an inherently cross-disciplinary challenge spanning domains of neuroscience, cognitive science, behavioral sciences, ethology, psychology, psychophysics, and medicine, as well as engineering and computer sciences. Naturally, the perspective of each of these disciplines puts the emphasis on different aspects of the problem and biases the computational theory to tackle the cocktail party problem at different levels of abstraction and granularity.

5.3 Principles of Modeling the Cocktail Party Problem

The cocktail party problem falls in the category of general information processing systems, which can be nicely framed in the context of Marrian models that emphasize different levels of granularity for understanding the underlying processes (Marr 1982). Although Marr's specific tri-level explanation may ultimately be incomplete (Poggio 2012), it nonetheless provides an integrated framework for understanding different levels of information processing. At the highest level, the *computational theory* describes the overall goal of the system and what a model of auditory scene analysis needs to achieve. In the case of the cocktail party problem, this remains one of the most challenging levels to describe. As highlighted in Sect. 5.2, the cocktail party effect is not a well-defined problem with an agreed-on objective. Most models strive to provide an informative mapping of a complex audio signal whether in the form of segregated streams, recognition of sound events, or synthesized variations of the same scene. At the next level of granularity, the *algorithm* describes the approach undertaken to achieve this goal. This level encompasses approaches based on analysis, recognition, or synthesis. At the lowest level, the *implementation* level details the practical realization of the algorithmic computation in terms of computational primitives or neural mechanisms at different levels of the hierarchy.

5.3.1 Algorithmic Strategies

The overall strategy undertaken by most models of the cocktail party problem focuses on invoking processes that extract “discriminative” cues from the incoming sensory input in such a way as to facilitate the differentiation of distinct sound streams or target selection. This is a particularly daunting task because these cues operate not only locally, but also globally, as sound streams evolve over time. These strategies have generally clustered into a few standard approaches, as outlined next.

5.3.1.1 The Population-Separation Theory

The premise of the “population-separation” theory and its related “peripheral channeling” account is that the perceptual organization of sounds into segregated streams is determined by the physical overlap between neural populations driven by sensory properties of the input. Van Noorden originally championed this principle in his doctoral work (van Noorden 1975), where he particularly emphasized the role of peripheral population separation. Specifically, sounds that activate separate peripheral channels (defined as tonotopic frequency channels or left–right lateral channels) would give rise to segregated stream percepts. A number of studies have in fact provided support for this observation confirming that formation of segregated auditory streams is strongest when sounds occupy separate peripheral channels (van Noorden 1977; Hartmann and Johnson 1991).

Subsequent experiments have contested the specific premise of peripheral channeling, showing that separate streams can in fact be formed even when sources share a common frequency range, as long as they differ along another acoustic dimension. Numerous psychoacoustic studies have shown that stream segregation can occur for sounds that differ in timbre (Cusack and Roberts 2000), bandwidth (Cusack and Roberts 1999), amplitude modulation rate (Grimault et al. 2002), binaural pitch (Akeroyd et al. 2005), unresolved pitch (Vliegen and Oxenham 1999), phase (Roberts et al. 2002), or perceived spatial location (Darwin and Hukin 1999; Gockel et al. 1999). Although most of these stimulus manipulations do not evoke peripheral channeling per se, they generate sound sources that activate separate neural channels at the brainstem or higher levels of auditory processing. In this way, these findings still support the more general population separation premise that activation of distinct neural populations (whether at peripheral or central nuclei of the auditory pathway) is a prerequisite for their perceptual segregation into distinct streams.

The population separation theory is supported by a number of neurophysiological studies that corroborate the role of feature selectivity in the auditory system in mediating the organization of sensory cues into segregated perceptual streams. Evidence of a correlation between responses at individual neuronal sites and perceptual judgments of streaming has been reported in animal models at various levels of processing from the cochlear nucleus (Pressnitzer et al. 2008) all the way to auditory cortex (Micheyl et al. 2007; Itatani and Klump 2011). Neural correlates of stream formation have also been explored in humans, using electroencephalography (EEG), magnetoencephalography (MEG), and functional magnetic resonance imaging (fMRI) (Simon, Chap. 7). Overall, human studies corroborate the role of feature selectivity and tonotopic organization along the auditory pathway in facilitating stream segregation.

Computationally, the role of population separation in the organization of auditory streams can be interpreted as providing a discriminable representation of acoustic cues that allows mapping the stimulus into a separable space. By projecting sensory information into a new feature space that provides non- or minimally overlapping manifolds of the data, the neural representation enhances

discriminability between different auditory streams in the scene, allowing them to be separated. This operation is reminiscent of classification and regression techniques such as support vector machines and kernel-based classifiers (Duda et al. 2000; Herbrich 2001).

5.3.1.2 The Temporal Coherence Theory

The general population-separation theory accounts for a number of perceptual findings about stream segregation induced by sufficiently salient differences across sound dimensions (Moore and Gockel 2002). However, it does not account for crucial aspects of stream segregation that relate to the relative timing between sound events. Specifically, as sounds evolve over time, the relative timing between individual components in a complex scene plays a crucial role in dictating whether these components will segregate as separate streams or group together as a single stream. For instance, frequency components that start together (i.e., share a common onset) are likely to be perceived as grouped together (Darwin and Carlyon 1995), while delays of a few tens of milliseconds can suffice to induce a segregated percept (Sheft 2008). Similarly, frequency channels that evolve together in time over hundreds of milliseconds are likely to be perceived as one group, whereas elements that are out of phase relative to each other are likely to segregate (Micheyl et al. 2013). These longer time constants over which sound features evolve directly influence the nature of the stimulus-induced neural response. Indeed, sound components—if sufficiently far apart, for example, in frequency—will activate clearly distinct frequency-selective neural populations regardless of whether there are perceived as segregated or grouped (Elhilali et al. 2009), hence violating the population separation premise.

The *temporal coherence* theory has been proposed to complement the population-separation theory by addressing its main shortcoming, notably by incorporating information about the relative timing across neural responses to sounds over longer time constants (Shamma et al. 2011). This concept emphasizes the notion of temporal coherence whereby neural populations whose responses are in phase relative to each other over long time windows (hundreds of milliseconds) should be treated as if they represent a perceptually coherent stream; conversely, neural populations whose responses are asynchronous should be treated as representing sounds that probably belong to different streams.

By combining together the ideas of feature selectivity (which is at the core of the population-separation theory) and grouping by temporal coherence, one obtains a general model of auditory stream formation as illustrated in Fig. 5.1. This model includes two main bottom-up stages: a feature analysis stage followed by a coherence analysis stage. The analysis of sound features begins with a frequency mapping, which simulates the spectral analysis performed at the level of the cochlea. The output of this initial frequency analysis is used to extract a variety of spectral and temporal sound features, including spectral shape and bandwidth, harmonicity, temporal periodicity, and interaural time and level differences. For

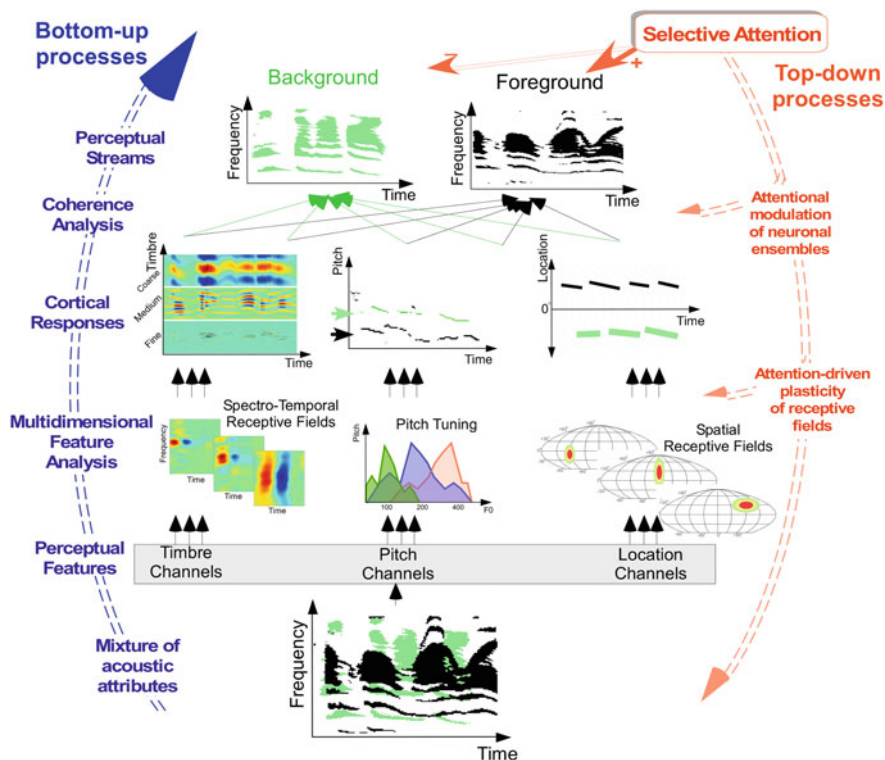


Fig. 5.1 Schematic of the temporal coherence strategy for modeling the cocktail party problem. An incoming signal (bottom of figure) consisting of a mixture of acoustic waveforms emanating from multiple sources is first analyzed through an array of auditory feature channels. These features extract cues (e.g., spatial location, pitch, etc.) that enable the segregation of sound attributes onto different perceptual streams. This process projects the low-dimensional acoustic signal onto a multidimensional space where different sound components occupy separate subspaces of the feature space, effectively segregating common sets of elements of the input and facilitating the process of formation of auditory objects or streams. This process takes advantage of the intricate time–frequency–space selectivity of neurons along the auditory pathway up to the level of auditory cortex. A coherence process tracks the trajectory of this feature space over “cortical” time constants of the order of few hundred milliseconds and binds together the elements that covary together, hence forming a representation of the foreground stream away from the background (top of figure). Top-down processes, particularly selective-attention (arrows on right-hand side) can modulate this entire process by exerting feedback projections that can reshape selectivity of cortical neurons or modulate ensemble of neurons. This process facilitates figure/ground segregation. [Figure adapted from Shamma et al. (2011).]

computational convenience, and illustration, these various feature detectors are assumed to be organized in “maps.” However, it is important to note that an orderly topographic representation of sound features is not required for the general model to operate. The key point is that the model includes neurons selective to different sound features, or different values of a particular feature. Temporal coherence then

operates on these neural outputs to bind together elements that covary over time, while segregating those that are out of synchrony relative to each other (Krishnan et al. 2014).

It is worth noting that the principle of temporal coherence falls in the general category of correlational models that have been proposed many decades ago to address the cocktail party problem (von der Malsburg 1994; Wang and Brown 1999). The correlation output is generated by an autonomous process via neural coupling that allows neurons to synchronize together if driven by temporally bound features, forming a topographic map. This concept has been formalized in computational scene analysis models using oscillatory networks, where each population of synchronized oscillators represents an auditory stream (Wang and Brown 1999; Brown et al. 2001). In the majority of these models, correlation is defined as pairwise instantaneous temporal coincidence between temporal trajectories along different acoustic features.

The concept of “temporal coherence” takes a different view than instantaneous associations across sound elements (Krishnan et al. 2014). It emphasizes correlations among slow-varying temporal outputs of feature-selective neurons over longer time scales—of the order of hundreds of milliseconds (Elhilali et al. 2009, 2010). These time scales are commensurate with dynamics observed in the mammalian primary auditory cortex. The contrast between the variable time scales of correlations between an oscillatory model and a temporal coherence model is highlighted in Eq. (5.1):

$$Cr_{ij} = \frac{1}{\Gamma} \int r_i(t)r_j(t)dt \text{ vs. } Ch_{ij} = \frac{1}{\Gamma} \int [r_i(t) *_t h_{\tau_k}(t)][r_j(t) *_t h_{\tau_k}(t)]^* dt \quad (5.1)$$

where $r_i(t)$ is the stimulus-driven response in the i th feature channel, Γ is an appropriately chosen normalization constant, $*_t$ represents convolution over time t , and $h_{\tau_k}(t)$ is the impulse response of a modulation-selective filter with time constant τ_k . $*$ is the conjugate symmetry operator that accounts for the fact that the filter $h_{\tau_k}(t)$ is modeled as a complex-valued system that reflects both the magnitude and phase alignment of the stimulus with the time integration channel τ_k . So, although both correlation and coherence are computing a coincidence across different feature channels; they are operating at different time scales. The former is an instantaneous correlation across pairs of feature channels, whereas the latter is an operator that tracks longer-term correlations, parameterized by filters $h_k(\cdot)$ over time constants τ_k . The coherence operator therefore is effectively tracking the trajectory of activity across feature channels, which results in a different tracing of coincidence across feature channels.

It is essential to note that the term temporal coherence used in the literature in the context of feature binding (Shamma et al. 2011) refers to stimulus-induced temporal coherence of neural activity and should not be confused with intrinsically generated temporal coherence, for example, oscillations in the gamma band. The current chapter refers specifically to the former. However, stimulus-induced neural

responses may interact with (and enhance or suppress) intrinsically generated temporal patterns of neural activity (Lakatos et al. 2005).

The role of temporal coherence in providing a framework for feature binding is not unique to the auditory modality, but has been advanced in other contexts and in other sensory modalities. It has been suggested that a similar principle operates in the visual modality (Alais et al. 1998; Blake and Lee, 2005). In addition, it has also been speculated that temporal coherence between cortical areas corresponding to different sensory modalities can, in principle, support cross-modal binding, for example, lip-reading, though not much is known about the exact role of cross-modal interactions in auditory stream formation (Almajai and Milner 2011; Mirbageri et al. 2012).

5.3.1.3 The Inference Theory

The concept of temporal coherence reviewed in Sect. 5.3.1.2 is based on a notion of tracking the temporal evolution of sound elements. A closely related strategy, posited as the underlying neural process for organizing a complex acoustic scene, is that of prediction-based or inference models (Winkler et al. 2009). Inference-based computation provides a framework for integrating all available cues (e.g., sensory, contextual, cognitive) to derive likely interpretations of the soundscape. Initially, this process maps the acoustic input onto a high dimensional representation or onto feature maps (akin to processes underlying population separation). This mapping parameterizes the acoustic environment along dimensions that represent an estimate of the likelihood of a particular decomposition of the soundscape, based on acoustic attributes. This representation can further be integrated with priors that represent sensory statistics or dynamics of the acoustic features, as well as potential contextual information and any additional prior knowledge. This evidence is then integrated using an optimal Bayesian framework or alternative strategies to infer knowledge about the state of the auditory scene and its constituent streams (Friston 2010; Elhilali 2013).

This inference process can take many forms. Arguably, one of the most biologically plausible implementations invokes predictive coding, which processes sensory information in terms of predictive interpretations of the underlying events in the scene (Mumford 1992; Rao and Ballard, 1999). The circuitry underlying such processing has been studied at various hierarchical levels and has been speculated to include microcircuitry spanning sensory, parietal, and frontal cortex (Bastos et al. 2012). In the context of the cocktail party problem, such mechanisms have been linked to the concept of regularity tracking as an underlying mechanism for perception in auditory scenes (Winkler et al. 2009). In this scheme, the brain's strategy is to capture the behavior of sound sources in the scene and their time-dependent statistics by inferring the evolution of sound streams: constantly generating new expectations that reflect the fidelity of the sensory evidence, and matching these predictions with the ongoing dynamics of the scene. This strategy has led to successful computational models of auditory scene analysis, framed either as discovery

of predictable patterns in the scene (Mill et al. 2013; Schroger et al. 2014) or as a tracking operator that follows the evolution of states in the auditory scene and integrates past behavior of sound sources with their expected trajectories (Elhilali and Shamma 2008). In many regards, the predictive tracking algorithm can be related to temporal coherence analysis, provided the temporal dynamics of both processes operate at similar “slow” time scales (4–20 Hz) commensurate with the neuronal dynamics at the level of primary auditory cortex (Krishnan et al. 2014).

5.3.1.4 Spatial Models

The spatial location of sound sources is one of the strong cues that facilitate the process of auditory scene analysis (Culling and Stone, Chap. 3). Acoustic events that emanate from the same location in space tend to be perceived as belonging to the same stream whereas events that originate from different locations tend to be assigned to different streams (Gilkey and Anderson, 2014). The effect of interferers on the perception of a target is greatly reduced when the signal and masker are perceived to be at different spatial locations, in a phenomenon referred to as spatial release from masking (Arbogast et al. 2002). The extent to which spatial separation of sound sources supports bottom-up stream segregation is an active topic of research (Middlebrooks, Chap. 6). Nevertheless, there is no doubt that spatial cues are crucial components in sound lateralization as well as object selection in complex soundscapes. As such, they have featured in a prominent role in a number of computational models of auditory scene analysis that operate with two or multiple microphones.

Models of the cocktail party for stereo and multimicrophone applications have indeed taken advantage of the spatial layout of the scene, either in conjunction with other acoustic cues or based solely on spatial processing. Bio-inspired models rely on binaural cues represented by interaural level, phase, or timing differences to facilitate the separation of sound components that originate from different locations (Stern et al. 2005). Central to these bio-inspired spatial models is the mechanism of cross-correlation or coincidence detection which allows a direct comparison of signals from the two ears. Building on a theory put forth by Jeffress (1948), an interaural cross-correlation is computed across different channels that often represent frequency-selective neural populations. A central processing stage generally follows to integrate cross-correlation responses across frequency and time (Colburn and Kulkarni 2005; Trahiotis et al. 2005).

In more engineering-centric models, binaural cues are used in conjunction with more probabilistic methods as complementary priors or to inform constraints on the location of sound sources (Marin-Hurtado et al. 2012; Alinaghi et al. 2014). In this body of work, the statistical structure of the sources or space itself plays a more prominent role in facilitating the segregation of the different signals. The most popular approach in this literature is blind source separation (BSS) which refers to a family of techniques that exploit the statistical structure of sources to separate their signals in a blind (i.e. unsupervised) manner (Bell and Sejnowski 1995; Naik and

Wang 2014). Generally, these algorithms are very effective at separating the sound sources under certain conditions that are gradually being relaxed by ongoing research efforts (Jutten and Karhunen 2004; Jadhav and Bhalchandra 2008).

Many engineering applications leverage the spatial analysis of a scene using multiple microphones. The rich sampling of the soundscape at multiple pick-up points opens the door to alternative techniques such as spatial sampling and beamforming (Van Veen and Buckley 1988; Krim and Viberg 1996). Such techniques aim at extracting a target source situated at a specific spatial direction using the sensor array. They focus on determining direction-of-arrival of sounds of interest, and are effectively filtering methods that operate in three-dimensional space to boost signals from a direction of interest. Although these techniques fall short of capitalizing on merits of spatial hearing, some have in fact benefited from human sound-source localization by employing adaptive beamformers that can judge the direction of target sounds, or take advantage of head-related transfer functions to reproduce out-of-head localization, or even incorporate simulations of room acoustics (Doclo and Moonen 2003; Farmani et al. 2015).

5.3.2 *Neural Infrastructure*

Most of the strategies discussed in Sect. 5.3.1 rely on intricate machinery or physical computations to achieve the required analysis of the complex scene. It is generally accepted that the pathway traveled by incoming acoustic information along the auditory system carries out the task of decomposing the sensory signal into its constituting elements and mapping them into perceptual streams (Nelken 2004). This neural transformation aims at extracting various acoustic features such as frequency, spectral profile, amplitude and frequency modulations, and interaural cues (Middlebrooks et al. 1980; Schreiner 1998). This feature representation is a canonical scheme for a discriminative representation of the scene that mediates the organization of features into segregated streams (Bizley and Cohen 2013).

Computationally, the incoming signal can be modeled as undergoing a series of mappings from acoustic space to a new feature space whose dimensionality facilitates the segregation or grouping of sound components into corresponding perceptual streams. At the core of this transformation is the concept of a receptive field, which has been instrumental in providing a functional descriptor of sensitivity of auditory neurons, as well as offering a computational medium for parameterizing the auditory feature space (Eggermont 2013). A receptive field can be thought of as a two-dimensional descriptor of the time-frequency sound features that best drive an auditory neuron, hence the name spectrotemporal receptive field (STRF) (Elhilali et al. 2013). It can be viewed as a time-dependent spectral transfer function, or a frequency-dependent dynamical filter (deCharms et al. 1998; Klein et al. 2000). In other words, if one views a neuron as a dynamical system, the STRF provides a descriptor of the linearized system function along both time and frequency, which maps the values of an input s at different time instants to a value of the output

(or response) r at the current time t (Korenberg and Hunter 1996), as described in Eq. (5.2):

$$r(t) = \sum_f \int \text{STRF}(\tau, f) s(t - \tau, f) d\tau \quad (5.2)$$

Receptive field descriptors have been successfully approximated at subcortical (Escabi and Schreiner 2002; Bandyopadhyay and Young 2013), as well as cortical stages (Depireux et al. 2001; Miller et al. 2002). By and large, convergent evidence suggests that the accumulation of transformations through these diverse receptive fields from the periphery up to auditory cortex is instrumental in providing the rich high-dimensional space necessary for segregating components of an acoustic scene (Sharpee et al. 2011; Christison-Lagay et al. 2015).

Indeed, a number of studies suggest that the organization of sound elements into mental representations of auditory objects may reside as early as primary auditory cortex (A1) (Nelken and Bar-Yosef 2008; Bizley and Cohen 2013). The neural representation of sounds as viewed through cortical receptive fields covers a rich feature space that spans at least three key dimensions (Fig. 5.2b): (1) *best frequencies* (BF) that cover the entire auditory range (Schreiner 1998; Klein et al. 2003); (2) *bandwidths* that span a wide range from very broad (2–3 octaves) to narrowly tuned (<0.25 octave) (Schreiner and Sutter 1992; Versnel et al. 1995); (3) *dynamics* that range from very slow to relatively fast (1–30 Hz) (Lu et al. 2001; Miller et al. 2002). This variability along different acoustic attributes is at the core of a multidimensional neural representation of sound mixtures, which in turn

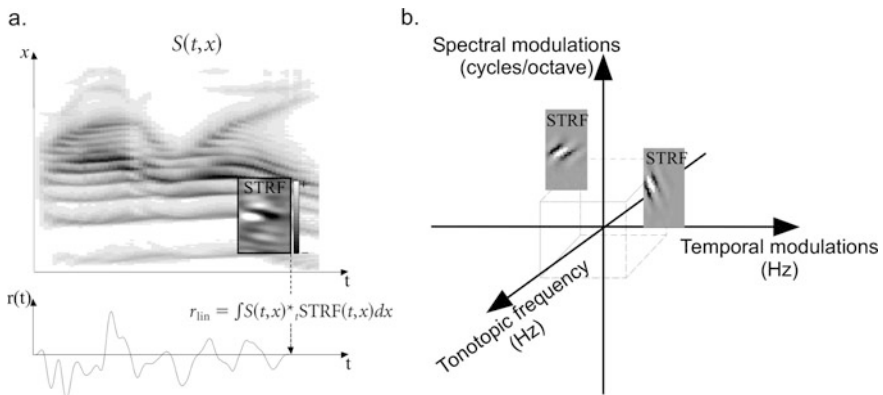


Fig. 5.2 Schematic of the concept of receptive field. **(a)** A spectrotemporal receptive field (STRF) operates as a two-dimensional filter that integrates stimulus information across time and frequency that best matches its selectivity. The corresponding neural response reflects the signal components that best drive the filter itself. **(b)** The selectivity of STRFs spans a high-dimensional space that spans tonotopic frequency, temporal modulations, and spectral modulations. Each STRF can be thought of as a mapping to a small portion of this space. The collection of responses through a network of neurons would correspond to a mapping onto a high-dimensional space

facilitates the execution of a number of strategies for modeling the cocktail party problem (Cooke and Ellis 2001; Elhilali and Shamma 2008). State-of-the-art models of auditory scene analysis also build on the same foundation, of a rich feature space extended to nonlinear manifolds. Current techniques using deep belief architectures, convolutional neural networks, and multivariate analysis have also been shown to exploit a rich time–frequency mapping similar to that observed in neural receptive fields to facilitate tasks of source separation (Le Roux et al. 2015; Simpson 2015).

5.4 Bottom-up Models of the Cocktail Party Problem

Together, the strategies driving modeling efforts of the cocktail party problem draw on viewpoints prompted by multidisciplinary efforts spanning the engineering, psychology, and neuroscience communities. On one end of the spectrum, numerous studies have attempted strict engineering approaches such as the successful application of blind source separation techniques (Roweis 2001; Jang and Lee 2003), statistical speech models (Varga and Moore 1990; Yoon et al. 2009), and other machine learning algorithms (Ming et al. 2013; Souden et al. 2013). Most of these approaches construct systems that exploit statistical knowledge about the target of interest (e.g., existing database of the target speaker’s voice), mine data about the physical or source characteristics of a target (e.g., knowledge about sources of noise), or utilize spatial characteristics of the receivers (usually in a multimicrophone setting) to hone in on desired signals to be segregated (Kristjansson et al. 2006; Madhu and Martin 2011). The statistical characteristics and possibly independence or uniqueness of the different sources (or at least the sound class of interest) are at the core of these approaches.

Despite their undeniable success, these algorithms often violate fundamental aspects of the manner in which humans and animals perform this task. They are generally constrained by their own mathematical formulations, are mostly applicable and effective in multisensor configurations, and/or require prior knowledge and training on the task at hand. By design, these systems target particular configurations of the sensory environment or require existing training databases or general knowledge about the task or target of interest. This reliance on training data or task-specific prior knowledge generally limits the applicability of these algorithms to general-purpose tasks. In this regard, the gap between these computational approaches and biological audition is still wide. A major effort in such engineering-centric systems deals with which patterns to extract from the scene and how to best capitalize on existing knowledge to perform the segregation, recognition, or synthesis task.

The best success stories in the category of engineering-centric systems are automatic speech recognition systems (Waibel and Lee 1990; Rabiner and Juang 1993) that focus on recognition of speech sounds even in the presence of unknown interferers and background noise. Although these systems are not immune to noise,

they have made great strides in improving the recognition accuracy by combining acoustic and language models that represent statistical representations of the sounds that make up each word and sequence of words as dictated by the grammatical rules of the language. This training knowledge is often combined by powerful machine learning tools such as convolutional systems and deep learning techniques (Hinton et al. 2012; Deng et al. 2013). These powerful tools, combined with abundance of training data, distance the challenge from the details of the feature analysis and compensate any weaknesses in the chosen signal representations by the strength of the statistical structure of the models. Unfortunately, these formulations limit any progress in truly understanding the strengths of the multiscale and parallel processing underlying sound processing in the auditory system and limit translating successes from these engineering approaches into cocktail party models that can truly mimic brain functions.

On the other end of the spectrum are perceptually driven studies that focus on factors influencing auditory scene analysis, in particular the segregation/binding cues that govern the simultaneous and sequential integration of sound patterns into objects emanating from a same environmental event (Bregman 1990; Carlyon 2004). These efforts have triggered a lot of interest in constructing *biologically inspired systems* that can perform intelligent processing of complex sound mixtures. Early instantiations of these models were strongly focused on the peripheral representations of sound. These models focused on peripheral selectivity, possibly allowing competition between different channels to result in a dominant foreground stream (Beauvois and Meddis 1996; McCabe and Denham 1997).

Other studies took more pragmatic approaches to modeling the cocktail party problem; particularly capitalizing on the salient acoustic attributes that can be tracked for individual sources to segregate them from competing backgrounds. Early work by Parsons (1976) and Weintraub (1985) focused on tracking the fundamental frequency of concurrent speakers. The role of a particular auditory feature (e.g., pitch) was later extended to additional acoustic cues and grouping dimensions following the basic premise of Gestalt principles and population separation theory, but with different computational implementations of the binding and integration stage (Brown and Cooke 1994).

The extraction of acoustic features has also been a cornerstone of correlation-based models mentioned in Sect. 5.3.1, by exploiting synchrony between different oscillators as a reflection of a grouped perceptual stream (Brown and Cooke 1998; Wang and Brown 1999). Synchrony of individual oscillators is initiated by regularity in the sound's spectrotemporal elements, and hence lateral connections between oscillators are implemented to encode harmonicity and proximity in time and frequency. A similar concept of feature extraction is also at the core of coherence-based models that emphasize the role of temporal integration over relatively long time scales; hence viewing feature analysis through the lens of temporal properties at the level of the mammalian primary auditory cortex (Shamma et al. 2011; Krishnan et al. 2014).

By and large, biomimetic models of auditory analysis of complex scenes have universally invoked the extraction of acoustic features as a foundation of any

subsequent processing. However, these implementations largely favor a bottom-up processing view (Fig. 5.1), relying on the salience of stimulus events. The models—with few exceptions—often abstract away intricate and indispensable contributions of goal-directed top-down processing and shy away from incorporating truly adaptive and task-dependent neural processing under top-down control.

5.5 Top-Down Processes and the Cocktail Party Problem

Along with the physical properties of sounds in the environment, listeners exploit learned knowledge from their recent and lifelong experiences to further complement processing of complex auditory scenes (Bregman 1990; Ciocca 2008). These learned “schemas” encompass a listener’s familiarity with the statistical structure of sound sources (e.g., natural sounds), recent and long-term memories about specific sources, expectation about the state of the world (e.g., speech sounds produced through a human vocal tract), as well as their attentional state which helps steer brain processes towards targets of interest while ignoring background interferers. These processes are believed to play a crucial role in tackling the cocktail party problem because they impose constraints on the space of possible solutions. They can be viewed as top-down or feedback projections that control the system’s performance to meet desired behaviors.

Of all schema-based processes, attention is one of the most widely studied top-down mechanisms affecting the cocktail party problem (Shinn-Cunningham, Chap. 2). It is a crucial component in the scene analysis process because it dictates what the targets of interest are, and orients the listener to the desired sound source or sources. It ultimately acts as a processing bottleneck that appropriately allocates neural resources to informative events in the acoustic scene and selectively filters the most relevant sensory inputs (Whiteley and Sahani 2012). While clearly behaviorally crucial, the specific roles of attention in auditory scene analysis remain an unsettled question in the field. It is certainly true that attention can strongly affect stream segregation. For instance, the ability to switch at will between hearing certain tone sequences as one or two streams can be thought of as an effect of attention, but the question of whether attention is *necessary* for streaming remains a matter of debate (Carlyon et al. 2001; Macken et al. 2003).

The bulk of the current literature suggests that at least some forms of stream segregation occur in the absence of attention, in what is termed “primitive” stream segregation (Bregman 1990; Sussman et al. 2007). As outlined in Sect. 5.3, the vast majority of cocktail party models have indeed implemented successful renditions of the problem solution in absence of any role of selective attention. Stream segregation may also be thought of as a process that facilitates attention (rather than only vice versa) in that it becomes possible to pay exclusive attention to tones of a single frequency only if they are successfully segregated from other tones in the sequence (Shinn-Cunningham 2008).

In the case of alternating tone sequences, early work by Van Noorden provided a useful distinction by defining two boundaries, the fission boundary and the coherence boundary (van Noorden 1975). The fission boundary defines the frequency difference (or other dimension) below which segregation is not possible, while the coherence boundary defines the point above which integration is not possible. The area in between these two boundaries can be thought of as the region in which attention can play a particularly important role in determining whether one or two streams are heard.

Though some computational models of the cocktail party problem have attempted to reproduce these effects (Beauvois and Meddis 1996; Wang and Chang 2008), they have not truly incorporated any mechanisms manipulating the attentional state of listener/model in a way that mimics the presumed feedback control exerted by attentional projections on feedforward sensory processing.

At the physiological level, a growing body of literature has established that auditory experience throughout adulthood can have profound global effects by reshaping cortical maps and significant local effects by transforming receptive field properties of neurons in primary auditory cortex (Suga et al. 1997; Weinberger 2001). The exact form of this remarkable plasticity is determined by the salience or task relevance of the spectral and temporal characteristics of the acoustic stimuli (Kilgard et al. 2001). Recent findings have also shown that cortical responses are heavily modulated by the attentional state of the brain and undergo rapid, short-term, and task-dependent changes that reflect not only the incoming sensory cues but also behavioral goals (Fritz et al. 2007; Mesgarani and Chang 2012). In this kind of adaptive plasticity, selective functional reconfiguration or resetting of the underlying cortical circuitry leads to changes in receptive field properties that may enhance perception in a cocktail party (Shamma and Fritz 2014).

Unfortunately, there is a notable lack of the incorporation of cognitive or adaptive mechanisms into mathematical models of auditory cortical processing and, ultimately, implementations of cocktail party models. This deficiency is itself motivated by lack of information and ignorance of the neural mechanisms underlying the ability of cortical circuits to adapt online to changing behavioral demands. In contrast, active and adaptive processing has more commonly been explored in models of the visual system. These implementations typically model parallels of predictive coding in the visual thalamus (LGN), contextual modulation in primary visual cortex (V1), attentional modulation in higher cortical areas (V2 and V4, and area MT), as well as decision making in parietal and frontal cortex. A commonly used formulation for such systems is that of generative models, whereby sensory input can be explained as being caused by hidden “causes” or “states” in the world (Duda et al. 2000). The model then estimates the probability of these causes based on inputs incoming up to a certain point in time. Modeling based on hidden causes or states is amenable to predictive coding, similar to concepts discussed in Sect. 5.3.1.3. In other words, the models employ a probabilistic formulation where optimization functions can then be defined as maximizing posterior probabilities, which is equivalent to minimizing the prediction error generated by this model. Some studies have presented successful implementations of these models as

hierarchical systems of early and higher visual cortical processing (Rao and Ballard 1999; Lee and Mumford 2003). This body of work has often relied on a linear formulation of the generative model, hence benefiting from existing linear hidden state estimation techniques such as Kalman filtering. The tracking of these latent states was also formulated to adapt the model parameters continuously to the statistics in the visual scene, hence giving the system a desired plastic behavior. Other techniques have also been explored to go beyond the generative model approach. Systems based on belief propagation, graphical models, as well as inference in recurrent networks have shown variable success in interpreting top-down feedback as prior probabilities (Rao 2005).

Recent models and frameworks for modeling the cocktail party effect and its biological bases have begun focusing on the role of schema-based processes, particularly attention in both its bottom-up and top-down forms in biasing selection and organization of sensory events (Shamma et al. 2011; Kaya and Elhilali 2014). Ultimately, progress in the integration of top-down processes in cocktail party models is closely tied to progress in unraveling neural mechanisms underlying cognitive effects on sensory processing, as well as models of feedback loops in shaping auditory processing of complex scenes.

5.6 Summary

The challenge of auditory scene analysis is a problem facing biological and engineering systems alike. Computational auditory scene analysis is a young field that aims at providing theoretical insights and solutions to the cocktail party problem that can inform neuroscience research as well as benefit audio applications. Though a lofty goal, translating perceptual phenomena related to the cocktail party problem to exact mathematical formulations requires more concise definitions of the problem, well-defined constraints on the desired system, as well as clear measurable outcomes and behaviors. Indeed, the cocktail party problem is a phenomenological description of multiple tasks related to processing complex soundscapes. These range from detection and recognition to tracking, description, and audio resynthesis. Translating these problems into computational models leaves the field somewhat fragmented.

Nonetheless, a rich body of computational models has offered insights into how the brain might tackle the cocktail party challenge. These invoke the rich feature selectivity that underlies neural processing through the auditory pathway from the periphery all the way to auditory cortex. The neural transformations up to sensory cortex offer part of the solution to the segregation of sound mixtures along informative dimensions for further processing. Additional processes such as temporal coherence play a role in the binding process that combines relevant acoustic cues onto perceptual streams corresponding to perceived objects. Computational models also capitalize on the structure of sound sources to track the regularities or dynamics of sound events over time.

All in all, models inspired from brain processes have laid the conceptual groundwork for interpreting the transformation from an acoustic space of a mixture of sound sources to a perceptual space with segregated streams. Translating this foundation into practical engineering applications and evaluating its effectiveness remains one of the big challenges in the field. In conjunction, additional factors, particularly with regard to schema-based processes (e.g., attention, learning), add extra hurdles in developing full solutions to the cocktail party problem that could come close to emulating the biological system. As the growing yet limited knowledge of the neural underpinnings of schema-based processes sheds light on their role in cocktail parties, truly intelligent systems will undoubtedly emerge that can mimic the complex processing exhibited by the brain when dealing with the cocktail party problem.

Acknowledgements Dr. Elhilali's work is supported by grants from The National Institutes of Health (NIH: R01HL133043) and the Office of Naval Research (ONR: N000141010278, N000141612045, and N000141210740).

Compliance with Ethics Requirements

Mounya Elhilali declares that she has no conflict of interest.

References

- Akeroyd, M. A., Carlyon, R. P., & Deeks, J. M. (2005). Can dichotic pitches form two streams? *The Journal of the Acoustical Society of America*, *118*(2), 977–981.
- Alais, D., Blake, R., & Lee, S. H. (1998). Visual features that vary together over time group together over space. *Nature Neuroscience*, *1*(2), 160–164.
- Alinaghi, A., Jackson, P. J., Liu, Q., & Wang, W. (2014). Joint mixing vector and binaural model based stereo source separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *22*(9), 1434–1448.
- Almajai, I., & Milner, B. (2011). Visually derived wiener filters for speech enhancement. *IEEE Transactions on Audio, Speech and Language Processing*, *19*(6), 1642–1651.
- Anemuller, J., Bach, J., Caputo, B., Havlena, M., et al. (2008). The DIRAC AWEAR audio-visual platform for detection of unexpected and incongruent events. In *International Conference on Multimodal Interaction*, (pp. 289–293).
- Arbogast, T. L., Mason, C. R., & Kidd, G. (2002). The effect of spatial separation on informational and energetic masking of speech. *The Journal of the Acoustical Society of America*, *112*(5 Pt 1), 2086–2098.
- Aubin, T. (2004). Penguins and their noisy world. *Annals of the Brazilian Academy of Sciences*, *76*(2), 279–283.
- Bandyopadhyay, S., & Young, E. D. (2013). Nonlinear temporal receptive fields of neurons in the dorsal cochlear nucleus. *Journal of Neurophysiology*, *110*(10), 2414–2425.
- Barchiesi, D., Giannoulis, D., Stowell, D., & Plumbley, M. D. (2015). Acoustic scene classification: Classifying environments from the sounds they produce. *IEEE Signal Processing Magazine*, *32*(3), 16–34.
- Bastos, A. M., Usrey, W. M., Adams, R. A., Mangun, G. R., et al. (2012). Canonical microcircuits for predictive coding. *Neuron*, *76*(4), 695–711.

- Beauvois, M. W., & Meddis, R. (1996). Computer simulation of auditory stream segregation in alternating-tone sequences. *The Journal of the Acoustical Society of America*, *99*(4), 2270–2280.
- Bell, A. J., & Sejnowski, T. J. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, *7*(6), 1129–1159.
- Bizley, J. K., & Cohen, Y. E. (2013). The what, where and how of auditory-object perception. *Nature Reviews Neuroscience*, *14*(10), 693–707.
- Blake, R., & Lee, S. H. (2005). The role of temporal structure in human vision. *Behavioral and Cognitive Neuroscience Review*, *4*(1), 21–42.
- Bregman, A. S. (1981). Asking the ‘what for’ question in auditory perception. In M. Kubovy & J. Pomerantz (Eds.), *Perceptual organization* (pp. 99–118). Hillsdale, NJ: Lawrence Erlbaum Associates.
- 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*(2), 244–249.
- Brown, G. J., & Cooke, M. (1994). Computational auditory scene analysis. *Computer Speech & Language*, *8*(4), 297–336.
- Brown, G. J., & Cooke, M. (1998). Temporal synchronization in a neural oscillator model of primitive auditory stream segregation. In D. L. Wang & G. Brown (Eds.), *Computational auditory scene analysis* (pp. 87–103). London: Lawrence Erlbaum Associates.
- Brown, G. J., Barker, J., & Wang, D. (2001). A neural oscillator sound separator for missing data speech recognition. In *Proceedings of International Joint Conference on Neural Networks, 2001 (IJCNN '01)* (Vol. 4, pp. 2907–2912).
- Buxton, H. (2003). Learning and understanding dynamic scene activity: A review. *Image and Vision Computing*, *21*(1), 125–136.
- 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.
- Chen, F., & Jokinen, K. (Eds.). (2010). *Speech technology: Theory and applications*. New York: Springer Science+Business Media.
- Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *The Journal of the Acoustical Society of America*, *25*(5), 975–979.
- Cherry, E. C. (1957). *On human communication*. Cambridge, MA: MIT Press.
- Christison-Lagay, K. L., Gifford, A. M., & Cohen, Y. E. (2015). Neural correlates of auditory scene analysis and perception. *International Journal of Psychophysiology*, *95*(2), 238–245.
- Ciocca, V. (2008). The auditory organization of complex sounds. *Frontiers in Bioscience*, *13*, 148–169.
- Cisek, P., Drew, T., & Kalaska, J. (Eds.). (2007). *Computational neuroscience: Theoretical insights into brain function*. Philadelphia: Elsevier.
- Colburn, H. S., & Kulkarni, A. (2005). Models of sound localization. In A. N. Popper & R. R. Fay (Eds.), *Sound source localization* (pp. 272–316). New York: Springer Science+Business Media.
- Collins, N. (2009). *Introduction to computer music*. Hoboken, NJ: Wiley.
- Cooke, M., & Ellis, D. (2001). The auditory organization of speech and other sources in listeners and computational models. *Speech Communication*, *35*, 141–177.
- Cusack, R., & Roberts, B. (1999). Effects of similarity in bandwidth on the auditory sequential streaming of two-tone complexes. *Perception*, *28*(10), 1281–1289.
- Cusack, R., & Roberts, B. (2000). Effects of differences in timbre on sequential grouping. *Perception and Psychophysics*, *62*(5), 1112–1120.
- Darwin, C. J., & Carlyon, R. P. (1995). Auditory grouping. In B. C. J. Moore (Ed.), *Hearing* (pp. 387–424). Orlando, FL: Academic Press.

- 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.
- deCharms, R. C., Blake, D. T., & Merzenich, M. M. (1998). Optimizing sound features for cortical neurons. *Science*, 280(5368), 1439–1443.
- Deng, L., Li, J., Huang, J., Yao, K., et al. (2013). Recent advances in deep learning for speech research at Microsoft. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, May 26–31, 2013 (pp. 8604–8608).
- 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.
- Doclo, S., & Moonen, M. (2003). adaptive. *EURASIP Journal of Applied Signal Processing*, 11, 1110–1124.
- Duda, R. O., Hart, P. E., & Stork, D. G. (2000). *Pattern classification*. Hoboken, NJ: Wiley.
- Eggermont, J. J. (2013). The STRF: Its origin, evolution and current application. In D. Depireux & M. Elhilali (Eds.), *Handbook of modern techniques in auditory cortex* (pp. 1–32). Hauppauge, NY: Nova Science Publishers.
- Elhilali, M. (2013). Bayesian inference in auditory scenes. In *Proceedings of the 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Osaka, Japan, (pp. 2792–2795).
- Elhilali, M., & Shamma, S. A. (2008). A cocktail party with a cortical twist: How cortical mechanisms contribute to sound segregation. *The Journal of the Acoustical Society of America*, 124(6), 3751–3771.
- Elhilali, M., Ma, L., Micheyl, C., Oxenham, A. J., & Shamma, S. A. (2009). Temporal coherence in the perceptual organization and cortical representation of auditory scenes. *Neuron*, 61(2), 317–329.
- Elhilali, M., Ma, L., Micheyl, C., Oxenham, A., & Shamma, S. (2010). Rate vs. temporal code? A spatio-temporal coherence model of the cortical basis of streaming. In E. Lopez-Poveda, A. Palmer & R. Meddis (Eds.), *Auditory physiology, perception and models* (pp. 497–506). New York: Springer Science+Business Media.
- Elhilali, M., Shamma, S. A., Simon, J. Z., & Fritz, J. B. (2013). A linear systems view to the concept of STRF. In D. Depireux & M. Elhilali (Eds.), *Handbook of modern techniques in auditory cortex* (pp. 33–60). Hauppauge, NY: Nova Science Publishers.
- Escabi, M. A., & Schreiner, C. E. (2002). Nonlinear spectrotemporal sound analysis by neurons in the auditory midbrain. *The Journal of Neuroscience*, 22(10), 4114–4131.
- Farmani, M., Pedersen, M. S., Tan, Z. H., & Jensen, J. (2015). On the influence of microphone array geometry on HRTF-based sound source localization. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (pp. 439–443).
- Friston, K. J. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11(2), 127–138.
- Fritz, J. B., Elhilali, M., David, S. V., & Shamma, S. A. (2007). Auditory attention—focusing the searchlight on sound. *Current Opinion in Neurobiology*, 17(4), 437–455.
- Gilkey, R., & Anderson, T. R. (Eds.). (2014). *Binaural and spatial hearing in real and virtual environments*. New York: Psychology Press.
- Gockel, H., Carlyon, R. P., & Micheyl, C. (1999). Context dependence of fundamental-frequency discrimination: Lateralized temporal fringes. *The Journal of the Acoustical Society of America*, 106(6), 3553–3563.
- Grimault, N., Bacon, S. P., & Micheyl, C. (2002). Auditory stream segregation on the basis of amplitude-modulation rate. *The Journal of the Acoustical Society of America*, 111(3), 1340–1348.
- Hartmann, W., & Johnson, D. (1991). Stream segregation and peripheral channeling. *Music Perception*, 9(2), 155–184.
- Haykin, S., & Chen, Z. (2005). The cocktail party problem. *Neural Computation*, 17(9), 1875–1902.

- Herbrich, R. (2001). *Learning kernel classifiers: Theory and algorithms*. Cambridge, MA: MIT Press.
- Hinton, G., Deng, L., Yu, D., Dahl, G. E., et al. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *Signal Processing Magazine, IEEE*, 29(6), 82–97.
- Hyvarinen, A., Karhunen, J., & Oja, E. (2001). *Independent component analysis*. Hoboken, NJ: Wiley.
- Itatani, N., & Klump, G. M. (2011). Neural correlates of auditory streaming of harmonic complex sounds with different phase relations in the songbird forebrain. *Journal of Neurophysiology*, 105(1), 188–199.
- Izumi, A. (2002). Auditory stream segregation in Japanese monkeys. *Cognition*, 82(3), B113–B122.
- Jadhav, S. D., & Bhalchandra, A. S. (2008). Blind source separation: Trends of new age—a review. In *IET International Conference on Wireless, Mobile and Multimedia Networks, 2008*, Mumbai, India, January 11–12, 2008 (pp. 251–254).
- Jang, G. J., & Lee, T. W. (2003). A maximum likelihood approach to single-channel source separation. *Journal of Machine Learning Research*, 4(7–8), 1365–1392.
- Jeffress, L. A. (1948). A place theory of sound localization. *Journal of Comparative and Physiological Psychology*, 41(1), 35–39.
- Jutten, C., & Karhunen, J. (2004). Advances in blind source separation (BSS) and independent component analysis (ICA) for nonlinear mixtures. *International Journal of Neural Systems*, 14(5), 267–292.
- Kaya, E. M., & Elhilali, M. (2013). Abnormality detection in noisy biosignals. In *Proceedings of the 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Osaka, Japan (pp. 3949–3952).
- Kaya, E. M., & Elhilali, M. (2014). Investigating bottom-up auditory attention. *Frontiers in Human Neuroscience*, 8(327), doi:10.3389/fnhum.2014.00327
- Kilgard, M. P., Pandya, P. K., Vazquez, J., Gehi, A., et al. (2001). Sensory input directs spatial and temporal plasticity in primary auditory cortex. *Journal of Neurophysiology*, 86(1), 326–338.
- Klein, D. J., Depireux, D. A., Simon, J. Z., & Shamma, S. A. (2000). Robust spectrotemporal reverse correlation for the auditory system: Optimizing stimulus design. *Journal of Computational Neuroscience*, 9(1), 85–111.
- Klein, D. J., Konig, P., & Kording, K. P. (2003). Sparse spectrotemporal coding of sounds. *EURASIP Journal on Applied Signal Processing*, 2003(7), 659–667.
- Korenberg, M., & Hunter, I. (1996). The identification of nonlinear biological systems: Volterra kernel approaches. *Annals of Biomedical Engineering*, 24(4), 250–268.
- Krim, H., & Viberg, M. (1996). Two decades of array signal processing research: The parametric approach. *IEEE Signal Processing Magazine*, 13(4), 67–94.
- Krishnan, L., Elhilali, M., & Shamma, S. (2014). Segregating complex sound sources through temporal coherence. *PLoS Computational Biology*, 10(12), e1003985.
- Kristjansson, T., Hershey, J., Olsen, P., Rennie, S., & Gopinath, R. (2006). Super-human multi-talker speech recognition: The IBM 2006 speech separation challenge system. In *International Conference on Spoken Language Processing*, Pittsburgh, PA, September 17–21, 2006.
- Lakatos, P., Shah, A. S., Knuth, K. H., Ulbert, I., et al. (2005). An oscillatory hierarchy controlling neuronal excitability and stimulus processing in the auditory cortex. *Journal of Neurophysiology*, 94(3), 1904–1911.
- Lee, T. S., & Mumford, D. (2003). Hierarchical bayesian inference in the visual cortex. *Journal of the Optical Society of America*, 20(7), 1434–1448.
- Le Roux, J., Hershey, J. R., & Weninger, F. (2015). Deep NMF for speech separation. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, April 19–24, 2015 (pp. 66–70).
- Lewicki, M. S., Olshausen, B. A., Surlykke, A., & Moss, C. F. (2014). Scene analysis in the natural environment. *Frontiers in Psychology*, 5, 199.

- Loizou, P. C. (2013). *Speech enhancement: Theory and practice* (2nd ed.). Boca Raton, FL: CRC Press.
- Lu, T., Liang, L., & Wang, X. (2001). Temporal and rate representations of time-varying signals in the auditory cortex of awake primates. *Nature Neuroscience*, 4(11), 1131–1138.
- Macken, W. J., Tremblay, S., Houghton, R. J., Nicholls, A. P., & Jones, D. M. (2003). Does auditory streaming require attention? Evidence from attentional selectivity in short-term memory. *Journal of Experimental Psychology: Human Perception and Performance*, 29(1), 43–51.
- Madhu, N., & Martin, R. (2011). A versatile framework for speaker separation using a model-based speaker localization approach. *IEEE Transactions on Audio, Speech and Language Processing*, 19(7), 1900–1912.
- Marin-Hurtado, J. I., Parikh, D. N., & Anderson, D. V. (2012). Perceptually inspired noise-reduction method for binaural hearing aids. *IEEE Transactions on Audio, Speech and Language Processing*, 20(4), 1372–1382.
- Marr, D. (1982). *Vision*. San Francisco: Freeman and Co.
- McCabe, S. L., & Denham, M. J. (1997). A model of auditory streaming. *The Journal of the Acoustical Society of America*, 101(3), 1611–1621.
- Mesgarani, N., & Chang, E. F. (2012). Selective cortical representation of attended speaker in multi-talker speech perception. *Nature*, 485(7397), 233–236.
- Micheyl, C., Carlyon, R. P., Gutschalk, A., Melcher, J. R., et al. (2007). The role of auditory cortex in the formation of auditory streams. *Hearing Research*, 229(1–2), 116–131.
- Micheyl, C., Hanson, C., Demany, L., Shamma, S., & Oxenham, A. J. (2013). Auditory stream segregation for alternating and synchronous tones. *Journal of Experimental Psychology: Human Perception and Performance*, 39(6), 1568–1580.
- Middlebrooks, J. C., Dykes, R. W., & Merzenich, M. M. (1980). Binaural response-specific bands in primary auditory cortex (AI) of the cat: Topographical organization orthogonal to isofrequency contours. *Brain Research*, 181(1), 31–48.
- Mill, R. W., Bohm, T. M., Bendixen, A., Winkler, I., & Denham, S. L. (2013). Modelling the emergence and dynamics of perceptual organisation in auditory streaming. *PLoS Computational Biology*, 9(3), e1002925.
- 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(1), 516–527.
- Ming, J., Srinivasan, R., Crookes, D., & Jafari, A. (2013). CLOSE—A data-driven approach to speech separation. *IEEE Transactions on Audio, Speech and Language Processing*, 21(7), 1355–1368.
- Mirbagheri, M., Akram, S., & Shamma, S. (2012). An auditory inspired multimodal framework for speech enhancement. In *Proceedings of the 13th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Portland, OR.
- Moore, B. C. J., & Gockel, H. (2002). Factors influencing sequential stream segregation. *Acta Acustica*, 88, 320–333.
- Mumford, D. (1992). On the computational architecture of the neocortex. II. The role of cortico-cortical loops. *Biological Cybernetics*, 66(3), 241–251.
- Naik, G., & Wang, W. (Eds.). (2014). *Blind source separation: Advances in theory, algorithms and applications*. Berlin/Heidelberg: Springer-Verlag.
- Nelken, I. (2004). Processing of complex stimuli and natural scenes in the auditory cortex. *Current Opinion in Neurobiology*, 14(4), 474–480.
- Nelken, I., & Bar-Yosef, O. (2008). Neurons and objects: The case of auditory cortex. *Frontiers in Neuroscience*, 2(1), 107–113.
- Parsons, T. W. (1976). Separation of speech from interfering speech by means of harmonic selection. *The Journal of the Acoustical Society of America*, 60(4), 911–918.
- Patil, K., & Elhilali, M. (2013). Multiresolution auditory representations for scene recognition. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, October 20–23, 2013.

- Poggio, T. (2012). *The levels of understanding framework, revised*. Computer Science and Artificial Intelligence Laboratory Technical Report MIT-CSAIL-TR-2012-014. Cambridge, MA: Massachusetts Institute of Technology.
- Pressnitzer, D., Sayles, M., Micheyl, C., & Winter, I. M. (2008). Perceptual organization of sound begins in the auditory periphery. *Current Biology*, 18(15), 1124–1128.
- Rabiner, L., & Juang, B. (1993). *Fundamentals of speech recognition*. Englewood Cliffs, NJ: Prentice Hall.
- Rao, R. P. (2005). Bayesian inference and attentional modulation in the visual cortex. *NeuroReport*, 16(16), 1843–1848.
- Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1), 79–87.
- Riesenhuber, M., & Poggio, T. (2002). Neural mechanisms of object recognition. *Current Opinion in Neurobiology*, 12(2), 162–168.
- Roberts, B., Glasberg, B. R., & Moore, B. C. (2002). Primitive stream segregation of tone sequences without differences in fundamental frequency or passband. *The Journal of the Acoustical Society of America*, 112(5), 2074–2085.
- Roweis, S. T. (2001). One microphone source separation. *Advances in Neural Information Processing Systems*, 13, 793–799.
- Schreiner, C. E. (1998). Spatial distribution of responses to simple and complex sounds in the primary auditory cortex. *Audiology and Neuro-Otology*, 3(2–3), 104–122.
- Schreiner, C. E., & Sutter, M. L. (1992). Topography of excitatory bandwidth in cat primary auditory cortex: Single-neuron versus multiple-neuron recordings. *Journal of Neurophysiology*, 68(5), 1487–1502.
- Schroger, E., Bendixen, A., Denham, S. L., Mill, R. W., et al. (2014). Predictive regularity representations in violation detection and auditory stream segregation: From conceptual to computational models. *Brain Topography*, 27(4), 565–577.
- Shamma, S., & Fritz, J. (2014). Adaptive auditory computations. *Current Opinion in Neurobiology*, 25, 164–168.
- Shamma, S. A., Elhilali, M., & Micheyl, C. (2011). Temporal coherence and attention in auditory scene analysis. *Trends in Neurosciences*, 34(3), 114–123.
- Sharpee, T. O., Atencio, C. A., & Schreiner, C. E. (2011). Hierarchical representations in the auditory cortex. *Current Opinion in Neurobiology*, 21(5), 761–767.
- Sheft, S. (2008). Envelope processing and sound-source perception. In W. A. Yost, A. Popper, & R. R. Fay (Eds.), *Auditory perception of sound sources* (pp. 233–280). New York: Springer Science+Business Media.
- Shinn-Cunningham, B. G. (2008). Object-based auditory and visual attention. *Trends in Cognitive Sciences*, 12(5), 182–186.
- Simpson, A. J. (2015). Probabilistic binary-mask cocktail-party source separation in a convolutional deep neural network. *arXiv Preprint arXiv:1503.06962*.
- Souden, M., Araki, S., Kinoshita, K., Nakatani, T., & Sawada, H. (2013). A multichannel MMSE-based framework for speech source separation and noise reduction. *IEEE Transactions on Audio, Speech and Language Processing*, 21(9), 1913–1928.
- Stern, R., Brown, G., & Wang, D. L. (2005). Binaural sound localization. In D. L. Wang & G. Brown (Eds.), *Computational auditory scene analysis: Principles, algorithms and applications* (pp. 147–186). Hoboken, NJ: Wiley-IEEE Press.
- Suga, N., Yan, J., & Zhang, Y. (1997). Cortical maps for hearing and egocentric selection for self-organization. *Trends in Cognitive Sciences*, 1(1), 13–20.
- Sussman, E. S., Horvath, J., Winkler, I., & Orr, M. (2007). The role of attention in the formation of auditory streams. *Perception and Psychophysics*, 69(1), 136–152.
- Trahiotis, C., Bernstein, L. R., Stern, R. M., & Buel, T. N. (2005). Interaural correlation as the basis of a working model of binaural processing: An introduction. In A. N. Popper & R. R. Fay (Eds.), *Sound source localization* (pp. 238–271). New York: Springer Science+Business Media.

- van der Kouwe, A. W., Wang, D. L., & Brown, G. J. (2001). A comparison of auditory and blind separation techniques for speech segregation. *IEEE Transactions on Speech and Audio Processing*, 9(3), 189–195.
- van Noorden, L. P. A. S. (1975). *Temporal coherence in the perception of tone sequences*. Ph.D. dissertation. Eindhoven, The Netherlands: Eindhoven University of Technology.
- van Noorden, L. P. A. S. (1977). Minimum differences of level and frequency for perceptual fission of tone sequences ABAB. *The Journal of the Acoustical Society of America*, 61(4), 1041–1045.
- Van Veen, B. D., & Buckley, K. M. (1988). Beamforming: A versatile approach to spatial filtering. *IEEE ASSP Magazine*, 5(2), 4–24.
- Varga, A. P., & Moore, R. K. (1990). Hidden Markov model decomposition of speech and noise. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Albuquerque, NM, April 3–6, 1990 (pp. 845–848).
- Versnel, H., Kowalski, N., & Shamma, S. A. (1995). Ripple analysis in ferret primary auditory cortex. III. Topographic distribution of ripple response parameters. *Journal of Auditory Neuroscience*, 1, 271–286.
- Virtanen, T., Singh, R., & Bhiksha, R. (Eds.). (2012). *Techniques for noise robustness in automatic speech recognition*. Hoboken, NJ: Wiley.
- Vliegen, J., & Oxenham, A. J. (1999). Sequential stream segregation in the absence of spectral cues. *The Journal of the Acoustical Society of America*, 105(1), 339–346.
- von der Malsburg, C. (1994). The correlation theory of brain function. In E. Domany, L. Van Hemmen, & K. Schulten (Eds.), *Models of neural networks* (pp. 95–119). Berlin: Springer.
- Waibel, A., & Lee, K. (1990). *Readings in speech recognition*. Burlington, MA: Morgan Kaufmann.
- Wang, D., & Chang, P. (2008). An oscillatory correlation model of auditory streaming. *Cognitive Neurodynamics*, 2(1), 7–19.
- Wang, D. L., & Brown, G. J. (1999). Separation of speech from interfering sounds based on oscillatory correlation. *IEEE Transactions on Neural Networks*, 10(3), 684–697.
- Wang, D. L., & Brown, G. J. (Eds.). (2006). *Computational auditory scene analysis: Principles, algorithms and applications*. Hoboken, NJ: Wiley-IEEE Press.
- Weinberger, N. M. (2001). Receptive field plasticity and memory in the auditory cortex: Coding the learned importance of events. In J. Steinmetz, M. Gluck, & P. Solomon (Eds.), *Model systems and the neurobiology of associative learning* (pp. 187–216). Mahwah, NJ: Lawrence Erlbaum Associates.
- Weintraub, M. (1985). *A theory and computational model of auditory monaural sound separation*. Ph.D. dissertation. Stanford University.
- Whiteley, L., & Sahani, M. (2012). Attention in a bayesian framework. *Frontiers in Human Neuroscience*, 6(100), doi:[10.3389/fnhum.2012.00100](https://doi.org/10.3389/fnhum.2012.00100)
- Winkler, I., Denham, S. L., & Nelken, I. (2009). Modeling the auditory scene: Predictive regularity representations and perceptual objects. *Trends in Cognitive Sciences*, 13(12), 532–540.
- Xu, Y., & Chun, M. M. (2009). Selecting and perceiving multiple visual objects. *Trends in Cognitive Sciences*, 13(4), 167–174.
- Yoon, J. S., Park, J. H., & Kim, H. K. (2009). Acoustic model combination to compensate for residual noise in multi-channel source separation. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, Taipei, Taiwan, April 19–24, 2009 (pp. 3925–3928).

