Springer Handbook of Auditory Research

Lori L. Holt · Jonathan E. Peelle
Allison B. Coffin · Arthur N. Popper
Richard R. Fay *Editors*

Speech Perception

With 33 Illustrations

ASA PRESS

Springer

Springer Handbook of Auditory Research

Volume 74

Series Editors

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

Editorial Board

Karen Avraham, Ph.D., Tel Aviv University, Israel
Andrew Bass, Ph.D., Cornell University
Lisa Cunningham, Ph.D., National Institutes of Health
Bernd Fritzsch, 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

ASA Press, which represents a collaboration between the Acoustical Society of America and Springer Nature, is dedicated to encouraging the publication of important new books as well as the distribution of classic titles in acoustics. These titles, published under a dual ASA Press/Springer imprint, are intended to reflect the full range of research in acoustics. ASA Press titles can include all types of books that Springer publishes, and may appear in any appropriate Springer book series.

Editorial Board

Mark F. Hamilton (Chair), University of Texas at Austin
James Cottingham, Coe College
Timothy F. Duda, Woods Hole Oceanographic Institution
Robin Glosemeyer Petrone, Threshold Acoustics
William M. Hartmann (Ex Officio), Michigan State University
Darlene R. Ketten, Boston University
James F. Lynch (Ex Officio), Woods Hole Oceanographic Institution
Philip L. Marston, Washington State University
Arthur N. Popper (Ex Officio), University of Maryland
Christine H. Shadle, Haskins Laboratories
G. Christopher Stecker, Boys Town National Research Hospital
Stephen C. Thompson, Pennsylvania State University
Ning Xiang, Rensselaer Polytechnic Institute

Lori L. Holt • Jonathan E. Peelle
Allison B. Coffin • Arthur N. Popper
Richard R. Fay
Editors

Speech Perception

ASA PRESS

Springer

Editors

Lori L. Holt
Department of Psychology
Carnegie Mellon University
Pittsburgh, PA, USA

Allison B. Coffin
Integrative Physiology and Neuroscience
Washington State University
Vancouver, WA, USA

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

Jonathan E. Peelle
Department of Otolaryngology
Washington University in St. Louis
Saint Louis, MO, USA

Arthur N. Popper
Department of Biology
University of Maryland
Silver Spring, MD, USA

ISSN 0947-2657 ISSN 2197-1897 (electronic)
Springer Handbook of Auditory Research
ISBN 978-3-030-81541-7 ISBN 978-3-030-81542-4 (eBook)
https://doi.org/10.1007/978-3-030-81542-4

© Springer Nature Switzerland AG 2022
This work is subject to copyright. All rights are solely and exclusively licensed 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, expressed 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.

This Springer imprint is published by the registered company Springer Nature Switzerland 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 on 10–11 May 1929 with a charter membership of about 450. Today, ASA has a worldwide membership of about 7000.

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 Standards 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

Springer Handbook of Auditory Research

The following preface is the one that we published in Volume I 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 75 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 co-editors 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 Sierra, 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, Nathaniel, Evan, and Stella Fay, and Sebastian Sierra.

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, post-doctoral 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.

Richard R. Fay, Chicago, IL, USA
Arthur N. Popper, College Park, MD, USA

SHAR logo by © Mark B. Weinberg, Potomac, Maryland, used with permission

Volume Preface

Speech is crucial in guiding human behavior. As a conspecific communication signal, speech is perhaps the most ubiquitous class of acoustic signals encountered by the human auditory system. Moreover, the acoustic complexity of speech and the extent to which it draws upon sensory encoding, prediction, hierarchical representation, attention, learning, and cognitive processing make it an ideal testbed for understanding auditory perceptual challenges, very generally.

In order to explore speech, this volume considers the neuroscience of speech perception in the broadest sense. Accordingly, the book is organized such that interested readers can dip into individual chapters of interest, or read the handbook cover to cover. Although it would be impossible to review the auditory cognitive neuroscience of speech perception in its entirety in a single volume, the chapters included here survey a broad range of theoretical perspectives, methodological approaches, and listening contexts that highlight current successes, challenges, and controversies in the field.

In Chap. 2, Bharath Chandrasekaran, Rachel Tessmer, and G. Nike Gnanateja provide an overview of the subcortical processing of speech sounds. This is followed by Chap. 3, in which Yulia Oganian, Neal P. Fox, and Edward F. Chang review contributions of human intracranial recordings to our understanding of speech perception, focusing on the superior temporal gyrus. Continuing on the theme of neural function in Chap. 4, Sarah Tune and Jonas Obleser explore the role of neural oscillations, the rhythmic or repetitive patterns of neural activity in the central nervous system that generally arise from feedback connections among neurons that result in synchronization of firing patterns, in speech perception.

In Chap. 5, Laura Gwilliams and Matthew H. Davis introduce an information-based approach to speech communication, grounded in statistical properties of speech content and the linguistic information conveyed by speech. In Chap. 6, Stephen C. Van Hedger and Ingrid S. Johnsrude explore how listeners understand speech in adverse listening conditions as well as provide a systematic review of behavioral and neurobiological evidence demonstrating that even minor challenges to listening demand interactions across perceptual, cognitive, and linguistic processes.

In Chap. 7, Shruti Ullas, Milene Bonte, Elia Formisano, and Jean Vroomen review evidence that the mappings from acoustics to phonetic categories representing the speech sounds of a native language are flexible, rather than fixed. Following from this, in Chap. 8 Judit Gervain reviews the development of speech perception, showing that infants begin learning about the patterns of speech in their native language even before birth and, by their first birthday, exhibit substantial experience-dependent reorganization of auditory processing of speech that accommodates the sound patterns of the native language(s). Finally, in Chap. 9, Chad S. Rogers and Jonathan E. Peelle extend the considerations of age and discuss interactions between audition and cognition in hearing loss and aging