Chapter 6

Spatial Stream Segregation

John C. Middlebrooks

Abstract “Stream segregation” refers to a listener’s ability to disentangle interleaved sequences of sounds, such as the ability to string together syllables from one talker in the presence of competing talkers. Spatial separation of sound sources is a key factor that enables the task of segregation. Psychophysical tasks that require listeners to integrate sounds across locations demonstrate that listeners can overcome spatial separation of sources, suggesting that space is a relatively weak segregating factor. Contrary to that suggestion tasks that require listeners to isolate a sound sequence within a complex background demonstrate robust benefits of spatial separation of the target from other sources. This chapter reviews psychophysical studies that show weak versus strong spatial effects on streaming and shows that the spatial acuity of stream segregation can approach the limits of acuity of spatial hearing. Responses from auditory cortex in anesthetized animals are presented demonstrating that single neurons can exhibit spatial stream segregation by synchronizing selectively to one or the other of two interleaved sound sequences. The results from animals imply that perceptually segregated sound sequences are represented in auditory cortex by discrete mutually synchronized neural populations. Human magneto- and electroencephalographic results then are described showing selective enhancement of cortical responses to attended versus unattended sounds. Available results lead to a picture showing bottom-up segregation of sound sources by brainstem mechanisms on the basis of spatial and other cues, followed by top-down selection of particular neural populations that could underlie perceptual auditory objects of attention.

Keywords Auditory cortex • Rhythmic masking release • Scene analysis • Spatial hearing • Spatial streaming • Stream integration • Spatial release from masking

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

6.1 Introduction

“Stream segregation” refers to a listener’s ability to disentangle temporally interleaved sequences of sounds from multiple sources. It may be regarded as an element of auditory scene analysis (Bregman 1990) and/or as a part of the solution to the “cocktail party problem” (Cherry 1953). In speech, normal-hearing listeners do this when they attend to sequences of syllables from one talker in the presence of a crowd of other talkers. In music, a listener can pick out a single musical line from an ensemble of multiple instruments, or a composer can exploit tricks of pitch and rhythm to create from a single instrument the impression of multiple segregated lines. As described in Chap. 2 by Shinn-Cunningham, Best, and Lee, the perceptual correlate of utterances from a specific talker or a musical line or any other distinct sound source can be referred to as an “auditory object” (Woods and Colburn 1992; Griffiths and Warren 2004). Stream segregation is a major element of auditory object formation.

The individual elements of sound sequences from multiple sources might overlap partially or completely in time, or the elements might interleave with no temporal overlap. Temporal and/or spectral overlap of sounds from multiple sources can result in energetic or informational masking, which are the topics of Chaps. 3 (Culling and Stone) and 4 (Kidd and Colburn). Even in the case of sequential interleaving of sound elements, in which there is no temporal overlap, it is a challenge for a listener to construct one or more discrete auditory objects when exposed to multiple competing sequences of sounds.

Sound features that enable stream segregation include differences in fundamental frequencies (corresponding to pitches), spectra (corresponding to timbre), and temporal envelopes, particularly differences in onset times (reviewed by Moore and Gockel 2002, 2012). The present chapter focuses on another key factor in stream segregation, the *spatial* differences among multiple sources. Spatial separation of sound sources has long been appreciated to aid in formation of auditory objects (Cherry 1953). Cherry, for example, wrote that “the voices come from different directions” (p. 976) was a key factor in segregating competing talkers. He simulated “different directions” by presenting two spoken messages dichotically, one to each ear, and noted that recognition of one or the other message improved dramatically compared to a condition in which the messages were presented diotically (i.e., mixed and both presented to both ears) (Cherry 1953). Surprisingly, objective measures of spatial effects on stream segregation have yielded a wide variety of results, ranging from “weak-to-no” effect of space to “robust spatial streaming.” Those conflicting results seemingly can be reconciled by considering the requirements of specific psychophysical tasks, as discussed in Sect. 6.2.

Regardless of the particular sound feature, the brain substrates of stream segregation likely involve brainstem and thalamocortical mechanisms for bottom-up formation of auditory objects combined with cortical mechanism for top-down selection among those objects. At least one study has suggested that neuronal stream segregation based on tonal frequencies is accomplished as early in the

auditory pathway as the cochlear nucleus (Pressnitzer et al. 2008). Most other physiological studies, however, have focused on forebrain levels. Correlates of stream segregation based on frequencies of pure tones have been demonstrated in neural recordings in primary auditory cortex (area A1) of macaque monkeys (*Macaca fascicularis*: Fishman et al. 2001; *Macaca mulatta*: Micheyl et al. 2005) and ferrets (*Mustela putorius furo*: Elhilali et al. 2009). In humans, correlates of streaming based on fundamental frequencies or interaural time differences (ITDs) have been demonstrated in nonprimary auditory cortex using event-related potentials, magnetoencephalography (MEG), and functional magnetic resonance imaging (fMRI) (Snyder and Alain 2007; Schadwinkel and Gutschalk 2010; Carl and Gutschalk 2012). Evidence of cortical streaming of high-level auditory objects, including speech streams, has been observed in humans with MEG techniques (Ding and Simon 2012a, b) and with recordings from the cortical surface (Mesgarani and Chang 2012); these topics are reviewed in Chap. 7 by Simon. Correlates of spatial stream segregation by single neurons in cortical area A1 are demonstrated by the study reviewed in Sect. 6.3. A model resulting from that study posits that spatial stream segregation arises in auditory cortex as a product of brainstem spatial processing that is then sharpened by forward suppression in the thalamocortical projection. Results from other studies suggest that spatially sensitive neurons are ubiquitous in auditory cortex but that various auditory cortical areas differ in their relative contributions to spatial stream segregation and to other aspects of spatial hearing.

6.2 Psychophysics of Spatial Stream Segregation

Psychophysical studies have evaluated the conditions under which interleaved sequences of sounds elicit perceptions either of single integrated streams or of two or more segregated streams. An oft-cited example is the dissertation work by van Noorden (1975). van Noorden presented listeners with sequences of tones, denoted here by A and B, that differed in frequency. When sequence ABA_ABA_ABA... was presented at a slow rate or with closely spaced frequencies, listeners reported hearing a succession of “gallops” consisting of the ABA triplets. At a higher rate or with wider frequency separation, however, two perceptually distinct streams emerged, one consisting of a rapid sequence of the A tones and the other consisting of a slower sequence of the B tones. van Noorden wrote of “fusion” (i.e., integration) of the A and B tones into a single stream of gallops and “fission” (segregation) of two segregated A and B streams.

Psychophysical measures of the importance of spatial cues for stream segregation have yielded quite disparate results depending on the design of the experiment. Studies reviewed in Sect. 6.2.1 have required listeners to *integrate* information across multiple source locations. Listeners’ generally good performance in such tasks seems to show that location is a weak segregation cue that can be defeated easily when task performance demands integration. In contrast, studies reviewed in

Sects. 6.2.2 and 6.2.3 required listeners to *segregate* multiple competing sounds. Those studies demonstrate that spatial separations of target and masker(s) are potent cues that a listener may exploit when attempting to segregate a particular target from other distracters, like the task of hearing out a particular talker amid a background of other voices.

6.2.1 *Weak Disruption of Stream Integration by Spatial Cues*

A number of psychophysical studies have tested the ability of a listener to integrate sequences of sounds that vary in spatial or other parameters. Such studies have been referred to variously as measures of fusion (van Noorden 1975), primitive streaming (Bregman 1990), obligatory or involuntary streaming (Vliegen et al. 1999), or integration (Micheyl and Oxenham 2010). Information needed for performance of integrative streaming tasks is distributed among two or more potentially segregated streams that the listener must fuse in order to make a correct judgment. In integrative tasks, the magnitude of stream segregation can be inferred by the degree to which a putative streaming factor *impairs* task performance by forcing signal components into differing perceptual streams. One commonly used test of stream integration is a so-called temporal asymmetry task. Sequences of sounds differing in spectral or spatial parameters, denoted here as A and B, are presented as sequences of ABA_ABA_..., and listeners are required to detect sequences in which the B sound falls asymmetrically between the two A markers; that is, when the AB time interval differs from the BA interval. Performance on such a task is impaired when the A and B sounds diverge into differing perceptual streams as, for example, when the A and B sounds differ in spectrum (Vliegen et al. 1999) or ear of entry (Boehnke and Phillips 2005); those are conditions in which the A and B sounds are assumed to activate distinct neural populations. van Noorden (1975) noted that it was easy to achieve a subjective experience of fission (i.e., segregation) when sounds were presented to opposite ears.

Surprisingly, differences in spatial cues in the A and B sounds, specifically ITDs or interaural level differences (ILDs), result in little or no impairment of temporal asymmetry detection when the ITDs and ILDs fall within the ranges produced by natural sound sources. Boehnke and Phillips (2005) found no significant effect of ITD on temporal asymmetry detection when A and B noise bursts had ITDs of 600 μ s, opposite in sign between A and B; ± 600 μ s correspond approximately to the ITDs produced by free-field sound sources located to the extreme right and left of the listener (Kuhn 1977; Middlebrooks and Green 1990). Similarly, Füllgrabe and Moore (2012) found that ITDs up to 500 μ s in tonal stimuli had only weak effects on temporal asymmetry detection. Both of those groups reported little or no subjective experience of stream segregation based on ITD using a procedure similar to that employed by van Noorden (1975) to evaluate tonal stream segregation.

Boehnke and Phillips (2005) also tested stream segregation based on ILDs, presenting stimuli with 12-dB ILDs differing in sign between the A and B noise bursts (corresponding roughly to free-field sound sources located $>30^\circ$ to the right and left of the frontal midline; Shaw 1974). That condition produced statistically significant but weak disruption of temporal asymmetry detection although there was a clear subjective experience of stream segregation.

Fusion of sounds sharing a common onset can be highly resistant to degradation by conflicting spatial cues. Several groups have tested perception of components of speech sounds presented dichotically. Cutting (1976) constructed two-formant syllables, /ba/ and /ga/, and presented various mismatched pairs of formants to listeners, one formant to each ear. In many instances, listeners presented with a lower formant from /ba/ in one ear and an upper /ga/ formant in the other reported hearing a single fused /da/ sound, even though there were no /da/ components in the stimulus. Broadbent and Ladefoged (1957) constructed stimuli consisting of brief sentences in which odd-numbered formants were presented to one ear and even-numbered formants to the other ear. Nearly all of the listeners experienced fusion of the dichotic stimuli such that they reported hearing only a single voice from a single (midline) location. Hukin and Darwin (1995) used vowel formant boundaries as a measure of listeners' ability to fuse vowel components that differed in spatial cues. Presentation of the 500-Hz component of a vowel at the ear opposite from the other components was equivalent to reducing the level of the 500-Hz component by only 5 dB. Displacement of the 500-Hz component from the other vowel components with 666- μ s ITDs differing in sign had an even smaller effect on the formant boundary. In a study using free-field stimuli from multiple loudspeakers, Takanen and colleagues (2013) studied recognition of concurrent vowels, with odd-numbered formants presented from one location and even-numbered formants from another. Vowel recognition by their listeners showed essentially no influence of spatial separation of sources of odd and even formants. All of these split-formant speech tasks show the capacity of common onsets to bind together elements of auditory objects; see Chap. 5 by Elhilali and Elhilali et al. (2009) for related animal physiological studies. Also, fusion was experienced only when the various components shared a common fundamental frequency or when the formants were excited by noise (i.e., were aperiodic; Takanen et al. 2013). Introduction of differences in the fundamental frequencies at the two ears could disrupt fusion (Broadbent and Ladefoged 1957).

In summary, published studies of spatial stream segregation measured with tasks that demanded integration have demonstrated minimal disruption of integration by spatial cues and only for cues corresponding to extreme spatial separation, as in the opposite-ear condition. More realistic differences in spatial cues between two sounds apparently are insufficient to disrupt fusion, especially when sounds are bound by common onset and/or common fundamental frequency. The failure of integrative tasks to demonstrate a strong effect of spatial cues may seem to conflict with the classic result by Cherry (1953), who showed that listeners readily segregated conflicting speech streams presented to opposite ears. In Cherry's study, however, listeners were encouraged to segregate the messages at the two ears, and

segregation *enhanced* performance, as in the stream segregation tasks described in the following section.

6.2.2 *Robust Stream Segregation by Spatial Cues*

Stream segregation can be measured directly by requiring a listener to segregate two or more sound streams and to make a judgment based on information in one stream while rejecting distraction by the other streams. Such tasks have been referred to as measures of fission (van Noorden 1975), segregation (Micheyl and Oxenham 2010), and voluntary segregation (Stainsby et al. 2011). This is the task of a person attempting to follow a particular conversation amid a crowd of other talkers. The magnitude of stream segregation is quantified by the degree to which it *improves* performance. Direct measures of stream segregation have demonstrated robust effects of space or of spatial cues. Hartmann and Johnson (1991) demonstrated that two interleaved melodic lines could be segregated when the two melodies were presented to the two ears with ITDs of ± 500 μ s. Identification of the melodies in that condition was nearly as accurate as when the signals were presented to opposite ears. In a study by Saupé and colleagues (Saupé et al. 2010), listeners heard musical phrases played by synthesized instruments differing in timbre and were instructed to report a large descending pitch interval played by a particular target instrument. Performance was enhanced substantially when the sources were separated in location in the free field by 28° compared to a co-located condition. Sach and Bailey (2004) asked listeners to distinguish rhythmic patterns of 500-Hz tone pips localized to the perceptual midline in the presence of interleaved masker pulses. Performance improved significantly when the masker was lateralized by introduction of a 100- to 200- μ s ITD or a 4-dB interaural level difference (ILD). The preceding three studies demonstrate that spatial features of sounds can enhance perceptual segregation of target and masker and can thereby enhance target recognition.

Spatial stream segregation can contribute substantially to recognition of speech in the presence of competing speech or other sounds. Most real-world efforts to recognize speech in the presence of other sounds are confounded by some combination of energetic masking, in which signal and masker overlap in time and spectrum, and sequential masking, in which there is no spectrotemporal overlap. Spatial cues are particularly important for segregating interleaved sequences of sounds from competing talkers and for linking together sequential sounds from the same talker. That phenomenon was illustrated by work by Ihlefeld and Shinn-Cunningham (2008a, b). In their experiments, energetic masking was minimized by restricting target and masker to multiple interleaved nonoverlapping spectral bands. A 90° separation of target and masker sources substantially improved the rate of correct identification of words, particularly by reducing the instances in which target words were replaced by words from the masker string. A cue to the location (but not the timbre) of the target enhanced the spatial effect. Kidd and colleagues evaluated the importance of spatial cues for linkage of

successive words in an utterance (Kidd et al. 2008). Listeners heard pairs of five-word sentences spoken by two talkers in which successive words alternated between the two talkers. The speech sources could be co-located or could be separated in perceived interaural location by introduction of ITDs. Along with talker identity and correct syntactic structure, interaural location improved word recognition by linking together words from the target talker.

6.2.3 Spatial Acuity of Stream Segregation

A study of the ability of human listeners to form perceptual streams based on source location utilized interleaved sequences of target and masker noise bursts having identical spectral envelopes and differing only in source location (Middlebrooks and Onsan 2012). A nonverbal objective task was adopted to facilitate comparison of human psychophysical results with animal psychophysical and physiological results. The sound bursts had no temporal overlap, thereby isolating the phenomenon of stream segregation and eliminating any energetic masking. Success in the task required a listener to segregate otherwise identical sequences of sounds into distinct streams on the basis of source location and to discriminate rhythmic patterns within one of those streams. The schematic in Fig. 6.1 shows, in solid bars,

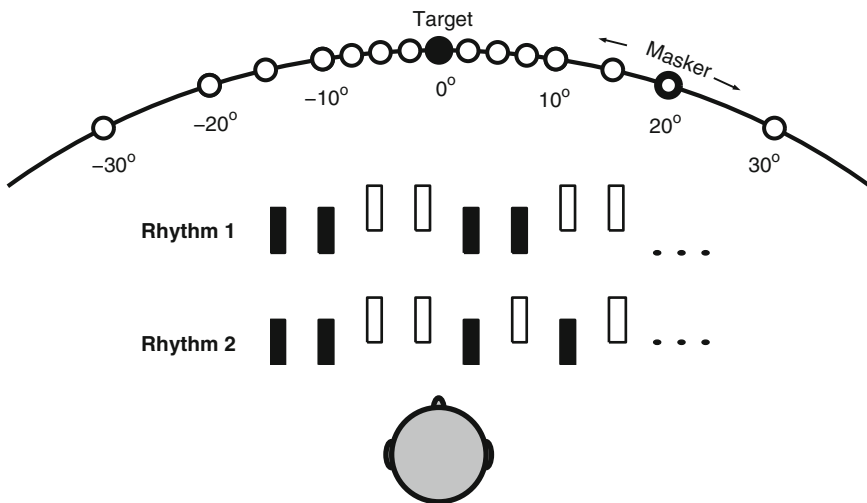


Fig. 6.1 Schematic of a psychophysical measure of spatial stream segregation using rhythmic masking release (RMR). The listener heard sequences of noise bursts presented from loudspeakers positioned in the horizontal plane. The target source was fixed at 0 or 40° , and the masker source, shown here at 20° , varied in location between trials. Target and masker noise bursts (indicated by solid and open bars, respectively) were interleaved in time and were identical except for their source locations. On each trial, one or the other illustrated rhythm was repeated four times without interruption, and the listener indicated whether the target sequence was rhythm 1 or 2

the two target rhythms that were to be discriminated along with, in open bars, the complementary masking sequences; the component broadband or band-passed noise bursts were 20 ms in duration and presented at an aggregate rate of 10/s. In this single-interval design, listeners reported by button press whether they heard rhythm 1 or rhythm 2. When target and masker sources were co-located, the stimulus was by design an undifferentiated sequence of noise bursts. In that condition, target and masker were heard as a single stream, and discrimination of the target rhythm was impossible. Hypothetically, spatial separation of target and masker sources could lead to perception of target and maskers sequences as distinct streams, thereby permitting analysis of the temporal pattern within the target stream and recognition of the target rhythm. This is referred to as rhythmic masking release.

The performance of one listener is shown in Fig. 6.2, with 6.2A and B representing results for target sources fixed respectively at 0 and 40° azimuth in the horizontal plane; the locations of masker sources are plotted as the horizontal axes. The accuracy in performance of the task is given by the discrimination index, d' , for discrimination of rhythm 1 from rhythm 2, where d' near zero indicates random-chance performance and $d' = 1$ was taken as the criterion for threshold rhythmic masking release. The expected near-zero d' values were obtained when the masker source location coincided with the target location. Even small displacements of the masker source, however, resulted in emergence of perceptually segregated streams, which resulted in unmasking of the target sequence and rapid improvement in rhythm discrimination. Dashed lines in the figure indicate the crossings of $d' = 1$ that indicate threshold target-masker displacements. In the broadband stimulus condition shown in Fig. 6.2, the median threshold for rhythmic masking release across seven listeners was 8.1° when the target source was at 0° and was 11.2° when the target was at 40°. When asked to report their subjective experiences, these listeners tended to report hearing two distinct streams when the target-masker separation was at or wider than the listeners' masking release thresholds and single streams when the separation was narrower.

Rhythmic masking release thresholds were significantly wider for the 40° than for the 0° target location, although the difference in medians was only 3.1°. A previous physiological study in auditory cortex of domestic cats (*Felis catus*) demonstrated proof of concept of a model of spatial hearing based on comparison of summed activity of left- and right-tuned neural populations in auditory cortex (Stecker et al. 2005). Models invoking only two or three spatial channels have gained some favor in regard to human psychophysics (Phillips 2008; Dingle et al. 2010) and to human neurophysiology using far-field magnetic and electric recordings (Salminen et al. 2009; Magezi and Krumbholz 2010; Briley et al. 2013). Those models, however, predict a rapid fall-off in spatial acuity with increasing distance of the target to the left or right of the midline, contrary to the spatial stream segregation results presented here. That is, in rhythmic masking release tested with a target at 40°, the target and all the tested masker locations would have been within the receptive fields of putative right-tuned neuronal populations, and a left-vs-right channel model would have predicted low (i.e., poor) spatial acuity. The observed

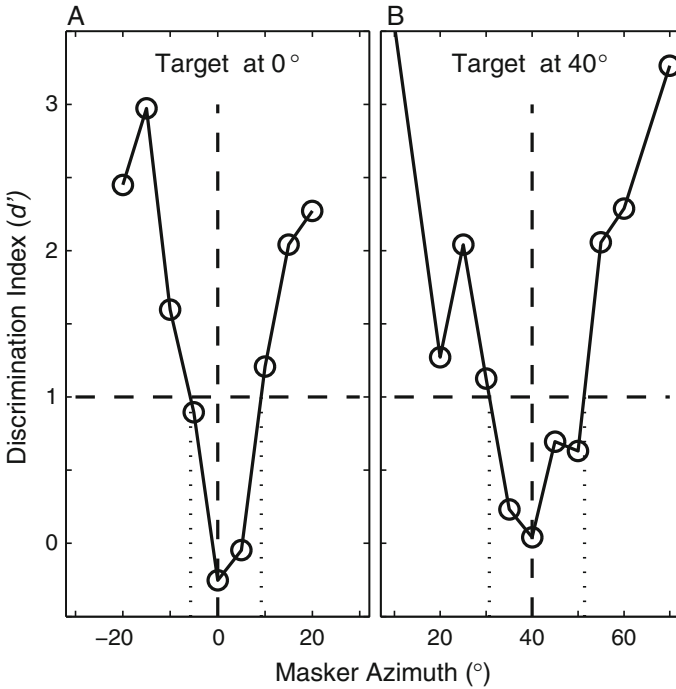


Fig. 6.2 Spatial stream segregation by one listener. Performance in the rhythmic masking release (RMR) task is represented by an index of the discrimination of rhythm 1 versus rhythm 2 (d') as a function of location of the masker source in the horizontal plane. (A, B) Conditions in which the target was fixed at 0° and 40° , respectively. RMR thresholds were given by the minimum interpolated target/masker separations at which performance exceeded a criterion of $d' = 1$, indicated by dashed lines. Two thresholds, indicated by dotted lines for masker locations to the left and right of the target, were determined for each listener and condition. (From Middlebrooks and Onsan 2012)

high acuity for the 40° target conflicts with that prediction and is more consistent with spatial-hearing models that incorporate the spatial sensitivity of single neurons (Middlebrooks et al. 1994; Lee and Middlebrooks 2013).

Thresholds for rhythmic masking release approached the thresholds measured in the same listeners for discrimination of a right-to-left from a left-to-right sequence of two sounds, their minimum audible angles (MAAs). The distributions of masking release thresholds overlapped with those of MAAs, but masking release thresholds of individual listeners generally were somewhat wider than their MAAs.

The high level of performance and fine spatial acuity obtained in this test of spatial stream segregation using rhythmic masking release contrast markedly with the weak, low-acuity spatial effects observed with tests of stream integration. Again, it appears that a listener can overcome spatial separation in tasks in which segregation is a liability for integrating information from multiple sources but,

alternatively, can take advantage of the spatial arrangement of an auditory scene when the goal is to attend to one of several interleaved sound streams.

6.2.4 *Acoustic Cues for Spatial Stream Segregation*

The locations of sound sources in space are computed within the central auditory system from acoustical cues that result from interaction of the incident sound wave with the head and external ears (Middlebrooks and Green 1991). The dominant cues for localization of broadband or low-pass sounds in the horizontal dimension (i.e., in azimuth) are ITDs in the ongoing temporal fine structure of sounds (Wightman and Kistler 1992), and the dominant cues for horizontal localization of high-pass sounds are ILDs (Macpherson and Middlebrooks 2002). One can distinguish the relative contribution of fine-structure ITDs or of ILDs to spatial stream segregation in the horizontal dimension by testing with low- or high-pass sounds. In the vertical dimension, the primary spatial cues are spectral shapes that result from the direction-depending filtering properties of the external ears. One can isolate spectral shape cues by testing locations in the vertical midline, where ITDs and ILDs are essentially uninformative.

Spatial stream segregation in the horizontal dimension was tested for broadband (0.4–16 kHz), low-band (0.4–1.6 kHz), and high-band (4–16 kHz) stimuli using the rhythmic masking release task (Middlebrooks and Onsan 2012); in each condition, pass-bands for target and masker stimuli were identical. Performance in the low-band condition was not significantly different from that in the broadband condition (Fig. 6.3). In contrast, performance was substantially worse in the high-band condition in which low-frequency cues were eliminated. Those results suggest that low-frequency ITD cues provided the highest spatial acuity for stream segregation in the horizontal dimension. A separate test demonstrated that the spatial stream segregation in the absence of low-frequency ITD cues (i.e., high-pass sounds) was derived primarily from ILD cues, with little or no contribution from better-ear level cues or from ITDs in envelopes of high-frequency sounds (Middlebrooks and Onsan 2012).

The demonstration that spatial stream segregation in the horizontal dimension relies on ITD and (with lesser spatial acuity) on ILD cues raises the question of whether the observed stream segregation is a property of spatial hearing in general or whether it is specifically a binaural process. That question was addressed by testing for spatial stream segregation with target and masker sources both located in the vertical midline, where binaural cues to target and masker separation are negligible and where spectral shapes are the primary spatial cue (Middlebrooks and Onsan 2012). The performance of one listener in such a task is shown in Fig. 6.4; in this figure, the horizontal axis depicts the vertical location of the masker above or below the horizontal plane (at 0° elevation). An unanticipated result was that sensitivity to target–masker separation depended rather strongly on the durations of the broadband sound bursts that formed the target and masker sequences. When

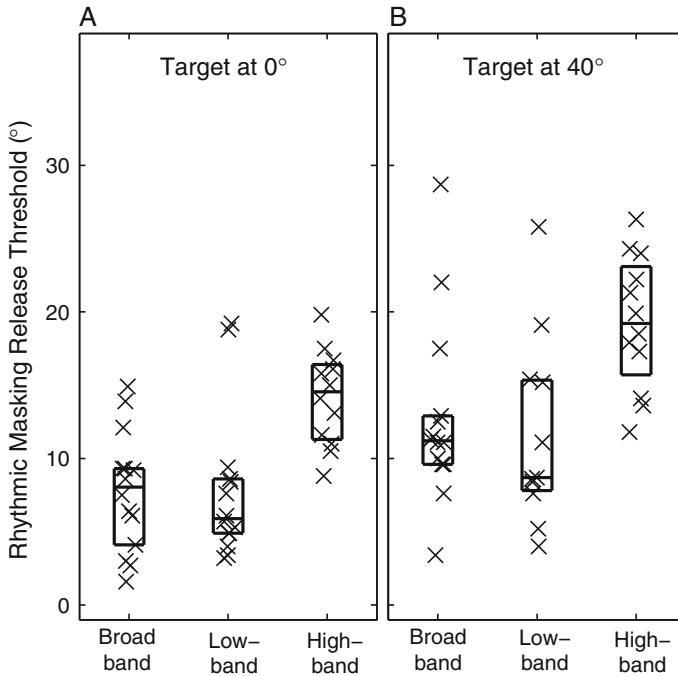


Fig. 6.3 Distributions of rhythmic masking release (RMR) thresholds as a function of stimulus band. Boxes indicate 25th, 50th, and 75th percentiles of distributions across seven listeners and across maskers to left and right of targets. Broadband pass bands were 0.4–16 kHz, low-pass bands were 0.4–1.6 kHz, and high-pass bands were 4.0–16 kHz. Thresholds were significantly wider for 40° (**right**) compared to 0° (**left**) target locations ($p < 0.0005$, paired signed rank test, in the broadband condition) and were significantly wider in the high-band condition than in the broadband or low-band conditions ($p < 0.005$ at 0° and $p < 0.05$ at 40°, Bonferroni-adjusted paired comparisons) (Middlebrooks and Onsan 2012)

burst durations were 10 ms (Fig. 6.4A), this listener and most others were unable to reach criterion sensitivity at any tested target–masker separation. Sensitivity improved for 20-ms bursts (Fig. 6.4B), although in this example the sensitivity hovered around $d' = 1$ for most of the masker locations below the horizontal plane. When the bursts were lengthened to 40 ms, however, sensitivity improved to levels comparable with those observed in the horizontal dimensions; the median of thresholds across listeners was 7.1° for 40-ms bursts in the vertical dimension compared with a median of approximately 4° for 40-ms bursts in the horizontal dimension. The observation that spatial stream segregation sensitivity varied with sound-burst duration is somewhat parallel to observations of impaired vertical localization of brief noise bursts (Hartmann and Rakerd 1993; Hofman and Van Opstal 1998; Macpherson and Middlebrooks 2000), although the impaired vertical localization in the previous studies was associated particularly with high sound

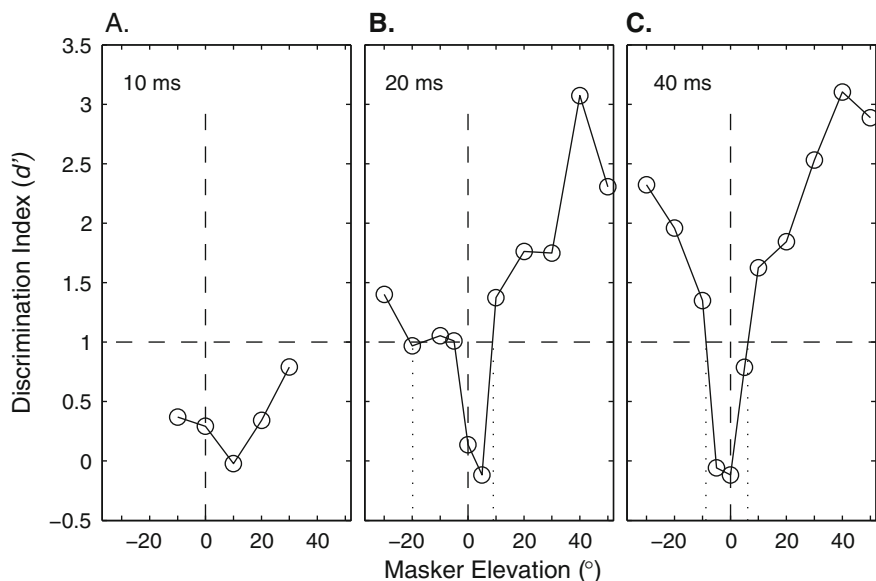


Fig. 6.4 Spatial stream segregation in the vertical midline by one listener. Target sources were fixed at 0° (i.e., in the horizontal plane containing the listener's ears), and the horizontal axes plot the location of the masker above or below the horizontal plane. The three panels show conditions in which the broadband noise bursts constituting the sound sequences were 10, 20, or 40 ms in duration (Middlebrooks and Onsan 2012)

levels whereas sounds were presented at moderate levels in the stream segregation experiments.

The distributions of rhythmic masking release thresholds observed in various experimental conditions are summarized in Fig. 6.5A. Minimum audible angles (Fig. 6.5B) also were measured for the same listeners as an indicator of their localization acuity independent of the complexity of the rhythmic masking release task; note the difference in the vertical scale between Fig. 6.5A and B. There is a striking difference between the two panels. Rhythmic masking release thresholds varied markedly in the horizontal dimension as a function of pass band and in the vertical dimension as a function of burst duration. In contrast, there was little variation in MAAs across those stimulus conditions. The difference in stimulus dependence between spatial stream segregation and MAAs suggests that the two spatial phenomena might result from differing brain structures or mechanisms. That issue is considered further in Sect. 6.4. That spatial stream segregation is observed in both the horizontal and vertical planes confirms, however, that spatial differences among competing sounds can support stream segregation irrespective of whether the spatial differences are processed by binaural or by other spatial mechanisms.

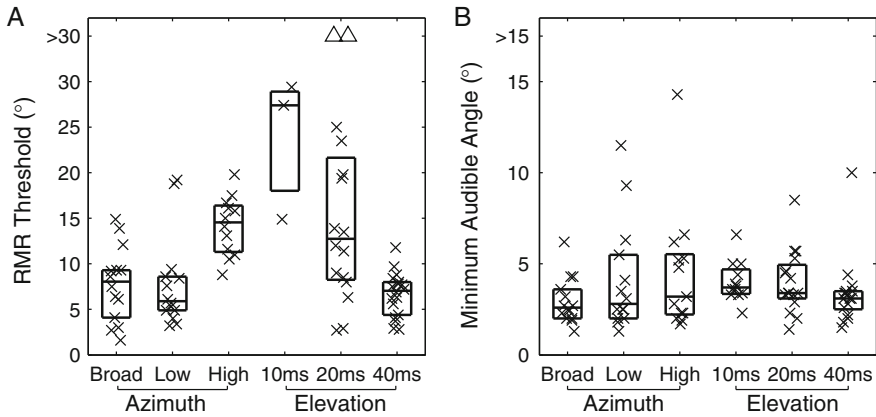


Fig. 6.5 Distributions of rhythmic masking release (RMR, **A**) and minimum audible angles (MAAs, **B**) thresholds in various conditions of stimulus pass-band and spatial dimension. The vertical scales differ between the two panels. Stimulus locations in the horizontal plane (azimuth) were tested in broadband, low-band, and high-band conditions, all with 20-ms sound bursts and with the target at 0° azimuth. RMR thresholds in the horizontal dimension varied significantly with stimulus pass-band ($p < 0.0005$, Kruskal–Wallis test), whereas there was no significant variation in MAAs across those conditions ($p > 0.05$). Locations in the vertical midline were tested with broadband sounds having 10 ms, 20 ms, and 40 ms durations, with the target at elevation 0°. RMR thresholds in the vertical dimension varied significantly with burst duration ($p < 0.0001$), and all pairwise differences between durations were significant ($p < 0.05$, Bonferroni-adjusted paired comparisons). MAA thresholds also varied significantly with burst duration ($p < 0.005$) but that was due entirely to the difference between 10- and 40-ms conditions; no other paired comparisons were statistically significant (Middlebrooks and Onsan 2012)

6.3 A Bottom-Up Substrate for Spatial Stream Segregation

Neural correlates of stream segregation based on tone frequency have been demonstrated in cortical area A1 of the macaque monkey (Fishman et al. 2001; Micheyl et al. 2005). When presented with sequences of tone pips that alternate in frequency, cortical neurons tend to synchronize to tones of one or the other frequency. Tonal stream segregation operates within a substrate of tonotopic organization in which frequency-selective single neurons are organized into orderly maps of tone frequency onto cortical place. Proposed mechanisms of tonal stream segregation have involved inhibitory interactions among loci along the cortical tonotopic axis (Fishman et al. 2001).

Neural pathways for spatial hearing begin with analysis of acoustic spatial cues in the auditory brainstem. Results of that analysis are conveyed to the level of auditory cortex, where responses of single neurons vary in magnitude and timing according to sound-source location (Middlebrooks et al. 1994). The spatial sensitivity of single neurons, however, is far less precise than is the ability of an animal

to localize a sound by, for instance, orienting to a sound source to receive a food reward (May and Huang 1996; Tollin et al. 2005). Neural localization performance comparable in precision to behavior has been demonstrated only in the coordinated activity of populations of neurons (Furukawa et al. 2000; Miller and Recanzone 2009). There is substantial evidence *contrary* to the presence of orderly maps of sound-source location onto cortical place (King and Middlebrooks 2011). That raises the question of whether or not single cortical neurons could exhibit spatial stream segregation analogous to the tonal stream segregation shown by Fishman and colleagues (2001). A recent study addressed that question and showed that, indeed, responses of single cortical neurons can segregate competing streams of sounds from differing locations (Middlebrooks and Bremen 2013). Section 6.3.1 reviews that study, showing that the acuity of spatial stream segregation by single cortical neurons is substantially greater than the acuity for locations of single sound sources and approaches that of humans in psychophysical tests. Section 6.3.2 considers the evidence that spatial stream segregation reflects bottom-up processing within the auditory pathway at or below the level of the thalamocortical projection.

6.3.1 *Spatial Stream Segregation in Primary Auditory Cortex*

A putative substrate of spatial stream segregation was studied in primary auditory cortex of anesthetized cats (Middlebrooks and Bremen 2013). The use of general anesthesia almost certainly influenced cortical responses, and any failure to find evidence of spatial stream segregation might have been blamed on anesthetic effects. Contrary to that concern, however, spatial stream segregation was observed in that anesthetized preparation, suggesting that at least some basal level of segregation arises from bottom-up processes that do not require an animal's attention.

Stimuli consisted of sequences of brief noise bursts alternating in location from two sources in the horizontal plane in a free sound field; the source locations were varied parametrically. As discussed in Sect. 6.2.3, human listeners report hearing such stimuli as two distinct streams when target–masker source separations are approximately 10° or wider. In the cat cortical experiment, the base rate of noise-burst presentation (i.e., the aggregate of both sources) was 5 or 10/s. Cortical neurons synchronized closely to noise burst presented at half the base rate from only one of the sources, as shown in the representative post-stimulus time (PST) histograms in the left column of panels in Fig. 6.6. The example neuron showed little sensitivity to the location of a single source, with essentially equal responses to sounds presented from straight ahead (0° , Fig. 6.6C) or from 40° contralateral (Fig. 6.6A) or ipsilateral (Fig. 6.6E) with respect to the side of the recording site. When one source was held at 0° and a second source was added from the same location (equivalent to simply raising the rate to the full aggregate rate), there was a reliable response to the first noise burst, but responses to later bursts

were sparse and irregular (Fig. 6.6D). In the figure, red and blue bars indicate spikes that were synchronized to the A or B source, respectively, although the A and B designation is arbitrary in the co-located condition shown in Fig. 6.6D. When the A source was held at 0° and the B source was shifted to ipsilateral 40° (Fig. 6.6F), the response to the B source largely disappeared and the neuron responded reliably to the A source. In that configuration, the response of the neuron could be said to segregate the A sound sequence from the B sequence. A largely symmetrical response was observed when the A source was held at 0° and the B source was shifted to contralateral 40° (Fig. 6.6B). In that configuration, the B sound sequence dominated the response of the neuron.

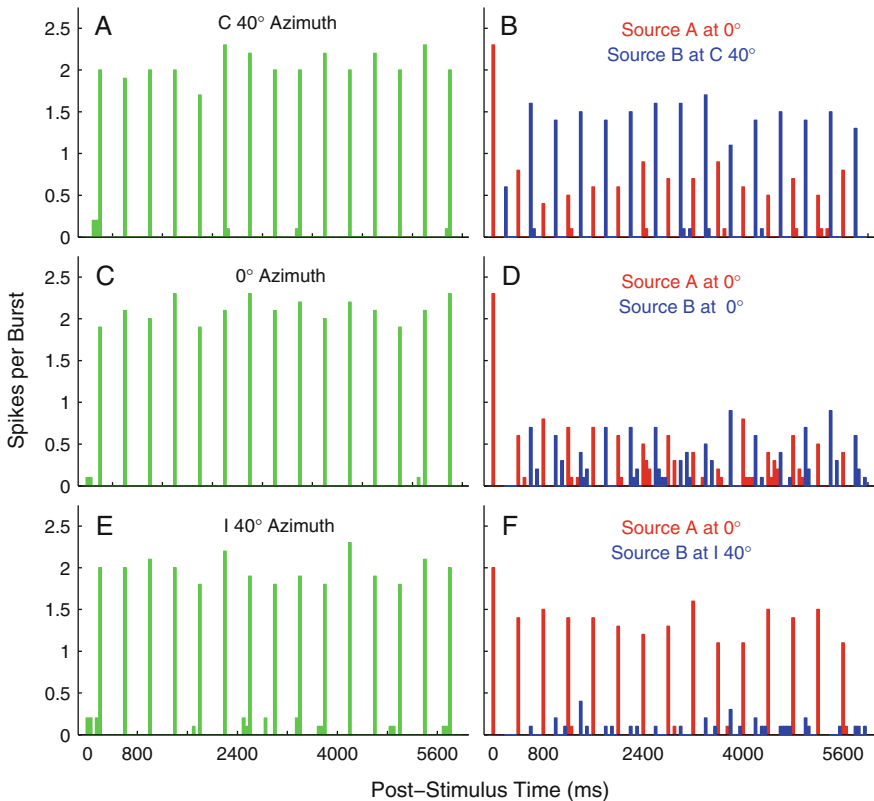


Fig. 6.6 Post-stimulus time histograms of a single neuron in cortical area A1 of an anesthetized cat. Bars indicate the neural spikes per stimulus burst in 50-ms bins. In the left panels, the stimuli were sequences of 5-ms noise bursts at a rate of 2.5/s presented from 40° contralateral with respect to the side of the recording site (A), straight ahead (C), or 40° ipsilateral (E). In the right panels, red bars indicate responses synchronized to source A, fixed at 0° azimuth, and blue bars indicate responses synchronized to source B located at contralateral 40° (B), straight ahead (D), or ipsilateral 40° (F). The aggregate rate of A and B sound bursts in the right panels was 5/s (Middlebrooks and Bremen 2013)

Responses of the unit represented in Fig. 6.6 are plotted in the left column of Fig. 6.7 as spike counts synchronized to A or B sound sources as a function of the location of the B source; Fig. 6.7A, C, and E represent conditions in which the A source was fixed in location at contralateral 40°, 0°, or ipsilateral 40°, respectively. As noted in the preceding text, the response to the B source alone (green line, duplicated in each panel) showed little sensitivity to the source location. Spatial sensitivity sharpened, however, in conditions of competing sources. In panels A, C, and E the neural response was suppressed, compared to the B-alone condition, in configurations in which the B source location coincided with the A location. In each case, however, the response that was synchronized to one or the other source, and the difference between the responses to the two sources, increased dramatically as the two sources were moved apart. In the right column of panels in Fig. 6.7, the blue line plots a measure of discrimination of spike counts synchronized to the B versus the A source; the discrimination index, d' , was computed from a receiver operating characteristic (ROC) analysis of trial-by-trial spike counts. The dashed black lines indicate criteria of $d' = \pm 1$. In nearly every case in this illustration, the neural spikes segregated the A and B sources with d' larger than 1 when they were separated by the minimum tested distance, either 10° or 20°. In contrast, a comparison of spike counts elicited by a single source at varying locations compared to a single source fixed at contralateral 40°, 0°, or ipsilateral 40° (green line) indicates minimal sensitivity of the neuron to the location of a single source.

The responses of the single neuron shown in Figs. 6.6 and 6.7 were representative of the sample from primary auditory cortex in the Middlebrooks and Bremen (2013) study in several regards. First, spike rates of the majority of neurons could reliably segregate two interleaved sequences of noise bursts. When the base stimulus rate was 10/s, for instance, spike rates of 78% of single- and multiple-unit recordings could segregate with $d' \geq 1$ sequences from one or more pairs of source locations separated by only 20°. Second, like the illustrated responses, the majority of neurons tended to synchronize preferentially to the more contralateral of the two sound sources. Nevertheless, a sizeable minority of neurons (not illustrated) preferred the more ipsilateral source. Third, nearly every neuron showed greater spatial sensitivity in conditions of two competing sound sources compared to conditions of a single source. The majority of spike rates were modulated by no more than 50% as a function of the location of a single source, as shown by the green line in Fig. 6.7. Other neurons showed contralateral hemifield tuning to single sources in that they responded strongly for a single source in the spatial hemifield contralateral to the recording site; there also were a few examples of ipsilateral or frontal spatial tuning for single sources. In nearly every case, however, tuning widths were narrower, modulation of spike rates by source location was deeper, and discrimination of locations by trial-by-trial spike counts was greater in competing-source conditions compared to the single-source condition ($p < 10^{-6}$, all pairwise comparisons).

Neurons that synchronized preferentially to the more contra- or ipsilateral of two competing sources tended to occupy distinct modules within the cortex such that an electrode track through the cortex would encounter uninterrupted sequences of neurons showing only one laterality followed by sequences showing the other

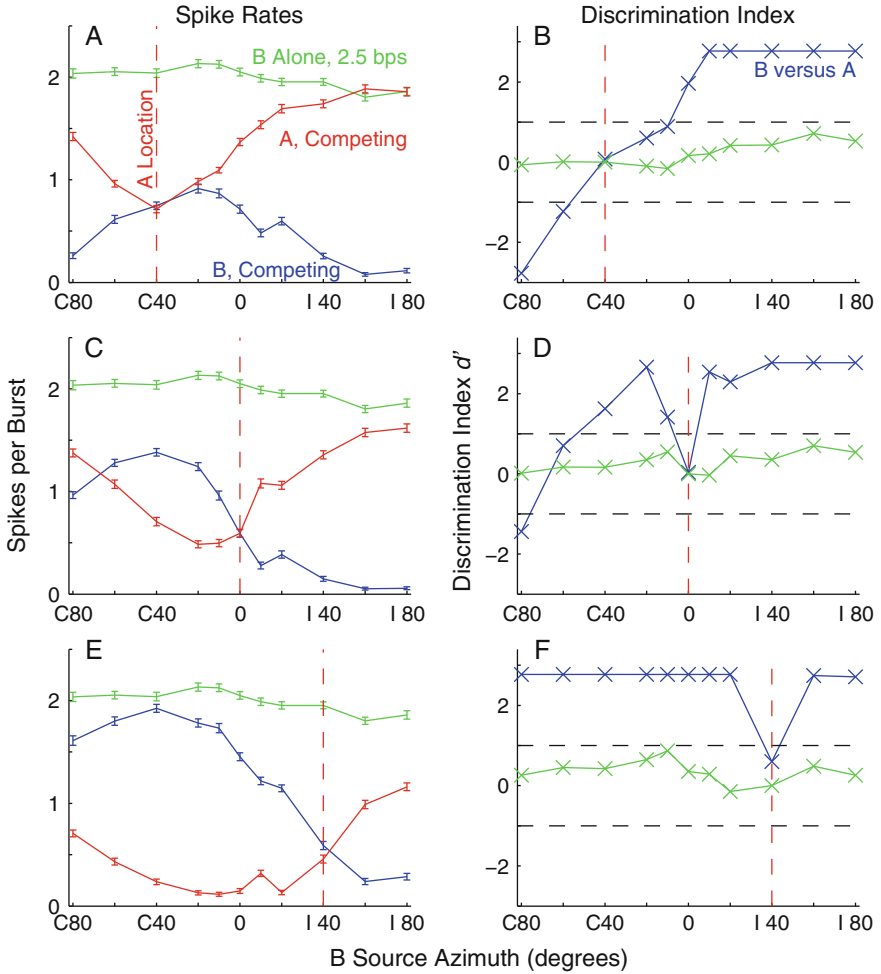


Fig. 6.7 Spike rates and stream segregation by the neuron shown in Fig. 6.6. In the left panels, lines indicate mean spikes per sound burst for a single source (green) or for spikes synchronized to the A (red) or B (blue) source when A and B sound sequences were interleaved. In each panel the location of the A source was fixed at the location indicated by the vertical dashed red line. The right panels show the discrimination index, d' , for discrimination of spike rates synchronized to A versus B sources (blue) or for discrimination of spike rates elicited by a single source varied in location compared to the source fixed at the fixed location (green). Positive values of d' indicate stronger responses to the more contralateral of two sound sources (Middlebrooks and Bremen 2013)

laterality; a permutation test showed a probability $<10^{-5}$ that such nonrandom distribution of laterality preferences could have arisen by chance. Sequences of neurons having constant laterality preference often were elongated along cortical columns, but there also were examples of such sequences extending across

columns, that is, spanning a range of characteristic frequencies. In anesthetized conditions, A or B sound sequences have no special significance as target or masking sounds. In awake conditions, however, the listener might identify one or the other sequence as an auditory object of interest. In such conditions, one might hypothesize that some as-yet-unidentified top-down mechanism facilitates activity of cortical modules synchronized to the target and/or suppresses activity of modules tuned to the masker.

6.3.2 Spatial Rhythmic Masking Release by Cortical Neurons

The degree of spatial stream segregation shown by single neurons in auditory cortex in anesthetized cats was sufficient to support spatial rhythmic masking release comparable to that demonstrated in human psychophysics (Middlebrooks and Onsan 2012). Rhythmic sequences of noise bursts such as those used in the human study (Fig. 6.1) were presented, and responses of single neurons were studied in primary auditory cortex of anesthetized cats (Middlebrooks and Bremen 2013). Responses of a single neuron are represented by PST histograms in Fig. 6.8. The sequence of open and filled squares across the top of each panel represents the rhythms of target (open) and masker (filled) noise bursts; top and bottom rows of panels represent rhythm 1 and rhythm 2. When target and masker sources were co-located, as in Fig. 6.8B and E, human listeners reported hearing a single stream and the cat cortical neuron synchronized equally to both sources. When the masker was shifted to contralateral 80° (Fig. 6.8A and D) or ipsilateral 80° (Fig. 6.8C and F), however, human listeners reported hearing two segregated streams, and the neuron responded with distinctive PST patterns. Within each stimulus sequence, the neuron tended to respond strongly to a change from target to masker location or vice versa. That resulted in two strong responses to target bursts for each repetition of rhythm 1 (Fig. 6.8A and C) and three strong responses per repetition of rhythm 2 (Fig. 6.8D and F).

A linear-classifier analysis was used to test whether distinctive PST histogram patterns of single neurons such as that shown in Fig. 6.8 could reliably distinguish target rhythms and thereby perform the rhythmic masking release task. That analysis used linear regression with terms given by spike counts in 50-ms time bins and coefficients optimized to yield outputs of 1 or 2 depending on the stimulus rhythm. An ROC analysis of the distribution of outputs resulting from stimulus rhythms 1 or 2 yielded d' for discrimination of the rhythms by single neurons; a one-out procedure was used so that test trials differed from the trials that were used to compute regression coefficients. Discrimination results are shown in Fig. 6.9A for the neuron represented in Fig. 6.8, and Fig. 6.9B shows the distribution of discrimination results across the neuron sample. More than 25% of the isolated single neurons could isolate target and masker streams adequately at a 10°

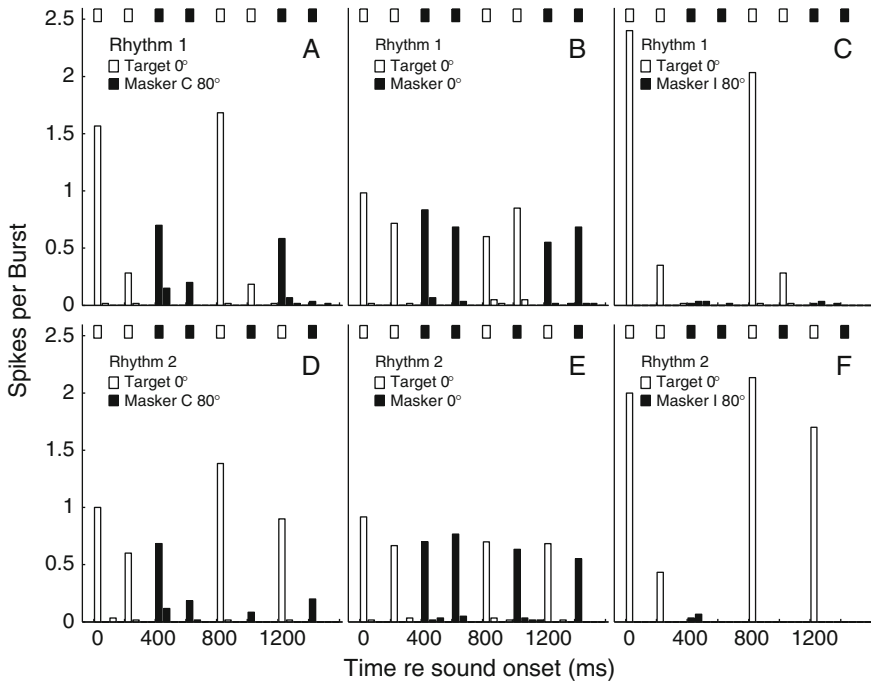


Fig. 6.8 Synchrony of a cortical neuron to rhythmic masking release stimuli. Top and bottom rows of panels represent responses to rhythms 1 and 2, respectively. Stimulus rhythms are represented by the row of squares at the top of each panel. The time axis is folded on the 1600-ms rhythm duration, so that each panel represents means of three repetitions per trial times 20 trials. Open and solid response bars indicate mean spike rates per target (open) and masker (solid) burst in 50-ms bins. Target and masker source locations were as indicated in each panel (Middlebrooks and Bremen 2013)

separation to achieve $d' \geq 1$. That compares favorably with median rhythmic masking release thresholds of 8.1° in human psychophysics reviewed in Sect. 6.2.3 (see also Middlebrooks and Onsan 2012). Task performance by the human listeners required judgments of the perceived rhythmic patterns, whereas in the neurophysiological study the rule for identifying the two rhythms was effectively programmed into the computer as a result of the feedback inherent in the regression analysis. For that reason, the results do not address discrimination of rhythms by neurons. Nevertheless, one can conclude that neural segregation of A and B sound sequences was sufficient to discriminate the rhythms and that segregation could be accomplished by brain mechanisms that were active even under anesthesia.

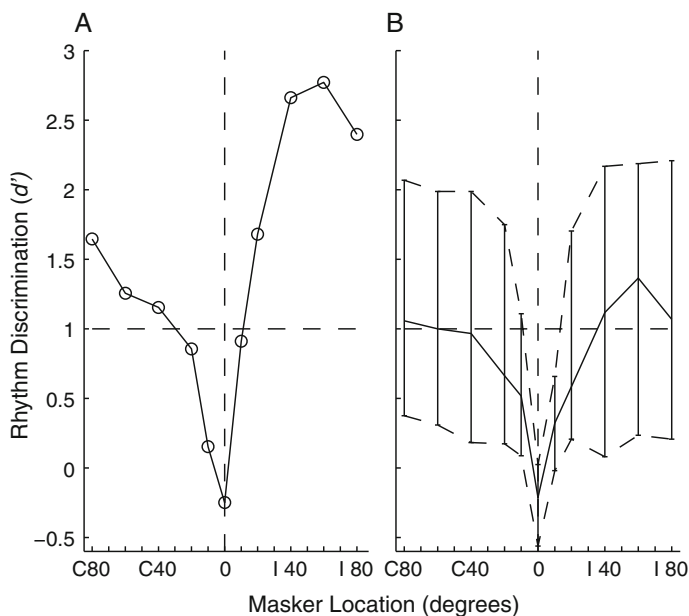


Fig. 6.9 Identification of masked rhythms based on synchronized neural responses. **(A)** The d' for discrimination of rhythm 1 versus 2 by a single neuron as a function of masker source location; the target was fixed at 0° . **(B)** The average d' across 57 single units. The curves indicated 25th, 50th, and 75th percentiles of the distribution (Middlebrooks and Bremen 2013)

6.3.3 A Mechanism for Bottom-Up Spatial Stream Segregation

The spatial stream segregation by single neurons observed in auditory cortex of anesthetized cats (Middlebrooks and Bremen 2013) could be predicted quantitatively by a model invoking (1) relatively weak spatial sensitivity inherited from the ascending brainstem auditory pathway; and (2) forward suppression tentatively localized somewhere in the projection from the medial geniculate body (MGB) and primary auditory cortex. Figure 6.10 shows recorded spike counts (symbols) and model predictions (lines) for one neuron tested with a 5/s stimulus rate (top row of panels) and another neuron tested at a 10/s rate (bottom row). Components of the model are discussed in the following.

The spatial sensitivity of cortical neurons for single sound sources presumably reflects sensitivity inherited from brainstem inputs plus any further sharpening that might occur at the cortical level. In the quantitative model, the responses of cortical neurons to single sources are taken as surrogates for the spatial sensitivity of the thalamocortical projection. As noted in Sect. 6.3.1, the spatial sensitivity of most cortical neurons was fairly broad, with the majority of neurons showing less than 50% modulation of their spike rates by varying sound-source location.

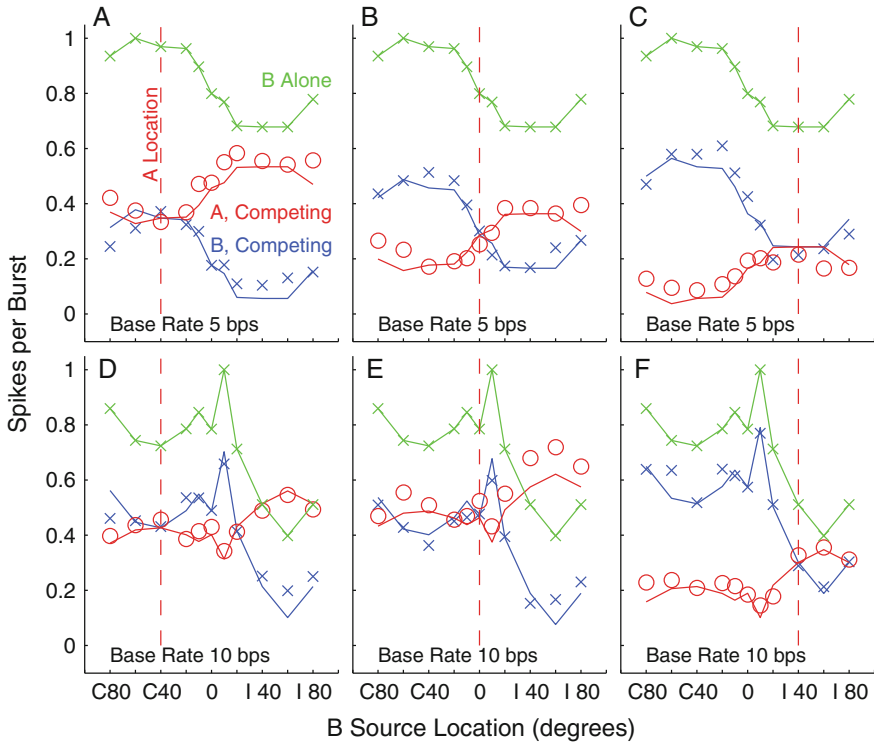


Fig. 6.10 Model predictions of neural stream segregation. The panels represent responses of one cortical neuron to stimuli at a base rate of 5/s (**A, B, C**) and of a different neuron to stimuli at a base rate of 10/s (**D, E, F**). Symbols represent mean spike bursts of the neurons, and curves represent model predictions based on responses to noise bursts from a single source (green) scaled by a forward suppression term. The neural spike count synchronized to source A (red) as a function of A and B locations θ_A and θ_B was given by $RA(\theta_A, \theta_B) = RSgl(\theta_A) - FS \times RSgl(\theta_B)$, and the response synchronized to source B (blue) was $RB(\theta_A, \theta_B) = RSgl(\theta_B) - FS \times RSgl(\theta_A)$, where $RSgl(\theta_A)$ and $RSgl(\theta_B)$ were the responses to sources A and B presented alone. The forward suppression term, FS , was estimated from the spike rates synchronized to trains of noise bursts at 5/s (top row) or 10/s (bottom row) divided by rates synchronized to trains at half those rates; ratios of spike rates were averaged across sources at contralateral 40°, 0°, and ipsilateral 40° (Middlebrooks and Bremen 2013)

Nevertheless, the spatial tuning for single sources tended to predict the spatial preference of the usually sharper tuning seen in the presence of a competing sound source. In the two examples in Fig. 6.10, the “B, competing” source plots (shown in blue) lie parallel to the single-source “B-alone” plots (green). The similarity between single- and competing-source spatial sensitivity was less obvious in the example in Fig. 6.7, but plots for the two conditions shared approximately the same peak locations and same signs of slopes. Across all the sampled units, the d' for discrimination of interleaved sound sequences from sources separated by 20° correlated highly ($r = 0.82$) with the d' for discrimination of single sources at the

equivalent locations, although d' in the competing-source condition tended to be about twice that in the single-source condition.

The response to a single source consistently was suppressed by addition of a competing sound. That is evident in Fig. 6.10 by the downward shift of the blue (competing-source) lines compared to the green (single-source) lines. The shifts tended to be a linear offset; that is, responses were attenuated by subtraction by a value that was constant within each panel, rather than by multiplication by a gain factor. The offset tended to be greater when the competing sound was at a location that would elicit a strong response to a single source (e.g., the A source fixed at 0° ; Fig. 6.10E) than when the competing sound was located at a less favored location (e.g., the A source fixed at ipsilateral 40° ; Fig. 6.10F). The response to sound A in the presence of competing sound B could be predicted by the response to a single source at the A location minus a forward suppression term times the response to a single source at the B location. That expression yielded the blue and red model lines in Fig. 6.10. The goodness of fit of the model (R^2) averaged 0.64 across 382 units tested with 5/s stimulus rates and averaged 0.46 across 295 units tested with 10/s rates.

The forward-suppression term in the Middlebrooks and Bremen model reflects the inability of cortical neurons to respond to rapid stimulus rates. Modulation transfer functions of primary auditory cortical neurons tend to peak around 10–30 Hz (Schreiner and Urbas 1988), which is consistent with the stimulus rates at which stream segregation is observed in the cortex (Fishman et al. 2001; Middlebrooks and Bremen 2013). Also, the relevant interstimulus times are on the same scale as forward masking that has been demonstrated in auditory cortex (Calford and Semple 1995; Brosch and Schreiner 1997). In contrast, neurons in the MGB respond well to stimulus rates in excess of 100 Hz (Creutzfeldt et al. 1980), considerably faster than the time scale of perceptual stream segregation and of the stream segregation demonstrated in the cortex. One possibility to consider for the failure of cortical neurons to follow rapid stimulus rates is that the minimum interspike time might be limited by the refractoriness of cortical neurons. That possibility was rejected in the case of stream segregation by Middlebrooks and Bremen (2013), who showed that the probability of a single neuron firing a spike to a particular noise burst was independent of whether or not that neuron had fired a spike to the immediately preceding noise burst. That indicates that rate-limiting step must be prior to the spiking activity of neurons in primary auditory cortex. The most likely alternative explanation is that forward suppression arises somewhere in the thalamocortical projection, possibly due to presynaptic inhibition or to synaptic depression in the thalamocortical synapses.

6.4 “Common” Versus “Dedicated” Spatial Representations for Localization and Spatial Stream Segregation

Basic bottom-up mechanisms for spatial stream segregation almost certainly are shared with pathways for sound localization *per se*; that is, for identification of the locations of sound sources. Those mechanisms include analysis of sound magnitude and phase spectra in the cochlea, interaural comparison of magnitude and phase in the superior olivary complex, identification of spectral cues for vertical locations, and some level of convergence leading to the spatial sensitivity of neurons in primary auditory cortex and in other cortical areas. It is less clear, however, whether the ultimate levels of stream segregation and localization occur within common cortical areas or whether there are particular cortical areas dedicated specifically to segregation and others to localization.

One might think of a hypothetical “common” cortical spatial representation that could be accessed for spatial segregation, localization, and possibly other spatial functions. In favor of such a common representation are the observations that spatial sensitivity of neurons in early levels of auditory cortex can both segregate interleaved sound sequences (Middlebrooks and Bremen 2013) and can identify sound-source locations; localization by cortical neurons with accuracy comparable to behavioral performance, however, can be accomplished only by integrated activity of multiple neurons in that cortical area (e.g., Mickey and Middlebrooks 2003; Miller and Recanzone 2009). These results suggest that, even if primary auditory cortex is not the ultimate locus of segregation and localization functions, it likely serves as a common pathway.

Seemingly a key prediction of a hypothesis of a common spatial representation is that the spatial acuity of all spatial functions (including localization and stream segregation) would vary in parallel as stimulus conditions are varied to favor spatial cues that differ in acuity (e.g., ITDs for azimuth compared to spectral shape cues for elevation). That prediction clearly was violated in the psychophysical results by Middlebrooks and Onsan (Middlebrooks and Onsan 2012, reproduced in Fig. 6.5 of the present chapter). That study used MAA as a measure of localization acuity. Distributions of MAA were largely constant across broadband, low-band, and high-band conditions in the horizontal dimension and across three pulse durations in the vertical dimension (Fig. 6.5B). In contrast, spatial acuity of stream segregation varied dramatically across stimulus pass-band and pulse-duration conditions (Fig. 6.5A). The difference between stream segregation compared to localization with respect to stimulus conditions suggests that spatial cues are utilized differently by dedicated pathways for spatial segregation and for localization, most likely in differing cortical areas.

Another indication that spatial segregation and localization involve differing brain structures, or at least differing mechanisms, comes from work by Edmonds and Culling (2005a, b). Speech targets were presented with maskers consisting of noise or competing speech. Speech reception was improved by introduction of

differences in ITD and/or ILD between target and masker. A summation of effects was observed for target and masker differing in both ITD and ILD (Edmonds and Culling 2005a) or in ITD in two distinct frequency bands (Edmonds and Culling 2005b). Surprisingly, that summation was observed whether the spatial direction was consistent across the interaural cues for a particular target or masker sound or whether the cues for a particular sound pointed in opposite directions. That is, in the opposite-direction case, spatial unmasking was possible even though the interaural cues for target and/or masker did not correspond to a plausible location in space. Localization was not requisite for spatial segregation.

At least some level of sensitivity to sound-source location is ubiquitous among auditory cortical neurons studied in animals. Although there are quantitative differences, qualitatively similar spatial sensitivity has been observed in every cortical area that has been studied in cats (Harrington et al. 2008), ferrets (Bizley et al. 2009), and nonhuman primates (Woods et al. 2006). One cannot yet say whether neurons in all those cortical areas also show stream segregation. Despite the widespread presence of cortical spatial sensitivity, behavioral studies show that cortical areas vary in their importance for localization and, presumably, other spatial tasks. For instance, Lomber and Malhotra (2008) compared the roles in behavior of two cortical areas in cat, the posterior auditory field (PAF) and anterior auditory field (AAF). Although differing in detail, neurons in both of those cortical areas are known to exhibit spatial sensitivity (Harrington et al. 2008). In the Lomber and Malhotra study, cats learned two tasks: they could identify the locations of sound sources and they could discriminate temporal patterns. Temporary inactivation of PAF disrupted performance of the localization task while preserving performance of the temporal pattern discrimination, whereas temporary inactivation of AAF disrupted performance of the temporal task and preserved localization. Those results suggest that the spatially sensitive neurons in PAF and AAF participate in dedicated networks that support localization in the case of PAF and temporal pattern analysis in the case of AAF. Hypothetically, the temporal pattern analysis by spatially sensitive neurons in AAF might participate in spatial stream segregation. Those dedicated networks might exist within PAF and AAF themselves and/or might reflect differential anatomical projections from those areas.

Human clinical results provide evidence for dedicated cortical substrates for particular auditory spatial functions. Thiran and Clarke (2003) and Duffour-Nikolov and colleagues (2012) evaluated 13 patients having unilateral cortical lesions varying in etiology. Of those patients, three showed pronounced deficits both in a lateralization task and in spatial release from masking, five showed lateralization deficits with preserved spatial release, and one showed intact lateralization with impaired spatial release. The dissociation of lateralization and spatial release from masking in 6 of 13 patients supports the view that these auditory spatial functions involve distinct cortical substrates.

6.5 Selection of Objects of Attention

Previous sections have demonstrated the importance of spatial cues for segregation of competing sounds (Sect. 6.2) and have demonstrated that interleaved sound sequences that human listeners would hear as segregated streams activate distinct neural populations in auditory cortex (Sect. 6.3). In real-world listening situations, however, humans or other animals must not only segregate multiple sounds but, from those segregated streams, must also select particular sound objects for attention and action. In Chap. 2, Shinn-Cunningham, Best, and Lee provide a broad overview of object selection. The present section considers task-dependent sharpening of spatial or spectral sensitivity of cortical neuron and presents other animal and human results that, although not specifically spatial, might be extrapolated to selection of objects specifically on the basis of location.

6.5.1 *Task-Dependent Modulation of Stimulus Specificity in Behaving Animals*

Two research groups have demonstrated that responses of cortical neurons can adapt on a rapid time scale to optimize performance when an animal is engaged in a sensory task. Neither of these groups has evaluated selection among simultaneous or interleaved sounds, but both have shown modulation of stimulus tuning during presentation of a reference sound that would enhance detection of a change from reference to a target sound.

Fritz and Shamma trained ferrets to detect the change from a broadband reference sound to a single- or multitone target (Fritz et al., 2003, 2007) or to discriminate between the directions of frequency shifts of two-tone sequences (Yin et al. 2014). The broadband reference probed the spectrotemporal receptive fields (STRFs) of neurons. Neurons in primary auditory cortex showed rapid changes in STRFs when the animal engaged in a task compared to during passive sound exposure. The STRF changes indicated changes in neuronal tuning that were adaptive in the sense that they would enhance discrimination of the target tone from the broadband reference or would enhance discrimination between two targets presented in separate time windows. The task-dependent modulation of stimulus tuning could be interpreted as a mechanism for enhancing the response to a particular object of attention, although the experiments tested only single targets, not simultaneous or interleaved target sequences.

Studies in trained cats have demonstrated task-dependent changes in selectivity for stimulus locations in space (Lee and Middlebrooks 2011, 2013). Cats pressed a pedal to initiate presentation of reference sounds consisting of a succession of broadband noise bursts from varying locations in the horizontal plane. The cat could release the pedal to receive a food reward when the sound changed to one of two targets. In “periodicity” trial blocks the target was a periodic click train,

whereas in “localization” blocks the target shifted in location to a higher elevation. Compared to the condition with unattended sound exposure, location specificity of neurons sharpened during performance of the periodicity task, which required attention to an auditory task. Further sharpening was observed during the performance of the localization task, which demanded evaluation of the location of each stimulus. The most common observed change in location specificity was that of increased suppression of responses to sounds from nonfavored locations, generally narrowing responses from omnidirectional sensitivity to responses restricted to locations contralateral to the recording site. Changes in location tuning were evident as soon as they could be evaluated, a few tens of seconds after the onset of task engagement. Again, that study did not test conditions of competing sounds, but the changes in neuronal spatial tuning such as those that accompanied engagement in that single-source task presumably would serve to enhance segregation and/or to selection of a sound sequence from one source among interleaved sequences from multiple sources.

In both the ferret and cat studies, the act of listening for a target resulted in changes in stimulus tuning to the reference sound, either to the broadband STRF probe or to the broadband noise bursts in the horizontal plane. A study in ferrets looked specifically at neural responses to the target sound. Atiani et al. (2009) evaluated the contrast between responses to reference and target sounds between on- and off-task conditions. Neurons developed robust, sustained responses to targets and suppressed their responses to reference sounds during task performance whereas they showed relatively little contrast between responses to target and reference during passive sound exposure. The task-dependent contrast in response to target and reference was moderate in the primary auditory area, stronger in a nonprimary area, and essentially binary in prefrontal cortex.

6.5.2 Object Selection in Human Neurophysiology

Recent studies of nonprimary auditory cortical areas in humans have demonstrated neurophysiological correlates of object selection by demonstrating enhanced synchrony of neural activity to the one of two competing speech streams that receives a listener’s attention. We consider here work from two research groups (Ding and Simon 2012a; Mesgarani and Chang 2012) that attempted to reconstruct features of speech stimuli from patterns of activity in neural populations. Both groups utilized auditory stimuli consisting of speech utterances from two talkers, and both found that the reconstructions varied markedly depending on which of the talkers received the listener’s attention. Neither group addressed specifically the influence of locations of sound sources, the principal topic of this chapter, although Ding and Simon (2012b) evaluated a condition of competing sounds delivered to the two ears. Nevertheless, one might take these studies as examples of selection of objects of attention, regardless of whether the objects are segregated by speech pitch and timbre or by location.

Ding and Simon (2012a) analyzed patterns of far-field neural activity recorded with MEG. Stimuli were 1-minute narratives uttered by two talkers mixed into a single audio signal. In successive blocks the listeners were instructed to attend to one or the other talker. The envelope of the attended talker was reconstructed from the neural activity by optimal integration across time and across MEG sensors. The reconstruction of attended speech generally correlated more closely with the envelope of the attended speech than with the envelope of the competing narrative. The unattended speech envelope also could be reconstructed, by optimization of different MEG sensors, although those correlations were lower than for the reconstruction of attended speech. That both attended and unattended signals could be reconstructed suggests that both signals are represented by neural populations, with the superior reconstruction of the attended compared to the unattended envelope reflecting relative facilitation of the representation of the attended signal. The STRFs reconstructed from MEG activity demonstrated that the modulation of neural responses by attention was largely limited to a magnetic component having approximately 100-ms latency and localized on the planum temporale. There was no significant attention effect on an earlier component having 50-ms latency and localized to Heschl's gyrus, presumably a primary auditory area.

Ding and Simon also tested a condition in which two speech narratives from the same talker were presented dichotically to the two ears (Ding and Simon 2012b). This was an MEG counterpart of the classic study by Cherry (1953), in which listeners could attend to the narrative at one or the other ear. Again, the envelope of the attended speech signal could be reconstructed from the MEG recordings. In this particular case, the selection of the object of attention was given by the ear of entry, which can be regarded as a coarse spatial cue, rather than by the identity of the talker. The results from these MEG studies indicate that selection of an auditory object of attention arises at a cortical level beyond the primary area, and that selection can be based on spectrotemporal cues (Ding and Simon 2012a) or on an approximation of a spatial cue (Ding and Simon 2012b).

A subsequent study by that group showed that the envelope of an attended speech signal presented in a background of speech-spectrum noise could be reconstructed from MEG recordings, but that both cortical synchrony and perceptual intelligibility were lost when the temporal fine structure of the speech was degraded with a four- or eight-channel vocoder (Ding et al. 2013). That result indicates that cortical neural populations in the planum temporale synchronize to attended auditory objects, defined in this case by the fine spectral and temporal characteristics needed to recognized speech sounds and likely analyzed at sub-cortical levels, rather than simply to the low-resolution temporal envelope of the stimulus.

Mesgarani and Chang (2012) recorded from arrays of cortical electrodes on the surface of the posterior superior temporal lobe in patients who were being evaluated for epilepsy surgery. Speech reconstruction filters were derived during passive listening conditions from responses to a corpus of sentences distinct from those used as test stimuli. Then, those filters were used to estimate spectrograms based on test utterances of single talkers and of mixtures of two talkers. When tested with the

mixed stimuli, the estimated spectrograms captured spectral and temporal features that correlated well with the spectrograms of the utterance of the attended talker, with substantially lower correlation with the unattended spectrogram. The listeners were asked to report words spoken by one of the two talkers. Reconstructions of the attended spectrogram were successful on trials in which the listeners answered correctly, and the reconstructions were degraded on trials in which the reports were incorrect. That result suggests a trial-by-trial correspondence between the patterns of cortical activity and task performance. Activity at single recording sites showed tuning for particular spectral features, but responses to those features were greater when contained in the speech stream from the attended talker than when the same features were present in the stream from the unattended talker. That observation supports the notion that neural responses in this nonprimary auditory cortex represent attended auditory objects rather than just particular acoustic features.

6.6 Summary, Synthesis, and Future Directions

It is clear from the available perceptual and physiological data that the locations of targets and competing sounds are key factors in parsing the auditory scene. Spatial separation of sources turns out to have remarkably little effect on the perception of multiple sound components that are otherwise bound by common onset time, fundamental frequency, and even visual cues, as shown in tests of obligatory streaming, of concurrent vowels, and of the ventriloquism effect (Stein and Meredith 1993). Spatial separations of signals and maskers, however, clearly are potent cues for voluntary stream segregation and object selection. Studies of primary auditory cortex in anesthetized cats demonstrate that distinct neural populations synchronize to sound sequences that presumably would be segregated perceptually on the basis of differences in target and interferer locations. That segregation as distinct neural populations is analogous to segregation that has been demonstrated previously on the basis of spectral differences. Studies in behaving ferrets and cats show that stimulus selectivity is modulated during task performance to enhance detection and discrimination of single targets. Neurophysiological studies in humans demonstrate enhanced synchrony of neurons in nonprimary cortical areas to attended speech streams.

The results lead to a general hypothetical organization for the neural substrates for spatial stream segregation and object selection. Spatial stream segregation appears to be a largely bottom-up phenomenon beginning with basic brainstem analysis of spatial cues, including interaural time and level differences and spectral shape. Forward suppression at the level of the thalamocortical projection leads to the first appearance of spatial stream segregation on temporal and spatial scales similar to those in perception. Task-dependent sharpening of spatial tuning for single sources could contribute to sharpening of segregation of sounds from multiple sources. Distinct neural populations in primary and possibly higher-order auditory cortical areas appear to represent both attended and competing sounds.

The evidence in human neurophysiology for enhanced cortical synchrony to attended sounds suggests that a top-down executive mechanism in some way facilitates activity of neural populations that represent attended sounds and/or suppresses populations that represent competitors. That is, selection of objects of attention could correspond to selection among simultaneously active neural populations. This hypothetical neural substrate for object selection invites future studies designed to confirm or reject the notion of selection among distinct neural populations in low-level auditory cortical areas, to explore how neural populations distinguished by spatial sensitivity might integrate spectral and other cues for segregation, and to identify sources of the executive signal(s) that could accomplish such selection.

Acknowledgements I thank Georg Klump, Lauren Javier, and Justin Yao for their helpful suggestions on the manuscript. This chapter was completed while the author was a resident fellow at the Hanse-Wissenschaftskolleg in Delmenhorst, Germany. The author's work is supported by the National Institutes of Health grant R01 DC000420.

Compliance with Ethics Requirements

John Middlebrooks has no conflicts of interest.

References

- Atiani, S., Elhilali, M., David, S. V., Fritz, J. B., & Shamma, S. A. (2009). Task difficulty and performance induce diverse adaptive patterns in gain and shape of primary auditory cortical receptive fields. *Neuron*, *61*, 467–480.
- Bizley, J. K., Walker, K. M., Silverman, B. W., King, A. J., & Schnupp, J. W. (2009). Interdependent encoding of pitch, timbre, and spatial location in auditory cortex. *The Journal of Neuroscience*, *29*, 2064–2075.
- Boehnke, S. E., & Phillips, D. P. (2005). The relation between auditory temporal interval processing and sequential stream segregation examined with stimulus laterality differences. *Perception and Psychophysics*, *67*, 1088–1101.
- Bregman, A. S. (1990). *Auditory scene analysis: The perceptual organization of sound*. Cambridge, MA: MIT Press.
- Briley, P. M., Kitterick, P. T., & Summerfield, A. Q. (2013). Evidence for opponent process analysis of sound source location in humans. *Journal of the Association for Research in Otolaryngology*, *14*, 973–983.
- Broadbent, D. E., & Ladefoged, P. (1957). On the fusion of sounds reaching different sense organs. *The Journal of the Acoustical Society of America*, *29*, 708–710.
- Brosch, M., & Schreiner, C. E. (1997). Time course of forward masking tuning curves in cat primary auditory cortex. *Journal of Neurophysiology*, *77*, 923–943.
- Calford, M. B., & Semple, M. N. (1995). Monaural inhibition in cat auditory cortex. *Journal of Neurophysiology*, *73*, 1876–1891.
- Carl, D., & Gutschalk, A. (2012). Role of pattern, regularity, and silent intervals in auditory stream segregation based on inter-aural time differences. *Experimental Brain Research*, *224*, 557–570.
- 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.
- Creutzfeldt, O. D., Hellweg, F. C., & Schreiner, C. (1980). Thalamocortical transformations of responses to complex auditory stimuli. *Experimental Brain Research*, *39*, 87–104.

- Cutting, J. E. (1976). Auditory and linguistic processes in speech perception: Inferences from six fusions in dichotic listening. *Psychological Review*, 2, 114–140.
- Ding, N., Chatterjee, M., & Simon, J. Z. (2013). Robust cortical entrainment to the speech envelope relies on the spectro-temporal fine structure. *NeuroImage*, 88, 41–46.
- Ding, N., & Simon, J. (2012a). Emergence of neural encoding of auditory objects while listening to competing speakers. *Proceedings of the National Academy of Sciences of the USA*, 109, 11854–11859.
- Ding, N., & Simon, J. (2012b). Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. *Journal of Neurophysiology*, 107, 78–89.
- Dingle, R. N., Hall, S. E., & Phillips, D. P. (2010). A midline azimuthal channel in human spatial hearing. *Hearing Research*, 268, 67–74.
- Duffour-Nikolov, C., Tardif, E., Maeder, P., Thiran, A. B., et al. (2012). Auditory spatial deficits following hemispheric lesions: Dissociation of explicit and implicit processing. *Neuropsychological Rehabilitation*, 22, 674–696.
- Edmonds, B. A., & Culling, J. F. (2005a). The role of head-related time and level cues in the unmasking of speech in noise and competing speech. *Acta Acustica united with Acustica*, 91, 546–553.
- Edmonds, B. A., & Culling, J. F. (2005b). The spatial unmasking of speech: Evidence for within-channel processing of interaural time delay. *The Journal of the Acoustical Society of America*, 117, 3069–3078.
- Elhilali, M., Ma, L., Micheyl, C., Oxenham, A., & Shamma, S. (2009). Temporal coherence in the perceptual organization and cortical representation of auditory scenes. *Neuron*, 61, 317–329.
- Fishman, Y., Reser, D., Arezzo, J., & Steinschneider, M. (2001). Neural correlates of auditory stream segregation in primary auditory cortex of the awake monkey. *Hearing Research*, 151, 167–187.
- Fritz, J. B., Elhilali, M., & Shamma, S. A. (2007). Adaptive changes in cortical receptive fields induced by attention to complex sounds. *Journal of Neurophysiology*, 98, 2337–2346.
- Fritz, J. B., Shamma, S. A., Elhilali, M., & Klein, D. J. (2003). Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex. *Nature Neuroscience*, 6, 1216–1223.
- Füllgrabe, C., & Moore, B. C. J. (2012). Objective and subjective measures of pure-tone stream segregation based on interaural time differences. *Hearing Research*, 291, 24–33.
- Furukawa, S., Xu, L., & Middlebrooks, J. C. (2000). Coding of sound-source location by ensembles of cortical neurons. *The Journal of Neuroscience*, 20, 1216–1228.
- Griffiths, T. D., & Warren, J. D. (2004). What is an auditory object? *Nature Review of Neuroscience*, 5, 887–892.
- Harrington, I. A., Stecker, G. C., Macpherson, E. A., & Middlebrooks, J. C. (2008). Spatial sensitivity of neurons in the anterior, posterior, and primary fields of cat auditory cortex. *Hearing Research*, 240, 22–41.
- Hartmann, W. M., & Johnson, D. (1991). Stream segregation and peripheral channeling. *Music Perception*, 9, 155–184.
- Hartmann, W. M., & Rakerd, B. (1993). Auditory spectral discrimination and the localization of clicks in the sagittal plane. *The Journal of the Acoustical Society of America*, 94, 2083–2092.
- Hofman, P. M., & Van Opstal, J. A. (1998). Spectro-temporal factors in two-dimensional human sound localization. *The Journal of the Acoustical Society of America*, 103, 2634–2648.
- Hukin, R. W., & Darwin, C. J. (1995). Effects of contralateral presentation and of interaural time differences in segregating a harmonic from a vowel. *The Journal of the Acoustical Society of America*, 98, 1380–1387.
- Ihlefeld, A., & Shinn-Cunningham, B. (2008a). Spatial release from energetic and informational masking in a selective speech identification task. *The Journal of the Acoustical Society of America*, 123, 4369–4379.
- Ihlefeld, A., & Shinn-Cunningham, B. (2008b). Disentangling the effects of spatial cues on selection and formation of auditory objects. *The Journal of the Acoustical Society of America*, 124, 2224–2235.

- Kidd, G., Jr., Best, V., & Mason, C. R. (2008). Listening to every other word: Examining the strength of linkage variables in forming streams of speech. *The Journal of the Acoustical Society of America*, *124*, 3793–3802.
- King, A. J., & Middlebrooks, J. C. (2011). Cortical representation of auditory space. In J. Winer & C. Schreiner (Eds.), *The auditory cortex* (pp. 329–341). New York: Springer Science+Business Media.
- Kuhn, G. F. (1977). Model for the interaural time differences in the azimuthal plane. *The Journal of the Acoustical Society of America*, *62*, 157–167.
- Lee, C.-C., & Middlebrooks, J. (2011). Auditory cortex spatial sensitivity sharpens during task performance. *Nature Neuroscience*, *14*, 108–114.
- Lee, C.-C., & Middlebrooks, J. (2013). Specialization for sound localization in fields A1, DZ, and PAF of cat auditory cortex. *Journal of the Association for Research in Otolaryngology*, *14*, 61–82.
- Lomber, S., & Malhotra, S. (2008). Double dissociation of ‘what’ and ‘where’ processing in auditory cortex. *Nature Neuroscience*, *11*, 609–616.
- Macpherson, E. A., & Middlebrooks, J. C. (2000). Localization of brief sounds: Effects of level and background noise. *The Journal of the Acoustical Society of America*, *108*, 1834–1849.
- Macpherson, E. A., & Middlebrooks, J. C. (2002). Listener weighting of cues for lateral angle: The duplex theory of sound localization revisited. *The Journal of the Acoustical Society of America*, *111*, 2219–2236.
- Magezi, D. A., & Krumbholz, K. (2010). Evidence of opponent-channel coding of interaural time differences in human auditory cortex. *Journal of Neurophysiology*, *104*, 1997–2007.
- May, B. J., & Huang, A. Y. (1996). Sound orientation behavior in cats. I. Localization of broadband noise. *The Journal of the Acoustical Society of America*, *100*, 1059–1069.
- Mesgarani, N., & Chang, E. F. (2012). Selective cortical representation of attended speaker in multi-talker speech perception. *Nature*, *485*, 233–236.
- Micheyl, C., & Oxenham, A. J. (2010). Objective and subjective psychophysical measures of auditory stream integration and segregation. *Journal of the Association for Research in Otolaryngology*, *11*, 709–724.
- Micheyl, C., Tian, B., Carlyon, R. P., & Rauschecker, J. P. (2005). Perceptual organization of tone sequences in the auditory cortex of awake macaques. *Neuron*, *48*, 139–148.
- Mickey, B. J., & Middlebrooks, J. C. (2003). Representation of auditory space by cortical neurons in awake cats. *The Journal of Neuroscience*, *23*, 8649–8663.
- Middlebrooks, J. C., & Bremen, P. (2013). Spatial stream segregation by auditory cortical neurons. *The Journal of Neuroscience*, *33*, 10986–11001.
- Middlebrooks, J. C., Clock, A. E., Xu, L., & Green, D. M. (1994). A panoramic code for sound location by cortical neurons. *Science*, *264*, 842–844.
- Middlebrooks, J. C., & Green, D. M. (1990). Directional dependence of interaural envelope delays. *The Journal of the Acoustical Society of America*, *87*, 2149–2162.
- Middlebrooks, J. C., & Green, D. M. (1991). Sound localization by human listeners. *Annual Review of Psychology*, *42*, 135–159.
- Middlebrooks, J. C., & Onsan, Z. A. (2012). Stream segregation with high spatial acuity. *The Journal of the Acoustical Society of America*, *132*, 3896–3911.
- Miller, L. M., & Recanzone, G. H. (2009). Populations of auditory cortical neurons can accurately encode acoustic space across stimulus intensity. *Proceedings of the National Academy of Sciences of the USA*, *106*, 5931–5935.
- Moore, B. C. J., & Gockel, H. (2002). Factors influencing sequential stream segregation. *Acta Acustica*, *88*, 320–332.
- Moore, B. C. J., & Gockel, H. (2012). Properties of auditory stream formation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*, 919–931.
- Phillips, D. P. (2008). A perceptual architecture for sound lateralization in man. *Hearing Research*, *238*, 124–132.
- Pressnitzer, D., Sayles, M., Micheyl, C., & Winter, I. (2008). Perceptual organization of sound begins in the auditory periphery. *Current Biology*, *18*, 1124–1128.

- Sach, A. J., & Bailey, P. J. (2004). Some characteristics of auditory spatial attention revealed using rhythmic masking release. *Perception and Psychophysics*, *66*, 1379–1387.
- Salminen, N. H., May, P. J., Alku, P., & Tiitinen, H. (2009). A population rate code of auditory space in the human cortex. *PLoS ONE*, *26*, e7600.
- Saue, K., Keoelsch, S., & Rubsamen, R. (2010). Spatial selective attention in a complex auditory environment such as polyphonic music. *The Journal of the Acoustical Society of America*, *127*, 472–480.
- Schadwinkel, S., & Gutschalk, A. (2010). Activity associated with stream segregation in human auditory cortex is similar for spatial and pitch cues. *Cerebral Cortex*, *20*, 2863–2873.
- Schreiner, C., & Urbas, J. (1988). Representation of amplitude modulation in the auditory cortex of the cat. II. Comparison between cortical fields. *Hearing Research*, *32*, 49–63.
- Shaw, E. A. G. (1974). Transformation of sound pressure level from the free field to the eardrum in the horizontal plane. *The Journal of the Acoustical Society of America*, *56*, 1848–1861.
- Snyder, J., & Alain, C. (2007). Toward a neurophysiological theory of auditory stream segregation. *Psychological Bulletin*, *133*, 780–799.
- Stainsby, T. H., Fullgrabe, C., Flanagan, H. J., Waldman, S. K., & Moore, B. C. J. (2011). Sequential streaming due to manipulation of interaural time differences. *The Journal of the Acoustical Society of America*, *130*, 904–914.
- Stecker, G. C., Harrington, I. A., & Middlebrooks, J. C. (2005). Location coding by opponent neural populations in the auditory cortex. *PLoS Biology*, *3*, 520–528.
- Stein, B. E., & Meredith, M. A. (1993). *The merging of the senses.*, Cognitive Neuroscience Series Cambridge, MA: MIT Press.
- Takanen, M., Raitio, T., Santala, O., Alku, P., & Pulkki, V. (2013). Fusion of spatially separated vowel formant cues. *The Journal of the Acoustical Society of America*, *134*, 4508–4517.
- Thiran, A. B., & Clarke, S. (2003). Preserved use of spatial cues for sound segregation in a case of spatial deafness. *Neuropsychologia*, *41*, 1254–1261.
- Tollin, D. J., Populin, L. C., Moore, J. M., Ruhland, J. L., & Yin, T. C. (2005). Sound-localization performance in the cat: The effect of restraining the head. *Journal of Neurophysiology*, *93*, 1223–1234.
- van Noorden, L. P. A. S. (1975). *Temporal coherence in the perception of tone sequences.* PhD dissertation, Eindhoven: University of Technology.
- Vliegen, J., Moore, B. C., & Oxenham, A. J. (1999). The role of spectral and periodicity cues in auditory stream segregation, measured using a temporal discrimination task. *The Journal of the Acoustical Society of America*, *106*, 938–945.
- Wightman, F. L., & Kistler, D. J. (1992). The dominant role of low-frequency interaural time differences in sound localization. *The Journal of the Acoustical Society of America*, *91*, 1648–1661.
- 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*, 3323–3337.
- Woods, W. S., & Colburn, H. S. (1992). Test of a model of auditory object formation using intensity and interaural time difference discrimination. *The Journal of the Acoustical Society of America*, *91*, 2894–2902.
- Yin, P., Fritz, J. B., & Shamma, S. A. (2014). Rapid spectrotemporal plasticity in primary auditory cortex during behavior. *The Journal of Neuroscience*, *34*, 4396–4408.

Chapter 7

Human Auditory Neuroscience and the Cocktail Party Problem

Jonathan Z. Simon

Abstract Experimental neuroscience using human subjects, to investigate how the auditory system solves the cocktail party problem, is a young and active field. The use of traditional neurophysiological methods is very tightly constrained in human subjects, but whole-brain monitoring techniques are considerably more advanced for humans than for animals. These latter methods in particular allow routine recording of neural activity from humans while they perform complex auditory tasks that would be very difficult for animals to learn. The findings reviewed in this chapter cover investigations obtained with a variety of experimental methodologies, including electroencephalography, magnetoencephalography, electrocorticography, and functional magnetic resonance imaging. Topics covered in detail include investigations in humans of the neural basis of spatial hearing, auditory stream segregation of simple sounds, auditory stream segregation of speech, and the neural role of attention. A key conceptual advance noted is a change of interpretational focus from the specific notion of attention-based neural gain, to the general role played by attention in neural auditory scene analysis and sound segregation. Similarly, investigations have gradually changed their emphasis from explanations of how auditory representations remain faithful to the acoustics of the stimulus, to how neural processing transforms them into new representations corresponding to the percept of an auditory scene. An additional important methodological advance has been the successful transfer of linear systems theory analysis techniques commonly used in single-unit recordings to whole-brain noninvasive recordings.

Keywords Attentional gain · Auditory scene analysis · Binaural integration · Electrocorticography · Electroencephalography · Functional magnetic resonance imaging · Heschl's gyrus · Human auditory system · Interaural level difference · Interaural time difference · Magnetoencephalography · Maskers · Planum temporale · Positron emission tomography · Selective attention · Speech · Superior temporal gyrus

J.Z. Simon (✉)

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

7.1 Introduction

The search for how the brain solves the “cocktail party problem” (Cherry 1953), or auditory scene analysis (Bregman 1990) in general, is often led by human psychophysical studies, as demonstrated by so much of the content in this volume. There has also been substantial progress in the underlying neuroscience, primarily in animals (Middlebrooks, Chap. 6). Investigation of the underlying neuroscience in humans, however, is still a relatively young field. Though human studies lag behind animal studies in many ways, it is a dynamic and vibrant area of neuroscience.

Progress in the area is dominated by a tension between two aspects of the problem. The use of invasive neurophysiological methods is very tightly constrained in human subjects because of obvious ethical concerns, and is almost entirely limited to patient populations suffering from debilitating illnesses. For this reason, studies that would be routine in an animal model are typically impossible in human subjects. In this sense, human neurophysiological neuroscience lags behind animal studies, and reasonably so.

At the same time, whole brain monitoring techniques are considerably more advanced for human subjects than for animal subjects. These methods allow the recording of neural responses from humans performing complex auditory tasks that would be very difficult, if not impossible, for animals to learn. This allows a subset of human neurophysiological and neuroimaging studies to actually outpace those possible in animals. Thus, it is often true both that methodological limitations dominate the field’s results, and yet at the same time, the field is also bristling with new and exciting finds.

7.1.1 *Common Experimental Methodologies*

The functional neurophysiological and neuroimaging methods used in human neuroscience fall into two broad categories, defined as much by their time scales as by the biology and physics that drives them. One category includes methods that directly measure the electromagnetic output of neurons. When used noninvasively, only the electromagnetic output summed over entire regions (typically on the centimeter scale) can be measured, which limits detailed analysis of the underlying neural sources. This category includes the techniques of electroencephalography (EEG) and magnetoencephalography (MEG), which directly measure neurally generated electric potentials and magnetic fields, respectively, outside the scalp. Fortunately, the *temporal* dynamics of extracranial electromagnetic fields are not obstructed by their passage through the brain, skull, and scalp, and so the neural time scales available to these techniques are relatively fast, from an upper limit in the hundreds of Hertz down to a few tenths of a Hertz.

Still in the same category, invasive electrophysiological techniques are allowable in patients whose clinical treatments already require similarly invasive methods.

These include the use of intracranial electrodes and, more commonly, electrocorticography (ECoG), the use of subdural surface electrodes (also known as intracranial EEG [iEEG]). In theory, these techniques allow electrophysiological measurements with the same fine-grained spatial and temporal resolution of similar experiments in animal models. In practice, though this theoretical limit can indeed be reached in a small number of subjects, the methods also have practical limitations that constrain their use. These constraints include restrictions on which brain areas can be recorded from (required to be consistent with the patients' clinical needs), being restricted to a small subject pool of patients in need of extraordinary treatments, and that the patients, by their very availability, suffer from neurological and often resultant cognitive problems.

The other broad category of functional methods typically used in experimental human neuroscience is that of hemodynamic measurements. These generally non-invasive methods do not measure neural responses directly, but indirectly through changes to blood flow (and/or blood content) that occur after metabolically intensive neural activity. Owing to the inherently slow hemodynamic changes that follow from neural activity, these methods are generally limited to time scales of a Hertz or slower. Nevertheless, because the spatial resolution of these techniques is generally superior to noninvasive electromagnetic techniques, they are used heavily. These methods include functional magnetic resonance imaging (fMRI), positron emission tomography (PET), single-photon emission computed tomography (SPECT), and functional near-infrared spectroscopy (fNIRS).

Investigations of the neural foundations of how the human brain solves the cocktail party problem, or of auditory scene analysis in general, date back to at least to the EEG work of Hillyard et al. (1973). In that foundational study, subjects listened to a simple auditory scene consisting of a tone pip stream presented to each ear (dichotically), but during which the subjects performed a difficult task that required they attend only to the tones in a single ear. The EEG responses, averaged over the responses to the individual tone pips, showed that responses to the attended auditory scene component (i.e., the task-relevant tone pips) were dramatically different from the responses to the unattended auditory scene component (i.e., the task-irrelevant tone pips). The difference manifested as a substantial enhancement of the (approximately) 100 ms latency negative response (N1), to the attended scene component over the unattended. Historically this enhancement has been depicted as the neural correlate of selective attention, but it might just as easily have been described as the neural correlate of successful auditory scene segregation.

7.1.2 Chapter Topics

The chapter is organized according to the levels of the auditory system, or level of auditory cognition, at which particular neural computations are thought to take place, irrespective of the response-recording modality (e.g., by fMRI vs. EEG). Additionally, an effort has been made to emphasize the specific topics most directly

connected to the other chapters of this volume. Consequently, not all areas of the human neuroscience of auditory scene analysis (which is quite broad) are covered here. The mismatch negativity (MMN) paradigm, for example, which exploits the change in an EEG or MEG response to a discrete stimulus when it is detected as a discriminable change in a stream of other discrete auditory stimuli, is not covered. The MMN is certainly used in investigations of auditory scene analysis (Sussman et al. 2014), and the omission is due to the sheer size of the MMN literature, which is well reviewed elsewhere (see, e.g., Naatanen et al. 2007). Other such blatant omissions have been made for reasons of both scope and space, and readers are directed to other reviews with different, but overlapping, scope, including of auditory scene analysis (Snyder et al. 2012; Gutschalk and Dykstra 2014), auditory selective attention (Lee et al. 2014), masked speech (Scott and McGettigan 2013), and sound localization (Ahveninen et al. 2014).

7.2 Neural Basis of Spatial Hearing in Humans

One of the key factors that can aid in parsing an auditory scene is spatial separation between an auditory target and the nontarget maskers. The principal acoustic cues for segregating sources in space include the differences in the time of arrival of sounds at the two ears (interaural time difference [ITD]) and differences in the sound levels (interaural level difference [ILD]). The human neural processing of these low-level binaural cues critical to sound localization is one area where some progress has been made. The human neuroscience of more general spatial hearing is reviewed by Ahveninen et al. (2014).

As in other mammals, the first auditory nuclei in the human brain to receive bilateral auditory input are in the superior olivary complex of the brainstem. The medial superior olive (MSO) and lateral superior olive (LSO) both play a prominent role in early neural computations of sound localization in mammals, computing and encoding the ITD and ILD produced from the spatial location of a sound's origin relative to the head and ears. It is typically assumed that the functional roles of MSO and LSO remain the same in humans, but little is known on this point (Kulesza 2007) and there are few studies investigating sound localization computations in the human superior olivary complex.

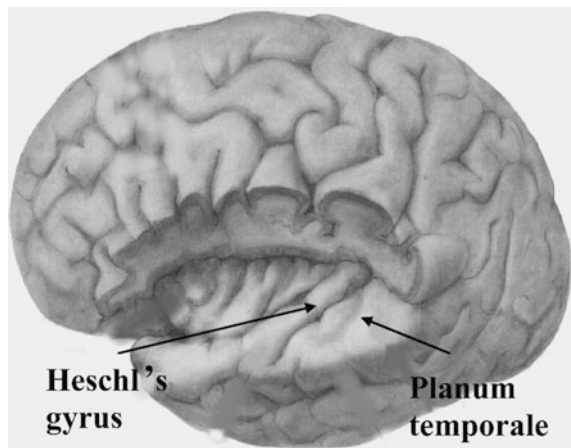
Proceeding along, the inferior colliculus (IC) of the midbrain is the first stage at which direct evidence of sound localization-related binaural processing in the human brain has been obtained. Thompson et al. (2006) used fMRI to demonstrate sensitivity to ITD in both left and right human IC. They presented low-frequency band-passed noise with both natural ITDs (0 or $\pm 500 \mu\text{s}$) and unnatural ITDs, $\pm 1500 \mu\text{s}$ (far longer than the approximately 700 ms maximum encountered naturally). The natural ITD condition produced more activity in the contralateral IC than the ipsilateral (with respect to the perceived lateralization), as predicted by many models. Additionally, less activity was found for unnatural ITD than for natural, as also predicted by many models. Perhaps counterintuitively, the unnatural

ITD produced more activity in the ipsilateral IC than the contralateral. This result is consistent only with a narrower range of models, those that compute ITD via a weighted cross correlogram (Thompson et al. 2006).

Von Kriegstein et al. (2008) analyzed the same dataset for evidence of cortical ITD processing. Concordant with the IC results, but in the higher-order auditory cortical region planum temporale (Fig. 7.1), the natural ITD condition again produced more contralateral activity than ipsilateral. In contrast to the IC results, however, there was little differentiation between contralateral and ipsilateral auditory activity for unnatural ITD in any auditory cortical area. Because perceptually the unnatural ITD is still lateralized, this result disconnects neural lateralization from perceptual lateralization. It also was found that activity in Heschl's gyrus (Fig. 7.1), the location of core auditory cortex (Kaas and Hackett 2000), was greater for unnatural ITD than natural.

The distribution of ITD representations in auditory cortex, which presumably inherits much of the spatial localization processing from subcortical levels, has been separately investigated by two groups using MEG (Salminen et al. 2010) and EEG (Magezi and Krumbholz 2010). The question addressed by both groups was whether the spatial distribution of ITD representations is better described by a fine-grained topographic array of ITD-tuned neurons, as would be employed in a Jeffress-like place-code model (Jeffress 1948), or by an opponent-channel model in which the distribution is tuned broadly to either the left or right hemifield (McAlpine 2005; Stecker et al. 2005). The two groups used distinct but related stimulus paradigms, both presenting an initial “adaptor” sound at one ITD and following it up with the “probe” sound at the ITD of interest. The results from both groups are consistent with an opponent-channel model but not a topographic model: both groups found the response to an outward ITD change (from zero ITD) to be greater than the response to the inward change (from strongly lateralized ITD), which is consistent only with the opponent-channel model. The fine-grained topographic model, which requires a greater density of neurons with best ITDs near

Fig. 7.1 A drawing of the brain (with a section of the upper posterior temporal lobe and parts of the frontoparietal area removed), revealing the left superior temporal gyrus, including the auditory core, lateral belt, and parabelt regions. Especially prominent are Heschl's gyrus and the triangularly shaped area just behind it, planum temporale. [From Hugdahl (2005), Fig. 3, with permission.]



zero, would instead predict the opposite. These results have been generalized using a broader range of similar stimuli (Briley et al. 2013), and more evidence for an opponent-channel model has also been obtained using MEG using slow binaural beats, an unrelated stimulus paradigm (Ross et al. 2014). These results, as integrated as they are, however, do not necessarily agree with the results of related psychophysical experiments (Middlebrooks, Chap. 6). For instance, there is growing behavioral evidence in favor of a three-channel model comprising left, right, and midline populations (Dingle et al. 2010, 2012) and even some EEG evidence (Briley et al. 2016).

The ability of the brain to compute ITDs, at all, depends on the ability of the ear and auditory nerve to accurately capture and convey the relative timing of signals between the ears. Chait et al. (2006), using MEG, measured this by recording cortical responses to binaurally generated Huggins pitch (Cramer and Huggins 1958). Huggins pitch is created by presenting almost identical white noise to both ears, where the only difference is that in one narrow frequency band the interaural phase is opposite instead of identical. The accompanying percept is of a continuous tone, with the same frequency as the center of the phase-inverted frequency band, in a noise background. Whenever this percept occurs, that alone is sufficient to prove that the binaural phase difference has been successfully integrated across the ears, as the monaural stimulus at each ear alone is ordinary white noise. In this study, robust onset responses were detected for Huggins pitch onset at all phase-inverse frequency bands (i.e., perceived pitches) employed, from 200 to 1000 Hz. Because MEG is insensitive to subcortical neural sources, the MEG responses observed arose from cortical areas that had already inherited the substantial binaural processing computed subcortically. Nevertheless they are a genuine neural measure of binaural integration occurring in the auditory system.

Also using MEG, Ross et al. (2007a) were able to measure the frequency range over which binaural neural integration occurs in greater detail. They employed binaural amplitude-modulated tones, with carrier frequencies ranging from 500 to 1500 Hz, for which, at specific moments of zero instantaneous amplitude, they reversed the interaural carrier phase. The sudden change of interaural carrier phase was found to evoke a cortical response, thus demonstrating successful subcortical binaural integration, but only for the frequencies at 1000 Hz and below. In contrast, none of the subjects evoked such a cortical response when the change of phase was at the higher carrier frequency of 1500 Hz. The authors estimated an upper frequency limit of 1250 Hz for successful binaural integration—consistent with their behavioral finding of a limit at 1200 Hz. When extended to a wider age range of subjects in related investigation, Ross et al. (2007b) also demonstrated that this threshold of binaural integration decreased progressively with age, down to 940 Hz for middle-aged adults and 760 Hz for older adults.

Measurements of direct ITD and ILD sensitivity in the human brain, analogous to those performed routinely in invasive animal studies, have been more difficult. McLaughlin et al. (2016), using fMRI, observed direct measurements of ILD response tuning in auditory cortex contralateral to the hemifield of positive ILD. The tuning is observed along the medial-to-lateral extent of Heschl's gyrus (i.e.,

core auditory cortex), and posterior sections of the superior temporal gyrus including planum temporale (i.e., higher order association auditory cortex). These results hold in both hemispheres, with greater effect size in the left hemisphere. The response-strength dependence on subject task engagement varies across cortical areas, with minimal effect in Heschl's gyrus but with significantly increasing activation in posterior superior temporal gyrus, at least in the right hemisphere.

In contrast to ILD, direct measurement in human cortex of ITD, often considered a much stronger behavioral cue than ILD for azimuthal localization in humans, shows only weak sensitivity. McLaughlin et al. (2016) found a small but significant dependence on ITD in posterior superior temporal gyrus (but not in Heschl's gyrus), and only in the left hemisphere (when contralateral to the indicated sound source). The response was also modulated by subjects' task engagement. It is possible that the difficulty seen in observing these direct dependencies follows the similar results observed from single-unit recordings comparing spatial tuning both with and without competing spatial sounds (Maddox et al. 2012; Middlebrooks and Bremen 2013). There the spatial tuning of individual neurons to a single sound is quite broad but narrows dramatically when in the presence of competing sounds from other spatial locations. The ITD results of McLaughlin and colleagues lend evidence to a model by Magezi and Krumbholz (2010), in which right hemisphere auditory cortex encodes both contralateral and ipsilateral ITD information, but left hemisphere auditory cortex encodes only contralateral ITD information.

In short, investigations of the neural processing of low-level binaural cues in humans still lag behind analogous studies in nonhuman animals. Nevertheless, substantial progress continues to be made. Some of it is in agreement with the relevant animal-based studies and human psychophysical literature, but far from all.

7.3 Neural Basis of Auditory Stream Segregation in Humans: Simple Sounds

This section and the next both cover investigations of the neural correlates of auditory stream segregation, using different classes of stimuli: simple sounds (e.g., tone pips) in this section, and speech in Sect. 7.4. Unlike in Sect. 7.2, however, the emphasis shifts to auditory scene analysis in the *absence* of informative spatial cues. Spatial separation of auditory scene elements, especially combined with binaural hearing, can greatly benefit the listener's ability to segregate those elements, but is not required (Brungart et al. 2001; Hawley et al. 2004). By employing stimuli that lack such spatial cues, it may be possible to uncover more general functional (and not specifically anatomically based) neural mechanisms that underlie the neural computations of auditory stream segregation. Of course such segregation requires *some* cue to differentiate different elements from each other, because without any such cue the common elements simply fuse into a single percept. For this reason there cannot be any single experiment that would find a

truly general functional mechanism underlying auditory stream segregation, and a variety of approaches are needed. For human neuroscience studies that more explicitly rely on spatial cues to differentiate auditory scene elements, readers are directed to Chap. 2 by Shinn-Cunningham, Best, and Lee.

In many of the studies reviewed here, the neurophysiological measure employed is the magnitude of the response to a brief sound (e.g., a tone pip), averaged over trials and/or serial presentations within a trial. Separate magnitudes are measured for different response components (e.g., different latencies in the same set of responses), which are taken to have different neural origins. Commonly used neural measures are the response component with post-stimulus latency of approximately 50 ms, called the P1 in EEG and P1m (or M50) in MEG, and the response component with post-stimulus latency of approximately 100 ms, called the N1 in EEG and N1m (or M100) in MEG. A wide variety of other components have been investigated as well, most with even longer latencies. The earlier latency P1/P1m has a spatial origin consistent with Heschl's gyrus (Makela et al. 1994), and therefore with core auditory cortex. The later latency N1/N1m has a spatial origin consistent with planum temporale (Lutkenhoner and Steinstrater 1998), and therefore with higher order auditory cortex.

7.3.1 Studies Using Limited Attentional Manipulation

7.3.1.1 Simple Tone Patterns

Some of the earliest evidence for a neural correlate of auditory scene segregation was obtained by Alain et al. (2001), who measured evoked EEG responses to harmonic complex tones with, or without, a mistuned harmonic. The mistuning of a harmonic results in the percept of two distinct (segregated) sound elements, in contrast to the unified (fused) percept when all harmonics are properly tuned. By comparing the responses to the segregated and fused conditions, and, critically, in both attended and unattended conditions, the investigators reported an additional electrically negative response component they named object-related negativity. In this case the “objects” are the segregated elements arising from the mistuned harmonics. Alain and colleagues later showed that this result also generalized beyond the specific use of mistuned harmonics to induce segregation of auditory scene elements (Alain et al. 2005). One problem with this general experimental approach, however, is the use of a stimulus change to induce the perceptual change. This creates a confound as to whether a neural response change is due to the stimulus change or the perceptual change. This confound is addressable, but only indirectly, by comparing responses to the same stimuli in separate conditions of auditory attention and auditory inattention. The confound can be avoided entirely, however, by a change of paradigm, as will be seen later in this section.

The next several investigations described all employ a common stimulus paradigm: the ABA tone-pip triplets of van Noorden (1975). Depending on the

frequency separation of the A- and B-tone pips, and depending on the interstimulus intervals, the percept of these stimuli can be that of a galloping pattern (in which the A- and B-tones fuse into a single stream with rising and falling frequency changes), or of a pair of separate regularly repeating patterns at a single frequency (one of the A-tones and the other of the B-tones), or, over longer time frames, a bistable condition in which the listener may slowly drift back and forth between those two percepts.

Gutschalk et al. (2005) performed two related experiments with these stimuli, while recording the listeners' cortical activity with MEG. In the first experiment the frequency separation varied over a range that allowed both fused and separated percepts. As expected, behaviorally, the likelihood of the stimulus being perceived as two segregated streams increased as the frequency separation increased. Neurally, the P1m and N1m amplitudes similarly increased in magnitude, in such a way that the P1m and N1m magnitude strongly correlated with the probability of stream segregation. This is consistent with the idea that this neural response arises from the same mechanism that allows the two streams to be separated perceptually, but still suffers from the confound that both the stimulus acoustics and the percept are changing simultaneously. The second experiment employed only the narrow frequency-separation regime. The stimulus parameters were kept constant, but the percept could take either mode, fused or segregated. In any given trial, the listener would report fusion or segregation, and the neural responses from each set were analyzed separately. Critically, it was seen that the P1m and N1m magnitudes covaried with the percept though the stimuli remained unchanged, in the same way as the first experiment: the neural response underlying the perceptually segregated tones was larger than that underlying the perceptually fused tones. In this way, the neural measure followed the (perceived) auditory scene analysis and not just the physical acoustics.

Snyder et al. (2006), using a similar stimulus paradigm but with EEG, found consistent and complementary results. Their ABA stimuli employed frequency separations that varied over the same range as did those of Gutschalk and colleagues and found similar behavioral results when their subjects were instructed to attend to the stimuli. In a second experiment, however, the subjects' attention was directed away from the acoustic stimuli by having them watch a subtitled silent movie. Their analysis of neural response magnitude was similar to that of Gutschalk and colleagues, but included an additional exploration of the response magnitude as a function of time within each trial (approximately 10 s), intended to analyze the response time course as the percept slowly developed over the many seconds of the entire trial ("buildup"). The neural buildup, which manifested as an increase over time of the neural response magnitude, occurred only in the auditory-attention condition, but not otherwise. Strongly differing spatial patterns of neural activation, indicative of different underlying neural sources, were also observed between the frequency-separation-dependent responses and the attention-dependent buildup responses. These complementary findings lend support to the idea of separate mechanisms for an attention-dependent buildup and a preattentive source segregation.

Hill et al. (2012), also using an ABA paradigm, recorded the EEG responses from listeners as their percept of the stimulus spontaneously switched back and forth from a single grouped stream to a pair of separate streams. Their analysis focused on the differences between stimulus-only related auditory processing and perception-only related processing. They found that stimulus-only related differences did indeed make a large and significant contribution to neural responses differences. These effects were relatively late in the EEG response pattern, between 200 and 300 ms latency, compared to what might be expected for acoustic differences that are well represented in the periphery. In contrast, the perception-only related processing differences were independent, and began early and remained late. The authors also make an important technical point in the analysis, that for this particular paradigm, the otherwise common EEG analysis step of baseline-correcting each epoch would have actually removed the basis of the across-condition analysis.

Using ECoG in epilepsy patients, and with a similar ABA stimulus paradigm, Dykstra et al. (2011) confirmed the aforementioned frequency separation-based behavioral results in these patients. They additionally found neural correlates of frequency separation in widespread brain areas (using electrode arrays that cover a broad range of cortical areas), most of which were far beyond auditory cortex, including frontal, parietal, and nonauditory temporal cortex. This observation of almost brain-wide auditory processing has also been found in other ECoG studies (see, e.g., the studies described in Sect. 7.4.1.2).

Investigations using fMRI with similar stimulus paradigms have unfortunately produced seemingly conflicting results. Cusack (2005) did not observe any effect in auditory cortical areas (only in the intraparietal sulcus) from varying the frequency separation within the ABA paradigm. In contrast, two other studies (Gutschalk et al. 2007; Wilson et al. 2007) did observe such an effect in auditory cortex, but only after modifying the stimulus paradigms to avoid habituation (i.e., using ABAB or ABBB tone pip patterns, rather than ABA).

7.3.1.2 Tones and Maskers

In contrast to the ABA (or related) stimulus, Gutschalk et al. (2008) used a more complex stimulus, based on one used originally in a specific investigation of informational masking (Kidd et al. 2003), but modified in a critical way. The stimulus (Fig. 7.2A), variants of which were also used in additional investigations discussed in Sect. 7.4, consists of several seconds of a regularly repeating rhythm of tone pips, added to a “cloud” of competing spectrally and temporally randomized, and so desynchronized, tone pips (the original Kidd stimulus employed a synchronized tone-pip cloud). The informational masking aspect of the stimuli derives from a use of a protected spectral zone, devoid of the competing tone pips, with enough bandwidth to ensure that the cloud cannot energetically mask the rhythmically regular tone stream. As masking of the rhythmic stream by the cloud does indeed occur perceptually despite the protection zone, the masking is necessarily informational rather than energetic (Culling and Stone, Chap. 3).

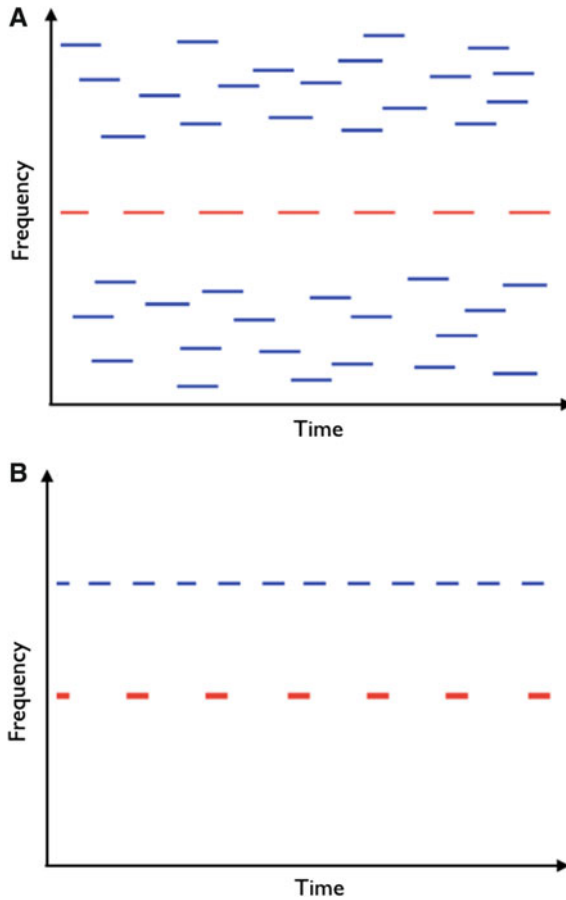


Fig. 7.2 Schematic illustrations of stimuli that consist of a pair of competing simple auditory streams. **(A)** In this case one auditory stream is a rhythmically repeating tone of constant pitch (red) and the other is a spectrotemporally random, arrhythmic cloud of unrelated tones (blue). The repetition rate of the rhythmic stream varies across experiments, including at approximately 1.25 Hz and 5 Hz (Gutschalk et al. 2008), at 4 Hz (Elhilali et al. 2009a), and at 7 Hz (Akram et al. 2014). **(B)** In this case, both auditory streams are rhythmically repeating streams at constant rates, but with incommensurate rates such that perceptual fusion is difficult. The rate pairs vary across experiments, including at 21 and 29 Hz (Bidet-Caulet et al. 2007) and at 4 and 7 Hz (Xiang et al. 2010)

Using these stimuli while scanning the subjects with MEG, the listeners were instructed to press a button when they detected the rhythmic component despite the interfering tone cloud. The task was sufficiently difficult that subjects detected the rhythmic tones in only 60% of the trials in which they were present. This allowed the investigators to separately analyze the MEG responses to the individual tones in the steady rhythm in two cases: when the rhythm tone pips were detected and when, even though still present, they were not detected (Fig. 7.3). The MEG response to

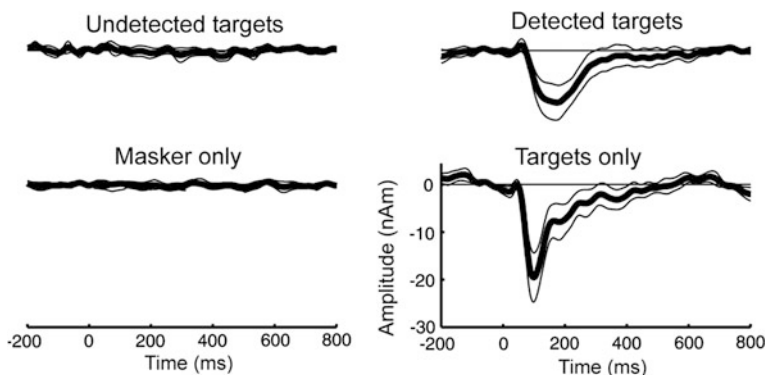


Fig. 7.3 (Top) Average responses to the individual rhythmic tones, but averaged separately according to whether or not they were detected behaviorally. **(Lower)** For comparison, averaged responses in the cases where tones were absent (*left*) or when the competing tone cloud was absent (*right*). [From Gutschalk et al. (2008), Fig. 1.]

the detected tones was similar in latency and neural origin to the N1m response (to tone pips presented in isolation), and so consistent with an origin in planum temporale, a higher order auditory cortical area. The MEG responses to the same tones when not detected were not even themselves detectable. As the stimuli were identical in both cases, the responses to the detected tones must arise from a representation linked to the auditory perception of the tones, rather than their mere acoustic presence in the auditory scene. A further control experiment confirmed that the rhythmic tones, whether perceived or not, are indeed always represented elsewhere in auditory cortex, but only at an earlier latency. This demonstrates a neural representation of the acoustic (not perceptual) presence of the tones that is separate from the later representation representing the listener's percept of their presence in the auditory scene.

Adding fMRI to a similar MEG study, Wiegand and Gutschalk (2012) took advantage of the enhanced spatial resolution of fMRI to better determine the source of these responses. Detected target tones resulted in widespread activation across auditory cortex, but the only area showing contrast between detected and undetected target tones was, perhaps surprisingly, in Heschl's gyrus. A succinct interpretation of this result is that primary auditory cortical areas are indeed important in the detection and processing of the targets, which is not restricted to higher order areas.

In summary, these investigations demonstrate that there are multiple cortical representations of the sounds present in an acoustic scene. The earlier representations in core auditory cortex are determined more by the acoustics of a sound than its percept. The later representations in higher-order auditory cortex tend to track (or perhaps more likely, precede) the percept of the sound. Earlier studies struggled with the confound of changing percept by changing acoustics, but studies since have been able to address this confound and work around it.

7.3.2 *Studies Using Explicit Attentional Manipulation*

The next set of investigations employ a different paradigm from Sect. 7.3.1, in which the attentional state of the listener is explicitly manipulated while keeping the acoustic scene unchanged. In this way, the neural representations of identical acoustic scenes can be probed under distinct, controllable parsings of the scene.

7.3.2.1 **Tones and Maskers**

Elhilali et al. (2009a), using a variation of the informational masking stimuli just described (Fig. 7.2A), had listeners attend to either the rhythmic tone pip stream (in this case at the faster rate of 4 Hz) or to the spectrotemporally random tone cloud, while being scanned by MEG. When attending to the rhythmic stream, the listeners' ability to perceive the rhythmic stream was measured by whether the listeners could detect deviants in the tone stream. Otherwise, the listeners were directed to attend to the random tone cloud by asking them to detect deviants in the tone cloud. The response to the rhythmic stream (alone) was measured and compared across the two tasks by examining the MEG response at the 4 Hz rate of the rhythmic tone stream. Notably, the listeners employed selective auditory attention in both tasks, but to different components of the same auditory scene.

Consistent with the findings of Gutschalk and colleagues, the representation of the rhythmic stream was consistent with the location of the N1m source in planum temporale, and was dramatically stronger when the listeners' attention was focused on that stream than when it was focused on the random tone cloud. Additional associations linking perception (via behavioral responses) and neural response were also noted, of both top-down and bottom-up natures. A critical top-down result was the finding of a positive association between the perceptual buildup of the stream over the course of each trial and the buildup of the neural representation over the same time course. A corresponding bottom-up result was a significant correlation between the listener's ability to detect the rhythmic stream as a function of its frequency (high-frequency tones were easier to detect) and the strength of its neural representation. A hemispheric and task-based asymmetry was noted, where the representation of the rhythmic stream was stronger in the left hemisphere when attended to (i.e., in the foreground), but stronger in the right hemisphere when unattended (i.e., in the background). This is consistent with the fMRI results of Deike and colleagues, who have also seen leftward-biased hemispheric asymmetry in directed attention streaming studies (Deike et al. 2004, 2010).

To investigate whether any of these results arose from the specific rhythmic rate of 4 Hz (the approximate frequency boundary between the delta and theta bands), Akram et al. (2014) conducted a parallel investigation using a 7-Hz rhythmic rate, and found that the majority of the original findings still held. This rate is almost twice as fast as the earlier rate, however, and there are also concomitant processing

and perceptual differences (Patel 2008), so not all the earlier results generalize to the faster rate. In particular, no hemispheric asymmetry was seen.

Ahveninen et al. (2011) also investigated the neural representations of simple sounds in a noisy acoustic scene, using EEG, MEG, and fMRI. Their stimuli were slow progressions of tone pips masked by notched noise, with a notch (i.e., protection zone) of 1/3 octaves, thus ensuring that the masking was dominantly informational rather than energetic. The N1/N1m responses to the masked tones demonstrated a release from adaptation, but only in situations in which the listener was attending to the tones (detecting frequency deviants among the tones). This was interpreted as an attentional sharpening of spectral tuning to the pure tone frequency, in addition to any attention-related strengthening of the neural representation of the tone.

7.3.2.2 Competing Simple Patterns

The next several investigations employed a different stimulus paradigm: competing simple rhythmic streams, each with a distinct rhythm and at incommensurate rates (Fig. 7.2B). Because of the incommensurate rates, the two streams always remain segregated and never fuse. The benefit of this stimulus paradigm is that the listener can be instructed to attend to only one of the streams, which becomes the foreground element in the auditory scene, at which point the remaining stream shifts the background. This allows for two conditions, both with identical simple stimuli and both with focused auditory attention, the only difference being where the focus of attention lies. In this sense, these studies are simpler precursors to the investigations described in Sect. 7.4, which use competing speech streams as more complex, but more natural, generalizations of the strictly rhythmic streams.

Bidet-Caulet et al. (2007) recorded from subjects listening to competing amplitude modulated stimuli, one modulated at 21 Hz and the other at 29 Hz. Each stream had a different (harmonic complex) carrier and subjects perceived the two streams as distinct. The subjects were patients with pharmacologically resistant epilepsy, who were stereotactically implanted with multicontact depth electrodes. The electrodes were used for planning surgery, but they could also be used to record local potentials in temporal cortex. Subjects were instructed to attend to one or the other of the two streams while their responses were recorded using the depth electrodes, the specific locations of which varied across the subjects. Responses time-locked to the amplitude modulations were observed all along Heschl's gyrus, but not in other auditory areas. The results demonstrated an upward modulation of the neural representation of the attended stream, and, correspondingly, a downward modulation of the unattended stream. Similar attentional modulation was observed in nonprimary areas, but only for transient and non-time-locked responses. These attentional effects in both primary and nonprimary areas are in broad agreement with the fMRI results of Wiegand and Gutschalk (2012) described in Sect. 7.2.

Similarly, but using much slower amplitude modulation rates of 4 Hz and 7 Hz, Xiang et al. (2010) investigated the effects of selective attention in healthy subjects

recorded with MEG. Just as in the previous case, the response to a stream was significantly larger when it was the foreground stream (i.e., when the subject was induced to attend to that stream) than when the same stream was in the background. Additionally, as seen in the related informational masking studies described in Sect. 7.3.2.1 (Elhilali et al. 2009a; Akram et al. 2014), the strength of the attended neural representation also correlated with the behavior of the subject performing a deviant detection task based on the attended stream. Unlike the results of either of those investigations, however, there was a right-sided hemispheric asymmetry in the neural responses, regardless of amplitude modulation rate or task. This result contrasts with the leftward-biased hemispheric asymmetry found using fMRI by Deike et al. (2004, 2010).

7.3.2.3 Suppressing Attention

In contradistinction to those studies that employ directed active attention to one auditory scene element or another, Chait et al. (2010) investigated the effects of actively ignoring an auditory scene element. Listening to tone pips arranged in the ABA paradigm, while being scanned with MEG, listeners were given a task (detecting deviants in the A stream) that is substantially easier when the B tones are not fused into the same stream as the A tones. In this manner listeners were motivated to suppress (ignore) tone pips in the B stream. A control task, with no incentive to either attend or ignore the B stream, was used for comparison. The main result was that the neural responses to the B stream's tone pips were significantly smaller when the B stream was actively ignored than during the control task. In other words, not only can active attention enhance (modulate upward) the neural representation of a particular sound, but also active ignoring can suppress (modulate downward) a sound's neural representation.

In summary, the paradigm of presenting identical stimuli in two contrasting behavioral and perceptual conditions allows one to cleanly separate the percept from the acoustics, and thus also separate the neural representation of a sound's percept as distinct from the neural representation of the sound's acoustics. The neural representation of a sound's percept is certainly modulated upward by attention, but, perhaps more importantly, it is modulated upward only when that particular sound is in the foreground, not just because selective attention has been applied at all. The distinction is important because, as demonstrated by the presence of the neural buildup, when the sound is not yet perceived (and so cannot be part of the foreground) it is *not* modulated upward by attention, even though the subject's attention is focused on acquiring the target. Additionally, it is seen that actively ignoring a sound suppresses the neural representation of that sound, even below baseline.

7.4 Neural Basis of Auditory Stream Segregation in Humans: Speech

In contrast to the simple signals used in the investigations just described in Sect. 7.3, continuous natural speech is a highly complex signal. Despite its intrinsic complexity, however, it has been surprisingly amenable to neural investigations of the cocktail party problem. Prominent characteristics of speech are its dynamic nature and characteristic rhythmicity. These properties both lend themselves to investigations of temporal encoding (de Cheveigne 2003), which have indeed proven very fruitful in several different modalities, including EEG (Lalor and Foxe 2010; Di Liberto et al. 2015), MEG (Luo and Poeppel 2007; Ding et al. 2016), and ECoG (Pasley et al. 2012; Mesgarani et al. 2014). As will be seen, the temporal encoding of speech in auditory cortex, manifesting as neural responses time locked to the rhythms of the speech, is robust to interference from other acoustic stimuli in the same auditory scene, including, but not limited to, other speech stimuli. These temporally based neural representations of speech can be characterized by both their ability to predict the neural responses to a given stimulus and also by their ability to reconstruct a stimulus feature from the neural responses (Fig. 7.4A). These representations are often investigated using methods from linear systems theory, due to the power and flexibility of such methods (despite the distinct shortcoming of being limited to the linear component of their relationships), but they are not restricted to such methods (see, e.g., Ahissar et al. 2001; Luo and Poeppel 2007). It should also be noted that results from the use of linear systems methods in this scenario are sufficiently reliable that the methodology has proven to be useful in brain–computer interface (BCI) applications (Dijkstra et al. 2015; O’Sullivan et al. 2015b).

7.4.1 Studies Using Speech in Stationary Noise

The simplest case of processing speech as one element in a larger auditory scene is when the speech is masked by stationary noise, since the speech is dynamic but the noise background is not. Ding and colleagues (Ding and Simon 2013; Ding et al. 2014) investigated this case using MEG.

In the first of these studies (Ding and Simon 2013), subjects listened to speech masked by spectrally matched stationary noise, over a wide range of linearly worsening signal-to-noise ratios (SNRs). The fidelity of the neural representation of the speech (despite the noise) was measured by the reconstructability of the speech envelope from the low-frequency time-locked MEG responses to the speech in noise. The speech ranged from highly intelligible to almost completely unintelligible. Intelligibility (self-reported on a scale of 0–100%) decreased in a very nonlinear manner, with almost no decrease of intelligibility until about -3 dB SNR, where the across-subject mean dropped to approximately 50% (with high variability over subjects), and after which it continued its drop to nearly 0%. The neural

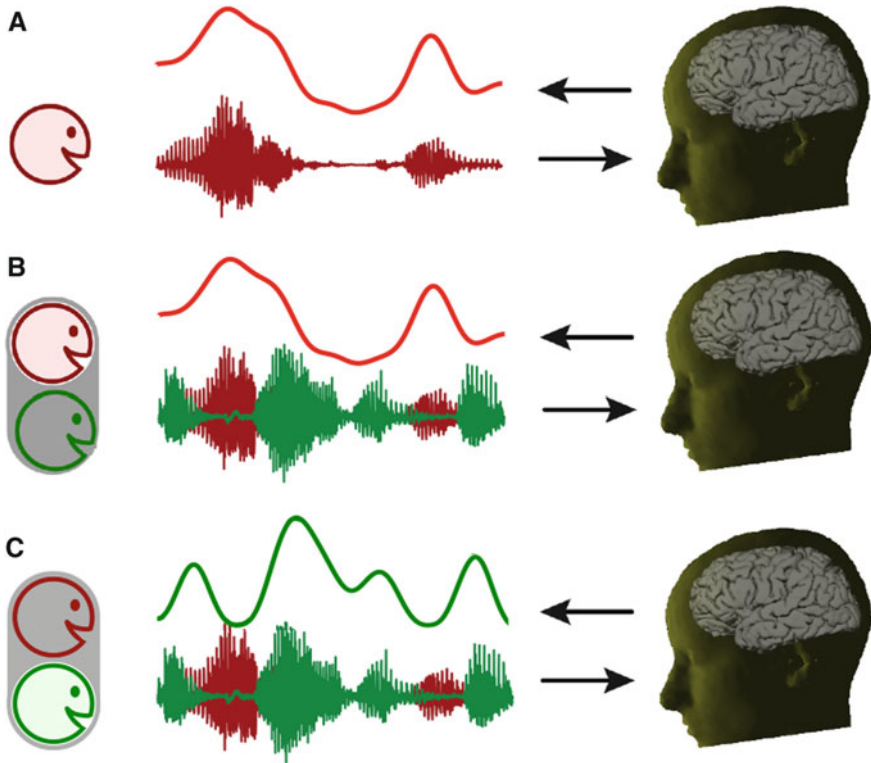


Fig. 7.4 (A) Schematic example of the neural representation of the speech of a single speaker. The (filtered version of a) neural response phase locks to the low-frequency envelope of the speech stream. (B, C) Schematic examples of the neural representation of an attended speech stream in the presence of a competing speech stream. The stimuli are identical in both cases; only the focus of attention of the listener has changed. In this case the example illustrates that the (filtered version of a) neural response is dominantly phase locked to the envelope of the attended speech stream, not the unattended. The same filter is applied in both cases; it is the neural response itself that changes between the two attentional states. [Based on Fig. 1 from Ding and Simon (2012b).]

representation of the speech, however, decreased only mildly with worsening SNR, dropping suddenly to close to floor level only at about -9 dB SNR. This result shows that the neural representation of the speech envelope is largely unaffected by stationary noise for moderate SNR, even for moderately poor SNR, but does eventually succumb when the SNR is very poor. Because the changeover SNR differs between the neural representation of the speech and its (linguistically based) intelligibility, the neural representation cannot directly reflect the linguistic basis of the speech. Furthermore, because the changeover occurs at a worse SNR for the neural representation than for intelligibility, it also follows that the neural representation is prelinguistic, as it successfully represents the speech at a noise level sufficiently high that the speech is audible but not intelligible. Importantly, there

was also a strong correlation between found between the fidelity of the speech neural representation (as measured by stimulus reconstructability), and intelligibility, for the single SNR level that showed enough variability across subjects to allow such analysis.

In a complementary study (Ding et al. 2014), the acoustic signal was further degraded using band vocoding. This acoustic manipulation both eliminates fine temporal acoustic structure and degrades the spectral structure in a controllable way, while keeping the slow acoustic envelope intact. The study showed that, while the neural representation of speech remained robust to noise, it was disrupted by the spectral distortions created by vocoding. This shows that the time locking of the neural responses to the speech envelope cannot be explained by mere passive envelope tracking mechanisms, as band vocoding does not directly affect the stimulus envelope. Rather, the neural representation, although mirroring only the stimulus envelope, specifically requires access to the spectrotemporal fine structure of the speech to neurally extract the speech from the noise. This study also reported a strong correlation between the robustness of the speech neural representation and intelligibility, this time for multiple stimulus conditions. Revealingly, this correlation between neural response and perception only held for the delta (1–4 Hz) band of the neural responses, and not for higher bands.

Using fMRI rather than MEG can give dramatically better spatial resolution at the cost of worse temporal resolution. Results from fMRI studies, however, using both energetic and modulation masking of speech, differ quite substantially in their findings as to how the cortical processing of speech-in-quiet differs from speech in noise (Scott and McGettigan 2013).

7.4.2 *Studies Using Competing Speech Streams*

The ability of human listeners to separate and segregate two simultaneous and competing speakers is profoundly different from the ability to separate and segregate a single speech speaker from noise, and is at the very core of how the brain solves the original cocktail party problem (Cherry 1953). This problem has been investigated using a variety of modalities, including EEG (Kerlin et al. 2010; Power et al. 2012), MEG (Ding and Simon 2012a, 2012b), ECoG (Mesgarani and Chang 2012; Zion Golumbic et al. 2013), PET (Scott et al. 2004), and fMRI (Scott et al. 2009). Studies using electromagnetically based scanning methods (EEG, MEG, and ECoG) typically emphasize the temporal representations of the acoustic signal, whereas the studies using hemodynamically based scanning methods (PET and fMRI) typically emphasize the anatomical locations of neural processing steps. The former, because of their emphasis on temporal representations, are especially well suited to investigate how different elements of an acoustic scene are sequentially represented in different areas of the brain, and are covered here in greater detail.

It also should also be noted that, although many of these studies' results are in strong agreement with each other, there are often noticeable differences in how the

results are interpreted. For instance, these studies typically find, for subjects listening to a mixture of two speech streams but attending to only one, that the neural representation of the attended speech is stronger than that of the unattended speech. Some studies interpret this as simple “attentional gain”: the neural representation of the object of attention has been amplified. This interpretation may be sufficient for simple auditory scenes, for instance, when the competing auditory streams are separated dichotically (Hillyard et al. 1973; Ding and Simon 2012a) or spectrally (Elhilali et al. 2009a; Ahveninen et al. 2011). It falls somewhat short, however, when the competing streams possess strongly overlapping acoustic properties. In these more general (and realistic) cases, because the competing speech streams are not isolatable at the periphery their representations must therefore each be reconstructed *ab initio* by the brain. Because the neural representation benefiting from attention does not even exist until after its “separation” (construction, really) from the rest of the auditory scene, describing it as having benefitted from attentional gain is questionable. In this case the attentional gain interpretation sidesteps the arguably more important questions of how the neural representation of the neurally separated object was created in the first place, and what role attention plays in this process (Elhilali, Chap. 5).

Kerlin et al. (2010) investigated the case of listening to two simultaneous and competing speech streams, using EEG, in subjects listening to single sentences presented from different virtual locations (via individually obtained head-related transfer functions, HRTFs). The neural responses to each speech stream were analyzed using response templates constructed from responses to individual sentences. The major finding was that the representation of the attended speech was stronger (as measured in electrical potential recordings) than that of the unattended speech. The strongest difference was found to be in the theta (4–8 Hz) band, as also seen by Hambrook and Tata (2014). The theta band is also known to be critical for the neural processing of intelligible speech (Luo and Poeppel 2007; Peelle et al. 2013).

Ding and Simon (2012a), using MEG and substantially longer dichotic speech stimuli (60 s duration), found similar results but instead emphasized the temporal properties of the attentional effects. Their analysis, using linear-systems methods routinely employed in auditory neurophysiology (Depireux et al. 2001), linked the dynamics of the auditory stimulus with the neural response using the spectrotemporal response function (STRF). The STRF may also be interpreted as the general response profile to any sound feature (Simon et al. 2007), and as such it can also be used to estimate more traditional response measures such as response component strengths and latencies. In this case, the speech-based STRF possessed a response peak at a latency of approximately 100 ms (i.e., with time course similar to the N1m) that was again substantially stronger for the attended speech stream than the unattended.

Power et al. (2012), using EEG with a similar dichotic paradigm and also using linear systems methods, also found attentional differences but substantially later (approximately 200 ms latency). Because EEG and MEG have differential sensitivities to sources with differing depth and orientation, however, the two results may

be interpreted as complementary rather than in conflict. Power and colleagues also found neural sources at earlier latencies (e.g., approximately 50 ms and 100 ms) that processed the different speech streams with undifferentiated strength. Horton et al. (2013), using EEG and a related stimulus paradigm, did find attentional differentiation as early as 100 ms latency, but using an analysis methodology that does not take into account the temporal autocorrelations of the speech envelope.

In a second investigation by Ding and Simon (2012b), the auditory scene changed to a pair of competing, individual, lengthy speech streams from different talkers (either same or different sex) but now mixed into a single acoustic channel, and presented diotically (i.e., identically to each ear). The listeners' task, attending to only one of the two speakers, is not difficult for young, normal hearing listeners. The neural mechanisms underlying this task, however, now have access only to derived (not explicit) auditory properties not encoded at the periphery, including instantaneous pitch, timbre, rhythmicity, and cadence. These derived properties, and how they change over time, are algorithmically nontrivial to track even in a single continuous speech signal, let alone for a mixture, and yet they are the only cues available in this cocktail party listening paradigm. The experimental benefit of using such a diotic signal is that it avoids several potential confounds that would arise from allowing spatial separation of the sources. For example, findings of hemispheric lateralization could be confused with findings of competing ipsilateral/contralateral processing. Even under these algorithmically difficult conditions, however, the neural representation of the attended speech stream (Fig. 7.4B, C) was found to be easily separable from that of the unattended speech stream (Fig. 7.5). The representation of the attended speech stream was again found to be stronger than that of the unattended, for both predicting the neural response from the stimulus, and for reconstructing the stimulus from the neural response. As in the case of spatially separated speech streams, this difference primarily arose from neural sources with post-stimulus latency of approximately 100 ms. Neural sources were also observed with approximately 50 ms post-stimulus latency, but that did not differentiate between the attended and unattended streams, in agreement with the early latency EEG results of Power et al. (2012). Neural source localization revealed that the later (100 ms latency) sources, which represented the attended speech so much more strongly than the unattended, had an origin consistent with that of the N1m, in planum temporale. The earlier (50 ms latency) sources, which did not distinguish between the attended and unattended speech, had an origin consistent with that of the P1m, in Heschl's gyrus. A reasonable interpretation is that primary auditory cortical areas process the entire auditory scene, only weakly sensitive to selective attention, but by the time that higher-order auditory areas in planum temporale receive their processed neural signals, the speech streams have since been segregated. At this level more neural resources are dedicated to processing the attended speech stream than the unattended, leading to the stronger neural signal.

Mesgarani and Chang (2012) also found enhanced representation of the speech of the attended speaker, taking advantage of the greatly enhanced spatial resolution of ECoG, which in turn allowed access to more finely grained neural

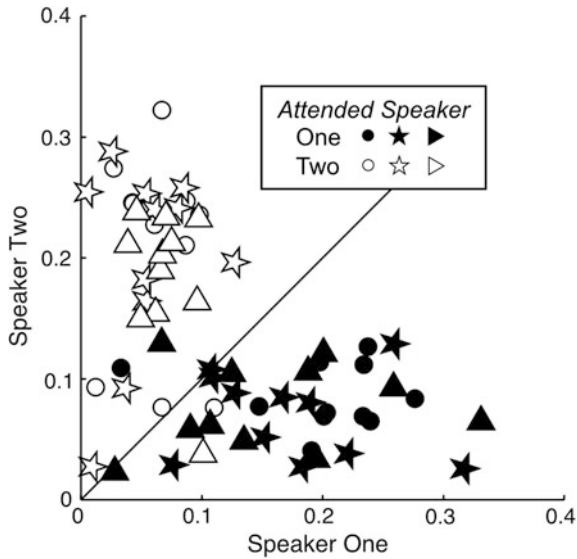


Fig. 7.5 Decoding of the speech representations from single trials. Scatterplot of the correlation coefficients measured from individual trials and individual subjects, between the individual decoded stimulus envelope and the actual individual envelope. The attentional focus of listeners is denoted by marker color and the separate trials are denoted by marker shapes. It can be seen that the speech of the attended speaker can be decoded separately from the response to both, even on a single trial basis. [From Ding and Simon (2012b), Supporting Information.]

representations of the speech. These representations are sufficiently detailed that the neural responses can be used not only to decode the global temporal envelope of the speech streams, but also to decode the detailed spectrotemporal envelope (i.e., the spectrogram). This allows reconstruction of the attended speech at much higher fidelity (see also Pasley et al. 2012). Furthermore, this reconstruction was successful only for trials during which the subjects could correctly answer questions about the attended speech, not for error trials, indicating that the subjects' attentional target was better identified by their neural responses than by the task-assigned target. The cortical locations investigated were constrained by clinical electrode placement, but included nonprimary auditory areas in the posterior superior temporal lobe. Within those areas, only the superior and middle temporal gyrus responded reliably to speech stimuli, and no spatial pattern was observed for attentional effects.

In that investigation, the neural representations of the separate speech streams were determined only from analysis of the high frequency (75–150 Hz) gamma band's low frequency envelope. Zion Golumbic et al. (2013), also using ECoG, observed two separate types of cortical representation of speech, one also from the gamma band envelope and the other from the low-frequency (delta and theta) band directly. As also seen by Mesgarani and Chang (2012), only the attended speech was measurable in the gamma band representation, and as seen by Ding and Simon

(2012a, 2012b) in MEG, both attended and unattended speech representations were seen in the low-frequency representations. The gamma band representations were again found nearest to auditory cortex. The lower frequency representations showed two different distributions, however, with theta band dominated representations found nearest to auditory cortex and delta-band dominated representations distributed far more widely. This difference in the spatial distributions parallels the theta/delta differences in links to perception described in Sect. 7.4.1 in MEG (Ding and Simon 2013; Ding et al. 2014) and EEG (Kayser et al. 2015), and may indicate that the more widely distributed delta band representations may be more diverse in nature (including, e.g., possible language-influenced representations).

7.4.3 Neuroanatomy of Speech-in-Noise Processing

ECoG studies have demonstrated that neural representations of speech in a noisy background (including other speech sources) are widespread throughout, and beyond, auditory cortex (Mesgarani and Chang 2012; Zion Golumbic et al. 2013; Dijkstra et al. 2015). Even so, from an anatomical perspective, those studies are limited in the brain regions they are sensitive to, both because of the limited coverage of the ECoG electrodes and because of their focus on dynamic evoked neural activity. PET and fMRI have access a much greater area, and are sensitive to any type of neural activity sufficiently large enough to produce a hemodynamic change. Scott and McGettigan (2013) review this literature, and although they find quite a wide variety of reports as to which cortical areas contribute to the processing of masked speech (especially for energetically masked speech), some common patterns have emerged. There is general agreement that there is considerable bilateral activation in the superior temporal lobe in general, and the superior temporal gyrus in particular (Nakai et al. 2005; Hill and Miller 2010). Additionally, the cortical processing of speech in noise occurs throughout prefrontal and parietal cortex (e.g., Scott et al. 2004, 2009).

EEG and MEG have significantly poorer spatial resolution than the other techniques discussed here, but can still contribute to neuroanatomical investigations. Their strongest contributions may be indirectly via latency: lower latency cortical representations, which are typically *less* sensitive to percept, are found in core/primary auditory cortex; longer latency cortical representations, which are typically *more* sensitive to percept, are found in higher order auditory cortex. Specifically, shorter latency representations typically localize to Heschl's gyrus and longer latency representations typically localize to more distant regions of the superior temporal gyrus (e.g., Ahveninen et al. 2011). There can be exceptions to this early/acoustic versus late/percept dichotomy, however, as statistical expectations regarding the stimulus set and stimulus context also affect early responses (Szalardy et al. 2013).

In summary, the use of speech as an element in a complex auditory scene has provided a surprisingly rich and consistent trove of experimental results. Despite its

acoustic complexity, speech has a robust temporal neural representation that is well suited for neural investigations of the cocktail party problem. There are different temporal representations of the speech as different levels of processing have been performed. Earlier representations are consistent with representations of the entire acoustic scene, relatively unaffected by selective attention, and are found in primary-like cortical areas located in Heschl's gyrus. Later representations are consistent with representations of specific elements of the entire acoustic scene, with attended elements generating substantially more neural activity than unattended. Additionally, representations dominated by rates in the theta band are more closely tied to the acoustics of the speech stream, whereas representations dominated by time-locking rates in the delta band are more closely tied to the perception of the speech stream, including intelligibility.

7.5 Other Aspects of the Human Auditory Neuroscience of Cocktail Party Processing

Not all investigations of human auditory neuroscience applied to the cocktail party problem easily fit into the specific categories of the earlier sections. This section covers a small selection of other such investigations.

7.5.1 Temporal Coherence

Temporal coherence is the name of a class of models of acoustic scene analysis that specify how the brain integrates common information from common physical sources and segregates sounds of interest from other sources. The theory posits that when neuronal populations share temporal coherence across various features of a sound source, they can be separated out from other sources and bound together as a unified perceptual object (Elhilali et al. 2009b; Shamma et al. 2011; Elhilali, Chap. 5). Temporal coherence models involve both feedforward components and feedback components using attentional selection and gating. O'Sullivan et al. (2015a) investigated the possible use of temporal coherence by recording EEG from subjects listening to a stochastic figure-ground stimulus (Teki et al. 2011). By modifying the stimuli so that the temporal coherence of the figure itself changed dynamically, linear systems methods could be used to determine the short-term time course of the neural processing of the temporal coherence. Use of both passive and active listening conditions allowed the feedforward and feedback components to be analyzed separately. Passive listening demonstrated a neural representation of temporal coherence from approximately 115 to 185 ms post-coherence latency, that is to say, with a later latency than N1. Active listening resulted in a larger neural representation of temporal coherence, beginning at the same time but lasting almost

100 ms longer than the passive case. These results demonstrate an early and preattentive component of temporal coherence processing, but that is enhanced and extended by active listening and selective attention.

7.5.2 Bottom-up Versus Top-Down Attention

Much of the selective attention occurring in the investigations described in this chapter is driven by “top-down” effects, for instance with the listeners performing a task designed by the experimenters. Some selective attention is “bottom-up,” driven by salient acoustic events that cause attention to become focused on the auditory object associated with that event. Some investigations, though not necessarily designed to do so, have seen interplay between the two effects (Elhilali et al. 2009a; Akram et al. 2014). Other investigations, by intentionally distracting the listener away from the auditory stimuli, are focused on bottom-up driven processing by design (e.g., Teki et al. 2011). Shuai and Elhilali (2014), using EEG, investigated the neural underpinnings of bottom-up salience and its role in selective attention as a whole. Subjects listened to rhythmic auditory streams with occasional salient events, under different attentional states, and in the presence or absence of a competing auditory stream. Two main effects were seen in response to the high-saliency events, the first being a strong evoked response to the salient sound events. The strength of this response was additionally enhanced under the influence of selective attention, and according to the saliency level of the events. The second was more far reaching, with increased amplitude for the representation of the *entire* rhythmic auditory stream that contained the occasional salient events. This increase did not depend on attentional state or complexity of the scene. Thus the role of bottom-up attention in the neural processing of a scene can be seen as twofold, both to secure attention to the salient events themselves, but also to secure attention to the auditory scene components that, as a group, contain not only the salient events but also other events related, but not identical, to them.

7.6 Summary

Investigation of the human neuroscience of the cocktail party problem, although not exactly in its infancy, is much less established than many of the other areas explored in this volume. One reason is the difficulty in carrying out traditional auditory neurophysiological studies in humans. Another is lack of experience in the field in how best to bring whole-head neuroimaging and noninvasive neurophysiological methods to bear on the questions of interest.

Nevertheless, the advances described in this chapter demonstrate that not only is the field vibrant and full of potential, but it is actually leading in several ways. One of the conceptual advances coming out of the fields is the progression from

investigations of auditory representations that are faithful to the *acoustics* of the stimulus, to investigations of auditory representations that are faithful to the *perception* of the stimulus, and of the mechanisms that allow this transition. The more perception diverges from acoustics (culminating in the classic cocktail party scenario), the more critical each stage of the auditory processing becomes, and the more important the neurophysiological processes that underlie auditory cognition become.

Other advances have come from usage of natural, long-duration speech as a stimulus. The use of speech as stimulus for nonhuman animals has obvious drawbacks (lack of direct behaviorally relevance, difficulties in stimulus quantification), and its use in human psychophysics experiments is also problematic (e.g., lack of agreement as to how to quantify intelligibility for long-duration speech). From the perspective of human auditory neurophysiology and behavior, however, it has remarkable properties. It is perhaps unmatched in behavioral relevance. It additionally drives auditory neural responses that are strong and reliable, and that covary with acoustic stimulus properties, and perceptual stimulus properties, and the behavioral state of the listener.

Still, the field is young, and future findings should give even more insight into the auditory system and its function. One limitation, slowly being chipped away at, is the paucity of investigators who are experts both in whole-brain neuroimaging/neurophysiology and, at the same time, in psychophysics and behavior. Although this is also a significant problem in nonhuman animal neurophysiology and behavior, the problem seems especially acute for human studies. There are high expectations of what perceptual details and complexity that should be experimentally obtainable from human subjects. Similarly, the ability to access the whole human brain at once, even constrained to the spatial resolution of fMRI and the temporal resolution of EEG and MEG, is extraordinary and its most important uses may still lie ahead.

Acknowledgements Support for the author's work was provided by the National Institute of Deafness and Other Communication Disorders Grant R01-DC-014085.

Compliance with Ethics Requirements

Jonathan Z. Simon declares that he has no conflict of interest.

References

- Ahissar, E., Nagarajan, S., Ahissar, M., Protopapas, A., et al. (2001). Speech comprehension is correlated with temporal response patterns recorded from auditory cortex. *Proceedings of the National Academy of Sciences of the USA*, 98(23), 13367–13372.
- Ahveninen, J., Hamalainen, M., Jaaskelainen, I. P., Ahlfors, S. P., et al. (2011). Attention-driven auditory cortex short-term plasticity helps segregate relevant sounds from noise. *Proceedings of the National Academy of Sciences of the USA*, 108(10), 4182–4187.
- Ahveninen, J., Kopco, N., & Jaaskelainen, I. P. (2014). Psychophysics and neuronal bases of sound localization in humans. *Hearing Research*, 307, 86–97.

- Akram, S., Englitz, B., Elhilali, M., Simon, J. Z., & Shamma, S. A. (2014). Investigating the neural correlates of a streaming percept in an informational-masking paradigm. *PLoS ONE*, 9(12), e114427.
- Alain, C., Arnott, S. R., & Picton, T. W. (2001). Bottom-up and top-down influences on auditory scene analysis: Evidence from event-related brain potentials. *Journal of Experimental Psychology: Human Perception and Performance*, 27(5), 1072–1089.
- Alain, C., Reinke, K., He, Y., Wang, C., & Lobaugh, N. (2005). Hearing two things at once: Neurophysiological indices of speech segregation and identification. *Journal of Cognitive Neuroscience*, 17(5), 811–818.
- Bidet-Caulet, A., Fischer, C., Besle, J., Aguera, P. E., et al. (2007). Effects of selective attention on the electrophysiological representation of concurrent sounds in the human auditory cortex. *The Journal of Neuroscience*, 27(35), 9252–9261.
- Bregman, A. S. (1990). *Auditory scene analysis: The perceptual organization of sound*. Cambridge, MA: MIT Press.
- Briley, P. M., Kitterick, P. T., & Summerfield, A. Q. (2013). Evidence for opponent process analysis of sound source location in humans. *Journal of the Association for Research in Otolaryngology*, 14(1), 83–101.
- Briley, P. M., Goman, A. M., & Summerfield, A. Q. (2016). Physiological evidence for a midline spatial channel in human auditory cortex. *Journal of the Association for Research in Otolaryngology*, 17(4), 331–340.
- Brungart, D. S., Simpson, B. D., Ericson, M. A., & Scott, K. R. (2001). Informational and energetic masking effects in the perception of multiple simultaneous talkers. *The Journal of the Acoustical Society of America*, 110(5 Pt 1), 2527–2538.
- Chait, M., Poeppel, D., & Simon, J. Z. (2006). Neural response correlates of detection of monaurally and binaurally created pitches in humans. *Cerebral Cortex*, 16(6), 835–848.
- Chait, M., de Cheveigne, A., Poeppel, D., & Simon, J. Z. (2010). Neural dynamics of attending and ignoring in human auditory cortex. *Neuropsychologia*, 48(11), 3262–3271.
- Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with 2 Ears. *The Journal of the Acoustical Society of America*, 25(5), 975–979.
- Cramer, E. M., & Huggins, W. H. (1958). Creation of pitch through binaural Interaction. *The Journal of the Acoustical Society of America*, 30(5), 413–417.
- Cusack, R. (2005). The intraparietal sulcus and perceptual organization. *Journal of Cognitive Neuroscience*, 17(4), 641–651.
- de Cheveigne, A. (2003). Time-domain auditory processing of speech. *Journal of Phonetics*, 31(3–4), 547–561.
- Deike, S., Gaschler-Markefski, B., Brechmann, A., & Scheich, H. (2004). Auditory stream segregation relying on timbre involves left auditory cortex. *NeuroReport*, 15(9), 1511–1514.
- Deike, S., Scheich, H., & Brechmann, A. (2010). Active stream segregation specifically involves the left human auditory cortex. *Hearing Research*, 265(1–2), 30–37.
- 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.
- Dijkstra, K. V., Brunner, P., Gunduz, A., Coon, W., et al. (2015). Identifying the attended speaker using electrocorticographic (ECoG) signals. *Brain-Computer Interfaces*, 2(4), 161–173.
- Di Liberto, G. M., O'Sullivan, J. A., & Lalor, E. C. (2015). Low-frequency cortical entrainment to speech reflects phoneme-level processing. *Current Biology*, 25(19), 2457–2465.
- Ding, N., & Simon, J. Z. (2012a). Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. *Journal of Neurophysiology*, 107(1), 78–89.
- Ding, N., & Simon, J. Z. (2012b). Emergence of neural encoding of auditory objects while listening to competing speakers. *Proceedings of the National Academy of Sciences of the USA*, 109(29), 11854–11859.
- Ding, N., & Simon, J. Z. (2013). Adaptive temporal encoding leads to a background-insensitive cortical representation of speech. *The Journal of Neuroscience*, 33(13), 5728–5735.

- Ding, N., Chatterjee, M., & Simon, J. Z. (2014). Robust cortical entrainment to the speech envelope relies on the spectro-temporal fine structure. *NeuroImage*, *88*, 41–46.
- Ding, N., Melloni, L., Zhang, H., Tian, X., & Poeppel, D. (2016). Cortical tracking of hierarchical linguistic structures in connected speech. *Nature Neuroscience*, *19*(1), 158–164.
- Dingle, R. N., Hall, S. E., & Phillips, D. P. (2010). A midline azimuthal channel in human spatial hearing. *Hearing Research*, *268*(1–2), 67–74.
- Dingle, R. N., Hall, S. E., & Phillips, D. P. (2012). The three-channel model of sound localization mechanisms: Interaural level differences. *The Journal of the Acoustical Society of America*, *131*(5), 4023–4029.
- Dykstra, A. R., Halgren, E., Thesen, T., Carlson, C. E., et al. (2011). Widespread brain areas engaged during a classical auditory streaming task revealed by intracranial EEG. *Frontiers in Human Neuroscience*, *5*, 74.
- Elhilali, M., Xiang, J., Shamma, S. A., & Simon, J. Z. (2009a). Interaction between attention and bottom-up saliency mediates the representation of foreground and background in an auditory scene. *PLoS Biology*, *7*(6), e1000129.
- Elhilali, M., Ma, L., Micheyl, C., Oxenham, A. J., & Shamma, S. A. (2009b). Temporal coherence in the perceptual organization and cortical representation of auditory scenes. *Neuron*, *61*(2), 317–329.
- Gutschalk, A., & Dykstra, A. R. (2014). Functional imaging of auditory scene analysis. *Hearing Research*, *307*, 98–110.
- Gutschalk, A., Micheyl, C., Melcher, J. R., Rupp, A., et al. (2005). Neuromagnetic correlates of streaming in human auditory cortex. *The Journal of Neuroscience*, *25*(22), 5382–5388.
- Gutschalk, A., Oxenham, A. J., Micheyl, C., Wilson, E. C., & Melcher, J. R. (2007). Human cortical activity during streaming without spectral cues suggests a general neural substrate for auditory stream segregation. *The Journal of Neuroscience*, *27*(48), 13074–13081.
- Gutschalk, A., Micheyl, C., & Oxenham, A. J. (2008). Neural correlates of auditory perceptual awareness under informational masking. *PLoS Biology*, *6*(6), e138.
- Hambrook, D. A., & Tata, M. S. (2014). Theta-band phase tracking in the two-talker problem. *Brain and Language*, *135*, 52–56.
- 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. *The Journal of the Acoustical Society of America*, *115*(2), 833–843.
- Hill, K. T., & Miller, L. M. (2010). Auditory attentional control and selection during cocktail party listening. *Cerebral Cortex*, *20*(3), 583–590.
- Hill, K. T., Bishop, C. W., & Miller, L. M. (2012). Auditory grouping mechanisms reflect a sound's relative position in a sequence. *Frontiers in Human Neuroscience*, *6*, 158.
- Hillyard, S. A., Hink, R. F., Schwent, V. L., & Picton, T. W. (1973). Electrical signs of selective attention in the human brain. *Science*, *182*(4108), 177–180.
- Horton, C., D'Zmura, M., & Srinivasan, R. (2013). Suppression of competing speech through entrainment of cortical oscillations. *Journal of Neurophysiology*, *109*(12), 3082–3093.
- Hugdahl, K. (2005). Symmetry and asymmetry in the human brain. *European Review*, *13*(Suppl. S2), 119–133.
- Jeffress, L. A. (1948). A place theory of sound localization. *Journal of Comparative and Physiological Psychology*, *41*(1), 35–39.
- Kaas, J. H., & Hackett, T. A. (2000). Subdivisions of auditory cortex and processing streams in primates. *Proceedings of the National Academy of Sciences of the USA*, *97*(22), 11793–11799.
- Kayser, S. J., Ince, R. A., Gross, J., & Kayser, C. (2015). Irregular speech rate dissociates auditory cortical entrainment, evoked responses, and frontal alpha. *The Journal of Neuroscience*, *35*(44), 14691–14701.
- Kerlin, J. R., Shahin, A. J., & Miller, L. M. (2010). Attentional gain control of ongoing cortical speech representations in a “cocktail party”. *The Journal of Neuroscience*, *30*(2), 620–628.
- Kidd, G., Jr., Mason, C. R., & Richards, V. M. (2003). Multiple bursts, multiple looks, and stream coherence in the release from informational masking. *The Journal of the Acoustical Society of America*, *114*(5), 2835–2845.

- Kulesza, R. J., Jr. (2007). Cytoarchitecture of the human superior olivary complex: Medial and lateral superior olive. *Hearing Research*, 225(1–2), 80–90.
- Lalor, E. C., & Foxe, J. J. (2010). Neural responses to uninterrupted natural speech can be extracted with precise temporal resolution. *European Journal of Neuroscience*, 31(1), 189–193.
- Lee, A. K., Larson, E., Maddox, R. K., & Shinn-Cunningham, B. G. (2014). Using neuroimaging to understand the cortical mechanisms of auditory selective attention. *Hearing Research*, 307, 111–120.
- Luo, H., & Poeppel, D. (2007). Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. *Neuron*, 54(6), 1001–1010.
- Lutkenhoner, B., & Steinstrater, O. (1998). High-precision neuromagnetic study of the functional organization of the human auditory cortex. *Audiology and Neuro-Otology*, 3(2–3), 191–213.
- Maddox, R. K., Billimoria, C. P., Perrone, B. P., Shinn-Cunningham, B. G., & Sen, K. (2012). Competing sound sources reveal spatial effects in cortical processing. *PLoS Biology*, 10(5), e1001319.
- Magezi, D. A., & Krumbholz, K. (2010). Evidence for opponent-channel coding of interaural time differences in human auditory cortex. *Journal of Neurophysiology*, 104(4), 1997–2007.
- Makela, J. P., Hamalainen, M., Hari, R., & McEvoy, L. (1994). Whole-head mapping of middle-latency auditory evoked magnetic fields. *Electroencephalography and Clinical Neurophysiology*, 92(5), 414–421.
- McAlpine, D. (2005). Creating a sense of auditory space. *Journal of Physiology*, 566(Pt 1), 21–28.
- McLaughlin, S. A., Higgins, N. C., & Stecker, G. C. (2016). Tuning to binaural cues in human auditory cortex. *Journal of the Association for Research in Otolaryngology*, 17(1), 37–53.
- Mesgarani, N., & Chang, E. F. (2012). Selective cortical representation of attended speaker in multi-talker speech perception. *Nature*, 485(7397), 233–236.
- Mesgarani, N., Cheung, C., Johnson, K., & Chang, E. F. (2014). Phonetic feature encoding in human superior temporal gyrus. *Science*, 343(6174), 1006–1010.
- Middlebrooks, J. C., & Bremen, P. (2013). Spatial stream segregation by auditory cortical neurons. *The Journal of Neuroscience*, 33(27), 10986–11001.
- Naatunen, R., Paavilainen, P., Rinne, T., & Alho, K. (2007). The mismatch negativity (MMN) in basic research of central auditory processing: A review. *Clinical Neurophysiology*, 118(12), 2544–2590.
- Nakai, T., Kato, C., & Matsuo, K. (2005). An fMRI study to investigate auditory attention: A model of the cocktail party phenomenon. *Magnetic Resonance in Medical Sciences*, 4(2), 75–82.
- O’Sullivan, J. A., Shamma, S. A., & Lalor, E. C. (2015a). Evidence for neural computations of temporal coherence in an auditory scene and their enhancement during active listening. *The Journal of Neuroscience*, 35(18), 7256–7263.
- O’Sullivan, J. A., Power, A. J., Mesgarani, N., Rajaram, S., et al. (2015b). Attentional selection in a cocktail party environment can be decoded from single-trial EEG. *Cerebral Cortex*, 25(7), 1697–1706.
- Pasley, B. N., David, S. V., Mesgarani, N., Flinker, A., et al. (2012). Reconstructing speech from human auditory cortex. *PLoS Biology*, 10(1), e1001251.
- Patel, A. D. (2008). *Music, language, and the brain*. New York: Oxford University Press.
- Peelle, J. E., Gross, J., & Davis, M. H. (2013). Phase-locked responses to speech in human auditory cortex are enhanced during comprehension. *Cerebral Cortex*, 23(6), 1378–1387.
- Power, A. J., Foxe, J. J., Forde, E. J., Reilly, R. B., & Lalor, E. C. (2012). At what time is the cocktail party? A late locus of selective attention to natural speech. *European Journal of Neuroscience*, 35(9), 1497–1503.
- Ross, B., Tremblay, K. L., & Picton, T. W. (2007a). Physiological detection of interaural phase differences. *The Journal of the Acoustical Society of America*, 121(2), 1017–1027.
- Ross, B., Fujioka, T., Tremblay, K. L., & Picton, T. W. (2007b). Aging in binaural hearing begins in mid-life: Evidence from cortical auditory-evoked responses to changes in interaural phase. *The Journal of Neuroscience*, 27(42), 11172–11178.

- Ross, B., Miyazaki, T., Thompson, J., Jamali, S., & Fujioka, T. (2014). Human cortical responses to slow and fast binaural beats reveal multiple mechanisms of binaural hearing. *Journal of Neurophysiology*, *112*(8), 1871–1884.
- Salminen, N. H., Tiittinen, H., Yrttiaho, S., & May, P. J. (2010). The neural code for interaural time difference in human auditory cortex. *The Journal of the Acoustical Society of America*, *127*(2), EL60–65.
- Scott, S. K., & McGettigan, C. (2013). The neural processing of masked speech. *Hearing Research*, *303*, 58–66.
- Scott, S. K., Rosen, S., Wickham, L., & Wise, R. J. S. (2004). A positron emission tomography study of the neural basis of informational and energetic masking effects in speech perception. *The Journal of the Acoustical Society of America*, *115*(2), 813–821.
- Scott, S. K., Rosen, S., Beaman, C. P., Davis, J. P., & Wise, R. J. S. (2009). The neural processing of masked speech: Evidence for different mechanisms in the left and right temporal lobes. *The Journal of the Acoustical Society of America*, *125*(3), 1737–1743.
- Shamma, S. A., Elhilali, M., & Micheyl, C. (2011). Temporal coherence and attention in auditory scene analysis. *Trends in Neurosciences*, *34*(3), 114–123.
- Shuai, L., & Elhilali, M. (2014). Task-dependent neural representations of salient events in dynamic auditory scenes. *Frontiers in Neuroscience*, *8*, 203.
- Simon, J. Z., Depireux, D. A., Klein, D. J., Fritz, J. B., & Shamma, S. A. (2007). Temporal symmetry in primary auditory cortex: Implications for cortical connectivity. *Neural Computation*, *19*(3), 583–638.
- Snyder, J. S., Alain, C., & Picton, T. W. (2006). Effects of attention on neuroelectric correlates of auditory stream segregation. *Journal of Cognitive Neuroscience*, *18*(1), 1–13.
- Snyder, J. S., Gregg, M. K., Weintraub, D. M., & Alain, C. (2012). Attention, awareness, and the perception of auditory scenes. *Frontiers in Psychology*, *3*, 15.
- Stecker, G. C., Harrington, I. A., & Middlebrooks, J. C. (2005). Location coding by opponent neural populations in the auditory cortex. *PLoS Biology*, *3*(3), e78.
- Sussman, E. S., Chen, S., Sussman-Fort, J., & Dinces, E. (2014). The five myths of MMN: Redefining how to use MMN in basic and clinical research. *Brain Topography*, *27*(4), 553–564.
- Szalardy, O., Bohm, T. M., Bendixen, A., & Winkler, I. (2013). Event-related potential correlates of sound organization: Early sensory and late cognitive effects. *Biological Psychology*, *93*(1), 97–104.
- Teki, S., Chait, M., Kumar, S., von Kriegstein, K., & Griffiths, T. D. (2011). Brain bases for auditory stimulus-driven figure-ground segregation. *The Journal of Neuroscience*, *31*(1), 164–171.
- Thompson, S. K., von Kriegstein, K., Deane-Pratt, A., Marquardt, T., et al. (2006). Representation of interaural time delay in the human auditory midbrain. *Nature Neuroscience*, *9*(9), 1096–1098.
- van Noorden, L. P. A. S. (1975). *Temporal coherence in the perception of tone sequences*. PhD dissertation, Eindhoven University of Technology.
- von Kriegstein, K., Griffiths, T. D., Thompson, S. K., & McAlpine, D. (2008). Responses to interaural time delay in human cortex. *Journal of Neurophysiology*, *100*(5), 2712–2718.
- Wiegand, K., & Gutschalk, A. (2012). Correlates of perceptual awareness in human primary auditory cortex revealed by an informational masking experiment. *NeuroImage*, *61*(1), 62–69.
- Wilson, E. C., Melcher, J. R., Micheyl, C., Gutschalk, A., & Oxenham, A. J. (2007). Cortical fMRI activation to sequences of tones alternating in frequency: Relationship to perceived rate and streaming. *Journal of Neurophysiology*, *97*(3), 2230–2238.
- Xiang, J., Simon, J., & Elhilali, M. (2010). Competing streams at the cocktail party: Exploring the mechanisms of attention and temporal integration. *The Journal of Neuroscience*, *30*(36), 12084–12093.
- Zion Golumbic, E. M., Ding, N., Bickel, S., Lakatos, P., et al. (2013). Mechanisms underlying selective neuronal tracking of attended speech at a “cocktail party”. *Neuron*, *77*(5), 980–991.

Chapter 8

Infants and Children at the Cocktail Party

Lynne Werner

Abstract The vast majority of children learn language despite the fact that they must do so in noisy environments. This chapter addresses the question of how children separate informative sounds from competing sounds and the limitations imposed on such auditory scene analysis by an immature auditory nervous system. Immature representation of auditory-visual synchrony, and possibly immature binaural processing, may limit the extent to which even school-age listeners can use those sources of information to parse the auditory scene. In contrast, infants have a relatively mature representation of sound spectrum, periodicity, and temporal modulation. Although infants and children are able to use these acoustic cues in auditory scene analysis, they are less efficient than adults at doing so. This lack of efficiency may stem from limitations of the mechanisms specifically involved in auditory scene analysis. However, the development of selective attention also makes an important contribution to the development of auditory scene analysis.

Keywords Attention · Auditory development · Children · Hearing · Infants · Masking

8.1 Introduction

Despite adults' remarkable ability to follow one conversation among several, few would endorse the cocktail party as a supportive environment for young listeners still in the process of learning about sound in general and about spoken language in particular. Nonetheless, recent studies suggest that children in developed societies are frequently exposed to acoustic environments rivaling the cocktail party in complexity. For example, the American Time Use Survey reports that the average

L. Werner (✉)

Department of Speech and Hearing Sciences, University of Washington,
1417 N.E. 42nd Street Seattle, Washington 98105-6246, USA
e-mail: lawerner@uw.edu

adult caregiver spends 1 or 2 h per day in direct care of a child younger than 6 years of age, but another 5 h per day keeping an eye on that child while engaged in other activities such as housework, conversing with another adult, listening to music, or watching television (Bureau of Labor Statistics 2014). In the classroom, competing sound, particularly “joint babble,” has been reported to have a detrimental effect on children’s speech perception (Prodi et al. 2013). Even day-care centers can be noisy places (Truchon-Gagnon 1988). Thus, the child is exposed to an ongoing background of sound during most waking hours.

Auditory scene analysis is the process by which we analyze complex sounds and group together and follow the spectral components coming from each source. This process likely involves specialized mechanisms for identifying coherence across auditory channels and over time, and immaturity of those mechanisms could limit infants’ or children’s ability to separate concurrent sounds. However, representation of spectral, temporal, and spatial properties of sound provides the basis for this process. Thus, the accuracy of auditory coding is a critical aspect of auditory scene analysis. Furthermore, although the precise role that selective attention plays in auditory scene analysis is still debated, that it plays a role is not. Development of selective attention, therefore, would be expected to contribute to the development of auditory scene analysis. The evidence bearing on the contributions of each of these processes is detailed in the sections that follow.

8.2 Development of Auditory Coding

8.2.1 *Spectral Resolution and Energetic Masking*

The representation of the amplitude spectrum of sound is fundamental to many aspects of hearing. With respect to the separation of competing sounds, it plays two major roles. First, spectral resolution determines how much energetic masking occurs. Energetic masking is the type of masking that occurs when the same peripheral neural elements respond to both the target sound and the competing sound (Culling and Stone, Chap. 3). Thus, the extent to which the representations of the spectra of two sounds overlap determines how much energetic masking occurs. Second, spectral resolution limits the accuracy of the representation of the shape of the amplitude spectrum of a sound. Spectral shape is a major determinant of timbre, and it contributes to sound localization. Both timbre and spatial location are potential cues for auditory scene analysis.

Beginning with the seminal study by Schneider et al. (1989), many studies have shown that masked thresholds decrease progressively between infancy and the school years (e.g., Berg and Boswell 1999; Buss et al. 1999). Six-month-olds’ masked thresholds are about 10–12 dB higher than young adults’. By 4 years, the adult–child difference is no more than 8 dB, and by 8 years, children’s thresholds are within 3 dB of adults’.

Of the possible explanations for the age-related improvement in masked threshold—maturation of spectral resolution, maturation of intensity resolution, or maturation of other “processing abilities”—immature spectral resolution was identified as a factor only in the youngest infants. Both psychophysical (Spetner and Olsho 1990) and auditory brainstem response (ABR; Abdala and Folsom 1995; Folsom and Wynne 1987) measures indicate that, at 3 months, spectral resolution is immature at frequencies above 4000 Hz, but not at lower frequencies. Brainstem immaturity appears to be the limiting factor in early spectral resolution (Eggermont and Salamy 1988; Ponton et al. 1996; Abdala and Keefe 2012).

In older infants and children, studies of psychophysical tuning curves (Olsho 1985; Spetner and Olsho 1990), critical bandwidth (Schneider et al. 1990), and auditory filter widths (Hall and Grose 1991) are in agreement that spectral resolution is adultlike by 6 months of age. Although there is some evidence that the maturation of intensity resolution may play a role in the development of masked thresholds (Buss et al. 2006), several studies support the idea that age-related changes in processes loosely described as “attention” are important as well. Both infants and children as old as 6 years demonstrate masking of a tone by a noise band centered three octaves higher in frequency (Werner and Bargones 1991; Leibold and Neff 2011), even though spectral resolution is mature at this age. The idea that infants and young children do not listen selectively to improve their detection of a signal at an expected frequency, as adults do, was supported by subsequent studies (Bargones and Werner 1994; Jones et al. 2015). The effect certainly involves a failure to separate a target from a competing sound and may reflect a sort of informational masking, discussed in detail in a subsequent section (see also Kidd and Colburn, Chap. 4).

Despite the fact that spectral resolution appears to develop early in life, several studies of spectral shape discrimination by infants and children report immature performance. Infants appear to be capable of discriminating differences in spectral tilt (Clarkson 1996), but preschool children and many school-age children demonstrate immature spectral shape discrimination (Allen and Wightman 1992; Peter et al. 2014). Mature spectral shape discrimination has been reported for 12- to 18-year-old listeners (Peter et al. 2014). Thus, the mechanisms required to extract spectral shape may be in place early in life, but require years to reach adult status. However, because studies of spectral shape discrimination require “roving” the level of the stimulus from trial to trial, poor performance on this task may be accounted for by the fact that many children are confused about what to listen for when stimuli vary along multiple dimensions simultaneously, rather than by immature spectral shape processing.

8.2.2 Fundamental Frequency

Although the mechanisms underlying the effect are not completely understood, it is well established that differences in fundamental frequency (F_0) between competing

complex sounds improve a listener's ability to separate those sounds (e.g., Culling and Darwin 1993). It is believed that the auditory system uses harmonicity to group the components coming from one source (e.g., Deroche and Culling 2011). In the classic "concurrent vowel" paradigm (Summerfield and Assmann 1991), listeners are better able to identify the two vowels with even very small F0 differences if they have access to the low-frequency resolved components of the vowels. A larger F0 difference is required to produce a benefit if only high-frequency unresolved components of the vowels are available (Culling and Darwin 1993). The parallels with complex pitch perception suggest a dependence on a common mechanism that likely involves a temporal representation of periodicity.

Although many studies had suggested that infants are sensitive to variations in pitch (e.g., Fernald and Kuhl 1987), Clifton and her colleagues (Clarkson 1992) were the first to show that infants perceive the hallmark of complex pitch, the pitch of the "missing fundamental." They showed that 7-month-old infants discriminated between tonal complexes on the basis of F0 when the fundamental component was not present in the complexes, in the face of spectral variation in harmonic composition from presentation to presentation.

Recent work examining pitch perception in younger infants has revealed an interesting contrast between electrophysiological and psychophysical measures. He et al. (2007) identified mismatch responses¹ (MMRs) to a change in the fundamental frequency of a piano tone in infants as young as 2 months of age, but subsequently reported that only infants older than 4 months of age exhibited an MMR to a change in the direction of a pitch shift carried by a missing fundamental (He and Trainor 2009). This result appears to be consistent with two previous observations: First, a pitch-specialized region exists in primate secondary auditory cortex (Bendor and Wang 2010; Hall and Plack 2009). Second, auditory cortex is markedly immature in the early months of infancy (Eggermont and Moore 2012). However, Lau and Werner (2012, 2014) found that infants as young as 3 months of age respond behaviorally to missing fundamental frequency changes in complexes, even in complexes with only unresolved harmonics. It is not clear why 3-month-olds would not demonstrate an MMR to such changes. It is possible that subcortical processing is sufficient for a pitch percept. There are also changes in the morphology of the mismatch response in early infancy that may have obscured the response in younger infants (Eggermont and Moore 2012). The psychophysical results, in any case, suggest that harmonicity is a grouping cue available to the youngest infants.

¹A mismatch response is the difference between the response to a sound when it is presented frequently and the response to the same sound when it is the "oddball" in a sound sequence. In adults, the difference waveform is referred to as the mismatch negativity (MMN). However, the polarity of the response is positive in young infants, and many authors refer to the difference waveform as the mismatch response (MMR) when it is recorded in young infants. For simplicity, this response is referred to as the MMR throughout this chapter.

8.2.3 *Temporal Resolution*

Temporal cues—relative onset and coherent temporal modulation among spectral components—are strong cues for auditory scene analysis (e.g., Micheyl et al. 2013). In fact, temporal coherence is at the heart of many models of auditory scene analysis (e.g., Shamma et al. 2013; see also Elhilali, Chap. 5).

Unlike pitch perception, which in infants is better than predicted from their electrophysiological responses, detection of changes in a sound over time by infants and young children is far worse than their electrophysiological responses would suggest. The auditory steady-state response (ASSR) is an evoked potential that follows the envelope of amplitude modulated (AM) tones. ASSR amplitude increases and response threshold decreases during infancy, approaching adult values at about 12 months (Casey and Small 2014). In contrast, early studies of gap detection by infants and children as old as 4 years of age reported poor gap detection performance (Wightman et al., 1989; Werner et al., 1992). Studies of AM detection showed similar results, with young listeners requiring greater modulation depth to detect AM than adults (Hall and Grose 1994; Werner 2006a).

Hall and Grose (1994), however, measured the perceptual temporal modulation transfer function (TMTF) of children as young as 4 years of age. The results showed that while 4-year-olds needed a greater modulation depth to detect AM than adults did, the effect of modulation frequency on their AM thresholds was no different from that seen in adults. The results of TMTF studies of infants have been less conclusive. Although there is no obvious difference between adults and 3- to 6-month-olds in the shape of the TMTF, infants' poor AM detection makes it difficult to assess (Werner 2006b). Ongoing work using a somewhat different threshold estimation technique, however, suggests that infants' temporal resolution is relatively adultlike at 3 months of age (Horn et al. 2013). Moreover, 6-month-olds appear to be able to use relatively high envelope frequencies to discriminate between speech tokens in vocoded speech (Cabrera et al. 2013).

8.2.4 *Spatial Hearing*

Spatial location is probably the most frequently mentioned segregation cue in treatments of the cocktail party problem. It features prominently in models of auditory scene analysis (e.g., Darwin and Hukin 1999). Of all the basic auditory capacities, one would think that spatial hearing would demonstrate the most dramatic developmental change: Because an infant's head is small, the available binaural cues are compressed into a narrow range, limiting the accuracy with which sound objects can be localized. As the head and ears grow, the binaural and monaural acoustic cues to sound location change continually. Thus, the auditory system is likely required to recalibrate the mapping of acoustic cues onto space as long as growth continues. Of course, it is possible that interaural differences can be

used for stream segregation, even if those interaural differences are not accurately associated with a spatial location (Bronkhorst 2015).

The accuracy of sound localization increases dramatically between birth and 5 years of age. Behavioral measures of the minimum audible angle (MAA), a measure of the ability to discriminate between sounds coming from different spatial locations, improves from more than 25° at birth to less than 5° at 2 years (summarized by Litovsky 2012). By 5 years of age, the MAA appears to be adultlike. It should be noted, however, that increasing the complexity of the task by roving sound level, adding a simulated reflection, or bandpass filtering the sound leads to disproportionate increases in the MAA for children compared to adults (Grieco-Calub and Litovsky 2012; Kuhnle et al. 2013).

The extent to which immature sensitivity to interaural time or level cues contributes to early immaturity of sound localization accuracy is far from clear. The data bearing on this question are sparse. Ashmead et al. (1991) measured interaural time difference (ITD) discrimination thresholds in infants between 16 and 28 weeks of age. ITD discrimination thresholds were between 50 and 75 μ s, similar to values reported for naïve adults (Wright and Fitzgerald 2001; Middlebrooks et al. 2013). More recently, Van Deun et al. (2009) reported immature ITD discrimination thresholds among 4- to 9-year-old children compared to adults, but no apparent improvement in threshold with age among the children. With no data with respect to ILD sensitivity and so little data with respect to ITD sensitivity, it is difficult to draw any conclusions with respect to the availability of interaural cues for auditory scene analysis.

8.2.5 Auditory-Visual Correspondence

The addition of visual information improves masked speech perception over auditory-only presentation (e.g., Grant and Seitz 2000; Grant et al. 2007). More important in the current context, this effect is greater under conditions that challenge auditory scene analysis (Helfer and Freyman 2005; Wightman et al. 2006). Visual information also influences auditory streaming (Rahne et al. 2007; Rahne and Bockmann-Barthel 2009). To use visual information in this way, people must be able to recognize the correspondence between auditory and visual information.

While infants as young as 2 months of age appear to notice desynchronization of auditory and visual displays (e.g., Lewkowicz 1992), the duration of the temporal window over which auditory-visual (AV) “binding” occurs decreases during development. Lewkowicz (1996) reported that whereas adults detected AV asynchronies in a simple display of 65 ms when the visual event preceded the auditory event and 112 ms when the auditory event followed the visual event, infants showed evidence of detecting asynchronies only on the order of 350 ms and 450 ms, respectively. Interestingly, no effect of age was observed between 2 months and 8 months. The results of electrophysiological studies of infants are consistent with these behavioral results (Kopp 2014; Kopp and Dietrich 2013). The

subsequent developmental course of the temporal AV binding window is quite prolonged, extending into adolescence (Hillock-Dunn and Wallace 2012; Lewkowicz and Flom 2014). Thus, immature sensitivity to AV asynchrony may well limit children's ability to use AV correspondence as a cue in auditory scene analysis.

8.3 Development of Auditory Scene Analysis

Current models of auditory scene analysis distinguish between simultaneous and sequential grouping processes (e.g., Darwin and Hukin 1999). Once the ear has broken a complex sound into its component frequencies, the components coming from the same source are grouped together. This process is referred to as simultaneous grouping; it can be thought of as the process by which an auditory object is formed. However, once an auditory object is identified, the listener can follow that object as the sounds comprising it change over time. This process is referred to as sequential grouping; it can be thought of as the process by which an auditory stream is formed. Because different acoustic cues and different processing mechanisms support the two aspects of auditory scene analysis, it is possible that they develop along different trajectories.

The literature reviewed in Sect. 8.2 suggests that the nervous system has access to a relatively mature representation of spectrum, periodicity, and temporal modulation in sounds by 6 months of age. A mature representation of spatial location is likely not available before 5 years of age, and cross-modality temporal coherence may be accurately represented no earlier than adolescence. Thus, it is possible that a lack of precision in the neural representation of some aspects of sound is one factor that limits auditory scene analysis.

In general, researchers tend to believe, however, that if auditory scene analysis is immature in early life, the limitations likely arise in the central processes involved in grouping the components of sound coming from a common source and following the sound from a single source over time. As will be discussed in this section, there are certainly observations of immature auditory scene analysis based on acoustic cues that are accurately represented in a child's auditory system, consistent with that belief. Unfortunately, the available data are quite limited and fail to address the possible contributions of immature sound representation to the development of auditory scene analysis.

It is not uncommon for developmental researchers to use cortical auditory evoked potentials (CAEPs) as a measure of perception in infants and children with the rationale that an evoked potential may be a better indication of what a child is hearing than behavior, because behavior is influenced by many nonsensory variables. However, the implications of a demonstration of a neural response in an infant or child to some stimulus manipulation are often not clear. Although some developmental changes in neural response likely reflect refinement of the neural circuitry that will subservise auditory scene analysis in adults, in at least some cases

the neurons generating the response and the inputs that lead to that response change over the course of development (Eggermont and Moore 2012). Thus, it is difficult to assert that a neural response to a change specifying an auditory object or auditory stream reflects operation of adultlike neural circuitry. For ease of exposition, studies of auditory scene analysis that have used CAEPs as a dependent measure are included with behavioral responses to the same stimulus manipulation here. However, caution in the interpretation of the results of such studies is advised.

8.3.1 *Listening to Speech in Speech*

Infants and children demonstrate immature detection of speech in noise. Infants need a target-to-masker ratio (TMR) about 10 dB higher than that required by adults to detect or discriminate syllables or words in a background of broadband noise (Nozza et al. 1988; Werner 2013). By preschool or school age, threshold TMRs improve to only 3–4 dB worse than those of adults (Elliott et al. 1979; Nozza et al. 1988). These values are about the same as those reported for detection of a tone in noise (Werner and Leibold 2017).

Infants' and children's speech perception is even more immature in a background of competing speech than it is in broadband noise. For example, Newman and Jusczyk (1996) reported that 7.5-month-old infants recognized a familiar noun produced by a woman at 5 or 10 dB TMR, but not at 0 dB TMR, in the presence of speech produced by a male talker. In a subsequent study, Newman (2009) reported that 5-month-old infants did not recognize their own name produced by a woman at 10 dB TMR in the presence of speech produced by a different female talker. Adults achieve better-than-chance performance at –5 to –10 dB TMR in similar tasks. Age-related immaturity of speech-in-speech processing persists into the early school years: The ability to recognize syllables or spondees in the presence of two-talker speech improves progressively from 5 years to 10 years, reaching adult levels by 11 years (Hall et al. 2002; Leibold and Buss 2013). In contrast, 5- to 10-year-olds are close to adultlike in speech recognition in a background of speech-spectrum noise (Hall et al. 2002; Leibold and Buss 2013).

Wightman and his colleagues have tested children's performance on the coordinate response measure (CRM; Bolia et al. 2000), which requires a listener to follow a speech stream produced by a target voice in the presence of a competing speech stream. Wightman and Kistler (2005) found that when both talkers were male, performance in that task develops along a trajectory similar to that observed in studies of syllable or word identification at positive TMR. The ability to perform the CRM task at negative TMR continues to improve progressively from 6 to about 13 years of age. Thus, the ability to follow the less intense of two speech streams may take longer to become adultlike.

Leibold and Buss (2013) made an interesting observation that suggests that the sources of immature masked speech processing in young children are

fundamentally different for competing noise and competing speech. They analyzed the consonant confusions made by 5- to 7-year-old children and by adults identifying syllables in noise and in two-talker speech at TMRs that led to equivalent levels of performance for each age group. In speech-spectrum noise, children and adults showed a similar pattern of confusions and information received across consonant features (voicing, manner, place). In two-talker babble, however, the patterns were much less distinct for children than for adults.

In contrast, an error analysis of children's and adults' errors in the CRM task led to the conclusion that children and adults are limited by the same processes, but that children are, for unknown reasons, less efficient at using those processes. Wightman and Kistler (2005), following Brungart (2001), argued that adult performance in the CRM task was not limited primarily by energetic masking on the basis that adults report target words produced by the distractor talker more often than expected by chance. One interpretation is that the limitation for adults was in sequential rather than simultaneous grouping. When 6- to 8-year-olds performed above chance in the CRM task (i.e., at positive TMR), they showed the same tendency to report what the distractor talker said. One could argue, therefore, that young school-age children are qualitatively similar to adults in the way that they perform this task. However, whereas almost none of adults' errors, for example, at 0 dB TMR, were words other than those produced by the distractor, more than 20% of the young children's errors were other words. Thus, children may have difficulty with both with simultaneous and sequential grouping.

8.3.2 *Cues Used in Auditory Scene Analysis*

Infants, children, and adults can analyze an acoustic scene in at least qualitatively similar ways. For example, using the high-amplitude sucking habituation procedure (Siqueland and Delucia 1969), McAdams and Bertoncini (1997) tested newborns' ability to distinguish the order of sounds in a sequence in two conditions: Adults perceived the sequence in one condition as two streams of notes separated by timbre, pitch, and spatial location; adults perceived the sequence in the other condition as a single stream of notes. In both conditions, the sounds in the sequence were synthetic vibraphone or trumpet notes played to the infant's left and right, respectively. Like adults, newborns responded to a change in the order of notes in the two-stream sequence, but not in the one-stream sequence. On the basis of the parallels between adult and infant performance, it was concluded that newborns organize auditory streams as adults do.

Studies such as these suggest that auditory scene analysis mechanisms are in place early in life. However, without knowing which acoustic cues support infants' and children's performance in tasks such as these, it is difficult to conclude that their auditory scene analysis skills are mature. Children and adults may use different cues in simultaneous and/or sequential grouping. Children may require greater acoustic

differences than adults. Examination of the acoustic cues that young listeners use in auditory scene analysis may provide some insight into possible age-related changes in the mechanisms involved.

8.3.2.1 Frequency Separation

The role of frequency cues in auditory scene analysis has been assessed in two general paradigms. The first is the classic auditory streaming paradigm, in which listeners are asked to identify, in some fashion, the number of auditory streams heard as the frequency separation between the elements of a sequence is varied (Van Noorden 1975). The second is the informational masking paradigm, in which the listener detects a target tone in the presence of a tonal complex with variable frequency components (Neff and Green 1987).

Nearly all studies of the development of auditory streaming demonstrate that the processes based on frequency are operative at an early age, but few direct quantitative comparisons between age groups have been made. Some studies have taken the approach of McAdams and Bertonicini (1997), demonstrating that infants' responses to changes in sound sequences are consistent with adult reports of streaming, in this case for streams separated by frequency (Demany 1982; Fassbender 1993; Smith and Trainor 2011). Electrophysiological evidence also suggests that auditory streaming based on frequency is operative in newborn infants. Winkler et al. (2003) showed that newborns' and adults' MMRs to an infrequent change in the level of a tone depended on whether or not the frequencies of leading and following tones biased adults toward hearing two streams. In a study of 7- to 10-year-old children, Sussman et al. (2001) compared the children's MMRs to their behavioral streaming judgments using the same tone sequences as Winkler and colleagues. MMR were observed in children only for the sequence identified as "two streams" by children of the same age.

Only one quantitative evaluation of children's ability to use frequency separation to form auditory streams has been reported, and its results are quite interesting. Sussman et al. (2007) examined auditory stream judgments over a range of frequency separations for school-aged children and for adults. In the standard ABA streaming paradigm (Van Noorden, 1975), 9- to 11-year-old children, like adults, nearly always reported that the two streams were separated with frequency separations of 11 semitones. Five- to 8-year-old children were much less likely to separate the two streams, even at frequency separations as great as 23 semitones. These findings suggest that although streaming is operative at birth, it is not mature until well into the school years.

The informational masking paradigm is another way of assessing a listener's ability to form auditory streams on the basis of frequency (Kidd and Colburn, Chap. 4). In this case, the listener's task is to detect a target tone at one frequency in the presence of a masker with multiple tonal components with frequencies that vary randomly from presentation to presentation. The masker component frequencies are chosen to fall outside the auditory filter centered on the target frequency.

Informational masking is generally thought of as a failure to perceptually separate the target from the masker. Manipulations of the stimulus that are known to promote auditory stream formation reduce informational masking (e.g., Neff 1995; Kidd et al. 2003).

Informational masking has been measured for infants and for children. These studies uniformly report that the difference between young listeners and adults is greater in informational masking than in energetic masking. The reported age difference in informational masking depends on the specifics of the masker, with estimates ranging from 20 to 40 dB in infants and preschool children (Oh et al., 2001; Leibold and Werner 2007), dropping to about 12 dB in school-age children. One interesting phenomenon in infants and young children is that the difference between young listeners and adults is about the same whether or not the components of a multitone masker vary in frequency (Leibold and Neff 2007; Leibold and Werner 2007).

As was the case in the auditory streaming paradigm, however, children's performance in the informational masking paradigm is qualitatively similar to that of adults in several respects. For example, the amount of informational masking observed is a nonmonotonic function of the number of masker components, with a peak around 10–20 components (Oh et al. 2001). Increasing the duration of the masker so that its onset precedes that of the target leads to a reduction in informational masking for both children and adults, although children derive less benefit than adults do from this manipulation (Hall et al. 2005; Leibold and Neff 2007). When the target tone is presented repeatedly in a background of repeated random-frequency masker bursts (Kidd et al. 2003), informational masking is reduced by about the same amount in children and adults (Leibold and Bonino 2009). Finally, large individual differences in the amount of informational masking are evident in both children and adults (e.g., Hall et al. 2005; Leibold and Neff 2007).

In summary, infants and children have greater difficulty separating a target tone from a multitone masker, even when there is no spectral overlap between the target and masker and when the frequencies in the masker do not vary. In fact, infants and young children may exhibit masking of a pure tone target by a fixed-frequency noise-band masker centered more than two octaves above the target frequency (Werner and Bargones 1991; Leibold and Neff 2011). Thus, both the auditory streaming and the informational masking studies indicate that the ability to separate auditory streams based on frequency separation develops slowly and well into the school years.

8.3.2.2 Timbre

Timbre, or spectral shape, is a cue that can be used in both simultaneous (e.g., Assman and Summerfield 1990) and sequential (Cusack and Roberts 2000; Bey and McAdams 2003) grouping. Because preschool and young school-age children seem to be less sensitive than adults to changes in spectral shape, whether they use this

cue in auditory scene analysis would be of some interest. Unfortunately, only one study has examined the development of the ability to use timbre as a cue in auditory scene analysis. Fassbender (1993) used a high-amplitude sucking habituation procedure to test 2- to 5-month-olds' ability to discriminate a reversal in the order of tones in one sequence when it was interleaved with a second sequence of tones in the same frequency range. When the tones in both sequences were pure tones, infants did not discriminate the change in order. When the tones in one sequence were complex tones, infants did discriminate the change in order. The implication is that infants can use timbre differences to form auditory streams. Whether they are able to use more subtle differences in timbre as a grouping cue remains to be shown.

8.3.2.3 Periodicity

Periodicity is a strong cue in both simultaneous (e.g., Assman and Summerfield 1990) and sequential (e.g., Darwin et al. 2003) grouping. Furthermore, even young infants seem to be as sensitive as adults to differences in periodicity (Lau and Werner 2012, 2014). One might predict, then, that the ability to use this cue in auditory scene analysis would appear at an early stage of development.

One measure of the use of periodicity in simultaneous grouping is the ability to detect mistuning of one harmonic in a complex tone. Folland et al. (2012) trained infants to detect 8% mistuning of the third harmonic of a complex with a 200-Hz F0, then tested their ability to detect a smaller mistuning. Infants' appeared to detect mistuning as small as 4%; many adults were able to detect mistuning as small as 1%. Similarly, Alain et al. (2003) reported that 8- to 12-year-old children were a little worse than adults in detecting a mistuning of the third harmonic of a complex with a 200-Hz F0. The extent to which these small age-related differences reflect performance differences rather than immature perceptual grouping is not clear. In any case, one might conclude that infants and children are capable of using harmonicity as a cue in simultaneous grouping.

Although children's ability to separate simultaneous sentences spoken by different talkers has been examined in a few studies (e.g., Wightman and Kistler 2005), to date, the relative contributions of fundamental frequency, vocal tract length, and their interaction (Darwin et al. 2003) to children's auditory scene analysis have not been evaluated.

8.3.2.4 Envelope Cues

Temporal envelopes—synchrony in onset time as well as in ongoing amplitude modulation—provide one of the strongest cues for simultaneous grouping (e.g., Culling and Stone, Chap. 3). As cues for auditory scene analysis go, temporal cues have been among the most studied developmentally.

Studies of infants have focused on the ability to take advantage of modulation of the masker to improve detection or discrimination. For example, Newman (2009)

compared 5- and 8.5-month-old infants' recognition of their own name in a competing single-talker masker to that in a competing 9-talker masker. Single-talker speech is more highly modulated than nine-talker speech, making it possible for a listener to give greater weight to information at low-amplitude portions of the single-talker masker where the TMR is higher. At the same time, spectral variability is greater in single-talker speech than in nine-talker speech, raising the possibility of greater informational masking by single-talker speech. In adults, speech reception thresholds are lower in single-talker than in multitalker maskers (e.g., Drullman and Bronkhorst 2004), suggesting that the advantage of modulation wins out over the detrimental effects of spectral variability. Infants show just the opposite pattern of results. Newman found that infants recognized their names at 10 dB TMR in the nine-talker masker, but in neither a single-talker masker nor a time-reversed version of the same single-talker masker. Newman suggested that masker modulation may distract infants, but whether it is the modulation envelope or the spectral variability of single-talker speech that is distracting is not clear. However, Werner (2013) found that 7-month-old infants' vowel detection or discrimination was poorer in a single-talker modulated noise than in an unmodulated noise when the vowels were presented at what should have been a clearly audible level. Because spectral variability could not be an issue in this case, this result suggests that modulation in competing sound distracts infants. Clearly, additional work will be required to determine how these effects translate into auditory scene analysis in more realistic acoustic environments.

Children as old as 10 years of age have also been reported to show less benefit of masker modulation in tone detection than adults do (Grose et al. 1993). Most studies, however, have examined children's speech perception in modulated noise (e.g., Stuart 2008; Hall et al. 2014). The reported performance differences between modulated and unmodulated masker conditions for children vary widely across studies; the reported age at which masking release is adultlike varies from before 6 years to after 11 years. The variability in results could be due to differences in the difficulty of the speech materials used, as well as the type and frequency of masker modulation. It is possible that the ability to use temporal information in auditory scene analysis is not limited by auditory capacity per se, but rather by cognitive processes such as working memory.

As noted previously, Leibold and Neff (2007) showed that a temporal offset between target and masker reduced informational masking for both children and adults. Children, however, demonstrated less of a release from masking than adults did. Similarly, Hall et al. (2005) found that the addition of leading masker bursts in Kidd and colleagues' (1994) informational masking paradigm reduced masking less for children than for adults. Thus, children seem to use temporal onset cues as adults do to separate target and masker, but again, they are less efficient at the process than adults are.

Comodulation masking release (CMR; Hall et al., 1984) is another paradigm that taps the processes underlying the use of temporal information in separating sounds. Briefly stated, CMR is the advantage in tone detection in modulated noise that results from the addition of off-frequency noise with modulation matching that in

the on-frequency masker. In general, children aged 5–11 years have been reported to have the same CMR as adults (e.g., Hall et al. 1997; but see Zettler et al. 2008). In fact, Hall and colleagues showed that adding two noise bands sharing the same modulation, but differing in modulation from the original “comodulated” noise bands, reduced CMR by the same amount for children and adults. However, creating an asynchrony between the on-frequency noise and the off-frequency noise reduced CMR for children much more than it did for adults. The implication may be that in complex environments in which auditory scene analysis requires the integration of several sorts of temporal information, school-aged children remain at a disadvantage relative to adults.

In summary, in many situations, infants and children appear to use temporal information to separate concurrent sounds, but they often do so less efficiently than adults. The degree to which children’s auditory scene analysis may benefit from temporal cues appears to depend on the nature of the material they are asked to process, as well as the complexity of the acoustic environment.

8.3.2.5 Spatial Cues

Spatial cues are an important source of information for auditory scene analysis, as discussed in previous chapters, particularly Chaps. 3 (Culling and Stone), 4 (Kidd and Colburn), and 6 (Middlebrooks). Spatial acuity matures over the course of infancy and the preschool years (Litovsky 2012), but under more difficult listening conditions may remain immature into the school years (Litovsky 1997). It is possible, then, that spatial acuity limits infants and children’s ability to use spatial information in auditory scene analysis. However, the available literature suggests that in at least some situations, children derive the same benefit from spatial separation between competing sounds as adults do. Children’s ability to use spatial cues in auditory scene analysis has been addressed in developmental studies of the binaural masking level difference (MLD), of masking by contralateral sounds, and of spatial unmasking in sound field.

Masking Level Difference

The MLD is the difference between two thresholds: The threshold for a tone in noise when tone and noise are presented in phase to the two ears, designated N_0S_0 , and the threshold for a tone in noise when either the tone or the noise is presented 180° out of phase in one ear relative to the other ear, designated N_0S_π and $N_\pi S_0$, respectively (Culling and Stone, Chap. 3). The largest MLD, about 15 dB, is observed for tones below 1000 Hz in frequency in the N_0S_π configuration. The MLD results primarily from a difference between the interaural correlation associated with the tone and the noise, respectively. The MLD has been examined in infants as young as 6 months of age as well as in school-aged children.

Nozza (1987) reported that 6- to 11-month-old infants had a smaller MLD for a 500-Hz tone than adults, about 5 dB compared to an adult value of 10 dB. One problem with the interpretation of this result is that it is difficult to match the level of sound in two earphones for infants; level differences between the ears reduce the size of the MLD (Egan 1965). However, in a sound-field paradigm Schneider et al. (1988) confirmed that 12-month-olds obtained less of a benefit of interaural differences than adults did in detecting a broadband noise.

Although no MLD data have been reported for children between the ages of 1 and 4 years, the MLD of older children has been estimated in several studies. In wideband noise, the MLD has been reported to be immature at 4 years, but adultlike at 5 years (Van Deun et al. 2009; Moore et al. 2011). Hall and colleagues have reported that in narrowband noise, children's MLD is not adultlike until 8–10 years (e.g., Hall et al. 2004). Their results suggest that, unlike adults, younger children are unable to take advantage of brief periods of interaural decorrelation that occur in minima of the envelope of narrow noise bands that allow adults to achieve better thresholds in the N_0S_π condition (Hall et al. 2004). Other results suggest that children's difficulty lies in immature binaural temporal resolution (Hall et al. 2007).

Contralateral Masking

If adults are asked to detect a tone or to report speech presented to one ear, presenting a masker simultaneously to the ear contralateral to the target sound has little effect on performance. That is true whether the masker is noise, speech, or a randomly varying multitone complex (e.g., Brungart and Simpson 2004; Wightman and Kistler 2005). Such effects have been examined in older preschool and school-aged children. Wightman et al. (2003) found that, in contrast to adults, 4- to 5-year-olds demonstrated nearly as much informational masking when a randomly varying multitone complex masker was presented to the ear contralateral to a target tone as they did when the masker was presented to the same ear as the target. Between 6 and 16 years, the number of children demonstrating informational masking with a contralateral masker decreased. None of the listeners showed contralateral masking by a broadband noise. Similarly, school-age children show substantial contralateral masking of speech by speech maskers, while adults do not (Wightman and Kistler 2005).

Spatial Release from Masking

The potentially most informative type of study of the ability to use spatial cues in auditory scene analysis is the spatial release from masking (SRM) study, which compares speech perception when the masker is co-located with the target speech with that when the masker comes from a different spatial location. The development of SRM, at least in preschoolers and older children, has been addressed in many

studies, with varying results. Recent articles provide good reviews and summaries of these studies (e.g., Ching et al. 2011; Yuen and Yuan 2014).

Given the variability in materials and procedures used in these studies, it is not surprising that the results are variable. Children have been asked to identify single words and sentences presented in pink noise, speech spectrum noise, noise matched to speech in spectrum and envelope, one-talker speech, two-talker speech, and two different four-talker maskers presented simultaneously. The ages and range of ages tested also vary widely across studies.

In general, studies in which listeners were asked to identify single words in noise or speech maskers tend to show about the same SRM in children and adults, for children as young as 18 months of age (e.g., Litovsky 2005; Murphy et al. 2011; cf. Yuen and Yuan 2014). Studies in which children were asked to repeat back sentences tend to show that the amount of SRM increases with age; children achieve adult values of SRM around 8 years in some cases (e.g., Cameron and Dillon 2007) but not until 12 years in others (e.g., Vaillancourt et al. 2008). One hypothesis, then, is that task difficulty moderates children's ability to use spatial information to separate speech targets from competing sounds.

An interesting trend in this literature is that under some conditions children actually get a greater SRM than do adults. For example, Litovsky (2005) found that 4- to 8-year-old children had about the same SRM as adults when asked to identify words in speech-shaped noise or in a one-talker masker but actually had greater SRM than adults in a two-talker masker. Similarly, Johnstone and Litovsky (2006) found that children in the same age range had greater SRM than adults when they were asked to identify words in one-talker speech or in time-reversed one-talker speech. A possible explanation is that spatial separation between target and masker is most beneficial under conditions in which informational masking predominates, and that because children are more susceptible than adults to informational masking, they show greater benefits of that spatial separation.

Summary

It appears that around the time that children achieve adultlike spatial acuity, they also achieve adultlike ability to use binaural cues to separate a tone from a noise, as indicated by the developmental studies of the MLD. It appears that even before this age children are also able to use spatial information to separate a word from a variety of competing sounds. However, when the task is a difficult one, either because segregation is difficult as in the CRM task (e.g., Wightman and Kistler 2005) or because they are required to report an entire sentence (e.g., Cameron and Dillon 2007; Vaillancourt et al. 2008), they derive less benefit than adults do from spatial information. In such cases, immaturity of processes such as selective attention and working memory may set the limit on how well children can perform.

8.3.2.6 Visual Cues

Access to visual as well as auditory speech significantly enhances speech perception for adults, especially in noisy environments (e.g., Sumby and Pollack 1954; Grant and Seitz 2000). Because the window of auditory-visual temporal integration continues to narrow with age into adolescence (e.g., Hillock-Dunn and Wallace 2012; Lewkowicz and Flom 2014), the ability to use visual information in auditory scene analysis might be expected to improve over a similarly long period. However, it is also possible that the mechanism responsible for an audiovisual advantage (AVA) changes over the course of development. Visual speech provides information about the timing and about the amplitude envelope of auditory speech, but it also carries articulatory information that can be integrated with the auditory signal at a phonetic level (Massaro, 1998). Temporal effects and phonetic integration effects may have different auditory-visual integration duration windows, and such differences have not been addressed developmentally.

It may be that infants, who have limited experience with speech, derive a benefit from knowing when to listen (Werner et al. 2009), but are unable to take advantage of the additional temporal and phonetic information provided by the visual signal. Lewkowicz (2010) found that infants were sensitive to the synchrony between auditory and visual sound onsets, but not to the ongoing temporal correlation between the auditory and visual events. Although there is evidence that infants as young as 2 months of age prefer to look at the talking head of a person saying a vowel that they also hear (Kuhl and Meltzoff 1982; Patterson and Werker 2003), the evidence that they match heard and seen consonant productions is weak (MacKain et al. 1983). Thus, it appears that at least early in infancy, infants' sensitivity to the correspondence of auditory and visual speech is limited. It is, therefore, not surprising that there is little in the way of convincing evidence for AV integration in speech perception by infants (e.g., Desjardins and Werker 2004).

Hollich et al. (2005) performed the only study of AVA in infants. Infants aged 7.5 months were familiarized with a target word in a background of running single-talker speech at 0 dB TMR and subsequently tested for recognition of that word in quiet. Infants showed evidence of recognizing the target word when visual information was available during familiarization, but only when the visual and auditory stimuli were synchronized. That result suggests that synchronized visual information helped them to segregate the target word from the masker speech. Based on the Lewkowicz (2010) results, it might be concluded that this effect was due primarily to a reduction in temporal uncertainty.

There is evidence, on the other hand, that older children integrate auditory and visual phonetic information to some extent. For example, Massaro (1984; Massaro et al. 1986) reported that 4-year-old children integrate auditory and visual information in identifying consonants in a qualitatively adultlike way, but that they were less influenced by the visual information than adults were. Studies of the McGurk effect, in which mismatched auditory and visual syllables give rise to a percept that matches neither the auditory nor visual syllable presented, are consistent with this view: Children reported the "unheard and unseen" syllable much less frequently

than adults and often reported that the auditory and visual stimuli didn't match (McGurk and MacDonald 1976). Some evidence suggests that experience producing speech is required to produce a strong McGurk effect (Desjardins et al. 1997). If the ability to integrate auditory and visual speech increases over the course of childhood, then the benefit derived from that integration in the case in which auditory speech is degraded by competing sounds should also increase.

There are surprisingly few studies of older children's AVA for speech perception in the presence of a competing sound. Holt et al. (2011) found that 3-, 4-, and 5-year-olds were able to repeat key words in sentences presented in speech-spectrum noise better with AV presentation than with auditory-only presentation. The AVA, in terms of percent correct words reported, increased between 3 years and 4 years. Using the same stimuli, Lalonde and Holt (2015) also reported that 3- and 4-year-olds were able to take advantage of visual information, although the advantage was significant only for visually salient speech contrasts.

Ross et al. (2011) compared 5- to 14-year-olds' and adults' audiovisual monosyllabic word identification in pink noise to that in an auditory-only and visual-only conditions. AVA increased with age, with 12- to 14-year-olds approaching adult levels, although auditory-only performance did not vary with age. The magnitude of the AVA correlated with performance in the visual-only condition in children, but not in adults, suggesting that the quality of the representation of visual speech is an important limitation on AVA during childhood. Ross et al. also reported an age-related change in the effect of TMR on the AVA advantage, a possible indication that maturation of the integration process is also involved.

The results of one study suggest that the AVA may be much more immature in children in more complex listening conditions. Wightman et al. (2006) compared adults to children ranging from 6 to 17 years in the CRM task with a single target talker and a single distractor in one ear, with and without video of the talker speaking the target sentence. The youngest children, 6–9 years old, showed no AVA. The size of the AVA increased with age; even the oldest children did not have an adultlike AVA. At least on a group basis, the AVA seemed to covary with the ability to perform the task with the visual stimulus alone.

8.3.3 Role of Selective Attention

Whether or not selective attention is necessary for the formation of auditory streams, it is clear that selective attention affects stream formation (e.g., Shamma et al. 2011; Akram et al. 2014). Thus, it is reasonable to ask how the development of selective auditory attention relates to the development of auditory scene analysis.

A casual search reveals many studies addressing the development of auditory attention. Interestingly, the tasks used in most of these studies are indistinguishable from those described in previous sections of this chapter: Multiple sound streams are presented simultaneously, and the listener is asked to report in some fashion what was heard in one of the streams. To the extent that listeners can perform such a

task, it is clear that they had a sufficient sensory representation of the sound, could correctly group the components from each source, were able to form the stream emanating from the target source, and could focus attention on the correct source. If a listener cannot perform such a task, the nature of the deficit is often unclear. Did a young child who reports what the wrong talker said in the CRM task get his streams crossed or was he unable to maintain attention on the correct talker?

Studies employing such methods almost universally report that performance improves with age, in some cases into adolescence (Leibold 2012). Most, however, report quantitative, but not qualitative, age-related differences: Children perform more poorly than adults, but manipulations of the stimuli and task tend to have the same effect on performance at each age (e.g., Doyle 1973; Cherry 1981). Such a pattern of results makes it difficult to distinguish immature attention from other immature processes.

Recent attempts to characterize the components of attention may provide a useful way to approach its development. Petersen and Posner (2012) summarize a model in which attention has three components, each subserved by its own neural network. The alerting network maintains an optimal level of arousal to perform a task. The orienting network prioritizes and selects sensory inputs. The executive control network maintains focal attention on a given input subject to top-down influences. On the basis of performance in visual tasks, it is argued that during infancy attention is based largely on the orienting network (Clohessy et al. 2001; Rothbart et al. 2011). Studies of older children indicate no change in the orienting network beyond 6 years. The executive control network comes online during the preschool period (Rothbart et al. 2003, 2011), develops rapidly first between 6 and 7 years and then again in later childhood (Rueda et al. 2004). The alerting network does not appear to be adultlike until sometime after 10 years of age (Rueda et al. 2004). Note, however, that the developmental trajectories of visual and auditory attention may differ (Gunther et al. 2014).

Studies of the development of auditory attention are consistent with an improvement in executive control around the transition to school age. For example, Bartgis et al. (2003) asked 5-, 7-, and 9-year-old children to detect a rapid sequence of five alternating-frequency tones in a series of longer duration, slower tones at a fixed frequency while an event-related potential (ERP) was recorded. Tones were presented to both ears, but children were instructed to respond only to the target sequence in one ear. Performance in the task improved with age, although even the oldest children were likely to respond to the target in the wrong ear. However, older children showed larger amplitude P300 responses to the target in the attended ear, whereas 5-year-olds showed equal responses in attended and unattended ears.

At least one study suggests that even school-age children have difficulty controlling allocation of attention. Choi et al. (2008) asked 7- to 14-year-old children to report words in noise while simultaneously remembering a sequence of four digits. Most children showed no evidence of being able to prioritize one task over the other when instructed to do so, consistent with a lack of executive control. However, 7- to 10-year-olds actually performed better in word recognition in the dual-task

condition than when they performed word recognition alone, suggesting some benefit of engaging the executive control network. Along the same lines, an ERP study by Gomes et al. (2000) found that 8- to 12-year-olds' MMR was larger when they attended to a deviant tone than when they listened passively, but only when the deviant was difficult to discriminate from the standard.

An interesting series of studies examined selective attention using ERP and behavioral responses of children and adults in a more naturalistic situation, in which auditory scene analysis would be involved (e.g., Sanders et al., 2006; Karns et al., 2015). The ERP to the probes embedded in an attended story is compared to the ERP to probes embedded in a simultaneously presented unattended story. In adults, the amplitude of the ERP elicited by the probes is greater in the attended story than in the unattended story. A similar effect is observed in children as young as 3 years of age, although in young children the polarity, scalp distribution, and the effects of manipulating the type of probe suggest differences between children and adults in the underlying neural circuitry. It is noteworthy that special efforts were made in these studies to provide additional cues, such as pictures that accompanied the to-be-attended story, to ensure that young children could perform the task. Thus, it may be that providing external orienting cues supports the young child's ability to attend selectively. A transition to a more adultlike attention-related response occurs between 10 and 13 years of age, but a complex pattern of change across age was observed during adolescence (Karns et al. 2015).

Sussman and Steinschneider (2009) made the most direct assessment of the role of attention on auditory streaming in school-aged children and adults. They recorded the MMR and P3b in response to an intensity deviant in one of two tone sequences, as the frequency difference between the two sequences was varied. Responses were recorded when subjects listened to the sequences passively and when they were asked to actively detect the intensity deviants. As in previous studies (Sussman et al. 2007), children required a somewhat larger frequency separation to detect the intensity deviant than adults did. In adults, the ERP to the deviants was the same in the passive and active listening conditions; moreover, the dependence of the ERP on frequency separation mirrored that observed behaviorally. In children, the ERP in the active listening condition also depended on frequency separation in a way that paralleled their behavioral response; however, the ERP appeared in the passive listening condition only when the frequency separation between sequences was very large (31 semitones). One interpretation of this result is that auditory scene analysis is more dependent on attention in children than it is in adults.

These studies suggest that selective attention may indeed be a limiting factor in the development of auditory scene analysis. Furthermore, it is likely that attentional effects will strongly interact with the salience of acoustic cues to sound source segregation, the availability of other (e.g., visual) cues, as well as task demands.

8.4 Summary, Conclusions, and Future Directions

Children generally have access to a sensory representation of sounds that is adequate to support auditory scene analysis, perhaps during infancy, but definitely by the time they are 5 years old. Nonetheless, in many situations, infants and young children have difficulty using those representations to separate competing sounds. While the acoustic cues that infants and children use to analyze the auditory scene have not been fully delineated, the existing literature indicates that they are sensitive to all of the cues that adults are known to use. However, when there are more than a couple of sound sources or when sounds are arranged in such a way that some acoustic cues are degraded, 5- or 6-year-olds may have tremendous difficulty processing a single auditory stream. The ability to deal with these more complex listening situations improves progressively over the course of childhood, and in some cases into adolescence. Taken together, the studies addressing the development of auditory scene analysis suggest that the neural circuitry underlying this ability is in place in some form at birth, although the maturation of the system extends through childhood.

It should be obvious, however, that this story is far from complete. First, although children appear to use the same acoustic cues that adults do to separate sounds, there has not been a systematic effort to examine a range of cues or to determine how the cues are weighted in the process of auditory scene analysis. The issue has not been well addressed in infancy, when immature sensory representations might be expected to limit the process. Second, systematic manipulations of the “complexity” of sound environments and of the difficulty of the tasks listeners are asked to perform would be extremely helpful in understanding the limitations children encounter in such environments. What makes a soundscape complex? What makes a perceptual task difficult? Finally, considerable information is now available about the development of visual attention in the context of current models. Extension of this approach to the development of auditory attention is most certainly warranted. A goal of research in this area might be to move beyond the statement, “Children have trouble hearing in noisy environments.”

References

- Abdala, C., & Folsom, R. C. (1995). Frequency contribution to the click-evoked auditory brain stem response in human adults and infants. *The Journal of the Acoustical Society of America*, *97*, 2394–2404.
- Abdala, C., & Keefe, D. H. (2012). Morphological and functional ear development. In L. A. Werner, A. N. Popper, & R. R. Fay (Eds.), *Human auditory development* (pp. 19–59). New York: Springer Science+Business Media.
- Akram, S., Englitz, B., Elhilali, M., Simon, J. Z., & Shamma, S. A. (2014). Investigating the neural correlates of a streaming percept in an informational-masking paradigm. *PLoS ONE*, *9*, e114427.
- Alain, C., Theunissen, E. L., Chevalier, H., Batty, M., & Taylor, M. J. (2003). Developmental changes in distinguishing concurrent auditory objects. *Cognitive Brain Research*, *16*, 210–218.

- Allen, P., & Wightman, F. (1992). Spectral pattern discrimination by children. *Journal of Speech Language and Hearing Research*, *35*, 222–233.
- Ashmead, D. H., Davis, D., Whalen, T., & Odom, R. (1991). Sound localization and sensitivity to interaural time differences in human infants. *Child Development*, *62*, 1211–1226.
- Assman, P. F., & Summerfield, Q. (1990). Modeling the perception of concurrent vowels: Vowels with different fundamental frequencies. *The Journal of the Acoustical Society of America*, *88*, 680–697.
- Bargones, J. Y., & Werner, L. A. (1994). Adults listen selectively; infants do not. *Psychological Science*, *5*, 170–174.
- Bartgis, J., Lilly, A. R., & Thomas, D. G. (2003). Event-related potential and behavioral measures of attention in 5-, 7-, and 9-year-olds. *Journal of General Psychology*, *130*, 311–335.
- Bendor, D., & Wang, X. Q. (2010). Neural coding of periodicity in marmoset auditory cortex. *Journal of Neurophysiology*, *103*, 1809–1822.
- Berg, K. M., & Boswell, A. E. (1999). Effect of masker level on infants' detection of tones in noise. *Attention, Perception, & Psychophysics*, *61*, 80–86.
- Bey, C., & McAdams, S. (2003). Postrecognition of interleaved melodies as an indirect measure of auditory stream formation. *Journal of Experimental Psychology-Human Perception and Performance*, *29*, 267–279.
- Bolia, R. S., Nelson, W. T., Ericson, M. A., & Simpson, B. D. (2000). A speech corpus for multitaler communications research. *The Journal of the Acoustical Society of America*, *107*, 1065–1066.
- Bronkhorst, A. W. (2015). The cocktail-party problem revisited: Early processing and selection of multi-talker speech. *Attention Perception & Psychophysics*, *77*, 1465–1487.
- Brungart, D. S. (2001). Informational and energetic masking effects in the perception of two simultaneous talkers. *The Journal of the Acoustical Society of America*, *109*, 1101–1109.
- Brungart, D. S., & Simpson, B. D. (2004). Within-ear and across-ear interference in a dichotic cocktail party listening task: Effects of masker uncertainty. *The Journal of the Acoustical Society of America*, *115*, 301–310.
- Bureau of Labor Statistics. (2014). *American time use survey*. Retrieved from <http://www.bls.gov/tus/>
- Buss, E., Hall, J. W., & Grose, J. H. (2006). Development and the role of internal noise in detection and discrimination thresholds with narrow band stimuli. *The Journal of the Acoustical Society of America*, *120*, 2777–2788.
- Buss, E., Hall, J. W., Grose, J. H., & Dev, M. B. (1999). Development of adult-like performance in backward, simultaneous, and forward masking. *Journal of Speech Language and Hearing Research*, *42*, 844–849.
- Cabrera, L., Bertocini, J., & Lorenzi, C. (2013). Perception of speech modulation cues by 6-month-old infants. *Journal of Speech Language and Hearing Research*, *56*, 1733–1744.
- Cameron, S., & Dillon, H. (2007). Development of the listening in spatialized noise-sentences test (lisn-s). *Ear and Hearing*, *28*, 196–211.
- Casey, K. A., & Small, S. A. (2014). Comparisons of auditory steady state response and behavioral air conduction and bone conduction thresholds for infants and adults with normal hearing. *Ear and Hearing*, *35*, 423–439.
- Cherry, R. S. (1981). Development of selective auditory attention skills in children. *Perceptual and Motor Skills*, *52*, 379–385.
- Ching, T. Y. C., van Wanrooy, E., Dillon, H., & Carter, L. (2011). Spatial release from masking in normal-hearing children and children who use hearing aids. *The Journal of the Acoustical Society of America*, *129*, 368–375.
- Choi, S., Lotto, A., Lewis, D., Hoover, B., & Stelmachowicz, P. (2008). Attentional modulation of word recognition by children in a dual-task paradigm. *Journal of Speech Language and Hearing Research*, *51*, 1042–1054.
- Clarkson, M. G. (1992). Infants' perception of low pitch. In L. A. Werner & E. W. Rubel (Eds.), *Developmental psychoacoustics* (pp. 159–188). Washington, DC: American Psychological Association.

- Clarkson, M. G. (1996). Infants' intensity discrimination; spectral profiles. *Infant Behavior and Development, 19*, 181–190.
- Clohessy, A. B., Posner, M. I., & Rothbart, M. K. (2001). Development of the functional visual field. *Acta Psychologica, 106*, 51–68.
- Culling, J. F., & Darwin, C. J. (1993). Perceptual separation of simultaneous vowels: Within and across-formant grouping by f₀. *The Journal of the Acoustical Society of America, 93*, 3454–3467.
- Cusack, R., & Roberts, B. (2000). Effects of differences in timbre on sequential grouping. *Attention, Perception, & Psychophysics, 62*, 1112–1120.
- Darwin, C. J., Brungart, D. S., & Simpson, B. D. (2003). Effects of fundamental frequency and vocal-tract length changes on attention to one of two simultaneous talkers. *The Journal of the Acoustical Society of America, 114*, 2913–2922.
- Darwin, C. J., & Hukin, R. W. (1999). Auditory objects of attention: The role of interaural time differences. *Journal of Experimental Psychology [Human Perception], 25*, 617–629.
- Demany, L. (1982). Auditory stream segregation in infancy. *Infant Behavior and Development, 5*, 261–276.
- Deroche, M. L., & Culling, J. F. (2011). Voice segregation by difference in fundamental frequency: Evidence for harmonic cancellation. *The Journal of the Acoustical Society of America, 130*, 2855–2865.
- Desjardins, R. N., Rogers, J., & Werker, J. F. (1997). An exploration of why preschoolers perform differently than do adults in audiovisual speech perception tasks. *Journal of Experimental Child Psychology, 66*, 85–110.
- Desjardins, R. N., & Werker, J. F. (2004). Is the integration of heard and seen speech mandatory for infants? *Developmental Psychobiology, 45*, 187–203.
- Doyle, A. B. (1973). Listening to distraction: A developmental study of selective attention. *Journal of Experimental Child Psychology, 15*, 100–115.
- Drullman, R., & Bronkhorst, A. W. (2004). Speech perception and talker segregation: Effects of level, pitch, and tactile support with multiple simultaneous talkers. *The Journal of the Acoustical Society of America, 116*, 3090–3098.
- Egan, J. P. (1965). Masking-level differences as a function of interaural disparities in intensity of signal and of noise. *The Journal of the Acoustical Society of America, 38*, 1043–1049.
- Eggermont, J. J., & Moore, J. K. (2012). Morphological and functional development of the auditory nervous system. In L. A. Werner, A. N. Popper, & R. R. Fay (Eds.), *Human auditory development* (pp. 61–105). New York: Springer Science+Business Media.
- Eggermont, J. J., & Salamy, A. (1988). Maturation time course for the abr in preterm and full term infants. *Hearing Research, 33*, 35–48.
- Elliott, L. L., Connors, S., Kille, E., Levin, S., et al. (1979). Children's understanding of monosyllabic nouns in quiet and noise. *The Journal of the Acoustical Society of America, 66*, 12–21.
- Fassbender, C. (1993). *Auditory grouping and segregation processes in infancy*. Norderstedt, Germany: Kaste Verlag.
- Fernald, A., & Kuhl, P. (1987). Acoustic determinants of infant perception for motherese speech. *Infant Behavior and Development, 10*, 279–293.
- Folland, N. A., Butler, B. E., Smith, N. A., & Trainor, L. J. (2012). Processing simultaneous auditory objects: Infants' ability to detect mistuning in harmonic complexes. *The Journal of the Acoustical Society of America, 131*, 993–997.
- Folsom, R. C., & Wynne, M. K. (1987). Auditory brain stem responses from human adults and infants: Wave v tuning curves. *The Journal of the Acoustical Society of America, 81*, 412–417.
- Gomes, H., Molholm, S., Ritter, W., Kurtzberg, D., et al. (2000). Mismatch negativity in children and adults, and effects of an attended task. *Psychophysiology, 37*, 807–816.
- Grant, K. W., & Seitz, P. F. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. *The Journal of the Acoustical Society of America, 108*, 1197–1208.
- Grant, K. W., Tufts, J. B., & Greenberg, S. (2007). Integration efficiency for speech perception within and across sensory modalities by normal-hearing and hearing-impaired individuals. *The Journal of the Acoustical Society of America, 121*, 1164–1176.

- Grieco-Calub, T. M., & Litovsky, R. Y. (2012). Spatial acuity in 2- to 3-year-old children with normal acoustic hearing, unilateral cochlear implants, and bilateral cochlear implants. *Ear and Hearing, 33*, 561–572.
- Grose, J. H., Hall, J. W., & Gibbs, C. (1993). Temporal analysis in children. *Journal of Speech Language and Hearing Research, 36*, 351–356.
- Gunther, T., Konrad, K., Haeusler, J., Saghraoui, H., et al. (2014). Developmental differences in visual and auditory attention: A cross-sectional study. *Zeitschrift für Neuropsychologie, 25*, 143–152.
- Hall, J. W., Buss, E., & Grose, J. H. (2005). Informational masking release in children and adults. *The Journal of the Acoustical Society of America, 118*, 1605–1613.
- Hall, J. W., Buss, E., & Grose, J. H. (2007). The binaural temporal window in adults and children. *The Journal of the Acoustical Society of America, 121*, 401–410.
- Hall, J. W., Buss, E., & Grose, J. H. (2014). Development of speech glimpsing in synchronously and asynchronously modulated noise. *The Journal of the Acoustical Society of America, 135*, 3594–3600.
- Hall, J. W., Buss, E., Grose, J. H., & Dev, M. B. (2004). Developmental effects in the masking-level difference. *Journal of Speech Language and Hearing Research, 47*, 13–20.
- Hall, J. W., & Grose, J. H. (1991). Notched-noise measures of frequency selectivity in adults and children using fixed-masker-level and fixed-signal-level presentation. *Journal of Speech Language and Hearing Research, 34*, 651–660.
- Hall, J. W., & Grose, J. H. (1994). Development of temporal resolution in children as measured by the temporal-modulation transfer-function. *The Journal of the Acoustical Society of America, 96*, 150–154.
- Hall, J. W., Grose, J. H., Buss, E., & Dev, M. B. (2002). Spondee recognition in a two-talker masker and a speech-shaped noise masker in adults and children. *Ear and Hearing, 23*, 159–165.
- Hall, J. W., Grose, J. H., & Dev, M. B. (1997). Auditory development in complex tasks of comodulation masking release. *Journal of Speech Language and Hearing Research, 40*, 946–954.
- Hall, J. W., Haggard, M. P., & Fernandes, M. A. (1984). Detection in noise by spectro-temporal pattern analysis. *The Journal of the Acoustical Society of America, 76*, 50–56.
- Hall, D. A., & Plack, C. J. (2009). Pitch processing sites in the human auditory brain. *Cerebral Cortex, 19*, 576–585.
- He, C., Hotson, L., & Trainor, L. J. (2007). Mismatch responses to pitch changes in early infancy. *Journal of Cognitive Neuroscience, 19*, 878–892.
- He, C., & Trainor, L. J. (2009). Finding the pitch of the missing fundamental in infants. *The Journal of Neuroscience, 29*, 7718–7722.
- Helfer, K. S., & Freyman, R. L. (2005). The role of visual speech cues in reducing energetic and informational masking. *The Journal of the Acoustical Society of America, 117*, 842–849.
- Hillock-Dunn, A., & Wallace, M. T. (2012). Developmental changes in the multisensory temporal binding window persist into adolescence. *Developmental Science, 15*, 688–696.
- Hollich, G., Newman, R. S., & Jusczyk, P. W. (2005). Infants' use of synchronized visual information to separate streams of speech. *Child Development, 76*, 598–613.
- Holt, R. F., Kirk, K. I., & Hay-McCutcheon, M. (2011). Assessing multimodal spoken word-in-sentence recognition in children with normal hearing and children with cochlear implants. *Journal of Speech Language and Hearing Research, 54*, 632–657.
- Horn, D., Werner, L. A., Rubinstein, J., & Won, J. H. (2013). Spectral ripple discrimination in infants: Effect of ripple depth and envelope phase randomization. *Abstracts of the Association for Research in Otolaryngology, 37*, 601.
- Johnstone, P. M., & Litovsky, R. Y. (2006). Effect of masker type and age on speech intelligibility and spatial release from masking in children and adults. *The Journal of the Acoustical Society of America, 120*, 2177–2189.
- Jones, P. R., Moore, D. R., & Amitay, S. (2015). Development of auditory selective attention: Why children struggle to hear in noisy environments. *Developmental Psychology, 51*, 353–369.

- Karns, C. M., Isbell, E., Giuliano, R. J., & Neville, H. J. (2015). Auditory attention in childhood and adolescence: An event-related potential study of spatial selective attention to one of two simultaneous stories. *Developmental Cognitive Neuroscience, 13*, 53–67.
- Kidd, G., Mason, C. R., Deliwala, P. S., Woods, W. S., & Colburn, H. S. (1994). Reducing informational masking by sound segregation. *The Journal of the Acoustical Society of America, 95*, 3475–3480.
- Kidd, G., Mason, C. R., & Richards, V. M. (2003). Multiple bursts, multiple looks, and stream coherence in the release from informational masking. *The Journal of the Acoustical Society of America, 114*, 2835–2845.
- Kopp, F. (2014). Audiovisual temporal fusion in 6-month-old infants. *Developmental Cognitive Neuroscience, 9*, 56–67.
- Kopp, F., & Dietrich, C. (2013). Neural dynamics of audiovisual synchrony and asynchrony perception in 6-month-old infants. *Frontiers in Psychology, 4*.
- Kuhl, P. K., & Meltzoff, A. N. (1982). The bimodal perception of speech in infancy. *Science, 218*, 1138–1140.
- Kuhnle, S., Ludwig, A. A., Meuret, S., Kuttner, C., et al. (2013). Development of auditory localization accuracy and auditory spatial discrimination in children and adolescents. *Audiology and Neurotology, 18*, 48–62.
- Lalonde, K., & Holt, R. F. (2015). Preschoolers benefit from visually salient speech cues. *Journal of Speech Language and Hearing Research, 58*, 135–150.
- Lau, B. K., & Werner, L. A. (2012). Perception of missing fundamental pitch by 3- and 4-month-old human infants. *The Journal of the Acoustical Society of America, 132*, 3874–3882.
- Lau, B. K., & Werner, L. A. (2014). Perception of the pitch of unresolved harmonics by 3- and 7-month-old human infants. *The Journal of the Acoustical Society of America, 136*, 760–767.
- Leibold, L. J. (2012). Development of auditory scene analysis and auditory attention. In L. A. Werner, A. N. Popper, & R. R. Fay (Eds.), *Human auditory development* (pp. 137–161). New York: Springer Science+Business Media.
- Leibold, L. J., & Bonino, A. Y. (2009). Release from informational masking in children: Effect of multiple signal bursts. *The Journal of the Acoustical Society of America, 125*, 2200–2208.
- Leibold, L. J., & Buss, E. (2013). Children's identification of consonants in a speech-shaped noise or a two-talker masker. *Journal of Speech Language and Hearing Research, 56*, 1144–1155.
- Leibold, L. J., & Neff, D. L. (2007). Effects of masker-spectral variability and masker fringes in children and adults. *The Journal of the Acoustical Society of America, 121*, 3666–3676.
- Leibold, L. J., & Neff, D. L. (2011). Masking by a remote-frequency noise band in children and adults. *Ear and Hearing, 32*, 663–666.
- Leibold, L. J., & Werner, L. A. (2007). The effect of masker-frequency variability on the detection performance of infants and adults. *The Journal of the Acoustical Society of America, 119*, 3960–3970.
- Lewkowicz, D. J. (1992). Infants' responsiveness to the auditory and visual attributes of a sounding/moving stimulus. *Attention, Perception, & Psychophysics, 52*, 519–528.
- Lewkowicz, D. J. (1996). Perception of auditory-visual temporal synchrony in human infants. *Journal of Experimental Psychology-Human Perception and Performance, 22*, 1094–1106.
- Lewkowicz, D. J. (2010). Infant perception of audio-visual speech synchrony. *Developmental Psychology, 46*, 66–77.
- Lewkowicz, D. J., & Flom, R. (2014). The audiovisual temporal binding window narrows in early childhood. *Child Development, 85*, 685–694.
- Litovsky, R. Y. (1997). Developmental changes in the precedence effect: Estimates of minimum audible angle. *The Journal of the Acoustical Society of America, 102*, 1739–1745.
- Litovsky, R. Y. (2005). Speech intelligibility and spatial release from masking in young children. *The Journal of the Acoustical Society of America, 117*, 3091–3099.
- Litovsky, R. (2012). Development of binaural and spatial hearing. In L. A. Werner, A. N. Popper, & R. R. Fay (Eds.), *Human auditory development* (pp. 163–195). New York: Springer Science +Business Media.

- MacKain, K., Studdert-Kennedy, M., Spieker, S., & Stern, D. (1983). Infant intermodal speech perception is a left-hemisphere function. *Science*, *219*, 1347–1349.
- Massaro, D. W. (1984). Children's perception of visual and auditory speech. *Child Development*, *55*, 1777–1788.
- Massaro, D. W. (1998). *Perceiving talking faces: From speech perception to a behavioral principle*. Cambridge, MA: The MIT Press.
- Massaro, D. W., Thompson, L. A., Barron, B., & Laren, E. (1986). Developmental-changes in visual and auditory contributions to speech-perception. *Journal of Experimental Child Psychology*, *41*, 93–113.
- McAdams, S., & Bertocini, J. (1997). Organization and discrimination of repeating sound sequences by newborn infants. *The Journal of the Acoustical Society of America*, *102*, 2945–2953.
- McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*, 746–748.
- Micheyl, C., Kreft, H., Shamma, S., & Oxenham, A. J. (2013). Temporal coherence versus harmonicity in auditory stream formation. *The Journal of the Acoustical Society of America*, *133*, EL188–EL194.
- Middlebrooks, J. C., Nick, H. S., Subramony, S. H., Advincola, J., et al. (2013). Mutation in the kv3.3 voltage-gated potassium channel causing spinocerebellar ataxia 13 disrupts sound-localization mechanisms. *PLoS ONE*, *8*, 6.
- Moore, D. R., Cowan, J. A., Riley, A., Edmondson-Jones, A. M., & Ferguson, M. A. (2011). Development of auditory processing in 6- to 11-yr-old children. *Ear and Hearing*, *32*, 269–285.
- Murphy, J., Summerfield, A. Q., O'Donoghue, G. M., & Moore, D. R. (2011). Spatial hearing of normally hearing and cochlear implanted children. *International Journal of Pediatric Otorhinolaryngology*, *75*, 489–494.
- Neff, D. L. (1995). Signal properties that reduce masking by simultaneous, random-frequency maskers. *The Journal of the Acoustical Society of America*, *98*, 1909–1920.
- Neff, D. L., & Green, D. M. (1987). Masking produced by spectral uncertainty with multicomponent maskers. *Attention, Perception, & Psychophysics*, *41*, 409–415.
- Newman, R. S. (2009). Infants' listening in multitalker environments: Effect of the number of background talkers. *Attention, Perception, & Psychophysics*, *71*, 822–836.
- Newman, R. S., & Jusczyk, P. W. (1996). The cocktail party effect in infants. *Attention, Perception, & Psychophysics*, *58*, 1145–1156.
- Nozza, R. J. (1987). The binaural masking level difference in infants and adults: Developmental change in binaural hearing. *Infant Behavior and Development*, *10*, 105–110.
- Nozza, R. J., Wagner, E. F., & Crandell, M. A. (1988). Binaural release from masking for a speech sound in infants, preschoolers, and adults. *Journal of Speech Language and Hearing Research*, *31*, 212–218.
- Oh, E. L., Wightman, F., & Lutfi, R. A. (2001). Children's detection of pure-tone signals with random multitone maskers. *The Journal of the Acoustical Society of America*, *109*, 2888–2895.
- Olsho, L. W. (1985). Infant auditory perception: Tonal masking. *Infant Behavior and Development*, *8*, 371–384.
- Patterson, M. L., & Werker, J. F. (2003). Two-month-old infants match phonetic information in lips and voice. *Developmental Science*, *6*, 191–196.
- Peter, V., Wong, K., Name, V. K., Sharma, M., et al. (2014). Assessing spectral and temporal processing in children and adults using temporal modulation transfer function (TMTF), iterated ripple noise (IRN) perception, and spectral ripple discrimination (SRD). *Journal of the American Academy of Audiology*, *25*, 210–218.
- Petersen, S. E., & Posner, M. I. (2012). The attention system of the human brain: 20 years after. *Annual Review of Neuroscience*, *35*(35), 73–89.
- Ponton, C. W., Moore, J. K., & Eggermont, J. J. (1996). Auditory brain stem response generation by parallel pathways: Differential maturation of axonal conduction time and synaptic transmission. *Ear and Hearing*, *17*, 402–410.

- Prodi, N., Visentin, C., & Feletti, A. (2013). On the perception of speech in primary school classrooms: Ranking of noise interference and of age influence. *Journal of the Acoustical Society of America*, *133*, 255–268.
- Rahne, T., Bochmann, M., von Specht, H., & Sussman, E. S. (2007). Visual cues can modulate integration and segregation of objects in auditory scene analysis. *Brain Research*, *1144*, 127–135.
- Rahne, T., & Bockmann-Barthel, M. (2009). Visual cues release the temporal coherence of auditory objects in auditory scene analysis. *Brain Research*, *1300*, 125–134.
- Ross, L. A., Molholm, S., Blanco, D., Gomez-Ramirez, M., et al. (2011). The development of multisensory speech perception continues into the late childhood years. *European Journal of Neuroscience*, *33*, 2329–2337.
- Rothbart, M. K., Ellis, L. K., Rueda, M. R., & Posner, M. I. (2003). Developing mechanisms temperamental effortful control. *Journal of Personality*, *71*, 1113–1143.
- Rothbart, M. K., Sheese, B. E., Rueda, M. R., & Posner, M. I. (2011). Developing mechanisms of self-regulation in early life. *Emotion Review*, *3*, 207–213.
- Rueda, M. R., Fan, J., McCandliss, B. D., Halparin, J. D., et al. (2004). Development of attentional networks in childhood. *Neuropsychologia*, *42*, 1029–1040.
- Sanders, L. D., Stevens, C., Coch, D., & Neville, H. J. (2006). Selective auditory attention in 3- to 5-year-old children: An event-related potential study. *Neuropsychologia*, *44*, 2126–2138.
- Schneider, B. A., Bull, D., & Trehub, S. E. (1988). Binaural unmasking in infants. *The Journal of the Acoustical Society of America*, *83*, 1124–1132.
- Schneider, B. A., Morrongiello, B. A., & Trehub, S. E. (1990). The size of the critical band in infants, children, and adults. *Journal of Experimental Psychology-Human Perception and Performance*, *16*, 642–652.
- Schneider, B. A., Trehub, S. E., Morrongiello, B. A., & Thorpe, L. A. (1989). Developmental changes in masked thresholds. *The Journal of the Acoustical Society of America*, *86*, 1733–1742.
- Shamma, S., Elhilali, M., Ma, L., Micheyl, C., et al. (2013). Temporal coherence and the streaming of complex sounds. In B. C. J. Moore, R. D. Patterson, I. M. Winter, R. P. Carlyon, & H. E. Gockel (Eds.), *Basic aspects of hearing: Physiology and perception* (Vol. 787, pp. 535–543). New York: Springer Science+Business Media.
- Shamma, S. A., Elhilali, M., & Micheyl, C. (2011). Temporal coherence and attention in auditory scene analysis. *Trends in Neurosciences*, *34*, 114–123.
- Siqueland, E. R., & Delucia, C. A. (1969). Visual reinforcement of nonnutritive sucking in human infants. *Science*, *165*, 1144.
- Smith, N. A., & Trainor, L. J. (2011). Auditory stream segregation improves infants' selective attention to target tones amid distractors. *Infancy*, *16*, 1–14.
- Spetner, N. B., & Olsho, L. W. (1990). Auditory frequency resolution in human infancy. *Child Development*, *61*, 632–652.
- Stuart, A. (2008). Reception thresholds for sentences in quiet, continuous noise, and interrupted noise in school-age children. *Journal of the American Academy of Audiology*, *19*, 135–146.
- Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *The Journal of the Acoustical Society of America*, *26*, 212–215.
- Summerfield, Q., & Assmann, P. F. (1991). Perception of concurrent vowels: Effects of harmonic misalignment and pitch-period asynchrony. *The Journal of the Acoustical Society of America*, *89*, 1364–1377.
- Sussman, E. S., Ceponiene, R., Shestakova, A., Naatanen, R., & Winkler, I. (2001). Auditory stream segregation processes operate similarly in school-aged children and adults. *Hearing Research*, *153*, 108–114.
- Sussman, E. S., & Steinschneider, M. (2009). Attention effects on auditory scene analysis in children. *Neuropsychologia*, *47*, 771–785.
- Sussman, E. S., Wong, R., Horvath, J., Winkler, I., & Wang, W. (2007). The development of the perceptual organization of sound by frequency separation in 5–11-year-old children. *Hearing Research*, *225*, 117–127.

- Truchon-Gagnon, C. (1988). Noise in day-care centers for children. *Noise Control Engineering Journal*, 39, 57–64.
- Vaillancourt, V., Laroche, C., Giguere, C., & Soil, S. D. (2008). Establishment of age-specific normative data for the Canadian French version of the hearing in noise test for children. *Ear and Hearing*, 29, 453–466.
- Van Deun, L., van Wieringen, A., Van den Bogaert, T., Scherf, F., et al. (2009). Sound localization, sound lateralization, and binaural masking level differences in young children with normal hearing. *Ear and Hearing*, 30, 178–190.
- van Noorden, L. P. A. S. (1975). *Temporal coherence in the perception of tone sequences*. PhD dissertation, Eindhoven University of Technology, The Netherlands.
- Werner, L. A. (2006a). Amplitude modulation detection by infants and adults. *The Journal of the Acoustical Society of America*, 119, 3234.
- Werner, L. A. (2006b). *Preliminary observations on the temporal modulation transfer functions of infants and adults*. Abstracts of the American Auditory Society Annual Meeting.
- Werner, L. A. (2013). Infants' detection and discrimination of sounds in modulated maskers. *The Journal of the Acoustical Society of America*, 133, 4156–4167.
- Werner, L. A., & Bargones, J. Y. (1991). Sources of auditory masking in infants: Distraction effects. *Attention, Perception, & Psychophysics*, 50, 405–412.
- Werner, L. A., & Leibold, L. J. (2017). Auditory development in normal-hearing children. In R. Sewald & A. M. Tharpe (Eds.), *Comprehensive handbook of pediatric audiology* (2nd ed., pp. 67–86). New York: Plural Publishing.
- Werner, L. A., Marean, G. C., Halpin, C. F., Spetner, N. B., & Gillenwater, J. M. (1992). Infant auditory temporal acuity: Gap detection. *Child Development*, 63, 260–272.
- Werner, L. A., Parrish, H. K., & Holmer, N. M. (2009). Effects of temporal uncertainty and temporal expectancy on infants' auditory sensitivity. *The Journal of the Acoustical Society of America*, 125, 1040–1049.
- Wightman, F., Allen, P., Dolan, T., Kistler, D., & Jamieson, D. (1989). Temporal resolution in children. *Child Development*, 60, 611–624.
- Wightman, F., Callahan, M. R., Lutfi, R. A., Kistler, D. J., & Oh, E. (2003). Children's detection of pure-tone signals: Informational masking with contralateral maskers. *The Journal of the Acoustical Society of America*, 113, 3297–3305.
- Wightman, F., & Kistler, D. J. (2005). Informational masking of speech in children: Effects of ipsilateral and contralateral distracters. *The Journal of the Acoustical Society of America*, 118, 3164–3176.
- Wightman, F., Kistler, D., & Brungart, D. (2006). Informational masking of speech in children: Auditory-visual integration. *The Journal of the Acoustical Society of America*, 119, 3940–3949.
- Winkler, I., Kushnerenko, E., Horvath, J., Ceponiene, R., et al. (2003). Newborn infants can organize the auditory world. *Proceedings of the National Academy of Sciences of the USA*, 100, 11812–11815.
- Wright, B. A., & Fitzgerald, M. B. (2001). Different patterns of human discrimination learning for two interaural cues to sound-source location. *Proceedings of the National Academy of Sciences of the USA*, 98, 12307–12312.
- Yuen, K. C. P., & Yuan, M. (2014). Development of spatial release from masking in mandarin-speaking children with normal hearing. *Journal of Speech Language and Hearing Research*, 57, 2005–2023.
- Zettler, C. M., Sevcik, R. A., Morris, R. D., & Clarkson, M. G. (2008). Comodulation masking release (CMR) in children and the influence of reading status. *Journal of Speech Language and Hearing Research*, 51, 772–784.

Chapter 9

Older Adults at the Cocktail Party

M. Kathleen Pichora-Fuller, Claude Alain, and Bruce A. Schneider

Abstract Successful communication and navigation in cocktail party situations depends on complex interactions among an individual's sensory, cognitive, and social abilities. Older adults may function well in relatively ideal communication situations, but they are notorious for their difficulties understanding speech in noisy situations such as cocktail parties. However, as healthy adults age, declines in auditory and cognitive processing may be offset by compensatory gains in ability to use context and knowledge. From a practical perspective, it is important to consider the aging auditory system in multitalker situations because these are among the most challenging situations for older adults. From a theoretical perspective, studying age-related changes in auditory processing provides a special window into the relative contributions of, and interactions among sensory, cognitive, and social abilities. In the acoustical wild, younger listeners typically function better than older listeners. Experimental evidence indicates that age-related differences in simple measures such as word recognition in quiet or noise are largely due to the bottom-up effects of age-related auditory declines. These differences can often be eliminated when auditory input is adjusted to equate the performance levels of listeners on baseline measures in quiet or noise. Notably, older adults exhibit enhanced cognitive compensation, with performance on auditory tasks being facilitated by top-down use of context and knowledge. Nevertheless, age-related differences can persist when tasks are more cognitively demanding and involve discourse comprehension, memory, and attention. At an extreme, older adults with

M.K. Pichora-Fuller (✉) · B.A. Schneider
Department of Psychology, University of Toronto, 3359 Mississauga Rd.,
Mississauga, ON L5L 1C6, Canada
e-mail: k.pichora.fuller@utoronto.ca

B.A. Schneider
e-mail: bruce.schneider@utoronto.ca

C. Alain,
Department of Psychology, The Rotman Research Institute, University
of Toronto, Baycrest, 3560 Bathurst Street, Toronto, ON M6A 2E1, Canada
e-mail: calain@research.baycrest.org

hearing loss are at greater risk for developing cognitive impairments than peers with better hearing.

Keywords Age-related hearing loss · Auditory scene analysis · Auditory spatial attention · Auditory temporal processing · Cognitive aging · Cognitive compensation · Communication ecology · Contextual support · Discourse comprehension · Event-related potentials · Listening effort · Presbycusis · Speech-in-noise listening · Voice fundamental frequency · Working memory

9.1 Introduction

The peripheral auditory system encodes the acoustic inputs that are used by the brain when listeners interact with the auditory world, monitor their own behaviors, and communicate with each other. Applying concepts from ecological biology, a communicative ecological system has been defined (Borg et al. 2008, p. S132) as “A system of communicating individuals in a social and physical background, who function together to circulate information and mental energy to create knowledge and emotions and a change in the system’s constitution and function over time.” From an ecological perspective, the importance of successful participation in social activities motivates listeners to allocate attentional resources to auditory and cognitive information processing in a range of everyday situations (Pichora-Fuller et al. 2016). The cocktail party situation is one of the most challenging of such situations, but it also offers one of the potentially most rewarding opportunities for social interaction.

At a cocktail party, sound provides information to listeners about their surroundings; for example, a doorbell ring alerts the host to the arrival of a guest and partygoers might hear rain against the window or music playing in the background. Sound provides feedback about an individual’s own actions; for example, the hostess hears her own footsteps while walking down the hall to open the door, crunching as she bites a piece of celery, or the clanking of glasses as she makes a celebratory toast. Interpersonal communication entails an exchange between a sender and a receiver of a message as they co-construct meaning in the social and physical setting of the party. Hearing is critical to spoken communication because it enables individuals to receive the speech signal sent by other communicators, monitor their own speech production, and assess the acoustical characteristics of the social (e.g., people laughing) and physical environments (e.g., reverberation in the concrete atrium of the art gallery) in which communication occurs at the party. For the most part, the goals of the listener determine how many and which sounds he or she intentionally samples from the auditory feast of the party soundscape, but sometimes highly salient sounds (e.g., hearing one’s own name or a phone ringing) may attract a listener’s attention to or distract it from an intended listening goal or task. At the cocktail party, listening will also be influenced by congruent or conflicting multisensory inputs and multitasking demands. Successful

communication at a cocktail party will depend on how the listener hears, attends to, comprehends, and remembers relevant information in the auditory scene.

The auditory and cognitive processing abilities that are needed at the cocktail party or in other complex auditory scenes mature over childhood and peak in young adulthood (Werner, Chap. 8). As described in other chapters, however, listening at a cocktail party challenges even young adult listeners with normal hearing because there are heavy demands on complex auditory and cognitive processing, including the formation and selection of auditory objects (Shinn-Cunningham, Best, and Lee, Chap. 2), general masking (Culling and Stone, Chap. 3), release from informational masking (Kidd and Colburn, Chap. 4), and stream segregation (Elhilali, Chap. 5; Middlebrooks, Chap. 6; Simon, Chap. 7). Chapter 10 by Litovsky, Goupell, Misurelli, and Kan describes the deleterious effects of hearing loss on listening at the cocktail party and how the use of technologies such as hearing aids or cochlear implants may restore or sometimes further disrupt functioning. The present chapter explores how age-related changes in auditory and cognitive processing may affect listening at the cocktail party by older adults, in particular those whose pure-tone audiometric hearing thresholds are normal or near-normal. From a practical perspective, it is important to consider the aging auditory system at the cocktail party because older adults who find such situations too demanding or stressful may cope by withdrawing from social interaction, with long-term negative effects on their quality of life and mental and physical health. From a theoretical perspective, age-related changes in auditory processing provide a special window into the relative contributions of sensory, cognitive, and social abilities during social interaction. Younger listeners typically function better than older listeners in the acoustical wild, and laboratory research helps to pinpoint the specific aspects of listening that are preserved or decline as adults age.

9.2 Auditory Aging

9.2.1 *Periphery*

Hearing loss is the third most common chronic health condition in older adults (Yueh et al. 2003). The symptoms of age-related hearing loss (ARHL) can begin in the fourth decade of life. Its prevalence increases with age, affecting roughly half of those older than the age of 65 years and up to 90% of those older than the age of 80 years (Cruikshanks et al. 2010). ARHL (sometimes called presbycusis) is commonly characterized by high-frequency sensorineural hearing loss defined in terms of audiometric thresholds (Kiessling et al. 2003). In standard clinical audiometric testing, pure-tone thresholds are measured in decibels referenced to normal human hearing levels (dB HL) at octave frequencies from 250 to 8000 Hz. Threshold elevations in ARHL begin at the highest frequencies and gradually progress to lower frequencies (ISO 7029 2000). In the earliest stages of auditory

aging, before clinically significant abnormal thresholds are observed, elevated thresholds (>25 dB HL) at frequencies above 8000 Hz may reduce the availability of interaural intensity cues to localization, including important pinna cues around 10,000 Hz. As ARHL progresses to lower frequencies (especially in the range from 500 to 4000 Hz), more of the speech signal becomes inaudible and speech perception worsens even in quiet environments. Amplification can restore audibility, in turn improving phoneme and word recognition accuracy, especially in quiet (Humes and Dubno 2010). Nevertheless, the difficulties that older adults have understanding speech in noise persist. Notably, when amplification is provided, speech-in-noise performance is not restored to normal levels, despite what would be predicted if the difficulties of older listeners were confined to reduced audibility. Speech-in-noise understanding depends on more than just making speech audible. It depends on nonaudiometric factors such as suprathreshold auditory temporal processing and cognitive processing (Humes 2007).

High-frequency sensorineural hearing loss, whether in younger or older adults, often involves damage to outer hair cells in the cochlea as a result of exposure to industrial and/or recreational noise. However, in ARHL, one or more structures in the cochlea or central auditory system can be damaged in ways that are not typical in younger adults who have high-frequency hearing loss (Schmiedt 2010). Specifically, high-frequency sensorineural hearing loss in older adults may be attributable to changes in the endocochlear potentials associated with changes to the cochlear blood supply in the stria vascularis (Mills et al. 2006; Saremi and Stenfelt 2013). There may also be neural changes that do not necessarily manifest in elevated audiometric thresholds. Mounting physiological evidence (Kujawa and Liberman 2009) and computational modeling (Lopez-Poveda 2014) point to neural degeneration and/or reductions in neural synchrony in the periphery that may underpin age-related differences in suprathreshold auditory and speech processing.

9.2.2 *Speech Understanding*

Importantly, the hearing abilities of older adults are heterogeneous. Their difficulties in understanding speech in noise vary considerably and are not well predicted from the audiogram (Füllgrabe et al. 2014). Indeed, difficulties understanding speech in noise often precede clinically significant elevation of audiometric pure-tone thresholds in quiet (Bergman 1980). Typically, older adults require higher signal-to-noise ratios (SNRs) to perform equivalently to younger adults on speech-in-noise tests, even if they have normal or near-normal audiograms. The SNR at which listeners reach 50% correct word recognition is the speech recognition threshold (SRT) in noise. A number of studies indicate that, over a broad range of conditions, older adults whose hearing thresholds in quiet are normal for their age have SRTs in noise from 2–4 decibels (dB) higher than those of younger adults (Schneider et al. 2010).

Age-related differences in speech understanding in noise could be due to declines in other auditory abilities that are unrelated to pure-tone threshold elevations and involve central auditory or cognitive processing (CHABA 1988). In addition to difficulties understanding speech in noise, age-related declines in melodic pitch perception (Russo et al. 2012), the identification of vocal emotion (Dupuis and Pichora-Fuller 2015), and the understanding of emotional speech in noise (Dupuis and Pichora-Fuller 2014) could also reduce an older listener's ability to participate at a cocktail party, where enjoying music and identifying emotions may be as or more important than recognizing words.

9.2.3 Psychoacoustics of Temporal Processing and Behavioral Measures of Speech Processing

Over the last 30 years, a large body of knowledge has accumulated to characterize human ARHL based on psychoacoustics and behavioral speech perception research (for a comprehensive review see Gordon-Salant et al. 2010). Of particular relevance to listening at the cocktail party are well-documented age-related differences in auditory temporal processing (Fitzgibbons and Gordon-Salant 2010; Walton 2010) and binaural hearing (Eddins and Hall 2010) that could undermine speech understanding in noise (Humes and Dubno 2010). Highlights of this research are provided to show how auditory aging might affect listening at the cocktail party.

It is important to differentiate among levels of auditory temporal processing (Phillips 1995), and to consider how aging might affect abilities at each level because they may have different consequences for listening to speech at the cocktail party. Monaural temporal cues are relevant to three main levels of speech processing in quiet (Greenberg 1996): subsegmental (phonetic), segmental (phonemic), and suprasegmental (syllabic and lexico-syntactic). Subsegmental speech processing relies on fine structure cues, including periodicity cues based on the fundamental frequency and harmonic structure of the voice. Some types of segmental information are provided by local gap and duration cues and properties of the speech envelope that contribute to phoneme identification (e.g., presence of a stop consonant, voice onset time). Suprasegmental processing depends on cues such as the pattern of fluctuations in the amplitude envelope of the time waveform that convey prosodic information related to the rate and rhythm of speech, and these cues also serve lexical and syntactic processing. Each level has been investigated in older adults using psychoacoustic and speech perception measures. The effects of age on some measures suggest losses in gap and duration coding or poorer use of envelope cues, while others implicate reductions in synchrony or periodicity coding.

9.2.3.1 Gap and Duration Detection

At the segmental level, gaps and duration cues provide temporal information about some phonemic contrasts, in particular contrasts based on distinctions in the manner of articulation for consonants (Gordon-Salant et al. 2006; Pichora-Fuller et al. 2006). The most common psychoacoustic measure of temporal processing is the gap detection threshold, the smallest gap that a listener can detect in a stimulus. Older adults with normal or near-normal audiograms do not detect gaps until they are significantly longer than the gaps that can be detected by younger adults, and their gap detection thresholds do not significantly correlate with audiometric thresholds (Schneider et al. 1994; Snell and Frisina 2000). Notably, age-related differences are more pronounced when the sound markers surrounding the gap are shorter than 10 ms (Schneider and Hamstra 1999), and when the location of the gap is near the onset or offset of the signal (He et al. 1999). When spectrally identical sounds precede and follow the gap (within-channel markers), gap detection thresholds are small (a few milliseconds). The perceptual operation required for within-channel gap detection is thought to involve relatively simple processing of activity in the neural channel representing the stimulus. In contrast, when there are spectral differences between the sounds that lead and lag the gap (between-channel markers), gap detection thresholds can be about 10 times larger than those obtained for within-channel markers. This suggests that more complex processing may be involved, such as a more central relative timing operation across different neural regions (Phillips et al. 1997). Importantly, speech processing likely relies on both within and between-channel processes, and age-related differences have been found for both types of markers.

The effect of age on gap detection thresholds is exacerbated when more complex stimuli are used, as illustrated in studies examining gap discrimination thresholds when the frequency of the leading marker was fixed and the frequency of the lagging marker was varied (Lister et al. 2002), or when synthetic speech stimuli with spectrally dynamic markers were compared to those with spectrally stable markers (Lister and Tarver 2004), or when the harmonic structure of the leading and lagging markers was manipulated (Heinrich et al. 2014). In a study investigating age-related differences in gap detection for both nonspeech and speech markers that were either spectrally symmetrical (within-channel condition) or spectrally asymmetrical (between-channel condition), gap detection thresholds were longer for both age groups and age-related differences were more pronounced when the markers were spectrally asymmetrical than when they were symmetrical (Pichora-Fuller et al. 2006). Notably, age-related differences for asymmetrical markers were less pronounced when the markers were speech sounds than when they were nonspeech sounds. Presumably, older listeners were able to compensate because of their familiarity with speech sequences in which gaps cue the presence of an unvoiced stop consonant (e.g., the silent gap for the stop consonant /p/ between /s/ and /u/ in the word *spoon*). Furthermore, the size of the gap needed to distinguish word pairs that differed in terms of whether or not an unvoiced stop consonant was present (e.g., *spoon* and *soon* or *catch* and *cash*) varied with the rate of speech (i.e., the

duration of the speech markers), but older listeners always needed larger gaps compared to younger listeners (Haubert and Pichora-Fuller 1999). Interestingly, patterns of scalp-related neuromagnetic activity during gap detection suggest that age-related differences are related to higher-level object formation rather than to lower-level registration of acoustical cues (Ross et al. 2009, 2010).

There is also abundant research on age-related differences in duration discrimination ability. This evidence converges with the findings on gap detection on three key points. First, age-related differences in duration discrimination do not significantly correlate with audiometric thresholds (Fitzgibbons et al. 2007). Second, age-related differences in ability to discriminate the duration of markers are more pronounced when the reference signal is shorter (20 ms) than when it is longer (200 ms) (Abel et al. 1990; Fitzgibbons et al. 2007). Third, age-related differences in duration discrimination can be exacerbated by increasing the complexity of the stimulus or task (Fitzgibbons and Gordon-Salant 2001). Similar findings using speech markers underscore the relevance of duration discrimination for the perception of phonemic contrasts serving word discrimination (Gordon-Salant et al. 2006). As with gap detection, different mechanisms may contribute to age-related deficits in duration discrimination depending on marker properties. Impaired coding of rapid onsets and offsets seems likely to be involved in deficits seen when brief markers are used, whereas higher-level auditory processing involving a central timing mechanism may be involved in the age-related differences observed for longer duration and more complex stimuli (Fitzgibbons et al. 2007).

9.2.3.2 Temporal Fluctuations in the Amplitude Envelope

The patterns of amplitude modulations in the speech time-waveform can be thought of as a sequence of gaps and durations that provide temporal information pertaining to the suprasegmental or prosodic level of speech processing required for lexical and syntactic analyses in the cortex (Pelle and Davis 2012). Significant effects of age have been found on psychoacoustic measures of modulation detection, and these behavioral results are correlated with electrophysiological envelope-following responses, suggesting the involvement of both brainstem and cortical subsystems in this level of temporal processing (Purcell et al. 2004). Envelope fluctuations in speech vary with a talker's speaking rate and rhythm. Older listeners have more difficulty understanding sentences when they are spoken at a fast rate or are time-compressed (Versfeld and Dreschler 2002; Wingfield et al. 2006). When speech is speeded, speech understanding may be hampered because acoustical speech cues are reduced and/or because the time available to process the speech information cognitively is reduced. For younger adults, the deleterious effects of speeding speech on word identification and sentence comprehension are explained by reduced availability of time for cognitive processing, whereas for older adults both cognitive and auditory factors seem to play a role (Wingfield et al. 1999; Vaughan et al. 2008). When speech is speeded, older listeners benefit more than younger listeners when prosody is congruent with syntactic structure, but they are

more disadvantaged when prosody and syntax are incongruent (Wingfield et al. 1992). Lexical decision reaction times are slower for older than for younger adults when the preceding sentence context is acoustically distorted by time compression, but reaction times are facilitated more for older than for younger listeners when the preceding sentence context is semantically congruent with the target item (Goy et al. 2013). In general, older listeners need to hear more speech information to identify words in a time-gating task, but they are as able as younger listeners to benefit from prosodic envelope information even when fine-structure cues are not available for phonemes identification (Wingfield et al. 2000). Furthermore, experiments using noise-vocoding with a varying number of bands have shown that older adults need a greater amount of temporal envelope information (i.e., more bands) to recognize word or syllables compared to younger adults (Souza and Boike 2006; Sheldon et al. 2008). Overall, it seems that older listeners have more difficulties understanding speeded speech and need more envelope information than younger listeners to understand syllables, words, and sentences in quiet. However, they can compensate by using semantic context and congruent prosody to linguistically parse the speech stream. At a noisy cocktail party, older adults may be well advised to converse with talkers who speak slowly and whose speech rhythm provides rich linguistic prosodic cues.

9.2.3.3 Synchrony or Periodicity Coding

Synchrony or periodicity coding involves phase locking to (quasi-)periodic, low-frequency sound inputs such as the fundamental frequency and lower harmonics of speech. These fine structure components of speech are relatively unimportant for word recognition in quiet, but listeners can use them to identify and follow the voice of a talker in a group. For instance, the continuity of pitch contours can help listeners to segregate the voices of competing talkers. Pitch cues contribute to linguistic prosody that helps listeners to identify word and sentence structures. These cues also contribute to affective prosody that is used to identify a talker's vocal emotion, and they contribute to the perception of musical melody or tonality.

Because the psychoacoustic frequency difference limen (DL) is thought to depend on phase locking at low frequencies, deficits in periodicity coding could explain why age-related increases in frequency DLs are greater for low frequencies than for high frequencies (e.g., Abel et al. 1990). Deficits in periodicity coding or loss of synchrony could also explain why age-related differences in the detection of FM modulation are larger at low frequencies than at high frequencies for older listeners (He et al. 2007), and why older listeners have larger intensity DLs for high-level low-frequency tones in noise compared to younger listeners (MacDonald et al. 2007). Furthermore, loss of synchrony might contribute to age-related declines in detection of a mistuned harmonic (Alain et al. 2001), melodic perception (Russo et al. 2012), or identification of concurrent vowels (Snyder and Alain 2005; Vongpaisal and Pichora-Fuller 2007). In addition, simulating a loss of synchrony in younger adults by introducing temporal jitter in the low frequencies (<1.2 kHz)

leads them to perform like older adults when the accuracy of word recognition is tested in babble (Pichora-Fuller et al. 2007; Smith et al. 2012). Note that these age-related differences affect the auditory processing of suprathreshold sounds in the lower frequencies where audiometric thresholds are in the normal range in typical cases of presbycusis.

9.2.3.4 Binaural Processing

In addition to the contributions of auditory temporal cues to speech processing in quiet listening conditions, auditory temporal processing abilities become even more important at the cocktail party where they can be used by the listener to unmask speech in noise, segregate concurrent speech streams, localize sounds, and direct spatial attention. Beyond age-related changes in monaural auditory temporal processing, age-related declines in binaural processing, even in older adults who have normal or near-normal audiograms, may contribute to the communication difficulties of older listeners at cocktail parties (Eddins and Hall 2010). Interestingly, age-related declines in the ability to detect a change in the interaural correlation of a noise presented to both ears (Wang et al. 2011), and in the ability to use interaural timing differences to unmask signals (Pichora-Fuller and Schneider 1992), have been shown to be consistent with age-related declines in neural synchrony. Such losses in neural synchrony would likely make it considerably more difficult for older adults to parse the auditory scene into its component sound sources, especially in multitalker situations where voice cues help to segregate the speech streams produced by different talkers.

9.3 Electrophysiological Measures of Auditory and Cognitive Aging

For the most part, psychoacoustic and speech understanding experiments measure the offline responses of listeners after auditory or speech processing has been completed. Other methods are needed to investigate the dynamic online changes in processing that occur over time, and to assess the brain operations and areas involved in processing incoming acoustic signals. Scalp recordings of neuroelectric brain activity or electroencephalography (EEG) make it possible to delineate normal and impaired systems at multiple stages of auditory processing (Alain et al. 2013). Notably, such recordings nicely complement behavioral assessments and allow scientists and clinicians to assess the activity in the auditory system with high temporal precision in the absence of overt behavioral responses (Simon, Chap. 7).

9.3.1 *Brainstem*

The brainstem frequency-following response (FFR) has been used to probe the neural registration and encoding of complex sounds (e.g., harmonic complex, vowels, or phonemes) at subcortical levels of processing (e.g., Bidelman and Krishnan 2009; Krishnan et al. 2010). Notably, FFRs have provided important insights into the early neural transcription of sound at subcortical levels, including how nascent sensory representations influence and contribute to the early formation of auditory percepts (Bidelman and Krishnan 2010; Bidelman et al. 2011). Compared to younger adults, older adults have reduced amplitude and delayed speech-evoked brainstem responses (Anderson et al. 2012). Such age-related declines in the temporal precision with which speech sounds are encoded at the subcortical level could negatively affect the cortical representation of speech (Bidelman et al. 2014).

9.3.2 *Cortex*

Auditory event-related potentials (ERPs) can be elicited by clicks, tone onsets, and speech sounds. The P1–N1–P2 complex occurs between 50 and 250 ms after sound onset. This complex represents the processing and encoding of acoustic information and is thought to reflect the activation of early forebrain structures including the thalamus and primary/secondary auditory cortices (Picton et al. 1999). Previous studies revealed that, like brainstem FFRs, these ERPs are sensitive to parametric changes in perceptual features related to the acoustic speech waveform, such as voice pitch, formant transitions, timbre, and harmonicity (Alain 2007; Chang et al. 2010). However, whereas brainstem responses appear to map acoustic details, cortical responses appear to reflect the perceptual organization of auditory objects. For example, in a study of categorical speech perception, activity from the brainstem was found to mirror properties of the speech waveform and changes in speech acoustics, whereas cortical evoked activity reflected distinct perceptual categories associated with abstract phonemic speech boundaries (Bidelman et al. 2013). These findings suggest a critical transformation in neural speech representations between brainstem and auditory cortex analogous to the acoustic-phonetic mapping necessary to generate categorical phoneme perception. In a study evaluating behavioral measures of categorical speech perception and both brainstem and cortical speech-evoked brain responses in the same younger and older listeners, older adults had slower and more variable speech classification performance than younger listeners, which coincided with reduced brainstem amplitude and increased, but delayed, cortical speech-evoked responses (Bidelman et al. 2014). The impoverished representation of speech sounds in older brainstems appears to be compensated by increased cortical responses in the aging brain, altering the acoustic-phonetic mapping necessary for robust speech understanding.

Older adults often generate larger cortical responses to speech stimuli compared to younger adults. Woods and Clayworth (1986) found an age-related increase in the amplitude and latency of early cortical evoked responses (approximately 30 ms after sound onset) that remained even after controlling for age-related differences in audiometric thresholds. The amplitude of the P1 wave is often larger for older than for younger adults (e.g., Ross et al. 2010; Lister et al. 2011). Some studies using pure tones or speech sounds during active or passive listening have also reported a larger N1 wave in older adults than in younger adults (e.g., Anderer et al. 1996; Chao and Knight 1997), while other studies have reported longer latencies (e.g., Iragui et al. 1993; Tremblay et al. 2003). For the P2 wave, studies using pure-tone or speech sounds have observed comparable amplitudes across age groups, but often the latencies of older adults are longer than those of younger adults (Alain and Snyder 2008; Lister et al. 2011). These age-related increases in latency could result from general slowing in perceptual and cognitive processing (Salthouse 1996), whereas age-related increases in auditory ERP amplitude may reflect impaired inhibitory functions at various levels within the afferent and efferent auditory pathways (Chao and Knight 1997; Alain and Woods 1999). Older adults may also have more difficulty filtering out task-irrelevant information such that they need to allocate more attentional resources to the processing of auditory stimuli compared to younger adults (Alain et al. 2004). Importantly, the difference between the amplitude of responses in attentive and nonattentive conditions is larger in older than in younger listeners, suggesting that attentional mechanisms are more often deployed by older than by younger listeners during listening. Such enhanced cortical evoked responses may also reflect a loss of stimulus specificity such that the older brain over-responds to incoming sounds (Leung et al. 2013). Larger N1 and P2 amplitudes may indicate that incoming sounds are processed at a deeper level of encoding, which could account for intrusions in subsequent memory tasks (Greenhut-Wertz and Manning 1995). That is, older adults may preserve representations in sensory memory, even when they are no longer relevant.

9.3.3 Reconciling Behavioral and Electrophysiological Findings Regarding Age-Related Changes

Behavioral studies have revealed numerous age-related declines in suprathreshold auditory processing, including declines in temporal processing at a number of different levels. However, notwithstanding the effects of age on neural activity in general, ERP studies that have incorporated a psychoacoustic design have shown that the rate of changes in neural activity as a function of signal duration (Ostroff et al. 2003), harmonicity (Alain et al. 2012), fundamental frequency (Snyder and Alain 2005), or first formant transition (Bidelman et al. 2014), is often comparable between younger and older adults. For example, in a study in which neuromagnetic auditory evoked responses were measured in young, middle-aged, and older healthy

participants who listened to sounds of various durations, age-related differences in absolute response magnitudes were found, but increases in sound duration resulted in comparable changes in cortical responses in all three age groups (Ross et al. 2009).

The results from these electrophysiological studies seem to be at odds with behavioral research suggesting that there are age-related declines in auditory processing. The results from studies measuring cortical evoked responses also appear to be inconsistent with those showing age-related differences in the amplitude and timing of brainstem responses to complex sounds in quiet. The apparent contradiction between the behavioral and electrophysiological data could be reconciled by assuming that there are age-related reductions in the ability of listeners to access or use sensory representations in short-term memory rather than a failure to initially encode temporal information. Another possibility is that there are age-related differences in attentional control during listening. For example, in a study comparing ERPs to gaps measured in controlled versus automatic listening conditions (either respond to the gap or watch a silent movie), when the gap sizes are chosen to equate younger and older listeners in terms of their behavioral performance, younger listeners detected gaps in either the automatic or controlled listening conditions, but older adults detected them only in the controlled condition (Alain et al. 2004). It is also possible that the apparent discrepancies between these neurophysiological findings and previously published behavioral data might be explained by differences between the experimental methods used in behavioral and EEG studies. Specifically, electrophysiological tests, especially brainstem tests, may be more immune than typical behavioral tests to the effects of cognitive factors such as attention and memory. Furthermore, EEG studies may not have used stimuli such as speeded speech or speech masking noise that reveal the most pronounced age-related differences in behavioral studies of auditory aging.

There is increasing evidence that difficulties understanding speech in noise may be related to problems in parsing the incoming acoustic signal into distinct representations of sound objects, especially when listening requires segregating concurrently or sequentially occurring streams of auditory objects. For instance, older adults have more difficulty than younger adults in using binaural cues, and this coincides with changes in neuromagnetic activity originating from the auditory cortices (Ross et al. 2007). Older adults also showed deficits in parsing and identifying two vowels presented simultaneously (Snyder and Alain 2005) and have more difficulty than younger adults in using first formant transitions to group speech sound that are presented sequentially (Hutka et al. 2013). Together, these results suggest that the speech in noise problems commonly observed in older adults could be related to deficits in perceptually organizing incoming acoustic signals into coherent concurrent and sequential sound objects (Alain et al. 2006). When there are multiple sound sources, the more similar the sound objects are acoustically, the more difficulty listeners, especially older listeners, will have segregating them and distinguishing foreground from background streams.

9.4 Age-Related Differences in Speech Understanding Depending on Masker Type

In addition to the many sounds that older adults may want to listen to at the cocktail party, there may also be many unwanted sounds that they would rather ignore. Listeners would experience a confusing jumble of sounds if they could not distinguish between different sounds and selectively attend to the one(s) of most importance to them. In general, older adults have more difficulty understanding speech in noise regardless of the type of masker. Importantly, depending on the type of masker, there may be shifts in the relative contributions of various auditory and cognitive factors to speech understanding, and the magnitude of age-related differences may also vary.

9.4.1 *Steady-State Maskers*

At the cocktail party, it is relatively easy for listeners to segregate speech from meaningless steady-state sounds (e.g., ventilation noise). Speech easily becomes the attended foreground sound and ventilation noise an ignored background sound. Understanding speech when there is primarily energetic masking depends heavily on peripheral and bottom-up auditory processing of the signals (Culling and Stone, Chap. 3). In this sort of noise background, age-related differences are minimal for older adults who have normal audiometric thresholds.

9.4.2 *Complex and Fluctuating Nonspeech Maskers*

More complex nonspeech sounds may be annoying (e.g., the sound of chairs scraping the floor, guests playing ping pong, the host demonstrating a new model train in the party room) or pleasant (e.g., music), but they are usually sufficiently dissimilar to speech that it is relatively easy to segregate them from a target speech stream and relegate them to the background. Informational masking will increase as the similarity between speech and the background sounds increases. As informational masking increases, the contribution of central auditory and cognitive abilities will also increase such that age-related differences may be observed to varying degrees depending on the specific nature of the masker. On the one hand, cognitive demands may increase as maskers become more complex. On the other hand, knowledge of the structures of complex nonspeech sounds or familiarity with them may help listeners to use expectations to efficiently allocate attention during listening. For example, accuracy in recognizing sentence-final words varies with knowledge of and familiarity with the background sound for younger adults, but not for older adults (Russo and Pichora-Fuller 2008). Specifically, the performance of

younger listeners was best when the background was familiar music, next best when the background was unfamiliar music, and worst when the background was multitalker babble. Interestingly, in a surprise memory test, the younger adults recalled the background music that they had been instructed to ignore whereas the older adults remembered that music had been in the background but they were unable to recall which specific pieces of music had been played. These findings suggest that the younger listeners processed the incoming speech and music streams efficiently and had ample cognitive capacity to listen to and remember both the target and background music. In contrast, more cognitive resources seem to be consumed by the older listeners who focused all of their attention on listening to the foreground speech, with little attention to or memory of even the familiar music in the background (Russo and Pichora-Fuller 2008).

9.4.3 *Speech Maskers*

Compared to nonspeech signals, the speech of another talker is not so easily dismissed because it is highly similar to the speech of the target talker in terms of its spectrum, temporal fluctuations, and linguistic structure and meaningfulness. Informational masking will be greatest when the masker is meaningful speech. Listening when there is competing speech will involve peripheral and central auditory processing and also draw heavily on cognitive processing. For older adults with normal audiograms, declines in temporal or central auditory processing may undermine performance when the masker is speech. However, if the incoming speech signal matches familiar and expected linguistic structures and has semantic meaning that is appropriate to the situation, then it should be easier for a listener to parse the auditory stream. Conversely, speech understanding will be more difficult if the acoustical properties of speech are somewhat unfamiliar, for example, if the talker has an accent (Van Engen and Peelle 2014). Notably, older adults are more susceptible to background noise and accented speech (Gordon-Salant et al. 2015), but they are often more skilled than younger adults in using knowledge to compensate in challenging listening conditions.

9.5 Behavioral Measures of Age-Related Differences in the Perceptual Organization of Foreground Versus Background Sounds

A number of behavioral experimental paradigms have been used to compare how younger and older adults understand speech in situations similar to cocktail parties. Typically, after experiments have been conducted to establish the abilities of younger adults, similar experiments are conducted to measure the abilities of older

adults and to determine if there are age-related differences in performance. Age-related differences have been studied using experiments to evaluate spatial release from masking, stream segregation, the allocation of auditory spatial attention, the comprehension of discourse, and memory.

9.5.1 Spatial Separation and Release from Masking

In one common experimental paradigm used to evaluate release from masking, word recognition is measured using short, syntactically correct, but semantically anomalous sentences such as “A *rose* can *paint* a *fish*” (keywords in italics) that are presented in a variety of masking conditions (Freyman et al. 1999, 2004). The listener’s task is to repeat the sentence verbatim. The number of keywords that are repeated correctly is scored. The SRT in noise can be calculated if testing is done over a range of SNRs. Release from informational masking is measured as the difference in performance between conditions in which the masker is primarily energetic in nature (e.g., steady-state noise) and conditions in which the masker has a high informational content (e.g., competing talkers) (Kidd and Colburn, Chap. 4). Similarly, spatial release from masking is measured as the difference in performance between conditions with and without spatial separation of the target speech and masker (Culling and Stone, Chap. 3). The effect of spatial separation on release from masking can be determined using either real or simulated spatial separation of the target and maskers. Importantly, different auditory cues enable listeners to achieve release from masking depending on the nature of the maskers and on whether or not there is real or simulated spatial separation between the target and masker(s). It is possible to assess age-related differences in how these cues are used by evaluating release from masking across conditions.

9.5.1.1 Real Spatial Separation

Real separation of the speech of a target talker from a competing masker is achieved in experiments by presenting the target from one loudspeaker and the masker from another loudspeaker at a different location. In anechoic environments, only the direct wave from each loudspeaker arrives at the two ears of a listener. When the target is presented from a loudspeaker in front of a listener and the masker is presented from a loudspeaker to the right, interaural intensity differences occur at high frequencies because the head casts a shadow on the masking sound coming from the right loudspeaker before it reaches the left ear of the listener. Thus, for higher frequencies, the SNR at the person’s left ear is markedly higher than the SNR at the person’s right ear. In addition, useful low-frequency interaural time difference cues occur because there is an interaural delay for the masker but not the target. Using a combination of these interaural difference cues, the listener perceives the target talker to be in front and the masker at the right. Thus, benefit from spatial

separation between the target and masker depends on high-frequency interaural intensity differences and low-frequency interaural time differences. In general, interaural intensity differences alone contribute more spatial release from masking (around 8 dB), interaural time differences alone contribute less (around 5 dB), and in combination they provide more spatial release from masking (about 10 dB), although the effects are not additive (Bronkhurst and Plomp 1988).

For older adults who do not have significantly elevated pure-tone thresholds, the interaural intensity cues resulting from head shadow remain available. For those who have high-frequency threshold elevations, however, the interaural cues conferred by head shadow at higher frequencies may be reduced or eliminated. Nevertheless, even if these cues are available, they may be more advantageous to younger adults than to older adults. Recall that, in general, older adults need a 2–4 dB better SNR to match the speech understanding performance of younger listeners (see Sect. 9.2.2), likely owing to age-related declines in temporal processing, especially periodicity coding. Age-related declines in temporal and binaural processing could also reduce the ability of older adults to segregate competing talkers based on interaural differences in the temporal fine structure of competing voices.

The relative contributions of high-frequency and low-frequency cues to spatial release from masking were assessed in a study of younger adults and older adults with normal or impaired hearing as defined by the audiogram (Dubno et al. 2002). For sentences in speech-shaped noise (primarily an energetic masker), spatial release from masking was 6.1 dB for younger listeners, 4.9 dB for older listeners with normal pure-tone thresholds, and 2.7 dB for older listeners with pure-tone hearing loss. Not surprisingly, older adults with high-frequency hearing loss benefitted little from high-frequency cues resulting from head shadow. Compared to younger listeners, older adults with normal audiometric thresholds achieved less benefit from spatial separation, possibly because of less effective use of both high- and low-frequency cues.

In a more recent study (Besser et al. 2015), younger and older adults with normal hearing for their age were tested on the Listening in Spatialized Noise–Sentences (LiSN-S) test (Cameron and Dillon 2007, 2009). In the LiSN-S test, SRTs are determined for target sentences in four informational masking conditions: The target speech and masking speech are spoken by the same female or by different females and they are co-located or spatially separated. Scores are also calculated for the advantage (release from masking) due to talker differences, spatial separation, and both factors combined. Younger adults outperformed older adults on all SRT and advantage measures. Notably, spatial release from masking was 14.1 dB for the younger group and 9.6 dB for the older group. For both age groups, spatial release from masking was predicted by high-frequency (6–10 kHz) pure-tone thresholds. In addition, linguistic factors contributed to individual differences in the performance of the younger listeners and cognitive factors contributed to individual differences in the performance of the older listeners.

9.5.1.2 Simulated Spatial Separation

In contrast to experiments in which conditions of real spatial separation are tested, most everyday listening environments are reverberant. If the cocktail party is held indoors, then a direct speech wave will be reflected from all of the surfaces of the room and multiple reflections may continue to occur over time. The sound-absorbing properties of the surfaces affect reverberation time in terms of how long it takes the series of reflections to dampen. Long reverberation times can have deleterious effects on speech understanding in noise, especially for older listeners when the intensity level of speech is relatively low or the rate of speech is fast (Helfer and Wilber 1990; Gordon-Salant and Fitzgibbons 1995).

The delays between the direct wave and the first reflections depend on the distance between the listener and the room surfaces. In typical rooms, the delays between the direct and the first reflected waves are relatively short (2–8 ms). In such rooms, the listener perceives a single sound source at the location that is the origin of the direct wave and no echoes are perceived. In other words, the direct wave takes precedence (precedence effect; Zurek 1987). A second source, or echo, would not be heard unless the delay between the direct and reflected waves became very long, as would be the case in a very large space. Interestingly, when the precedence effect was simulated under headphones using time-delayed 2-kHz tone-pips, no age-related differences were found in the time delay at which listeners transitioned from perceiving a single source to perceiving two sound sources (Schneider et al. 1994).

The presence of a reflective surface can be simulated in an anechoic room by introducing a time delay in the presentation of a stimulus from two loudspeakers. For example, a listener perceives the location of a stimulus to come from the right when it is presented over a loudspeaker to the right beginning 4 ms before the same stimulus starts to be presented over a loudspeaker at the front. Similar to echoes in everyday reverberant environments, the delayed copy of the stimulus from the front loudspeaker is not perceived as a sound from a second source. Notably, when spatial separation is simulated in this way, the high-frequency interaural intensity difference cues arising from head shadow are largely eliminated and the SNRs at the two ears are equalized. As in the real spatial separation condition, the low-frequency interaural time difference cues remain available for the direct waves of the target and masker, but there are additional interaural difference cues for the simulated reflections.

For all listeners, speech understanding is better and spatial release from masking is greater when there is real, rather than simulated, spatial separation between target and masker. In a seminal study of younger adults, 12 dB of spatial release from masking was achieved when a real spatial separation was introduced between the target and competing speech, but only 3–9 dB was achieved when spatial separation was introduced in a simulation based on the precedence effect (Freyman et al. 1999). The most likely explanation for spatial release from masking being at least 3 dB poorer is that it is not possible to benefit from high-frequency interaural intensity and SNR differences when spatial separation is simulated. The superior

ability of younger adults to use interaural intensity and SNR difference cues could account for the age-related differences observed in conditions of real spatial separation. If so, then when spatial separation is simulated and the SNRs at the two ears are equalized, age-related differences in spatial release from masking should be less pronounced than they are in conditions of real spatial separation.

Younger and older adults were tested in a study of release from masking conducted using the same basic method as had been used in the seminal study of younger adults (Freyman et al. 1999). For both age groups, release from informational masking and spatial release from masking was evaluated by comparing the results obtained in four conditions with the positions of the target and maskers simulated using the precedence effect: (1) sentence target and noise masker co-located; (2) sentence target and noise masker spatially separated; (3) sentence target and speech masker co-located; and (4) sentence target and speech masker spatially separated (Li et al. 2004). There were three noteworthy findings. First, SRTs were approximately 3 dB SNR higher in older than in younger adults in all four conditions. This result is consistent with the more general finding that older adults need a higher SNR to achieve an SRT equivalent to that of younger adults. Second, the release from masking achieved by spatially separating the target and masker was the same for both age groups when the masker was two-talker speech (about 5 dB) and when the masker was steady-state noise (about 1.8 dB). Third, both age groups demonstrated a similar degree of release from informational masking when the target and maskers were co-located. Neither group demonstrated much, if any, release from informational masking when the target and masker were spatially separated, presumably because masking release had already been optimized based on the advantage conferred by spatially separating the target and masker.

Importantly, although the SRTs of the older listeners were 3 dB higher in all conditions, no significant age-related differences were found when spatial locations are simulated using the precedence effect and interaural intensity and SNR differences are minimized. Taken together, it seems that age-related differences understanding speech in multitalker scenes is attributable primarily to difficulties in auditory processing of interaural intensity and SNR cues rather than to declines in cognitive processing (Li et al. 2004).

9.5.2 Speed of Buildup of Stream Segregation

Stream segregation refers to the ability to disentangle sequences of sounds from competing sources, such as the task of forming distinct streams of speech from two or more talkers. The perception of segregated streams tends to build up over time (Bregman 1978). Some experimental evidence suggests that the buildup of stream segregation may proceed more slowly in older than in younger adults. In younger adults, word recognition improves as the delay between masker onset and word onset increases, whether the masker is steady-state noise or multitalker babble.

When the masker is steady-state noise, younger and older listeners show similar improvements, but when the masker is multitalker babble, there is no observable improvement by older adults for word-onset delays up to 1 s (Ben-David et al. 2012). Such slowing is not surprising in light of the evidence that there are age-related differences in auditory temporal processing, and also age-related generalized perceptual and cognitive slowing in adults (Salthouse 1996).

To investigate if age-related slowing in the buildup of stream segregation could influence word recognition during sentence processing, performance on the release from masking paradigm described in Sect. 9.5.1 (Freyman et al. 1999; Li et al. 2004) was examined to take the position of the key word into account (Ezzatian et al. 2012). For younger adults, when syntactically correct but semantically anomalous sentences are masked by co-located two-talker speech, word recognition improves from the first to the last keyword in a sentence. In contrast, when there is simulated spatial separation between the target and masker, there is substantial improvement in overall performance, but there is no improvement as a sentence unfolds. Whether or not listeners perceive the target and masker to be spatially separated, when the masker is a steady-state noise, word recognition is relatively easy, and there is no evidence that performance improves over time. This pattern of results for word recognition in anomalous sentences suggests that speech stream segregation is relatively rapid (less a second) in easier listening conditions (spatial separation or energetic masking). Speech stream segregation may take longer (a couple of seconds) and continue to develop over the course of a sentence being spoken when listening conditions are more challenging (no spatial separation or informational masking).

Like younger adults, older adults do not improve from the first to the last keyword position when the masker is a steady-state energetic noise masker (Ezzatian et al. 2015). For younger adults, stream segregation is slowed only when both the target and masker are intact, highly similar, and co-located speech stimuli, but older adults are slowed in a wider range of informational masking conditions. For older adults, stream segregation builds up over the course of the sentence when there is a two-talker masker, including when it is made more dissimilar to the target by either vocoding the masker to diminish the availability of fine-structure cues or by spatially separating the target and masker. Of course, the degree to which target and masking sounds are perceived to be dissimilar, and therefore, the degree to which they can be segregated from one another, could be affected by ARHL (see Sect. 9.2). For instance, age-related declines in temporal processing may account for the finding that speech stream segregation is rapid for younger listeners but slowed for older adults when the two-talker masker is vocoded. Furthermore, stream segregation is slowed in older listeners even though they can achieve spatial release from masking (Sect. 9.5.1). Age-related losses in neural synchrony are likely to degrade the interaural timing cues that contribute to locating an auditory object in space, thereby slowing stream segregation in situations where there is a spatial separation (either virtual or real) between a target voice and masking voices.

9.5.3 *Auditory Spatial Attention*

Humes et al. (2006) explored the effects of the acoustic similarity between a speech target and a speech masker in a study in which listeners attended to and reported the content of one of two sentences presented monaurally. The sentences were taken from the corpus of the coordinated response measure (CRM; Bolia et al. 2000) and have the form “ready (call sign), go to (color, number) now.” Call signs were the names of individuals and the colors and numbers were from a closed set (e.g., “Ready Baron go to green 2 now.”) Before or after the sentences were presented, participants were informed that the target sentence would begin with a particular call sign. The percentage of correctly identified color–number pairs was higher when the listener was informed of the call sign before rather than after the trial, presumably because prior knowledge of the call sign helped listeners to focus attention and reduce memory load. Performance was also higher when there was a gender difference than when there was no gender difference between the target talker and the masking talker, with the benefit from the voice pitch contrast being larger for younger than for older adults. It is possible that age-related declines in auditory temporal processing at the level of periodicity coding hamper the ability of older listeners to take advantage of gender-related differences in the fundamental frequency and harmonic structure of the voices of the target and masking talker, thereby slowing the buildup of stream segregation and impeding the efficient allocation of attention to the target speech stream.

CRM sentences have also been used to study how spatial attention affects word recognition in a three-talker display with real or simulated spatial separation between the target and two competing talkers (Singh et al. 2008). For a block of trials, the probability that the target would appear in each of the three possible locations varied from certainty (100%) to chance (33%), with two intermediate probabilities (80% and 60%). In general, older adults performed worse than younger adults in all conditions. Importantly, however, age did not interact with (1) the probability that the target would appear at a specific location, (2) whether or not the listener had prior knowledge of the call sign, or (3) whether or not the separation of the three sentences was real (coming from three different loudspeakers) or simulated (using the precedence effect). A follow-up analysis investigated the cost incurred when the target sentences were presented at an unlikely location instead of the most likely position (Singh et al. 2008). As expected, the cost of reallocating attention from the likely to the unlikely position was substantial in all conditions, with the extent of the reduction in performance being the same for both younger and older adults.

At the cocktail party, the need to redirect auditory spatial attention could happen if Fred unexpectedly begins talking (the listener’s attention is directed to Fred) and announces that everyone should listen to Mary because she has some important news to tell (Fred cues the listener to switch attention to Mary). To introduce such realistic attentional demands into the CRM experiment (Singh et al. 2008), new task instructions were used (Singh et al. 2013). As before, when the call sign appeared at the expected center location, participants were asked to report the color and number

associated with it. However, when the call sign appeared in an unexpected location (to the left or right of center), participants were asked to report the color and number from the sentence presented at the opposite side (i.e., they had to redirect their attention). As expected, there was a significant interaction between age and the complexity of the instructions (older simple instructions versus the new more complex instructions), with the older adults performing significantly worse than younger adults when the instructions increased the complexity of the task. These results suggest that older adults are not as agile as younger adults in redirecting their attention when the listening task is more demanding, as it might be in everyday situations.

9.5.4 Discourse—Beyond Words and Sentences

9.5.4.1 Adjusting SNR to Study Comprehension of Monologues

The difficulties that older adults have understanding speech in noise are not well explained by their audiometric thresholds. It is possible that their difficulties might be explained better by their SRTs in noise. Experiments using more complex linguistic materials were conducted to investigate how SRTs in noise might affect tasks requiring comprehension rather than only word recognition. Younger and older participants answered a series of questions concerning a lecture that they had just heard when the lecture was masked by multitalker babble presented from the same spatial location (Schneider et al. 2000). When the SNR (level of the lecture/level of the babble in dB) was the same for both age groups, older adults answered fewer questions correctly compared to younger adults. However, when the SNR was individually adjusted to take into account the higher SRTs in noise of older individuals, both age groups performed equivalently. These findings suggest that apparent age-related differences in comprehension could be attributed to the higher SRTs in noise of older adults.

9.5.4.2 Adjusting Spatial Separation in Dialogues and Triologues

In the experiment described in Sect. 9.5.4.1, both the lecture and the babble masker were mixed and presented monaurally over the same earphone (co-located condition). In more realistic everyday listening situations, including cocktail parties, talkers would be spatially separated and there would likely be more than two talkers in a conversation. In a follow-up experiment (Murphy et al. 2006), younger and older participants were asked to answer questions concerning two-person conversations. The dialogues and masking babble were played over a single central loudspeaker, or there was a real spatial separation between the three sound sources. After adjusting the SNR for individual differences in SRTs, both age groups answered the same number of questions correctly in the co-located condition, but younger adults outperformed older adults in the condition with spatial separation.

As described previously, it seems that older listeners do not benefit as much as younger listeners do from the availability of binaural cues when there is real separation between the sources (Sect. 9.5.1.1). Reduced benefit from binaural cues would make it more difficult for the older listeners to segregate and allocate attention effectively to the three streams. The influence of possible age-related differences in ability to use binaural cues was supported by the finding of no age-related differences when the experiment was repeated using the precedence effect to control the perceived locations of the stimuli (Avivi-Reich et al. 2014). Furthermore, the inadequacy of either pure-tone thresholds or SRTs in noise to account fully for the everyday listening problems of older adults is consistent with their self-reports on a questionnaire (Banh et al. 2012). Given that listeners must function in conditions in which there is real separation in the location of talkers at the cocktail party, even if the level of the background noise were reduced to improve the SNR, older partygoers would still struggle more than younger partygoers when conversing in group situations.

9.5.5 *Memory*

The preserved ability of older adults to comprehend discourse in most conditions (see Sect. 9.5.4) seems to be at odds with research on cognitive aging suggesting that memory for heard material is poorer in older than in younger adults. In most memory experiments, however, no corrections are made for age-related differences in the ability to hear the words. In addition, often the words to be recalled are presented in random lists rather than in meaningful sentences or discourse. Older adults benefit more than younger adults from contextual support for both recognizing and remembering words in sentences that are presented in babble (Pichora-Fuller et al. 1995). The discourse materials used in the comprehension experiments provided rich and socially relevant context. Older adults' knowledge of language and culture is preserved and is often superior to that of younger adults. It is possible that they used their expert knowledge and were able to take advantage of the richness of contextual support provided in discourse to compensate for poorer basic memory abilities. Alternatively, their poorer memory for heard words presented in lists may have arisen because they were not able to perceptually encode the words as precisely as younger adults owing to age-related changes in auditory processing.

To investigate the extent to which auditory aging is responsible for age-related differences in memory, the ability of younger and older adults to recall words in a paired-associates memory task was measured when the words were masked by babble, but with the SNRs adjusted for individuals' SRTs in noise (Murphy et al. 2000; Heinrich and Schneider 2011a, b). Even after adjusting SNRs to equate for individuals' SRTs in noise, older adults were less able to recall the words than younger adults in a wide variety of masking conditions. Interestingly, age-related differences were greatest when the masker was gated on and off with the to-be-remembered words, but they were less pronounced when words were heard in

continuous masking. Slower buildup of stream segregation in older adults may contribute to their memory problems when background noise and the words have simultaneous onsets. Overall, age-related declines in auditory processing do seem to exacerbate the memory problems of older adults. In everyday discourse, however, contextual support is abundant and it can help older adults to bind words into meaningful sequences that are easier to remember (for a more detailed discussion, see Schneider et al. 2016a, b).

9.6 Cognitive Aging and Sensory-Cognitive Interactions

9.6.1 *Cognitive Aging*

Some aspects of cognition decline with age, but others continue to improve. In general, there are declines in dynamic or fluid processing of information, whereas static or crystallized linguistic and world knowledge are well preserved in healthy aging. Importantly, the ability of older adults to use knowledge and contextual support is a strength that they can use to compensate for weaknesses in rapid information processing (Craik and Bialystok 2006). Age-related declines in cognitive processing that could affect communication include slower speed of information processing, reduced working memory, and difficulty dividing attention or selectively attending to relevant information while inhibiting distractions (Pichora-Fuller and Singh 2006).

9.6.2 *Sensory-Cognitive Interactions*

9.6.2.1 **Cognitively Healthy Older Adults**

There is growing evidence that audition and cognition interact, even in healthy older communicators who have clinically normal or near-normal audiograms and no clinically significant cognitive impairment (Schneider et al. 2010; Humes et al. 2013). Furthermore, for older adults with clinically significant audiometric threshold elevations, even when amplification has been provided to restore audibility, individual differences in understanding speech in noise remain and are associated with auditory temporal processing cognitive processing abilities (Humes 2007).

On the one hand, declines in auditory processing may impose increased demands on cognitive processing capacity. On the other hand, increased allocation of cognitive resources and use of knowledge can be compensatory when tasks involving listening are challenging (Grady 2012). Furthermore, age-related changes in brain activity and how complex tasks are performed involve more than the effects of ARHL. For older adults, the cognitive demands of multitasking can affect posture

and gait (Woollacott and Shumway-Cook 2002). Multisensory integration may reduce cognitive demands when information across modalities is congruent, but increase demands when it is incongruent (Mozolic et al. 2012), including during speech reading (Tye-Murray et al. 2010). Furthermore, social factors such as self-efficacy (Wingfield and Tun 2007), stigma, and ageist stereotypes may affect and be affected by age-related declines in auditory and cognitive performance (Chasteen et al. 2015; Pichora-Fuller 2016).

The interactions of auditory and cognitive aging are seen in how listeners contend with the challenging listening conditions of the cocktail party. In addition to the demands of listening, older adults may have difficulty multitasking or processing conflicting multisensory inputs as they mingle among the guests. Despite these demands on their cognitive resources, they may be motivated to interact socially. They may even benefit from what seem to be distractions so long as information is sufficiently congruent to support the allocation of attention (Weeks and Hasher 2014). When cognitive compensation is insufficient, however, and demands outweigh the possible benefits of social interaction, older adults may cope by withdrawing from noisy social situations.

9.6.2.2 Older Adults with Cognitive Loss

Provocative epidemiological findings indicate that cognitive loss is more prevalent and may progress more quickly in people with hearing loss compared to peers with good hearing, although the mechanisms underpinning these correlations are not yet known (Gates et al. 2011; Lin et al. 2013). Nevertheless, the increasingly common co-occurrence of declines in sensory loss and cognitive loss as people get older suggests that there is not simply increasing prevalence of these conditions with age but that they are interrelated (Albers et al. 2015). When sensory inputs are diminished, there can be short-term consequences to brain functioning. Long-term deprivation or alternations in processing can affect brain neuroplasticity. One possibility is that, as ARHL progresses over decades, the effects of information degradation on memory may become permanent (Dupuis et al. 2015). It remains to be determined if these cognitive declines could be slowed or prevented by auditory exercise such as playing music (Parbery-Clark et al. 2011), or if cognitive training would help older adults compensate for sensory aging (Reuter-Lorenz and Park 2014).

9.6.3 Brain Plasticity and Compensation

There is emerging evidence that the neural networks engaged when people are processing speech differ between younger and older adults (Harris et al. 2009). There is also behavioral evidence that the extent to which younger and older adults engage top-down processes in listening to speech is modulated by the listening situation. As listening becomes more challenging, compensatory use of knowledge

increases. Such knowledge includes lexical-level information, sentence-level information, discourse-level information, and world knowledge.

9.6.3.1 Vocabulary

In an analysis of the results of two studies (Schneider et al. 2000; Murphy et al. 2006), no significant correlation was found between how well listeners comprehended a lecture in quiet and the size of their vocabulary (Schneider et al. 2016a, b). However, when the same participants were tested in noisy backgrounds, listening comprehension was strongly correlated with vocabulary scores. These results indicate that when auditory input is degraded, top-down processes involving the use of linguistic knowledge facilitate lexical access for both younger and older adults. However, older adults are more vulnerable than younger adults when contexts are misleading (Rogers et al. 2012).

9.6.3.2 Sentences and Discourse

Once lexical access is achieved, additional processing is needed to integrate words into meaningful sentences, match this information to stored knowledge, construct inferences, and store the information for later recall. It is reasonable to assume that the post-lexical processes subsuming these tasks would be similar and modality independent (e.g., listening versus reading). Indeed, when listening is easy (quiet), and reading is easy (large font), the number of questions correctly answered concerning a story is highly correlated across modalities for both younger and older. In contrast, when listening is difficult (co-located babble masker) and reading is easy, listening comprehension is no longer significantly correlated with reading comprehension for older adults, although the correlation remains high for younger adults (Avivi-Reich et al. 2015). Importantly, age-related differences in listening comprehension were eliminated in these experiments when the SNR was adjusted according to individual participants' SRTs in noise. Hence, even though there was no age-related difference in the comprehension outcome measure, the results suggest that there are differences in the ways in which younger and older adults engage cognitive processes to achieve speech understanding.

9.7 Summary

Overall, the performance of older adults in situations like a cocktail party is often, but not always, poorer than that of younger adults. In general, older adults need a 2–4 dB better SNR to perform as well as younger adults on tasks involving speech understanding in noise. When the SNR is adjusted according to individual participants' SRTs in noise, many but not all age-related differences are eliminated.

Younger adults are better able to take advantage of the rich interaural cues provided when there is real spatial separation between targets and informational maskers. However, both age groups achieve a similar release from masking when spatial separation is simulated using the precedence effect. Older adults underperform compared to younger adults when speech is speeded and they demonstrate a slower buildup of stream segregation in a wider range of informational masking conditions. However, both age groups demonstrate similar benefit from allocating spatial attention when targets are presented at expected locations and instructions are simple. Older adults have poorer recall, especially when context is minimal. However, when context is available, older adults are better at using it to compensate for difficulties in hearing during comprehension and recall tasks.

Over the last three decades, much has been learned about auditory aging. Behavioral research demonstrates age-related declines in speech processing related to declines in auditory temporal processing at various levels. Electrophysiological research has advanced knowledge of the similarities and differences in how the brains of younger and older adults are engaged in processing complex auditory and speech information. Future research will explore further how auditory aging interacts with age-related changes in cognition and across nonauditory aspects of sensorimotor function. The interactions of these multiple sensory, motor, cognitive, and social factors and how they change over the course of adult aging will need to be studied to understand fully how older adults listen a cocktail parties and in the other complex auditory scenes in everyday life.

Acknowledgements This work was supported by grants to M. Kathleen Pichora-Fuller from the Natural Sciences and Engineering Research Council of Canada (RGPIN 138472), to Bruce Schneider from the Canadian Institutes of Health Research (MOP-15359, TEA-1249) and the Natural Sciences and Engineering Research Council of Canada (RGPIN-9952-13), and to Claude Alain from the Canadian Institutes of Health Research (MOP 106619).

Compliance with Ethics Requirements

M. Kathleen Pichora-Fuller has no conflicts of interest.

Claude Alain has no conflicts of interest.

Bruce A. Schneider has no conflicts of interest.

References

- Abel, S. M., Krever, E. M., & Alberti, P. W. (1990). Auditory detection, discrimination and speech processing in ageing, noise-sensitive and hearing-impaired listeners. *Scandinavian Audiology*, 19(1), 43–54.
- Alain, C. (2007). Breaking the wave: Effects of attention and learning on concurrent sound perception. *Hearing Research*, 229(1–2), 225–236.
- Alain, C., Dyson, B. J., & Snyder, J. S. (2006). Aging and the perceptual organization of sounds: A change of scene? In M. Conn (Ed.), *Handbook of models for the study of human aging* (pp. 759–769). Amsterdam: Elsevier Academic Press.

- Alain, C., McDonald, K. L., Ostroff, J. M., & Schneider, B. A. (2001). Age-related changes in detecting a mistuned harmonic. *The Journal of the Acoustical Society of America*, *109*(5), 2211–2216.
- Alain, C., McDonald, K. L., Ostroff, J. M., & Schneider, B. A. (2004). Aging: A switch from automatic to controlled processing of sounds? *Psychology and Aging*, *19*(1), 125–133.
- Alain, C., McDonald, K., & Van Roon, P. (2012). Effects of age and background noise on processing a mistuned harmonic in an otherwise periodic complex sound. *Hearing Research*, *283*(1–2), 126–135.
- Alain, C., Roye, A., & Arnott, S. A. (2013). Middle and late auditory evoked responses: What are they telling us on central auditory disorders? In G. G. Celesia (Ed.), *Disorders of peripheral and central auditory processing* (pp. 177–199, Vol. 10: Handbook of clinical neurophysiology). Amsterdam, The Netherlands: Elsevier.
- Alain, C., & Snyder, J. S. (2008). Age-related differences in auditory evoked responses during rapid perceptual learning. *Clinical Neurophysiology*, *119*(2), 356–366.
- Alain, C., & Woods, D. L. (1999). Age-related changes in processing auditory stimuli during visual attention: Evidence for deficits in inhibitory control and sensory memory. *Psychology and Aging*, *14*(3), 507–519.
- Albers, M. W., Gilmore, G. C., Kaye, J., Murphy, C., et al. (2015). At the interface of sensory and motor dysfunctions and Alzheimer's disease. *Alzheimer's and Dementia*, *11*(1), 70–98.
- Anderer, P., Semlitsch, H. V., & Saletu, B. (1996). Multichannel auditory event-related brain potentials: Effects of normal aging on the scalp distribution of N1, P2, N2 and P300 latencies and amplitudes. *Electroencephalography and Clinical Neurophysiology*, *99*(5), 458–472.
- Anderson, S., Parbery-Clark, A., White-Schwoch, T., & Kraus, N. (2012). Aging affects neural precision of speech encoding. *The Journal of Neuroscience*, *32*(41), 14156–14164.
- Avivi-Reich, M., Daneman, M., & Schneider, B. A. (2014). How age and linguistic competence alter the interplay of perceptual and cognitive factors when listening to conversations in a noisy environment. *Frontiers in Systems Neuroscience*, *8*. doi:10.3389/fnsys.2014.00021
- Avivi-Reich, M., Jakubczyk, A., Daneman, M., & Schneider, B. A. (2015). How age, linguistic status, and the nature of the auditory scene alter the manner in which listening comprehension is achieved in multitalker conversations. *Journal of Speech Language and Hearing Research*, *58*(5), 1570–1591.
- Banh, J., Singh, G., & Pichora-Fuller, M. K. (2012). Age affects responses on the speech, spatial, and qualities of hearing scale (SSQ) for adults with minimal audiometric loss. *Journal of the American Academy of Audiology*, *23*(2), 81–91.
- Ben-David, B. M., Tse, V. Y. Y., & Schneider, B. A. (2012). Does it take older adults longer than younger adults to perceptually segregate a speech target from a background masker? *Hearing Research*, *290*(1–2), 55–63.
- Bergman, M. (1980). *Aging and the perception of speech*. Baltimore: University Park Press.
- Besser, J., Festen, J. M., Goverts, S. T., Kramer, S. E., & Pichora-Fuller, M. K. (2015). Speech-in-speech listening on the LiSN-S test by older adults with good audiograms depends on cognition and hearing acuity at high frequencies. *Ear and Hearing*, *36*(1), 24–41.
- Bidelman, G. M., Gandour, J. T., & Krishnan, A. (2011). Musicians demonstrate experience-dependent brainstem enhancement of musical scale features within continuously gliding pitch. *Neuroscience Letters*, *503*(3), 203–207.
- Bidelman, G. M., & Krishnan, A. (2009). Neural correlates of consonance, dissonance, and the hierarchy of musical pitch in the human brainstem. *The Journal of Neuroscience*, *29*(42), 13165–13171.
- Bidelman, G. M., & Krishnan, A. (2010). Effects of reverberation on brainstem representation of speech in musicians and non-musicians. *Brain Research*, *1355*, 112–125.
- Bidelman, G. M., Moreno, S., & Alain, C. (2013). Tracing the emergence of categorical speech perception in the human auditory system. *NeuroImage*, *79*, 201–212.
- Bidelman, G. M., Villafuerte, J. W., Moreno, S., & Alain, C. (2014). Age-related changes in the subcortical-cortical encoding and categorical perception of speech. *Neurobiology of Aging*, *35* (11), 2526–2540.

- Bolia, R. S., Nelson, W. T., Ericson, M. A., & Simpson, B. D. (2000). A speech corpus for multitalter communications research. *The Journal of the Acoustical Society of America*, *107*(2), 1065–1066.
- Borg, E., Bergkvist, C., Olsson, I.-S., Wikström, C., & Borg, B. (2008). Communication as an ecological system. *International Journal of Audiology*, *47*(Suppl. 2), S131–S138.
- Bregman, A. S. (1978). Auditory streaming is cumulative. *Journal of Experimental Psychology: Human Perception and Performance*, *4*, 380–387.
- Bronkhurst, A. W., & Plomp, R. (1988). The effect of head-induced interaural time and level differences on speech intelligibility in noise. *The Journal of the Acoustical Society of America*, *83*, 1508–1516.
- Cameron, S., & Dillon, H. (2007). Development of the listening in spatialized noise—sentences test. *Ear and Hearing*, *28*(2), 196–211.
- Cameron, S., & Dillon, H. (2009). *Listening in spatialized noise—sentences test (LiSN-S)*. Murten, Switzerland: Phonak Communications AG.
- CHABA. (Committee on Hearing, Bioacoustics and Biomechanics). (1988). Speech understanding and aging. *The Journal of the Acoustical Society of America*, *83*(3), 859–895.
- Chang, E. F., Rieger, J. W., Johnson, K., Berger, M. S., et al. (2010). Categorical speech representation in human superior temporal gyrus. *Nature Neuroscience*, *13*(11), 1428–1432.
- Chao, L. L., & Knight, R. T. (1997). Prefrontal deficits in attention and inhibitory control with aging. *Cerebral Cortex*, *7*(1), 63–69.
- Chasteen, A., Pichora-Fuller, M. K., Dupuis, K., Smith, S., & Singh, G. (2015). Do negative views of aging influence memory and auditory performance through self-perceived abilities? *Psychology and Aging*, *30*(4), 881–893.
- Craik, F. I. M., & Bialystok, E. (2006). Lifespan cognitive development: The roles of representation and control. In F. I. M. Craik & Salthouse, T. A. (Eds.), *The handbook of aging and cognition* (3rd ed., pp. 557–602). New York: Psychology Press.
- Cruikshanks, K. J., Zhan, W., & Zhong, W. (2010). Epidemiology of age-related hearing impairment. In S. Gordon-Salant, R. D. Frisina, A. Popper, & R. R. Fay (Eds.), *The aging auditory system: Perceptual characterization and neural bases of presbycusis* (pp. 259–274). New York: Springer Science + Business Media.
- Dubno, J. R., Ahlstrom, J. B., & Horwitz, A. R. (2002). Spectral contributions to the benefit from spatial separation of speech and noise. *Journal of Speech, Language, and Hearing Research*, *45*(12), 1297–1310.
- Dupuis, K., & Pichora-Fuller, M. K. (2014). Intelligibility of emotional speech in younger and older adults. *Ear and Hearing*, *35*(6), 695–707.
- Dupuis, K., & Pichora-Fuller, M. K. (2015). Aging affects identification of vocal emotions in semantically neutral sentences. *Journal of Speech, Language and Hearing Research*, *58*(3), 1061–1076.
- Dupuis, K., Pichora-Fuller, M. K., Marchuk, V., Chasteen, A., et al. (2015). Effects of hearing and vision impairments on the montreal cognitive assessment. *Aging, Neuropsychology, and Cognition*, *22*(4), 413–427.
- Eddins, D. A., & Hall III, J. W. (2010). Binaural processing and auditory asymmetries. In S. Gordon-Salant, R. D. Frisina, A. Popper, & R. R. Fay (Eds.), *The aging auditory system: Perceptual characterization and neural bases of presbycusis* (pp. 135–166). New York: Springer Science + Business Media.
- Ezzatian, P., Li, L., Pichora-Fuller, M. K., & Schneider, B. A. (2012). The effect of energetic and informational masking on the time-course of stream segregation: Evidence that streaming depends on vocal fine structure cues. *Language and Cognitive Processes*, *27*(7–8), 1056–1088.
- Ezzatian, P., Li, L., Pichora-Fuller, M. K., & Schneider, B. A. (2015). Delayed stream segregation in older adults: More than just informational masking. *Ear and Hearing*, *36*(4), 482–484.
- Fitzgibbons, P. J., & Gordon-Salant, S. (2001). Aging and temporal discrimination in auditory sequences. *The Journal of the Acoustical Society of America*, *109*(6), 2955–2963.
- Fitzgibbons, P. J., & Gordon-Salant, S. (2010). Behavioral studies with aging humans: Hearing sensitivity and psychoacoustics. In S. Gordon-Salant, R. D. Frisina, A. Popper, & R. R. Fay

- (Eds.), *The aging auditory system: Perceptual characterization and neural bases of presbycusis* (pp. 111–135). New York: Springer Science + Business Media.
- Fitzgibbons, P. J., Gordon-Salant, S., & Barrett, J. (2007). Age-related differences in discrimination of an interval separating onsets of successive tone bursts as a function of interval duration. *The Journal of the Acoustical Society of America*, *122*(1), 458–466.
- Freyman, R. L., Balakrishnan, U., & Helfer, K. S. (2004). Effect of number of masking talkers and auditory priming on informational masking in speech recognition. *The Journal of the Acoustical Society of America*, *115*(5), 2246–2256.
- Freyman, R. L., Helfer, K. S., McCall, D. D., & Clifton, R. K. (1999). The role of perceived spatial separation in the unmasking of speech. *The Journal of the Acoustical Society of America*, *106*(6), 3578–3588.
- Füllgrabe, C., Moore, B. C. J., & Stone, M. A. (2014). Age-group differences in speech identification despite matched audiometrically normal hearing: Contributions from auditory temporal processing and cognition. *Frontiers in Aging Neuroscience*, *6*, 347.
- Gates, G. A., Anderson, M. L., McCurry, S. M., Feeney, M. P., & Larson, E. B. (2011). Central auditory dysfunction as a harbinger of Alzheimer’s dementia. *Archives of Otolaryngology-Head and Neck Surgery*, *137*(4), 390–395.
- Gordon-Salant, S., & Fitzgibbons, P. J. (1995). Recognition of multiply degraded speech by young and elderly listeners. *Journal of Speech and Hearing Research*, *38*(5), 1150–1156.
- Gordon-Salant, S., Frisina, R. D., Popper, A. N., & Fay, R. R. (Eds.). (2010). *The aging auditory system: Perceptual characterization and neural bases of presbycusis*. New York: Springer Science + Business Media.
- Gordon-Salant, S., Yeni-Komshian, G. H., Fitzgibbons, P. J., & Barrett, J. (2006). Age-related differences in identification and discrimination of temporal cues in speech segments. *The Journal of the Acoustical Society of America*, *119*(4), 2455–2466.
- Gordon-Salant, S., Yeni-Komshian, G. H., Fitzgibbons, P. J., & Cohen, J. (2015). Effects of age and hearing loss on recognition of unaccented and accented multisyllabic words. *The Journal of the Acoustical Society of America*, *137*(2), 884–897.
- Goy, H., Pelletier, M., Coletta, M., & Pichora-Fuller, M. K. (2013). The effects of semantic context and the type and amount of acoustical distortion on lexical decision by younger and older adults. *Journal of Speech, Language and Hearing Research*, *56*(6), 1715–1732.
- Grady, C. L. (2012). The cognitive neuroscience of ageing. *Nature Reviews Neuroscience*, *13*(7), 491–505.
- Greenberg, S. (1996). Auditory processing of speech. In N. J. Lass (Ed.), *Principles of experimental phonetics* (pp. 362–407). St. Louis, MO: Mosby.
- Greenhut-Wertz, J., & Manning, S. K. (1995). Suffix effects and intrusion errors in young and elderly subjects. *Experimental Aging Research*, *21*(2), 173–190.
- Harris, K. C., Dubno, J. R., Keren, N. I., Ahlstrom, J. B., & Eckert, M. A. (2009). Speech recognition in younger and older adults: A dependency on low-level auditory cortex. *The Journal of Neuroscience*, *29*(19), 6078–6087.
- Haubert, N., & Pichora-Fuller, M. K. (1999). The perception of spoken language by elderly listeners: Contribution of auditory temporal processes. *Canadian Acoustics*, *27*(3), 96–97.
- He, N., Horwitz, R., Dubno, J. R., & Mills, J. H. (1999). Psychometric functions for gap detection in noise measured from young and aged subjects. *The Journal of the Acoustical Society of America*, *106*(2), 966–978.
- He, N., Mills, J. H., & Dubno, J. R. (2007). Frequency modulation detection: Effects of age, psychophysical method, and modulation waveform. *The Journal of the Acoustical Society of America*, *122*(1), 467–477.
- Heinrich, A., De la Rosa, S., & Schneider, B. A. (2014). The role of stimulus complexity, spectral overlap, and pitch for gap-detection thresholds in young and old listeners. *The Journal of the Acoustical Society of America*, *136*(4), 1797–1807.
- Heinrich, A., & Schneider, B. A. (2011a). The effect of presentation level on memory performance. *Ear and Hearing*, *32*(4), 524–532.

- Heinrich, A., & Schneider, B. A. (2011b). Elucidating the effects of aging on remembering perceptually distorted word-pairs. *Quarterly Journal of Experimental Psychology*, *64*(1), 186–205.
- Helfer, K. S., & Wilber, L. A. (1990). Hearing loss, aging, and speech perception in reverberation and noise. *Journal of Speech and Hearing Research*, *33*(1), 149–155.
- Humes, L. E. (2007). The contributions of audibility and cognitive factors to the benefit provided by amplified speech to older adults. *Journal of the American Academy of Audiology*, *18*(7), 590–603.
- Humes, L. E., Busey, T. A., Craig, J., & Kewley-Port, D. (2013). Are age-related changes in cognitive function driven by age-related changes in sensory processing? *Attention, Perception, & Psychophysics*, *75*(3), 508–524.
- Humes, L. E., & Dubno, J. R. (2010). Factors affecting speech understanding in older adults. In S. Gordon-Salant, R. D. Frisina, A. N. Popper, & R. R. Fay (Eds.), *The aging auditory system: Perceptual characterization and neural bases of presbycusis* (pp. 211–258). New York: Springer Science + Business Media.
- Humes, L. E., Lee, J. H., & Coughlin, M. P. (2006). Auditory measures of selective and divided attention in young and older adults using single-talker competition. *The Journal of the Acoustical Society of America*, *120*(5), 2926–2937.
- Hutka, S. A., Alain, C., Binns, M. A., & Bidelman, G. M. (2013). Age-related differences in the sequential organization of speech sounds. *The Journal of the Acoustical Society of America*, *133*(6), 4177–4187.
- Iragui, V. J., Kutas, M., Mitchiner, M. R., & Hillyard, S. A. (1993). Effects of aging on event-related brain potentials and reaction times in an auditory oddball task. *Psychophysiology*, *30*(1), 10–22.
- ISO. (International Organization for Standardization). (2000). *Acoustics: Statistical distribution of hearing thresholds as a function of age, ISO 7029*. Geneva: International Organization of Standards.
- Kiessling, J., Pichora-Fuller, M. K., Gatehouse, S., Stephens, D., et al. (2003). Candidature for and delivery of audiological services: Special needs of older people. *International Journal of Audiology*, *42*(Supp 2), S92–S101.
- Krishnan, A., Bidelman, G. M., & Gandour, J. T. (2010). Neural representation of pitch salience in the human brainstem revealed by psychophysical and electrophysiological indices. *Hearing Research*, *268*(1–2), 60–66.
- Kujawa, S. G., & Liberman, M. C. (2009). Adding insult to injury: Cochlear nerve degeneration after “temporary” noise-induced hearing loss. *The Journal of Neuroscience*, *29*(45), 14077–14085.
- Leung, A. W. S., He, Y., Grady, C. L., & Alain, C. (2013). Age differences in the neuroelectric adaptation to meaningful sounds. *PLoS ONE*, *8*(7), e68892.
- Li, L., Daneman, M., Qi, J., & Schneider, B. A. (2004). Does the information content of an irrelevant source differentially affect spoken word recognition in younger and older adults? *Journal of Experimental Psychology: Human Perception and Performance*, *30*(6), 1077–1091.
- Lin, F. R., Yaffe, K., Xia, J., Xue, Q. L., et al. (2013). Hearing loss and cognitive decline in older adults. *JAMA Internal Medicine*, *173*(4), 293–299.
- Lister, J., Besing, J., & Koehnke, J. (2002). Effects of age and frequency disparity on gap discrimination. *The Journal of the Acoustical Society of America*, *111*(6), 2793–2800.
- Lister, J. J., Maxfield, N. D., Pitt, G. J., & Gonzalez, V. B. (2011). Auditory evoked response to gaps in noise: Older adults. *International Journal of Audiology*, *50*(4), 211–225.
- Lister, J., & Tarver, K. (2004). Effect of age on silent gap discrimination in synthetic speech stimuli. *Journal of Speech, Language, and Hearing Research*, *47*(2), 257–268.
- Lopez-Poveda, E. A. (2014). Why do I hear but not understand? stochastic undersampling as a model of degraded neural encoding of speech. *Frontiers in Neuroscience*, *8*, 348.
- MacDonald, E., Pichora-Fuller, M. K., & Schneider, B. A. (2007). Intensity discrimination in noise: Effect of aging. In *Proceedings of the 23rd Annual Meeting of the International Society for Psychophysicists* (pp. 135–140), Tokyo.

- Mills, J. H., Schmiedt, R. A., Schulte, B. A., & Dubno, J. R. (2006). Age-related hearing loss: A loss of voltage, not hair cells. *Seminars in Hearing, 27*(4), 228–236.
- Mozolic, J. L., Hugenschmidt, C. E., Peiffer, A. M., & Laurienti, P. J. (2012). Multisensory integration and aging. In M. M. Murray & M. T. Wallace (Eds.), *The neural bases of multisensory processes*. Boca Raton, FL: CRC Press.
- Murphy, D. R., Craik, F. I. M., Li, K., & Schneider, B. A. (2000). Comparing the effects of aging and background noise on short-term memory performance. *Psychology and Aging, 15*(2), 323–334.
- Murphy, D. R., Daneman, M., & Schneider, B. A. (2006). Why do older adults have difficulty following conversations? *Psychology and Aging, 21*(1), 49–61.
- Ostroff, J. M., McDonald, K. L., Schneider, B. A., & Alain, C. (2003). Aging and the processing of sound duration in human auditory cortex. *Hearing Research, 181*(1–2), 1–7.
- Parbery-Clark, A., Strait, D. L., Anderson, S., Hittner, E., & Kraus, N. (2011). Musical experience and the aging auditory system: Implications for cognitive abilities and hearing speech in noise. *PLoS ONE, 6*(5), e18082.
- Peelle, J. E., & Davis, M. H. (2012). Neural oscillations carry speech rhythm through to comprehension. *Frontiers in Psychology, 3*, 320.
- Phillips, D. P. (1995). Central auditory processing: A view from auditory neuroscience. *American Journal of Otology, 16*(3), 338–352.
- Phillips, D. P., Taylor, T. L., Hall, S. E., Carr, M. M., & Mossop, J. E. (1997). Detection of silent intervals between noises activating different perceptual channels: Some properties of ‘central’ auditory gap detection. *The Journal of the Acoustical Society of America, 101*(6), 3694–3705.
- Pichora-Fuller, M. K. (2016). How social factors may modulate auditory and cognitive functioning during listening. *Ear and Hearing, 37*(Suppl.), 92S–100S.
- Pichora-Fuller, M. K., Kramer, S. E., Eckert, M., Edwards, B., et al. (2016). Consensus report on Eriksholm “Hearing Impairment and Cognitive Energy” workshop. *Ear and Hearing, 37* (Suppl.), 5S–S27.
- Pichora-Fuller, M. K., & Schneider, B. A. (1992). The effect of interaural delay of the masker on masking-level differences in young and elderly listeners. *The Journal of the Acoustical Society of America, 91*(4), 2129–2135.
- Pichora-Fuller, M. K., Schneider, B. A., Benson, N. J., Hamstra, S. J., & Storzer, E. (2006). Effect of age on detection of gaps in speech and nonspeech markers varying in duration and spectral symmetry. *The Journal of the Acoustical Society of America, 119*(2), 1143–1155.
- Pichora-Fuller, M. K., Schneider, B. A., & Daneman, M. (1995). How young and old adults listen to and remember speech in noise. *The Journal of the Acoustical Society of America, 97*(1), 593–608.
- Pichora-Fuller, M. K., Schneider, B. A., MacDonald, E., Brown, S., & Pass, H. (2007). Temporal jitter disrupts speech intelligibility: A simulation of auditory aging. *Hearing Research, 223*(1–2), 114–121.
- Pichora-Fuller, M. K., & Singh, G. (2006). Effects of age on auditory and cognitive processing: Implications for hearing aid fitting and audiological rehabilitation. *Trends in Amplification, 10* (1), 29–59.
- Picton, T., Alain, C., Woods, D. L., John, M. S., et al. (1999). Intracerebral sources of human auditory-evoked potentials. *Audiology and Neuro-Otology, 4*(2), 64–79.
- Purcell, D. W., John, S. M., Schneider, B. A., & Picton, T. W. (2004). Human temporal auditory acuity as assessed by envelope following responses. *The Journal of the Acoustical Society of America, 116*(6), 3581–3593.
- Reuter-Lorenz, P. A., & Park, D. C. (2014). How does it STAC up? Revisiting the scaffolding theory of aging and cognition. *Neuropsychology Review, 24*(3), 355–370.
- Rogers, C. S., Jacoby, L. L., & Sommers, M. S. (2012). Frequent false hearing by older adults: The role of age differences in metacognition. *Psychology and Aging, 27*(1), 33–45.
- Ross, B., Fujioka, T., Tremblay, K. L., & Picton, T. W. (2007). Aging in binaural hearing begins in mid-life: Evidence from cortical auditory-evoked responses to changes in interaural phase. *The Journal of Neuroscience, 27*(42), 11172–11178.

- Ross, B., Schneider, B., Snyder, J. S., & Alain, C. (2010). Biological markers of auditory gap detection in young, middle-aged, and older adults. *PLoS ONE*, *5*(4), e10101.
- Ross, B., Snyder, J. S., Aalto, M., McDonald, K. L., et al. (2009). Neural encoding of sound duration persists in older adults. *NeuroImage*, *47*(2), 678–687.
- Russo, F. A., Ives, D. T., Goy, H., Pichora-Fuller, M. K., & Patterson, R. D. (2012). Age-related difference in melodic pitch perception is probably mediated by temporal processing: Empirical and computational evidence. *Ear and Hearing*, *33*(2), 177–186.
- Russo, F., & Pichora-Fuller, M. K. (2008). Tune in or tune out: Age-related differences in listening when speech is in the foreground and music is in the background. *Ear and Hearing*, *29*, 746–760.
- Salthouse, T. A. (1996). The processing-speed theory of adult age differences in cognition. *Psychological Review*, *103*(3), 403–428.
- Saremi, A., & Stenfelt, S. (2013). Effect of metabolic presbycusis on cochlear responses: A simulation approach using a physiologically-based model. *Journal of Acoustical Society of America*, *134*(4), 2833–2851.
- Schmiedt, R. A. (2010). The physiology of cochlear presbycusis. In S. Gordon-Salant, R. D. Frisina, A. Popper, & R. R. Fay (Eds.), *The aging auditory system: Perceptual characterization and neural bases of presbycusis* (pp. 9–38). New York: Springer Science + Business Media.
- Schneider, B. A., Avivi-Reich, M., & Daneman, M. (2016a). How spoken language comprehension is achieved by older listeners in difficult listening situations. *Experimental Aging Research*, *42*(1), 40–63.
- Schneider, B. A., Avivi-Reich, M., Leung, C., & Heinrich, A. (2016b). How age and linguistic competence affect memory for heard information. *Frontiers in Psychology*, *7*, 618.
- Schneider, B. A., Daneman, M., Murphy, D. R., & Kwong See, S. (2000). Listening to discourse in distracting settings: The effects of aging. *Psychology and Aging*, *15*(1), 110–125.
- Schneider, B. A., & Hamstra, S. (1999). Gap detection thresholds as a function of tonal duration for younger and older listeners. *The Journal of the Acoustical Society of America*, *106*(1), 371–380.
- Schneider, B. A., Pichora-Fuller, M. K., & Daneman, M. (2010). The effects of senescent changes in audition and cognition on spoken language comprehension. In S. Gordon-Salant, R. D. Frisina, A. Popper, & R. R. Fay (Eds.), *The aging auditory system: Perceptual characterization and neural bases of presbycusis* (pp. 167–210). New York: Springer Science + Business Media.
- Schneider, B. A., Pichora-Fuller, M. K., Kowalchuk, D., & Lamb, M. (1994). Gap detection and the precedence effect in young and old adults. *The Journal of the Acoustical Society of America*, *95*(2), 980–991.
- Sheldon, S., Pichora-Fuller, M. K., & Schneider, B. A. (2008). Effect of age, presentation method, and learning on identification of noise-vocoded words. *The Journal of the Acoustical Society of America*, *123*(1), 476–488.
- Singh, G., Pichora-Fuller, M. K., & Schneider, B. A. (2008). The effect of age on auditory spatial attention in conditions of real and simulated spatial separation. *The Journal of the Acoustical Society of America*, *124*(2), 1294–1305.
- Singh, G., Pichora-Fuller, M. K., & Schneider, B. A. (2013). Time course and cost of misdirecting auditory spatial attention in younger and older adults. *Ear and Hearing*, *34*(6), 711–721.
- Smith, S. L., Pichora-Fuller, M. K., Wilson, R. H., & MacDonald, E. N. (2012). Word recognition for temporally and spectrally distorted materials: The effects of age and hearing loss. *Ear and Hearing*, *33*(3), 349–366.
- Snell, K. B., & Frisina, D. R. (2000). Relationships among age-related differences in gap detection and word recognition. *The Journal of the Acoustical Society of America*, *107*(3), 1615–1626.
- Snyder, J. S., & Alain, C. (2005). Age-related changes in neural activity associated with concurrent vowel segregation. *Cognitive Brain Research*, *24*(3), 492–499.
- Souza, P. E., & Boike, K. T. (2006). Combining temporal-envelope cues across channels: Effects of age and hearing loss. *Journal of Speech, Language, and Hearing Research*, *49*(1), 138–149.

- Tremblay, K. L., Piskosz, M., & Souza, P. (2003). Effects of age and age-related hearing loss on the neural representation of speech cues. *Clinical Neurophysiology*, *114*(7), 1332–1343.
- Tye-Murray, N., Sommers, M., Spehar, B., Myerson, J., & Hale, S. (2010). Aging, audiovisual integration, and the principle of inverse effectiveness. *Ear and Hearing*, *31*(5), 636–644.
- Van Engen, K. J., & Peelle, J. E. (2014). Listening effort and accented speech. *Frontiers in Human Neuroscience*, *8*, 577.
- Vaughan, N., Storzbach, D., & Furukawa, I. (2008). Investigation of potential cognitive tests for use with older adults in audiology clinics. *Journal of the American Academy of Audiology*, *19* (7), 533–541.
- Versfeld, N. J., & Dreschler, W. A. (2002). The relationship between the intelligibility of time-compressed speech and speech in noise in young and elderly listeners. *The Journal of the Acoustical Society of America*, *111*(11), 401–408.
- Vongpaisal, T., & Pichora-Fuller, M. K. (2007). Effect of age on F₀ difference limen and concurrent vowel identification. *Journal of Speech, Language, and Hearing Research*, *50*(5), 1139–1156.
- Walton, J. P. (2010). Timing is everything: Temporal processing deficits in the aged auditory brainstem. *Hearing Research*, *264*(1–2), 63–69.
- Wang, M., Wu, X., Li, L., & Schneider, B. A. (2011). The effects of age and interaural delay on detecting a change in interaural correlation: The role of temporal jitter. *Hearing Research*, *275* (1–2), 139–149.
- Weeks, J. C., & Hasher, L. (2014). The disruptive—and beneficial—effects of distraction on older adults' cognitive performance. *Frontiers in Psychology*, *5*, 133.
- Wingfield, A., Lindfield, K. C., & Goodglass, H. (2000). Effects of age and hearing sensitivity on the use of prosodic information in spoken word recognition. *Journal of Speech, Language, and Hearing Research*, *43*(4), 915–925.
- Wingfield, A., McCoy, S. L., Peelle, J. E., Tun, P. A., & Cox, L. C. (2006). Effects of adult aging and hearing loss on comprehension of rapid speech varying in syntactic complexity. *Journal of the American Academy of Audiology*, *17*(7), 487–497.
- Wingfield, A., & Tun, P. A. (2007). Cognitive supports and cognitive constraints on comprehension of spoken language. *Journal of the American Academy of Audiology*, *18*(7), 548–559.
- Wingfield, A., Tun, P. A., Koh, C. K., & Rosen, M. J. (1999). Regaining lost time: Adult aging and the effect of time restoration on recall of time-compressed speech. *Psychology and Aging*, *14*(3), 380–389.
- Wingfield, A., Wayland, S. C., & Stine, E. A. (1992). Adult age differences in the use of prosody for syntactic parsing and recall of spoken sentences. *Journals of Gerontology*, *47*(5), P350–P356.
- Woods, D. L., & Clayworth, C. C. (1986). Age-related changes in human middle latency auditory evoked potentials. *Electroencephalography and Clinical Neurophysiology*, *65*(4), 297–303.
- Woollacott, M., & Shumway-Cook, A. (2002). Attention and the control of posture and gait: A review of an emerging area of research. *Gait Posture*, *16*(1), 1–14.
- Yueh, B., Shapiro, N., MacLean, C. H., & Shekelle, P. G. (2003). Screening and management of adult hearing loss in primary care: Scientific review. *JAMA*, *289*(15), 1976–1985.
- Zurek, P. M. (1987). The precedence effect. In W. A. Yost & G. Gourevitch (Eds.), *Directional hearing* (pp. 85–105). New York: Springer-Verlag.

Chapter 10

Hearing with Cochlear Implants and Hearing Aids in Complex Auditory Scenes

Ruth Y. Litovsky, Matthew J. Goupell,
Sara M. Misurelli, and Alan Kan

Abstract One of the most important tasks that humans face is communication in complex, noisy acoustic environments. In this chapter, the focus is on populations of children and adult listeners who suffer from hearing loss and are fitted with cochlear implants (CIs) and/or hearing aids (HAs) in order to hear. The clinical trend is to provide patients with the ability to hear in both ears. This trend to stimulate patients in both ears has stemmed from decades of research with normal-hearing (NH) listeners, demonstrating the importance of binaural and spatial cues for segregating multiple sound sources. There are important effects due to the type of stimuli used, testing parameters, and auditory task utilized. The review of research in hearing impaired populations notes auditory cues that are potentially available to users of CIs and HAs. In addition, there is discussion of limitations resulting from the ways that devices handle auditory cues, auditory deprivation, and other factors that are inherently problematic for these patients.

Keywords Cochlear implants · Cocktail party · Hearing loss · Noise · Speech understanding

R.Y. Litovsky (✉) · A. Kan
Waisman Center, University of Wisconsin–Madison,
1500 Highland Ave., Madison, WI 53705, USA
e-mail: Litovsky@waisman.wisc.edu

A. Kan
e-mail: ahkan@waisman.wisc.edu

M.J. Goupell
Department of Hearing and Speech Sciences,
University of Maryland, College Park, MD 20742, USA
e-mail: goupell@umd.edu

S.M. Misurelli
Department of Communication Sciences and Disorders, University of Wisconsin–Madison,
1500 Highland Ave., Madison, WI 53705, USA
e-mail: smisurelli@wisc.edu

10.1 Introduction

One of the most important tasks that humans face is communication in complex, noisy acoustic environments. As the many chapters in this book focus on how normal-hearing (NH) listeners deal with the “cocktail party problem,” here the focus is on particular populations of listeners who suffer from hearing loss and are fitted with cochlear implants (CIs) and/or hearing aids (HAs) in order to hear. The clinical trend is to provide patients with the ability to hear in both ears and has stemmed from decades of research with NH listeners, demonstrating the importance of binaural and spatial cues for segregating multiple sound sources. There are important effects due to the type of stimuli used, testing parameters, and auditory task utilized. Although much of the research was originally conducted with adults, recently different investigators have adapted testing methods appropriate for children. Those studies are thus able to gauge the impact of hearing loss and the potential benefits of early intervention on the development of the ability to segregate speech from noise. The review of research in hearing impaired populations notes auditory cues that are potentially available to users of CIs and HAs. In addition, there is discussion of limitations due to the ways that devices handle auditory cues, auditory deprivation, and other factors that are inherently problematic for these patients.

The overall goal of clinically motivated bilateral stimulation is to provide patients with the auditory information necessary for sound localization and for functioning in cocktail party or other complex auditory environments. When adults with hearing loss are concerned, there is emphasis on minimizing social isolation and maximizing the ability to orient in the environment without exerting undue effort. In the case of children, although these aforementioned goals are also relevant, there are concerns regarding success in learning environments that are notoriously noisy, and significant discussion regarding the need to preserve hearing in both ears because future improved stimulation approaches will be most ideally realized in people whose auditory system has been stimulated successfully in both ears.

In the forthcoming pages, results from studies in adults and children are described, focusing on their ability to hear target speech in the presence of “other” sounds. In the literature those “other” sounds have been referred to, sometimes interchangeably, as maskers, interferers, or competitors. Here the term *interferers* is used because the context of the experiments is that of assessing the interference that background sounds have on the ability of hearing impaired individuals to communicate. It has long been known that spatial separation between target speech and interferers can lead to improved speech understanding. A common metric in the literature that is used here is “spatial release from masking” (SRM), the measured advantage gained from the spatial separation of targets and interferers; it can be quantified as change in percent correct of speech understanding under conditions of spatial coincidence versus spatial separation, or as change in speech reception thresholds (SRTs) under those conditions. The SRM can be relatively small when

target and interferers have already been segregated by other cues, such as differing voice pitches. However, the SRM is particularly large when the interferer is similar to the target (e.g., two same-sex talkers with approximately the same voice pitch) and few other segregation cues are available. In such conditions, the listener must rely on spatial cues to segregate the target from the interferer. SRM is also particularly large for NH listeners when the interferer consists of multiple talkers as would occur in a realistic complex auditory environment (Bronkhorst 2000; Hawley et al. 2004; Jones and Litovsky 2011). For example, SRM can be as high as a 12-dB difference in SRTs under binaural conditions, especially when multiple interferers are present and the interferers and target speech are composed of talkers that can be confused with one another. SRM can be as low as a 1- to 2-dB difference in SRT under conditions when binaural hearing is not available. An important note regarding SRM and “real world” listening conditions is that the advantage of spatial separation is reduced in the presence of reverberation, whereby the binaural cues are smeared and thus the locations of the target and interferers is not easily distinguishable (Lavandier and Culling 2007; Lee and Shinn-Cunningham 2008).

In this chapter, results from studies in adults with CIs and with HAs are reviewed first, followed by results from children with CIs and with HAs.

10.2 Adults at the Cocktail Party

10.2.1 Factors that Limit Performance

Listening in cocktail parties is generally more of a challenge for listeners who suffer from hearing loss than for listeners with an intact auditory system. The limitations experienced by hearing impaired individuals generally can be subdivided into two main categories. One set of factors is the biological nature of the auditory system in a patient who has had sensorineural hearing loss, suffered from auditory deprivation, and has been stimulated with various combinations of acoustic and/or electric hearing. A second set of factors is the nature of the device(s) that provide hearing to the patient, whether electric or acoustic in nature. Although CIs and HAs aim to provide sound in ways that mimic natural acoustic hearing as much as possible, the reality is that these devices have limitations, and those limitations are likely to play a role in the patients’ ability to function in complex listening situations.

10.2.2 Physiological Factors that Limit Performance in Hearing Impaired Individuals

Hearing impairment can arise from a number of different factors, some of which are hereditary, some that are acquired (such as ototoxicity, noise exposure, etc.), and

others that have unknown causes. In some patients, these factors can result in hearing loss that is single sided or asymmetric across the ears. In other patients, the hearing loss can be more symmetric. It is important to consider the extent of hearing loss and change over time, as many patients undergo progressive, continued loss of hearing over months or years. For each patient, access to useful auditory information will be affected by access to sound during development and by the health of the auditory system. Finally, the extent to which binaural cues are available to the patient is one major factor that will likely determine their success with performing spatial hearing tasks. Research to date has attempted to identify the exact level within the auditory system at which the availability of binaural cues is most important, but ongoing work is needed to address this important issue.

For patients who use CIs, additional factors need to be considered. Owing to their diverse and complex hearing histories, CI users are generally a more variable group than NH listeners. Longer periods of auditory deprivation between the onset of deafness and implantation can lead to atrophy, and possibly poor ability of the auditory neurons to process information provided by auditory input. The condition of the auditory neural pathway can be affected by several other factors as well, including age, etiology of deafness, and experience with HAs or other amplification. Another factor that applies to both CI users and other hearing impaired populations is that poor innervation can result in decreased audibility and frequency selectivity, and can manifest as spectral dead regions, or “holes in hearing” (e.g., Moore and Alcántara 2001). These holes that can lead to difficulty understanding certain speech sounds, a need for increased stimulation levels, and discrepancies in frequency information between the ears (Shannon et al. 2002). Neural survival is both difficult to identify and even harder to control for in CI users.

The delicate nature of CI surgery can introduce further variability and complications that affect binaural processing. The insertion depth of the electrode into the cochlea is known to be variable (Gstoettner et al. 1999) and can be difficult to ascertain. Because frequencies are mapped along the length of the cochlea, with low frequencies near the apex; shallow insertion depths can truncate middle or low frequency information. This has been shown to reduce speech recognition (Başkent and Shannon 2004). In the case of bilateral cochlear implants (BICIs), differing insertion depths between the ears can lead to an interaural mismatch in perceived frequencies. Using CI simulations, this has been shown to negatively affect binaural benefit with speech recognition (Siciliano et al. 2010) and to degrade the reliability of binaural cues such as interaural time differences (ITDs) and interaural level differences (ILDs) (Kan et al. 2013, 2015). In addition, the distance between the electrode array and the modiolus (central pillar of the cochlear) is not uniform along the cochlea or identical between the ears. For electrodes that are located further from the modiolus, high levels of stimulation are sometimes needed to elicit auditory sensation. These high stimulation levels excite a wider range of neurons, causing a broader spread of excitation along the length of the cochlea; again, using simulations of CIs, this has been shown to reduce frequency selectivity and word recognition (Bingabr et al. 2008) and may be detrimental to binaural processing abilities.

10.2.3 *Devices*

10.2.3.1 **Cochlear Implants**

For patients with a severe to profound hearing loss, the multichannel CI is becoming the standard of care for providing access to sound. For children, CIs have been particularly successful at providing access to acquisition of spoken language and oral-auditory communication. The CI is designed to stimulate the auditory nerve and to bypass the damaged cochlear hair cell mechanism. A full review of the CI speech processor and the internal components that are surgically implanted into the patient is beyond the scope of this chapter (Loizou 1999; Zeng et al. 2011). However, the important points are summarized as follows. In a CI sound processor, a bank of bandpass filters separates the incoming signal into a small number of frequency bands (ranging from 12 to 22), from which the envelope of each band is extracted. These envelopes are used to modulate the amplitudes of electrical pulse trains that are presented to electrodes at frequency-specific cochlear loci. In general, uniform pulse rates across the channels have been used in speech processing strategies, although some recent advances have led to mixed-rate approaches (Hochmair et al. 2006; Churchill et al. 2014) whose success is yet to be determined. By its basic design, current speech-coding strategies in clinical CI processors eliminate the temporal fine structure in the signal and focus on transmitting information provided by the time-varying envelope at each bandpassed channel. The result is loss of spectral resolution and temporal fine-structure cues, both of which are important for binaural hearing. In addition, CIs typically work independently of each other and act as unsynchronized monaural signal processors. This independent processing can distort the transmission of binaural cues, which are beneficial for sound localization and speech understanding in noise.

Several factors could limit the amount of SRM achieved by BICI listeners. First, the envelope encoding employed by most CI speech coding strategies removes usable fine-structure information (Loizou 2006) which limits the ability of CIs to convey the potent low-frequency ITDs (Wightman and Kistler 1992; Macpherson and Middlebrooks 2002) that are likely to be important for SRM (Ihlefeld and Litovsky 2012). Second, even if the fine-structure cues were available, the fact that CIs do not have obligatory synchronization of the stimulation between the ears results in poor or improper encoding of ITD cues. Third, it is difficult to represent binaural cues with fidelity for complex stimuli such as those that occur in natural everyday situations. Speech stimuli are dynamic, with ongoing dynamic changes in spectral and temporal information, rendering the encoding of binaural cues for speech sounds extremely difficult. An additional issue is that cochlear spread with monopolar stimulation, which is nearly 5 mm (Nelson et al. 2008) is likely to have further detrimental effects on binaural encoding of complex stimuli. Recent work has shown that when multiple binaural channels receive carefully controlled ITDs, CI users are able to extract this information with little interference across electrodes. The effectiveness of this approach for improving SRM remains to be better

understood. The modulations in speech envelopes may also distort binaural cues (Goupell et al. 2013; Goupell 2015; Goupell and Litovsky 2015); amplitude modulations affect the loudness of stimuli, and the range of perceived loudness is not necessarily similar at all the electrodes across the electrode array within each ear, across the ears. Currently, CI processors and mapping consider only threshold and comfortable levels, and apply a generic compression function for all electrodes. As a modulated stimulus, such as speech, varies in instantaneous amplitude, distorted ILDs are produced. In summary, binaural cues presented through clinical processors are likely distorted by multiple mechanisms, which could in turn reduce perceived separation between targets and interferers in situations in which BICI listeners might otherwise benefit from SRM.

10.2.3.2 Hearing Aids

For patients with some usable residual hearing, the standard of care has been the prescription of HAs. The purpose of a HA is to partially compensate for the loss in sensitivity due to cochlear damage by the amplification of select frequencies. For listening in noisy situations, the amount of amplification that can be provided by HAs is typically limited by feedback issues, patient comfort, and audible dynamic range. Most modern HAs are digital, which means that the amplification stage is done through digital signal processing rather than analog electronic circuits.

Up to the 1990s, most HAs worked as linear amplifiers. However, loudness recruitment and the reduced dynamic range of patients with hearing impairment limited the usefulness of these HAs. Automatic gain control (AGC) systems, or compression, are now used in HAs to reduce the range of incoming sound levels before amplification to better match the dynamic range of the patient. At low input levels, the gain (ratio between the output and input levels) applied to the incoming signal is independent of input level. At high input levels, the gain decreases with increasing input level; that is, the input signal is compressed. For a more comprehensive overview of compression systems, see Kates and Arehart (2005). The use of compression in HA has generally provided positive results compared to linear amplification. However, different implementations of compression can have different consequences on performance, and there is no consensus on the best way to implement compression in HAs (Souza 2002).

10.3 Adults with Cochlear Implants

10.3.1 Availability of Spatial Cues

When listening to speech in quiet environments, adults with CIs can perform relatively well. However, listening to speech in a quiet environment is highly unrealistic, especially for listeners who spend much of their day communicating in

environments with multiple auditory sources. Anecdotal reports, as well as research findings, clearly demonstrate that listening in noisy environments can be challenging for CI users even if they perceive themselves to be functioning very well with their devices.

In the context of controlled research environments, research has been aimed at simulating aspects of realistic auditory environments. The difficulty experienced by listeners in extracting meaningful information from sources of interest results in reduced access to information in the target auditory source. Auditory interference is often referred to as being accounted for by effects that are attributed to either the energy in the interferers or the information that the interferers carry (Culling and Stone Chap. 3; Kidd and Colburn, Chap. 4). The former is thought to occur at the level of the peripheral auditory system when the target and interfering stimuli overlap in the spectral and temporal domains. The latter is thought to occur more centrally within the auditory system and is due to auditory and nonauditory mechanisms. The definitions and auditory mechanisms involved in these phenomena are more controversial within the field of psychoacoustics, but informational effects are often attributed to uncertainty of which stimulus to attend to and/or similarity between the target and interfering stimuli (Durlach et al. 2003; Watson 2005).

One way to improve the perceptual clarity of a target signal and to provide subjects with greater access to the content of the speech is to separate the target spatially from the interfering sources. Figure 10.1 shows three configurations,

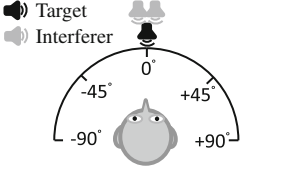
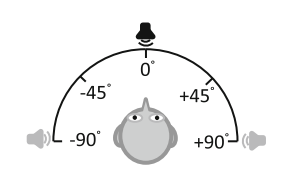
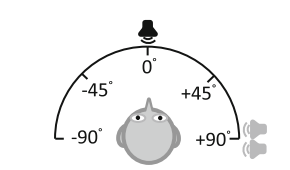
Condition	Cues Available	Loudspeaker Layout
Co-located	<ul style="list-style-type: none"> •Target-interferer vocal differences •Signal to noise ratio 	
Separated: Symmetrical	<ul style="list-style-type: none"> •Target-interferer vocal differences •Signal to noise ratio •Interaural timing differences •Interaural level differences 	
Separated: Asymmetrical	<ul style="list-style-type: none"> •Target-interferer vocal differences •Signal to noise ratio •Head shadow (better ear) •Interaural timing differences •Interaural level differences 	

Fig. 10.1 Stimulus configurations are shown that are commonly used for measuring SRM. The target is always at 0° in front, and the interfering sounds are either also in front or on the side

where the target is always at 0° in front, and the interfering sounds are either also in front or on the side. This figure is intended to provide a summary of the auditory cues that are typically available for source segregation in these listening situations. To quantify the magnitude of SRM, performance is measured (e.g., percent correct speech understanding) in conditions with the target and interferer either spatially separated or co-located. A positive SRM, for example, would be indicated by greater percent correct in the separated condition than the co-located condition. SRM can also be measured by comparing SRTs for the co-located versus separated conditions, and in that case positive SRM would be reflected by lower SRTs in the separated condition.

SRM relies on effects due to differences in the locations of the interferers relative to the location of the target, while holding constant unilateral or bilateral listening conditions. In contrast, other related effects that have been investigated result from comparing performance between unilateral and bilateral listening conditions. Figure 10.2 shows these three effects, widely known as the *head shadow*, *binaural*

Effects that contribute to SRM		
Head Shadow (better ear)		Attenuation of a sound that occurs as it passes through the head
Binaural Squelch		Benefit of adding the ear with the poorer signal-to-noise ratio
Binaural Summation		Boost in perceived loudness when both ears hear a signal

Fig. 10.2 The three most common effects due to bilateral hearing are shown here. **(Top)** Head Shadow. **(Middle)** Binaural Squelch. **(Bottom)** Binaural Summation

squelch, and *binaural summation*. The head shadow effect results from monaural improvement in signal-to-noise ratio (SNR) of the target speech. A measured improvement in speech understanding is due to the fact that the head casts an acoustic shadow on the interferer at the ear that is farther (on the opposite side of the head) from the interferer. In Fig. 10.2 (top), the left ear is shadowed, meaning that the intensity of the interferer is attenuated by the head before reaching the left ear. For this condition, the target speech in the left ear has a better SNR than that in the right ear. The second effect, known as squelch (Fig. 10.2, middle), refers to the advantage of adding an ear with a poorer SNR compared with conditions in which listening occurs only with an ear that has a good SNR. Binaural squelch is considered to be positive if performance is better in the binaural condition than in the monaural condition, despite having added an ear with a poor SNR. The third effect, binaural summation (Fig. 10.2, bottom), results from the auditory system receiving redundant information from both ears, and is effectively demonstrated when SRTs are lower when both ears are activated compared to when only one ear is activated. These effects will be considered in the following sections of the chapter in relation to the research conducted with hearing impaired individuals.

10.3.2 Binaural Capabilities of Adult BICI Users

The main goal of BICIs is to provide CI users with access to spatial cues. One way to evaluate success of BICIs is to determine whether patients who listen with two CIs gain benefit on tasks that measure SRM. Numerous studies have shown that head shadow and summation occur in BICI users. However, compared with NH listeners, there is a diminished amount of squelch, and unmasking due to binaural processing per se. Although this chapter focuses on SRM, it is worthwhile to pay attention to binaural capabilities in the subject population of interest, because diminished capacities in binaural sensitivity are likely to be related to reduction in unmasking of speech that occurs when binaural cues are readily available.

Much of the research on binaural capabilities of adults with BICIs has been performed with BICI listeners who have developed a typical auditory system because they were born with typical hearing and lost hearing after acquisition of language. One assumption that could be made is that these listeners have a fully intact and functioning central auditory system and that the problems of encoding the auditory stimulus are limited to the periphery. This is in contrast to prelingually deafened BICI listeners, in whom it is unclear if the correct development of the binaural system has occurred (Litovsky et al. 2010).

10.3.3 *Sound Localization*

A prerequisite for achieving SRM might be to perceive the multiple sound sources at different locations. This perception is produced by the ITDs and ILDs in the horizontal plane, and location-specific spectral cues in the vertical plane (Middlebrooks and Green 1990). When using clinical processors, adult BICI listeners can perceive different locations in the horizontal plane, and generally localize better when listening through two CIs versus one CI (Litovsky et al. 2012). In general, performance is better than chance when listening either in quiet (van Hoesel and Tyler 2003; Seeber and Fastl 2008) or in background noise (Kerber and Seeber 2012; Litovsky et al. 2012) though it seems that BICI listeners are relying more on ILDs than ITDs to localize sounds (Seeber and Fastl 2008; Aronoff et al. 2010). This is in contrast to NH listeners who weight ITDs more heavily than ILDs for sound localization (Macpherson and Middlebrooks 2002). Although BICI listeners do not appear to rely on ITDs for sound localization with their clinical processors, ITD sensitivity has been demonstrated in BICI listeners when stimulation is tightly controlled and presented using synchronized research processors (e.g., see Litovsky et al. 2012; Kan and Litovsky 2015 for a review). One notable important control is that ITD sensitivity is best when the electrodes that are activated in the two ears are perceptually matched by place of stimulation. The goal of using interaurally pitch-matched pairs of electrodes is to activate neurons with similar frequency sensitivity so that there may be a greater chance of mimicking the natural manner in which binaural processing occurs at the level of the brainstem of NH mammals. However, with monopolar stimulation, which produces substantial spread of excitation along the basilar membrane, BICI listeners appear to be able to tolerate as much as 3 mm of mismatch in stimulation between the right and left ears before showing significant decreases in binaural sensitivity (Poon et al. 2009; Kan et al. 2013, 2015; Goupell 2015). In contrast, ILDs seem even more robust than ITDs to interaural place-of-stimulation mismatch (Kan et al. 2013). Studies have shown that when interaurally pitch-matched pairs of electrodes are stimulated, ITD sensitivity varies with the rate of electrical stimulation. ITD sensitivity is typically best (discrimination thresholds are approximately 100–500 μ s) at low stimulation rates (<300 pulses per second [pps]) and tends to be lost at stimulation rates above 900 pps. However, ITD sensitivity is also observed when low modulation rates are imposed on high-rate carriers (van Hoesel et al. 2009; Noel and Eddington 2013). The ITD thresholds reported in many BICI users are considerably greater (i.e., worse) than the range of 20–200 μ s observed in NH listeners when tested with low-frequency stimulation or with high-rate carriers that are amplitude modulated (Bernstein and Trahiotis 2002); however, note that several BICI users can achieve ITD thresholds in this range, as low at about 40–50 μ s (Bernstein and Trahiotis 2002; Kan and Litovsky 2015; Laback et al. 2015).

10.3.4 Binaural Masking Level Differences

Another binaural prerequisite to achieve SRM in BICI listeners is binaural unmasking of a tone in noise, otherwise known as a binaural masking level difference (BMLD). This form of unmasking is similar to SRM, but experiments are performed with simpler signals. A BMLD is the degree of unmasking that is measured when detecting a tone-in-noise with dichotic stimuli (e.g., noise in phase, tone out of phase = N_0S_π) compared to a diotic stimuli (e.g., noise in phase, tone in phase = N_0S_0). Several studies have shown that BICI listeners can achieve BMLDs up to about 10 dB using single-electrode direct stimulation when the amplitude modulations are compressed as in a speech processing strategy (Goupell and Litovsky 2015) and can be quite large if amplitudes do not have a speech processing strategy amplitude compression (Long et al. 2006). Lu et al. (2011) took the paradigm a step further by also measuring BMLDs for multiple electrode stimulation, which is more akin to multielectrode stimulation needed to represent speech signals. Figure 10.3 shows a BMLD of approximately 9 dB for single-electrode stimulation; in that study there was a reduction of the effect size to approximately 2 dB when multielectrode stimulation was used (not shown in this figure). These findings suggested that spread of excitation along the cochlea in monopolar stimulation, which is known to produce masking and interference, also results in degraded binaural unmasking. Using auditory evoked potentials to examine this issue, the authors found that conditions with larger channel interaction correlated with psychophysical reduction in the BMLD. The studies on BMLDs offer insights into binaural mechanisms in BICI users. The fact that BMLDs can be elicited suggests that the availability of carefully controlled binaural cues could play an important role in the extent to which patients demonstrate SRM. In cases in which

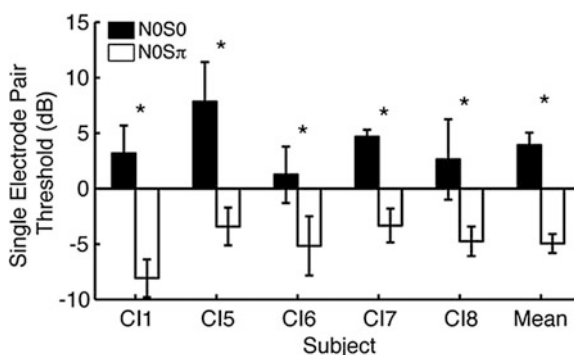


Fig. 10.3 Data are plotted for conditions with diotic stimuli (dark fills) or dichotic stimuli (white fills). In the former, noise in the right and left ears in phase for both signal and noise, hence referred to as N_0S_0 . In the latter, noise in the right and left ears is in phase, while the tone in the right and left ears is out of phase, hence referred to as N_0S_π . (Replotted with permission from Lu et al. 2011.)

BMLDs in BICI users are poorer than those observed in NH listeners, insights can be gained into the limitations that BICI users face. These limitations include factors such as neural degeneration and poor precision of binaural stimuli in the context of monopolar stimulation.

10.3.5 SRM in BICI Users

From a historical perspective, there has been increasing interest in examining SRM in CI users. The first study, by van Hoesel et al. (1993), tested only co-located conditions in one bilateral CI listener using unsynchronized clinical processors. After 12 months of use, bilateral presentation produced about 10–15% of a binaural summation advantage (see Fig. 10.2). Although it was a case study, the data suggested that the approach was valuable and numerous studies thereafter addressed similar questions. For instance, Buss et al. (2008) tested 26 postlingually deafened BICI users with a short duration of deafness before implantation, and with nearly all patients undergoing simultaneous bilateral implantation. Target and interferer were either co-located or spatially separated by 90°. Testing was repeated at 1, 3, 6, and 12 months after activation of the CIs. Significant head shadow and binaural summation benefits were observed at 6 and 12 months after CI activation. Binaural squelch became more apparent only 1 year after activation (Eapen et al. 2009). In this study, performance was also measured as change in percent correct. Litovsky et al. (2009) tested a very similar population of 37 postlingually deafened BICI users, and measured change in performance as change in SRT. Within 6 months after activation of the CIs, the largest and most robust bilateral benefit was due to the head shadow effect, averaging approximately 6 dB improvement in SRT. Benefits due to binaural summation and squelch were found in a small group of patients, where effect sizes were more modest, 1–2 dB change in SRT. These and numerous other studies have generally shown that the largest benefit of having BICIs is accounted for by the head shadow, or being able to attend to an ear at which the target has a good signal-to-noise ratio.

In addition to the studies using clinical processors, a few studies have investigated SRM using more controlled stimulation approaches that attempted to provide ITDs directly to the CI processors. These approaches were motivated by the idea that maximizing control over the ITDs presented to the BICIs would improve speech perception through increased SRM, mainly through the squelch mechanism. van Hoesel et al. (2008) imposed spatial separation by presenting the noise from the front and the target speech at one ear. One of the main purposes of this experiment was to compare three different types of speech coding strategies, two that were envelope based and one that explicitly encoded temporal fine-structure information. None of the speech coding strategies produced a significant binaural unmasking of speech. There are many possible reasons why squelch was not observed in this study. For one, the fine-structure ITDs were slightly different across electrode pairs, following that of the individual channel, and each channel had a high rate carrier, which could

have blurred the perceived spatial locations of targets and interferers. Recent research from Churchill et al. (2014) showed a larger benefit of sound localization and ITD discrimination when coherent fine-structure ITDs are presented on multiple electrodes redundantly, particularly when the low-frequency channels have lower rate stimulation. It is possible that with this newer type of fine-structure encoding, squelch would be achieved. Another problem with the van Hoesel et al. (2008) study is that they had only four subjects, who may have had long durations of deafness and were sequentially implanted, as compared to the 26 subjects with short duration of deafness and simultaneous implantations studied by Buss et al. (2008).

Another approach to tightening control over the binaural stimuli presented to BICI users was that of Loizou et al. (2009). In that study, a single binaural digital signal processor was used to present stimuli to both the right and left CIs. Stimuli were convolved through head-related transfer functions (HRTFs), and hence ITDs and ILDs were preserved at the level of the processor receiving the signals. This study was designed so as to replicate a previous study in NH listeners by Hawley et al. (2004) and compared results from BICI users directly with those of the NH listeners. One goal of this study was to examine whether BICI users experience informational masking similar to that seen in NH listeners; evidence of informational masking in these studies would be a larger amount of SRM with speech

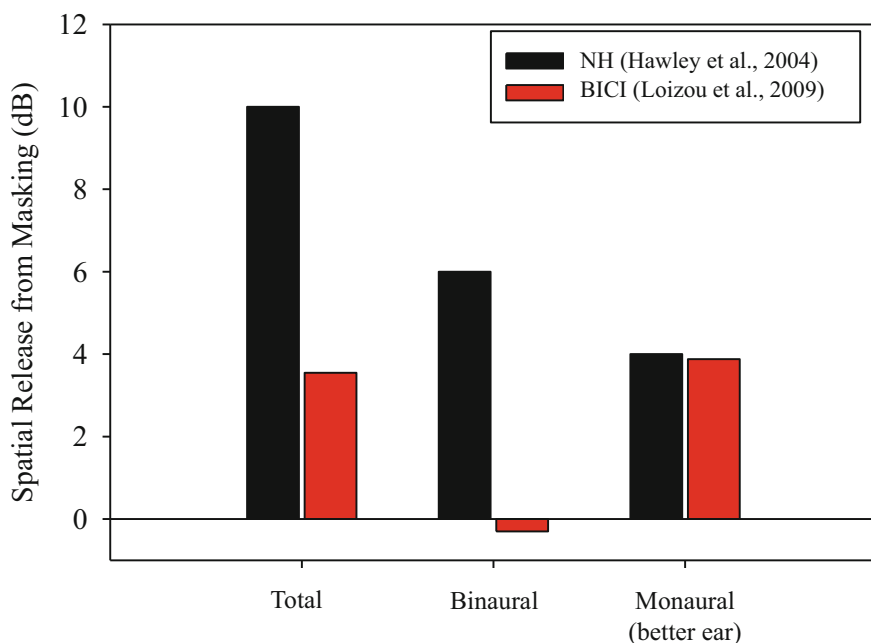


Fig. 10.4 SRM data are shown for subjects with normal hearing (NH) and bilateral cochlear implants (BICIs). Values are shown for the total amount of SRM; amount accounted for by binaural processing; and the remainder, which is attributed to monaural processing. (Replotted with permission from Loizou et al. 2009.)

versus noise interferers. Figure 10.4 shows summary data from the Loizou et al. (2009) study only when speech interferers were used. Unlike NH listeners, in BICI users the SRM was small, around 2–4 dB, regardless of condition. In only a few cases the noise and speech interferers produced different amounts of spatial advantage. In the NH data SRM was much larger, especially with speech interferers compared to noise interferers. This finding has been interpreted as evidence that NH listeners experience informational masking with speech maskers due to target/interferer similarity. Thus, when the target and interferers are co-located the ability to understand the target is especially poor, and spatial separation renders spatial cues particularly important. Finally, NH listeners showed effects due to both binaural and monaural spatial unmasking. In contrast, and as can be seen in Fig. 10.4, BICI users generally show monaural effects with weak or absent binaural unmasking effects. In summary, even with attempted presentation of binaural cues at the input to the processors, squelch was not observed. This is probably due to the fact that the ITDs and ILDs present in the signals were not deliberately sent to particular pairs of electrodes. Thus, ITDs and ILDs were likely to have been disrupted by the CI processors at the level of stimulation in the cochlear arrays.

Binaural unmasking of speech in BICI listeners was tested more recently by Bernstein et al. (2016) in an attempt to resolve many of the issues regarding the role of binaural cues in SRM. Conditions included a target monaurally compared to an interferer presented monaurally (the same ear) or diotically to both ears, thus attempting to produce the largest spatial separation with an effectively infinite ILD. Note that no head shadow occurs in such a paradigm; thus the effects observed are most likely related to the binaural squelch effect. A range of interferers were tested: noise and one or two talkers. The target and interferers came from the same corpus, known to produce high informational masking, and performance was measured at a range of levels that varied the target-to-masker (interferer) ratios (TMRs). Results from eight postlingually deafened BICI listeners showed 5 dB of squelch in the one-talker interferer condition, which is much larger than that in the previous studies, possibly because of the particular methodology that sought to maximize binaural unmasking of speech. Another interesting result from this study is that the amount of unmasking was larger for lower TMRs; many previous studies found only the 50% SRT and therefore may have missed larger squelch effects. Relatively smaller amounts of unmasking were observed for noise or two talkers; in NH listeners there is typically larger SRM for more interfering talkers. Interestingly, one BICI listener who had an early onset of deafness consistently showed interference (i.e., negative binaural unmasking) from adding the second ear. It may not be a coincidence that this listener was in the group that may not have a properly developed binaural system.

By way of highlighting some of the main effects, Fig. 10.5A shows summary data from this study, as well as from Loizou et al. (2009) and two other studies that measured the three effects (squelch, summation, and head shadow) as improvement in SRT. As can be seen from Fig. 10.5A, head shadow is the major contributor to the SRM improvement. Summation and squelch are about equal in their contribution to SRM.

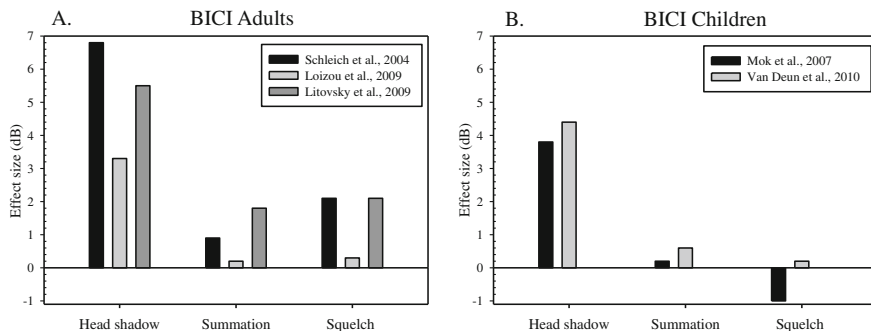


Fig. 10.5 Results are summarized from studies in which spatial separation was studied in adults (A) and children (B) with cochlear implants. Measurements of three effects are compared: head shadow, summation, and squelch

10.3.6 *Simulating Aspects of CI Processing for Testing in Normal-Hearing Listeners*

Research to date provides a window into the successes and shortcomings of the currently available CI technology, and underscores difficulties that CI users experience, especially when listening to speech in noisy environments. These studies do carry some inherent limitations, however, including high intersubject variability and lack of control over many aspects of signal processing. These limitations are somewhat overcome by the use of multichannel vocoders, which enable the manipulation and testing of CI signal processing algorithms in NH listeners. The application of this approach to NH listeners is particularly useful because of the presumed unimpaired peripheral auditory systems in these listeners. NH listeners also have less variance across the population, making them ideal subjects for studies that attempt to exclude individual variability as a factor. In addition, factors such as the effect of processor synchronization, spread of excitation along the cochlea with monopolar stimulation, and matched place of stimulation across the ears can be simulated and individually investigated under ideal conditions.

Multichannel vocoders mimic the same preprocessing steps as in CI processing; that is the acoustic input is divided into a number of frequency bands, and the envelope in each band is extracted for modulation of a carrier. The carriers can be sine tones or bandpass noise, depending on what aspects of CI electrical stimulation the vocoder is attempting to mimic. These modulated carriers are then recombined and presented to a NH listener via headphones or loudspeakers. Vocoders have been used to investigate many aspects of cochlear implant performance including the effect of mismatched frequency input between the ears on binaural processing (Siciliano et al. 2010; van Besouw et al. 2013). To date, few research studies have used vocoders as a tool to investigate the SRM performance gap between NH and BICI users. Garadat et al. (2009) used vocoders to investigate whether BICI users'

poor performance in binaural hearing tasks was an effect of the lack of fine structure ITD cues. They tested this hypothesis using two conditions. In the first, they processed stimuli through a vocoder and then convolved the vocoded stimuli through HRTFs. This order of processing replaced the fine structure with sine tones, but allowed for fine-structure ITD to remain in the carrier signal. In the second condition, the order of processing was reversed, eliminating fine structure ITDs and more accurately simulating CI processing. Listeners averaged 8–10 dB SRM in conditions with more degraded spectral cues. In addition, performance was not significantly different between orders of processing, which the authors interpreted to indicate that the removal of temporal fine structure cues is not a key factor in binaural unmasking. Freyman et al. (2008) also found that vocoder processing, with fine structure removed, allowed for SRM when stimuli were delivered via loudspeakers, as long as adequate informational masking was present. The authors interpreted the findings to indicate that, should there be ample coordination between the CIs, the signal processing would preserve adequate spatial information for SRM. Thus, poor performance from BICI users is likely due to other factors. One such factor was investigated by Garadat et al. (2010), who simulated spectral holes that extended to 6 or 10 mm along the cochlea, at either the basal, middle, or apical regions of the cochlea. The results of this study attest to the relative frailty of the binaural system and the importance of various portions of the frequency spectrum for binaural processing.

Bernstein et al. (2016) directly compared BICI listener performance to NH listeners presented with eight-channel noise vocoded stimuli. They found very similar trends between the two groups, demonstrating that such a comparison is useful in understanding expectations for the BICI listeners. However, there was much more squelch in the NH listeners than in the BICI listeners, which might be explained by the myriad of factors already outlined, such as deficiencies at the electrode–neural interface, differences in insertion depth, and so forth.

It should be acknowledged that there are limitations to the use of vocoders. To deliver spatialized stimuli to NH subjects, stimuli must be processed using HRTFs, and it is not well understood if the effects of vocoders processing on HRTF cues are similar to those that occur with CIs in actual three-dimensional space. In addition, acoustic stimulation with processed signals is fundamentally different from the direct electrical stimulation that occurs with CIs, and the systematic differences between acoustic and electrical stimulation are not well understood.

10.4 Adults with HAs

Similar to CIs, the effectiveness of HAs in helping speech understanding in noisy situations is affected by an interplay of individual differences, acoustical factors, and technological issues. However, the elements surrounding these factors are largely different between CIs and HAs. Whereas research in CIs has focused on the transmission of appropriate acoustic cues by electric hearing, HA research and

development has been concerned with audibility. In general, satisfaction toward modern HAs has increased (Bertoli et al. 2009; Kaplan-Neeman et al. 2012). Speech understanding has improved, in part due to improved SNR before presentation.

10.4.1 Unilateral Versus Bilateral Fitting

One of the important clinical questions when considering HAs is whether one or two HAs should be prescribed. For the case of a unilateral hearing impairment, the choice of aiding the poorer ear may appear to be the obvious choice. However, in the case of a bilateral hearing impairment, the choice is not as obvious, especially if the impairment is asymmetric in the two ears and the patient can afford only one HA. If one has to choose a single ear, which of the two ears should be amplified is an area of debate. In some listening conditions aiding the poorer ear generally leads to improved hearing ability and is preferred by a majority of patients (Swan and Gatehouse 1987; Swan et al. 1987), while in other listening situations, aiding the better ear was found to be more beneficial (Henkin et al. 2007). Even when two HAs are prescribed, the literature does not show a consistent benefit. Henkin et al. (2007) reported bilateral interference in speech understanding in about two-thirds of their subjects. In contrast, Kobler and Rosenhall (2002) reported significant improvement when two HAs were used compared to an aid in the better ear. These conflicted results typically arise from differences in test setups and the amount of hearing impairment, which highlights the fact that the benefit of having bilateral HA over unilateral is situation dependent, and typically more useful in more demanding environments (Noble and Gatehouse 2006). Despite conflicting results, there appears to be a general trend that the advantage of bilateral fitting for speech intelligibility increases with increasing degree of hearing loss (Mencher and Davis 2006; Dillon 2012), though this may not be necessarily be predictive of HA user preferences (Cox et al. 2011).

10.4.2 Bilateral Benefits

In theory, prescription of bilateral HAs should provide improvements for both sound localization and speech-in-noise performance. Both these abilities are important in a cocktail party-like situation. As previously described, sound localization ability helps with distinguishing the location talkers from one another in a private conversation, and from among the background chatter of the cocktail party. Being able to understand speech in the noisy environment is important for being able to carry on a conversation.

The effect of hearing impairment on sound localization ability is related to the audibility of cues important for sound localization. For high-frequency hearing loss,

common among those with hearing impairment, there is a decrease in vertical localization ability and an increase in front–back confusions, because high-frequency cues are important in helping discriminate these locations. In addition, for a sound located on the horizontal plane the ability to use ILDs for localization is also reduced. The ability to use low-frequency ITDs appears to be only mildly affected, and deteriorates only when low-frequency hearing loss exceeds about 50 dB. Bilateral aids increase the audibility of sounds in both ears, such that ITD and ILD cues can be heard and compared by the auditory system. Thus, the benefit of bilateral aids will likely be more significant for those with moderate to severe losses, compared to those with a mild hearing loss. Front–back confusions and vertical localization ability typically remain the same regardless of amount of hearing loss. This may be because the high-frequency cues necessary for the restoration of these abilities are not sufficiently audible, or because the microphone location of the HA is not in a location that maximally captures the spectral cues important for front–back and vertical location discrimination. Byrne and Nobel (1998) and Dillon (2012) provide excellent reviews on the effects of hearing impairment on sound localization ability and how HAs can provide benefit.

For speech-in-noise understanding, there are a number of binaural benefits arising from being able to hear with both ears, particularly in a cocktail party–like situation in which the target and maskers are spatially separated. These include binaural squelch and summation. Together with the head shadow effect, these benefits provide a spatial release from masking. In essence, the provision of bilateral HAs is to increase sensitivity to sounds at both ears in the hope that the same binaural advantages that are available in the NH listeners can occur, though not necessarily to the same degree. Hence, those with greater hearing loss are more likely to show a benefit from bilateral HAs (Dillon 2012). Bilateral HAs are also more likely to provide a benefit when the listening situation is more complex, such as in cocktail parties, though benefits of SRM have typically been much less than those observed in NH listeners (Festen and Plomp 1986; Marrone et al. 2008). However, the use of bilateral HA has been shown to lead to improved social and emotional benefits, along with a reduction in listening effort (Noble and Gatehouse 2006; Noble 2010).

10.4.3 Technological Advances

A known limitation faced by HA users (whether unilateral or bilateral) is whether the target signal can be heard above the noise. Hence, much of the progress in HA technology has focused on improving the SNR of the target signal before presentation to the HA user. This has included the introduction of directional microphone technology, noise reduction signal processing algorithms, and wireless links between HAs to provide bilateral processing capabilities.

Directional microphones attempt to improve the SNR of a target signal by exploiting the fact that target and noise signals are typically spatially separated. The

directivity of a microphone can be changed such that maximum amplification is provided to a target signal located in a known direction (usually in front of the listener), and sounds from other directions are given much less amplification. This is usually achieved by forming a directional beam by combining the signals from two or more microphones. Comprehensive overviews of the different ways an array of microphone signals can be combined to shape the directivity of a microphone can be found in Chung (2004) and Dillon (2012). In laboratory test settings, SNR benefits from directional microphones range from 2.5 to 14 dB, depending on test configuration (Bentler et al. 2004; Dittberner and Bentler 2007; Hornsby and Ricketts 2007). However, reported benefit in real life situations is much less than expected (Cord et al. 2002) because in real-world listening situations, environmental acoustics and the dynamics of the sound scene are more complex than those of laboratory conditions, which leads to poorer performance than expected.

In contrast to directional microphones, noise reduction algorithms attempt to exploit the time–frequency separation between target and noise signals. The aim of noise reduction algorithms is to identify which components of the incoming signal are from the target and provide greater amplification to these components, compared to the noise components. This is not an easy task and the different HA manufacturers each have their own algorithms for noise reduction. An overview of noise reduction algorithms can be found in Chung (2004) and Dillon (2012). In general, noise reduction algorithms may not necessarily improve speech intelligibility, but the HA user will find the background noise less troublesome (Dillon 2012).

A more recent development has been the development of wireless communication between the left and right HAs of a bilaterally fit HA user (Edwards 2007). Currently, the wireless link is typically used for linked volume control and a few other basic functions. However, connectivity between HAs opens up the opportunity for linked bilateral processing, such as sharing of computation cycles to increase computational speed and power (Edwards 2007), implementation of more advance directional microphone and noise reduction algorithms by combining the microphone inputs from both ears (e.g., Luts et al. 2010; Kokkinakis and Loizou 2010), and the linking of compression systems in both ears to improve speech intelligibility (Wiggins and Seeber 2013).

10.5 Introduction to Pediatric Studies

For the majority of children who currently receive CIs, the first time they are exposed to sound will be at their CI activation. The CI device provides children who are deaf with hearing via electrical stimulation to the auditory nerve, ultimately giving them the opportunity to develop spoken language and use speech as their primary mode of communication.

As discussed above in Sect. 10.1, SRM can provide an assessment of the extent to which listeners who are fitted with two CIs are able to use spatial cues for

segregation of target speech from background interferers. SRM in CI users is clearly observed when listeners can rely on monaural head shadow cues. However, if listeners must rely on binaural cues, SRM is small. One interesting question is whether auditory history has an impact on SRM in ways that would demarcate between children and adults. Studies reviewed in Sect. 10.5.2 were aimed at understanding if children who are generally congenitally deaf and implanted at a young age are able to utilize spatial cues for source segregation in ways that adults cannot.

10.5.1 Studies in Children with BICIs

Children who are diagnosed with a profound bilateral sensorineural hearing impairment, and do not benefit from a HA, can receive either unilateral CIs or BICIs. However, the standard of care in most countries is to provide children with bilateral CIs. The goal from a clinical perspective is to provide children with improved speech understanding in noise, access to spatial hearing, and to stimulate both the right and left auditory pathways. While a unilateral CI can clearly lead to the development of relatively good speech understanding in quiet, bilateral implantation has clear benefits for the spatial unmasking of speech that has been described above. For children, these benefits are also framed in the context of auditory models and theories on plasticity, which argue for better results with stimulation of the neural pathways early in life. Despite the many advances and large body of evidence to support the benefit of having two CIs versus one, there continues to be a gap in performance when listening to speech in noise between children with BICIs and their NH peers.

In a number of studies conducted by Litovsky and colleagues, children with BICIs or with a CI and a HA (bimodal hearing) were studied using similar approaches to evaluating SRM as those used with adults. The first study showed that children with bimodal hearing demonstrated small SRM or “negative SRM” (performance was worse with interferers on the side compared with in the front) (Litovsky et al. 2006). By comparison, children with BICIs showed SRM that was small but consistent. One possible reason for weak SRM is that the children had been bilaterally implanted at an older age (early to mid-childhood), and their auditory system may not have adapted well to spatial cues through the CIs. In a subsequent study (Misurelli and Litovsky 2012), SRM was investigated in children whose activation with BICIs occurred at younger ages. Results from conditions with interferers positioned to one side of the head (asymmetrical configuration) showed SRM to be small but consistent (see Fig. 10.6, left). In that study, a novel condition was added with “symmetrical” interferers, intended to reduce the available benefit of the head shadow as well as create a more realistic listening environment where children have to rely mostly on binaural cues to segregate the auditory sources. Results from the symmetrical condition (Fig. 10.6, right) showed that, on average, children with BICIs demonstrated little to no SRM, and in some

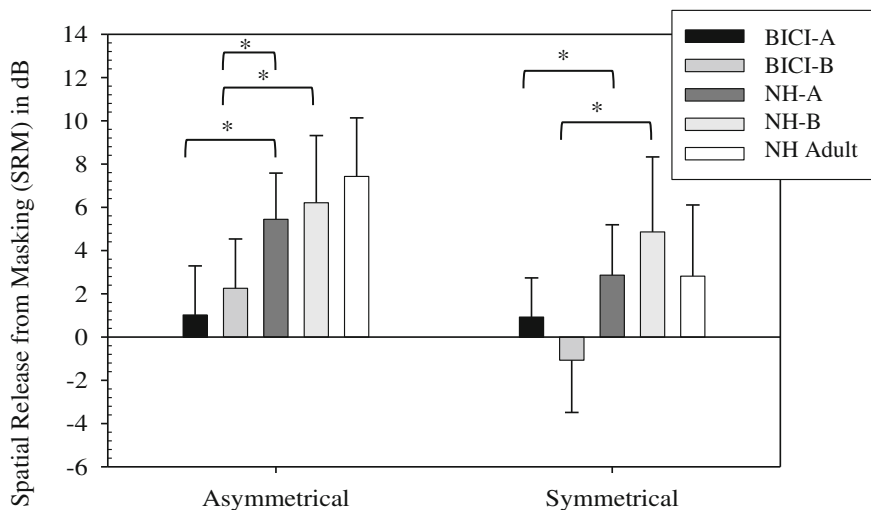


Fig. 10.6 Mean (\pm SD) SRM data are shown for children and adults with normal hearing and with bilateral cochlear implants. In this study, interferers were placed either asymmetrally around the head (at 90° to one side; see left side of graph) or symmetrically (at 90° to the right and left; see right side of graph). Statistically significant differences between groups are shown in brackets with an asterisk on top. (Replotted with permission from Misurelli and Litovsky 2012.)

cases SRM was negative or “anti-SRM” was seen, similar to the finding in the children with bimodal hearing (Litovsky et al. 2006). These findings suggest that children with BICIs do benefit from having two CIs, and that the benefit arises largely due to the head shadow effect. Figure 10.5B shows data from two studies (Mok et al. 2007; Van Deun et al. 2010) that estimated the three effects summarized for adults in Fig. 10.5A—head shadow, summation, and squelch—also computed from differences in SRTs. The children’s data, similarly to those from adults, show largest effects from head shadow, and similar, small effects for summation and squelch.

It has been shown that spatial cues are more useful for NH children in source segregation tasks in which there are large amounts of perceived informational masking. When the target and interfering stimuli comprise speech or speech-like stimuli (such as time-reversed speech), children with NH demonstrate robust SRM (Johnstone and Litovsky 2006). The authors interpreted the findings to suggest that similarity between the target and interferers produced informational masking; thus the use of spatial cues for segregation was enhanced. In a recent study (Misurelli and Litovsky 2015), the effect of informational masking was investigated by comparing SRM with target and interferers that were either same sex (both target and interferers male) or different sex (male target, female interferers). Children with BICIs did not demonstrate a significant benefit from spatial cues in conditions that were designed to create more informational masking. The reasons for small SRM found in children with BICIs are poorly understood. Future work could provide

insights into this issue by determining whether access to binaural cues through synchronized CI devices will be useful. Other factors may also be important, including the fact that the BICI pediatric population is not exposed to binaural cues during development, in the years when neural circuits that mediate SRM are undergoing maturation.

In an attempt to understand whether maturation of auditory abilities lead to better SRM, longitudinal SRM data were collected with BICI pediatric users (Litovsky and Misurelli 2016). Children participated in the same task that was used in the aforementioned studies; however, testing was repeated over a 2–4-year period at annual intervals. The goal of this testing was to determine whether SRM undergoes changes, possibly increasing in magnitude, as each child gained bilateral experience and acquired additional context for listening in noisy situations. The children were instructed to identify a target spondee (i.e., a two-syllable word with equal stress on both syllables) in the presence of competing speech. During testing, SRTs were measured in a front condition with the target and the interferers co-located at 0° azimuth, in an asymmetrical condition with the target at 0° and the two-talker interferers spatially separated 90° to the side of the first implant, and in a symmetrical condition with the target at 0° and one interferer at 90° to the right and one at 90° to the left. SRM was calculated by subtracting the SRTs either in the asymmetrical condition or in the symmetrical from the SRTs in the front condition. For the majority of children SRM did not improve as children gained bilateral experience, suggesting that the limitations of the CI devices are the primary factors that contribute to the gap in performance on spatial unmasking, rather than the amount of bilateral experience with the CI devices.

10.5.2 Sequential Versus Simultaneous BICIs

Although a gap in performance remains between children with NH and with BICIs, there is evidence (Litovsky et al. 2006; Peters et al. 2007) to suggest that children who have BICIs perform better than children with a unilateral CI on speech-in-noise tasks. As a result, many children currently undergo CI surgery around 1 year of age with the intention of bilateral implantation. Successful surgical implantation of BICI devices to very young children has led to the question of whether there is physiological and functional benefit to receiving two CIs within the same surgical procedure (simultaneously) versus receiving two CIs with a delay in between (sequentially). It is evident that considerable extended periods of unilateral auditory deprivation can negatively affect auditory development and outcomes of CI users.

For children receiving their BICIs sequentially, it appears that it is best to implant the second ear as early as possible, and that performance on speech-in-noise tasks with the second CI improves as the users gain experience listening with that device (Peters et al. 2007). Further, in children with sequentially implanted BICIs, more SRM is demonstrated when the interferers are directed toward the side of the

second CI than when they are directed toward the first CI (Litovsky et al. 2006; Van Deun et al. 2010; Chadha et al. 2011). This suggests that the dominant CI is the side that was activated first (and is also the side with which the child has the most listening experience). However, even when interferers are directed toward the side of the first CI, thereby establishing a more difficult listening environment than with the interferers directed toward the second CI, some BICI users as young as 5 years of age do show SRM (Misurelli and Litovsky 2012). In children who receive their CIs sequentially, though, the amount of SRM demonstrated with interferers directed toward the first CI (dominant) is variable.

Neurophysiological changes in the auditory brainstem and cortex can help to indicate neuronal survival and reorganization after the CI is activated. Nonbehavioral responses to change in sound can be made using electrically evoked responses from the auditory nerve (electrically evoked response action potential [ECAP]) and brainstem (electrically evoked auditory brainstem [EABR]). These studies have shown greater success, and specifically better speech perception abilities, for children who receive BICIs with only a small delay between the first and second CI (Gordon and Papsin 2009). A recent study showed that if a child receives the second CI within 1.5 years of the first CI, symmetrical neurological development of the auditory brainstem and cortex is more likely to occur. Functionally, this can be associated with better speech perception, especially in noise (Gordon et al. 2013).

A recent psychoacoustic study measured performance on a speech-in-noise task comparing children with sequential and simultaneous BICIs, and revealed that the simultaneously implanted group had significantly better performance when listening to speech in noise (Chadha et al. 2011). More research is needed regarding the functional and cortical benefits of simultaneous implantation versus sequential implantation with a minimal delay in time between the first and the second implant. There is currently not enough evidence to suggest a specific age limit, or time limit between CIs, in which a child would no longer benefit from bilateral implantation (Litovsky and Gordon 2016).

10.5.3 Children with HAs

As with the decision to implant children at an early age, the decision surrounding the prescription of HAs for children is motivated primarily by developmental concerns. For children born with, or who acquire, hearing impairment, early diagnosis is seen as essential and the prescription of HAs is considered of high importance for the child to have a typical educational and social development (Dillon 2012).

Although the importance of having a typical development for a child with hearing impairment cannot be denied for social, educational, and economic reasons, the data supporting whether unilateral or bilateral HAs should be prescribed is mixed, and recommendations are influenced primarily by the particular outcome measure that is considered important. Dillon (2012) provides a detailed summary of

the literature concerning the impact of a unilateral loss on different outcome measures. He argues that although the effect of unilateral loss on a child's language and educational development is mixed, an effect is still prevalent in every study, and it is likely that having a hearing impairment, especially with an increasing loss, will make it more difficult for a child to easily acquire language and social skill. However, whether the aiding of the ear with hearing impairment is important is left open for discussion. Although amplification of the poorer ear may increase the possibility of bilateral benefits, it may also introduce binaural interference that may have a negative impact on development. It may seem, however, that early fitting of an HA may provide sound localization benefits for children with a unilateral hearing impairment (Johnstone et al. 2010).

For children with a bilateral loss, binaural amplification may not provide much additional benefit in terms of language development and understanding over unilateral amplification (Grimes et al. 1981), and children who use HAs do not seem to demonstrate SRM (Ching et al. 2011). However, the provision of HAs may be of importance for promoting near-typical development of the binaural auditory system. Neural development occurs most rapidly during the first few years of life, and having access to binaural auditory input may be important for promoting near-normal-like development of the binaural auditory system (Gordon et al. 2011). Having access to binaural acoustic stimulation early in life may have an impact on the future success of cochlear implantation (Litovsky et al. 2010, 2012).

Finally, a growing population of children has been diagnosed with auditory neuropathy spectrum disorder (ANSD). A possible deficit with ANSD is in the temporal domain: perception of low- to mid-frequency sounds. A common treatment option for severe impairment in ANSD is unilateral cochlear implantation, and because the degree of impairment is unrelated to degree of hearing loss by audiometric thresholds, this population may have significant acoustic sensitivity in the unimplanted contralateral ear. Runge et al. (2011) recently tested a group of children with ANSD using the test protocols that Litovsky and colleagues have used, to investigate the effects of acute amplification. Results showed that children with ANSD who are experienced CI users tend to benefit from contralateral amplification, particularly if their performance with the CI is only moderate. This study opened up many questions, including the long-term impact of contralateral HA use in real-world situations. However, the one take-home message regarding the "cocktail party effects" in children with complex hearing issues is that electric + acoustic hearing may provide benefits that need to be closely tracked and evaluated.

10.5.4 Variability and Effects of Executive Function

Even when all demographic factors are accounted for (e.g., age at implantation, age at deafness, CI device type, etiology of hearing loss), children with CIs still demonstrate variability in outcomes (Geers et al. 2003; Pisoni and Cleary 2003). The large amount of variability is a hallmark of CI research and presents a

challenge to clinicians trying to counsel expected outcomes to parents who are considering CIs for their children.

Executive function is fundamental for development of language processing and functioning in complex environments. Previous work has shown that deficits in executive function (i.e., working and short-term memory, attention, processing speed) are associated with poor performance in noisy environments. Although executive function is not a direct measure of the ability to hear speech in complex auditory environments, it is clear that individuals must be able to extract the target signal, retain attention to the target, and manipulate incoming linguistic information when in complex environments. A gap in performance on measures of short-term and working memory exists for children with CIs and with NH, such that children with CIs generally perform lower on these measures than NH age-matched children (Cleary et al. 2001; Pisoni and Cleary 2003). Deficiency in specific aspects of executive function could negatively impact a child's ability to understand speech successfully in noisy environments, and therefore performance on these measures may help to predict some of the variability that is demonstrated in BICI groups. More work is necessary to define which specific aspects of executive function are related to performance on spatial hearing tasks.

10.5.5 Future Directions and Clinical Applications

It is well demonstrated that children with two CIs perform better on tasks of spatial hearing than children with one CI. Some recent evidence has shown that children with BICIs who are implanted simultaneously, or with a very minimal interimplant delay, may have a greater advantage in noisy environments than those with a long delay between the first and second CI. The gap that exists between BICI and NH listeners in the ability to segregate the target talker in cocktail party environments most likely reflects the limitations of the current clinically available CI devices. Improvements in the speech processing strategies and device synchronization must be made for children with BICIs to function more similarly to their NH peers in multisource acoustic environments. For BICI users, the lack of synchronization between the two CI devices greatly reduces binaural cues, or even prohibits the user from accessing any binaural cues, that have shown to benefit NH listeners in noisy environments. The development and implementation of enhanced speech processing strategies that take advantage of fine-structure spectral information in the acoustic signal would likely provide children with access to cues to aid in speech understanding in both quiet and noise.

The information presented in this section suggests that additional research and clinical advances are needed to narrow the gap in performance when listening to speech in noise between children with BICIs and with NH. Until the clinically available CI device allows the user to receive more fine-tuned signals, more like those a NH listener receives, the gap in performance between the two groups

remain. In the interim, it is important to increase the signal-to-noise ratio in noisy classroom settings where children are expected to learn.

10.6 Conclusions

Mammals have evolved with two ears positioned symmetrically about the head, and the auditory cues arising from the brain's analysis of interaural differences in the signals play an important role in sound localization and speech understanding in noise. Contributions of auditory experience and nonsensory processes are highly important, and less well understood. This chapter reviewed studies that focus on populations of listeners who suffer from hearing loss, and who are fitted with HAs and/or CIs in order to hear. Although fitting of dual devices is the clinical standard of care, the ability of the patient to benefit from stimulation in the two ears varies greatly. The common metric in the literature that was described here is spatial release from masking (SRM). A simplified version of SRM was described in the original study on the cocktail party by Cherry (1953), but the paradigm used in that study involved spatial separation of sources across the two ears, rather than in space. Later studies simulated realistic listening environments and captured more detailed knowledge about the many factors that augment or diminish unmasking of speech in noisy background. What is common to HA and BICI users are the limitations that are thought to contribute to the gap in performance when comparing performance in NH and hearing impaired populations. Several factors seem to play a role in the limitations, thus providing ample evidence to suggest that the devices do not analyze and present spatial cues to the auditory system with fidelity, and that patient histories related to auditory deprivation and diminished neural health are inherently problematic for these patients. Several studies provide evidence to suggest that, should there be ample coordination between the CIs in the right and left ears, the signal processing would preserve adequate spatial information for SRM.

In children the outcomes are not much different than in adults, possibly because the limitations of the devices that are used today create the same lack of access to important spatial unmasking cues. It is possible, however, that training the brain to utilize cues that are subtle or somewhat inconsistent will yield positive results, and this is fertile ground for future work.

Compliance with Ethics Requirements

Ruth Litovsky received travel support for a conference from Cochlear Ltd. and from MedEl.

Matthew Goupell had no conflicts of interest.

Alan Kan owns stocks in Cochlear Ltd. Sara Misurelli had no conflicts of interest.

References

- Aronoff, J. M., Yoon, Y. S., Freed, D. J., Vermiglio, A. J., et al. (2010). The use of interaural time and level difference cues by bilateral cochlear implant users. *The Journal of the Acoustical Society of America*, 127(3), EL87–EL92.
- Başkent, D., & Shannon, R. V. (2004). Frequency-place compression and expansion in cochlear implant listeners. *The Journal of the Acoustical Society of America*, 116(5), 3130–3140.
- Bentler, R. A., Egge, J. L., Tubbs, J. L., Dittberner, A. B., & Flamme, G. A. (2004). Quantification of directional benefit across different polar response patterns. *Journal of the American Academy of Audiology*, 15(9), 649–659.
- Bernstein, J., Goupell, M. J., Schuchman, G. I., Rivera, A. L., & Brungart, D. S. (2016). Having two ears facilitates the perceptual separation of concurrent talkers for bilateral and single-sided deaf cochlear implantees. *Ear and Hearing*, 37(3), 289–302.
- Bernstein, L. R., & Trahiotis, C. (2002). Enhancing sensitivity to interaural delays at high frequencies by using “transposed stimuli.” *The Journal of the Acoustical Society of America*, 112(3 Pt. 1), 1026–1036.
- Bertoli, S., Staehelin, K., Zemp, E., Schindler, C., et al. (2009). Survey on hearing aid use and satisfaction in Switzerland and their determinants. *International Journal of Audiology*, 48(4), 183–195.
- Bingabr, M., Espinoza-Varas, B., & Loizou, P. C. (2008). Simulating the effect of spread of excitation in cochlear implants. *Hearing Research*, 241(1–2), 73–79.
- 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(1), 117–128.
- Buss, E., Pillsbury, H. C., Buchman, C. A., Pillsbury, C. H., et al. (2008). Multicenter U.S. bilateral MED-EL cochlear implantation study: speech perception over the first year of use. *Ear and Hearing*, 29(1), 20–32.
- Byrne, D., & Noble, W. (1998). Optimizing sound localization with hearing AIDS. *Trends in Amplification*, 3(2), 51–73.
- Chadha, N. K., Papsin, B. C., Jiwani, S., & Gordon, K. A. (2011). Speech detection in noise and spatial unmasking in children with simultaneous versus sequential bilateral cochlear implants. *Otology & Neurotology*, 32(7), 1057–1064.
- Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *The Journal of the Acoustical Society of America*, 25, 975–979.
- Ching, T. Y., van Wanrooy, E., Dillon, H., & Carter, L. (2011). Spatial release from masking in normal-hearing children and children who use hearing aids. *The Journal of the Acoustical Society of America*, 129(1), 368–375.
- Chung, K. (2004). Challenges and recent developments in hearing aids. Part I. Speech understanding in noise, microphone technologies and noise reduction algorithms. *Trends in Amplification*, 8(3), 83–124.
- Churchill, T. H., Kan, A., Goupell, M. J., & Litovsky, R. Y. (2014). Spatial hearing benefits demonstrated with presentation of acoustic temporal fine structure cues in bilateral cochlear implant listeners. *The Journal of the Acoustical Society of America*, 136(3), 1246–1256.
- Cleary, M., Pisoni, D. B., & Geers, A. E. (2001). Some measures of verbal and spatial working memory in eight- and nine-year-old hearing-impaired children with cochlear implants. *Ear and Hearing*, 22(5), 395–411.
- Cord, M. T., Surr, R. K., Walden, B. E., & Olson, L. (2002). Performance of directional microphone hearing aids in everyday life. *Journal of the American Academy of Audiology*, 13(6), 295–307.
- Cox, R. M., Schwartz, K. S., Noe, C. M., & Alexander, G. C. (2011). Preference for one or two hearing AIDS among adult patients. *Ear and Hearing*, 32(2), 181–197.
- Dillon, H. (2012). *Hearing aids*. New York: Thieme.
- Dittberner, A. B., & Bentler, R. A. (2007). Predictive measures of directional benefit. Part 1: Estimating the directivity index on a manikin. *Ear and Hearing*, 28(1), 26–45.

- Durlach, N. I., Mason, C. R., Shinn-Cunningham, B. G., Arbogast, T. L., et al. (2003). Informational masking: Counteracting the effects of stimulus uncertainty by decreasing target-masker similarity. *The Journal of the Acoustical Society of America*, *114*(1), 368–379.
- Eapen, R. J., Buss, E., Adunka, M. C., Pillsbury, H. C., 3rd, & Buchman, C. A. (2009). Hearing-in-noise benefits after bilateral simultaneous cochlear implantation continue to improve 4 years after implantation. *Otology & Neurotology*, *30*(2), 153–159.
- Edwards, B. (2007). The future of hearing aid technology. *Trends in Amplification*, *11*(1), 31–46.
- Festen, J. M., & Plomp, R. (1986). Speech-reception threshold in noise with one and two hearing aids. *The Journal of the Acoustical Society of America*, *79*(2), 465–471.
- Freyman, R. L., Balakrishnan, U., & Helfer, K. S. (2008). Spatial release from masking with noise-vocoded speech. *The Journal of the Acoustical Society of America*, *124*(3), 1627–1637.
- Garadat, S. N., Litovsky, R. Y., Yu, G., & Zeng, F.-G. (2009). Role of binaural hearing in speech intelligibility and spatial release from masking using vocoded speech. *The Journal of the Acoustical Society of America*, *126*(5), 2522–2535.
- Garadat, S. N., Litovsky, R. Y., Yu, G., & Zeng, F.-G. (2010). Effects of simulated spectral holes on speech intelligibility and spatial release from masking under binaural and monaural listening. *The Journal of the Acoustical Society of America*, *127*(2), 977–989.
- Geers, A., Brenner, C., & Davidson, L. (2003). Factors associated with development of speech perception skills in children implanted by age five. *Ear and Hearing*, *24*(1 Suppl.), 24S–35S.
- Gordon, K. A., Jiwani, S., & Papsin, B. C. (2013). Benefits and detriments of unilateral cochlear implant use on bilateral auditory development in children who are deaf. *Frontiers in Psychology*, *4*, 719.
- Gordon, K. A., & Papsin, B. C. (2009). Benefits of short interimplant delays in children receiving bilateral cochlear implants. *Otology & Neurotology*, *30*(3), 319–331.
- Gordon, K. A., Wong, D. D. E., Valero, J., Jewell, S. F., et al. (2011). Use it or lose it? Lessons learned from the developing brains of children who are deaf and use cochlear implants to hear. *Brain Topography*, *24*(3–4), 204–219.
- Goupell, M. J. (2015). Interaural envelope correlation change discrimination in bilateral cochlear implantees: Effects of mismatch, centering, and onset of deafness. *The Journal of the Acoustical Society of America*, *137*(3), 1282–1297.
- Goupell, M. J., Kan, A., & Litovsky, R. Y. (2013). Mapping procedures can produce non-centered auditory images in bilateral cochlear implantees. *The Journal of the Acoustical Society of America*, *133*(2), EL101–EL107.
- Goupell, M. J., & Litovsky, R. Y. (2015). Sensitivity to interaural envelope correlation changes in bilateral cochlear-implant users. *The Journal of the Acoustical Society of America*, *137*(1), 335–349.
- Grimes, A. M., Mueller, H. G., & Malley, J. D. (1981). Examination of binaural amplification in children. *Ear and Hearing*, *2*(5), 208–210.
- Gstoettner, W., Franz, P., Hamzavi, J., Plenk, H., Jr., et al. (1999). Intracochlear position of cochlear implant electrodes. *Acta Oto-Laryngologica*, *119*(2), 229–233.
- 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. *The Journal of the Acoustical Society of America*, *115*(2), 833–843.
- Henkin, Y., Waldman, A., & Kishon-Rabin, L. (2007). The benefits of bilateral versus unilateral amplification for the elderly: Are two always better than one? *Journal of Basic and Clinical Physiology and Pharmacology*, *18*(3), 201–216.
- Hochmair, I., Nopp, P., Jolly, C., Schmidt, M., et al. (2006). MED-EL cochlear implants: State of the art and a glimpse into the future. *Trends in Amplification*, *10*(4), 201–219.
- Hornsby, B. W., & Ricketts, T. A. (2007). Effects of noise source configuration on directional benefit using symmetric and asymmetric directional hearing aid fittings. *Ear and Hearing*, *28*(2), 177–186.
- Ihlefeld, A., & Litovsky, R. Y. (2012). Interaural level differences do not suffice for restoring spatial release from masking in simulated cochlear implant listening. *PLoS ONE*, *7*(9), e45296.

- Johnstone, P. M., & Litovsky, R. Y. (2006). Effect of masker type and age on speech intelligibility and spatial release from masking in children and adults. *The Journal of the Acoustical Society of America*, *120*(4), 2177–2189.
- Johnstone, P. M., Nabelek, A. K., & Robertson, V. S. (2010). Sound localization acuity in children with unilateral hearing loss who wear a hearing aid in the impaired ear. *Journal of the American Academy of Audiology*, *21*(8), 522–534.
- Jones, G. L., & Litovsky, R. Y. (2011). A cocktail party model of spatial release from masking by both noise and speech interferers. *The Journal of the Acoustical Society of America*, *130*(3), 1463–1474.
- Kan, A., & Litovsky, R. Y. (2015). Binaural hearing with electrical stimulation. *Hearing Research*, *322*, 127–137.
- Kan, A., Litovsky, R. Y., & Goupell, M. J. (2015). Effects of interaural pitch matching and auditory image centering on binaural sensitivity in cochlear implant users. *Ear and Hearing*, *36*(3), e62–e68.
- Kan, A., Stoelb, C., Litovsky, R. Y., & Goupell, M. J. (2013). Effect of mismatched place-of-stimulation on binaural fusion and lateralization in bilateral cochlear-implant users. *The Journal of the Acoustical Society of America*, *134*(4), 2923–2936.
- Kaplan-Neeman, R., Muchnik, C., Hildesheimer, M., & Henkin, Y. (2012). Hearing aid satisfaction and use in the advanced digital era. *Laryngoscope*, *122*(9), 2029–2036.
- Kates, J. M., & Arehart, K. H. (2005). *A model of speech intelligibility and quality in hearing aids*. In IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, NY, October 16–19, 2005.
- Kerber, S., & Seeber, B. U. (2012). Sound localization in noise by normal-hearing listeners and cochlear implant users. *Ear and Hearing*, *33*(4), 445–457.
- Kobler, S., & Rosenhall, U. (2002). Horizontal localization and speech intelligibility with bilateral and unilateral hearing aid amplification. *International Journal of Audiology*, *41*(7), 395–400.
- Kokkinakis, K., & Loizou, P. C. (2010). Multi-microphone adaptive noise reduction strategies for coordinated stimulation in bilateral cochlear implant devices. *The Journal of the Acoustical Society of America*, *127*(5), 3136–3144.
- Laback, B., Egger, K., & Majdak, P. (2015). Perception and coding of interaural time differences with bilateral cochlear implants. *Hearing Research*, *322*, 138–150.
- Lavandier, M., & Culling, J. F. (2007). Speech segregation in rooms: Effects of reverberation on both target and interferer. *The Journal of the Acoustical Society of America*, *122*(3), 1713.
- Lee, A. K., & Shinn-Cunningham, B. G. (2008). Effects of reverberant spatial cues on attention-dependent object formation. *Journal of the Association for Research in Otolaryngology*, *9*(1), 150–160.
- Litovsky, R. Y., & Gordon, K. (2016). Bilateral cochlear implants in children: Effects of auditory experience and deprivation on auditory perception. *Hearing Research*. doi:[10.1016/j.heares.2016.01.003](https://doi.org/10.1016/j.heares.2016.01.003).
- Litovsky, R. Y., Goupell, M. J., Godar, S., Grieco-Calub, T., et al. (2012). Studies on bilateral cochlear implants at the University of Wisconsin's Binaural Hearing and Speech Laboratory. *Journal of the American Academy of Audiology*, *23*(6), 476–494.
- Litovsky, R. Y., Johnstone, P. M., & Godar, S. P. (2006). Benefits of bilateral cochlear implants and/or hearing aids in children. *International Journal of Audiology*, *45*(Suppl. 1), S78–891.
- Litovsky, R. Y., Jones, G. L., Agrawal, S., & van Hoesel, R. (2010). Effect of age at onset of deafness on binaural sensitivity in electric hearing in humans. *The Journal of the Acoustical Society of America*, *127*(1), 400–414.
- Litovsky, R. Y., & Misurelli, S. M. (2016). Does bilateral experience lead to improved spatial unmasking of speech in children who use bilateral cochlear implants? *Otology & Neurotology*, *37*(2), e35–e42.
- Litovsky, R. Y., Parkinson, A., & Arcaroli, J. (2009). Spatial hearing and speech intelligibility in bilateral cochlear implant users. *Ear and Hearing*, *30*(4), 419.
- Loizou, P. C. (1999). Introduction to cochlear implants. *IEEE Engineering in Medicine and Biology Magazine*, *18*(1), 32–42.

- Loizou, P. C. (2006). *Speech processing in vocoder-centric cochlear implants* (Vol. 64). Basel, Switzerland: Karger.
- Loizou, P. C., Hu, Y., Litovsky, R., Yu, G., et al. (2009). Speech recognition by bilateral cochlear implant users in a cocktail-party setting. *The Journal of the Acoustical Society of America*, *125*(1), 372–383.
- Long, C. J., Carlyon, R. P., Litovsky, R. Y., & Downs, D. H. (2006). Binaural unmasking with bilateral cochlear implants. *Journal of the Association for Research in Otolaryngology*, *7*(4), 352–360.
- Lu, T., Litovsky, R., & Zeng, F. G. (2011). Binaural unmasking with multiple adjacent masking electrodes in bilateral cochlear implant users. *The Journal of the Acoustical Society of America*, *129*(6), 3934–3945.
- Luts, H., Eneman, K., Wouters, J., Schulte, M., et al. (2010). Multicenter evaluation of signal enhancement algorithms for hearing aids. *The Journal of the Acoustical Society of America*, *127*(3), 1491–1505.
- Macpherson, E. A., & Middlebrooks, J. C. (2002). Listener weighting of cues for lateral angle: The duplex theory of sound localization revisited. *The Journal of the Acoustical Society of America*, *111*(5), 2219–2236.
- Marrone, N., Mason, C. R., & Kidd, G., Jr. (2008). Evaluating the benefit of hearing aids in solving the cocktail party problem. *Trends in Amplification*, *12*(4), 300–315.
- Mencher, G. T., & Davis, A. (2006). Bilateral or unilateral amplification: Is there a difference? A brief tutorial. *International Journal of Audiology*, *45*(Suppl. 1), S3–11.
- Middlebrooks, J. C., & Green, D. M. (1990). Directional dependence of interaural envelope delays. *The Journal of the Acoustical Society of America*, *87*(5), 2149–2162.
- Misurelli, S. M., & Litovsky, R. Y. (2012). Spatial release from masking in children with normal hearing and with bilateral cochlear implants: Effect of interferer asymmetry. *The Journal of the Acoustical Society of America*, *132*(1), 380–391.
- Misurelli, S. M., & Litovsky, R. Y. (2015). Spatial release from masking in children with bilateral cochlear implants and with normal hearing: Effect of target-interferer similarity. *The Journal of the Acoustical Society of America*, *138*(1), 319–331.
- Mok, M., Galvin, K. L., Dowell, R. C., & McKay, C. M. (2007). Spatial unmasking and binaural advantage for children with normal hearing, a cochlear implant and a hearing aid, and bilateral implants. *Audiology and Neuro-Otology*, *12*(5), 295–306.
- Moore, B. C., & Alcántara, J. I. (2001). The use of psychophysical tuning curves to explore dead regions in the cochlea. *Ear and Hearing*, *22*(4), 268–278.
- Nelson, D. A., Donaldson, G. S., & Kreft, H. (2008). Forward-masked spatial tuning curves in cochlear implant users. *The Journal of the Acoustical Society of America*, *123*(3), 1522–1543.
- Noble, W. (2010). Assessing binaural hearing: results using the speech, spatial and qualities of hearing scale. *Journal of the American Academy of Audiology*, *21*(9), 568–574.
- Noble, W., & Gatehouse, S. (2006). Effects of bilateral versus unilateral hearing aid fitting on abilities measured by the Speech, Spatial, and Qualities of Hearing Scale (SSQ). *International Journal of Audiology*, *45*(3), 172–181.
- Noel, V. A., & Eddington, D. K. (2013). Sensitivity of bilateral cochlear implant users to fine-structure and envelope interaural time differences. *The Journal of the Acoustical Society of America*, *133*(4), 2314–2328.
- Peters, B. R., Litovsky, R., Parkinson, A., & Lake, J. (2007). Importance of age and postimplantation experience on speech perception measures in children with sequential bilateral cochlear implants. *Otology & Neurotology*, *28*(5), 649–657.
- Pisoni, D. B., & Cleary, M. (2003). Measures of working memory span and verbal rehearsal speed in deaf children after cochlear implantation. *Ear and Hearing*, *24*(1 Suppl.), 106S–120S.
- Poon, B. B., Eddington, D. K., Noel, V., & Colburn, H. S. (2009). Sensitivity to interaural time difference with bilateral cochlear implants: Development over time and effect of interaural electrode spacing. *The Journal of the Acoustical Society of America*, *126*(2), 806–815.

- Runge, C. L., Jensen, J., Friedland, D. R., Litovsky, R. Y., & Tarima, S. (2011). Aiding and occluding the contralateral ear in implanted children with auditory neuropathy spectrum disorder. *Journal of the American Academy of Audiology*, 22(9), 567–577.
- Seeber, B. U., & Fastl, H. (2008). Localization cues with bilateral cochlear implants. *The Journal of the Acoustical Society of America*, 123(2), 1030–1042.
- Shannon, R. V., Galvin, J. J., III, & Baskent, D. (2002). Holes in hearing. *Journal of the Association for Research in Otolaryngology*, 3(2), 185–199.
- Siciliano, C. M., Faulkner, A., Rosen, S., & Mair, K. (2010). Resistance to learning binaurally mismatched frequency-to-place maps: Implications for bilateral stimulation with cochlear implants a. *The Journal of the Acoustical Society of America*, 127(3), 1645–1660.
- Souza, P. E. (2002). Effects of compression on speech acoustics, intelligibility, and sound quality. *Trends in Amplification*, 6(4), 131–165.
- Swan, I. R., Browning, G. G., & Gatehouse, S. (1987). Optimum side for fitting a monaural hearing aid. 1. Patients' preference. *British Journal of Audiology*, 21(1), 59–65.
- Swan, I., & Gatehouse, S. (1987). Optimum side for fitting a monaural hearing aid 2. Measured benefit. *British Journal of Audiology*, 21(1), 67–71.
- van Besouw, R. M., Forrester, L., Crowe, N. D., & Rowan, D. (2013). Simulating the effect of interaural mismatch in the insertion depth of bilateral cochlear implants on speech perception. *The Journal of the Acoustical Society of America*, 134(2), 1348–1357.
- Van Deun, L., van Wieringen, A., & Wouters, J. (2010). Spatial speech perception benefits in young children with normal hearing and cochlear implants. *Ear and Hearing*, 31(5), 702–713.
- van Hoesel, R., Bohm, M., Pesch, J., Vandali, A., et al. (2008). Binaural speech unmasking and localization in noise with bilateral cochlear implants using envelope and fine-timing based strategies. *The Journal of the Acoustical Society of America*, 123(4), 2249–2263.
- van Hoesel, R. J., Jones, G. L., & Litovsky, R. Y. (2009). Interaural time-delay sensitivity in bilateral cochlear implant users: Effects of pulse rate, modulation rate, and place of stimulation. *Journal of the Association for Research in Otolaryngology*, 10(4), 557–567.
- van Hoesel, R., Tong, Y., Hollow, R., & Clark, G. M. (1993). Psychophysical and speech perception studies: A case report on a binaural cochlear implant subject. *The Journal of the Acoustical Society of America*, 94(6), 3178–3189.
- van Hoesel, R. J., & Tyler, R. S. (2003). Speech perception, localization, and lateralization with bilateral cochlear implants. *The Journal of the Acoustical Society of America*, 113(3), 1617–1630.
- Watson, C. S. (2005). Some comments on informational masking. *Acta Acustica united with Acustica*, 91(3), 502–512.
- Wiggins, I. M., & Seeber, B. U. (2013). Linking dynamic-range compression across the ears can improve speech intelligibility in spatially separated noise. *The Journal of the Acoustical Society of America*, 133(2), 1004–1016.
- Wightman, F. L., & Kistler, D. J. (1992). The dominant role of low-frequency interaural time differences in sound localization. *The Journal of the Acoustical Society of America*, 91(3), 1648–1661.
- Zeng, F.-G., Popper, A., & Fay, R. R. (2011). *Auditory prostheses: New horizons*. New York: Springer Science & Business Media.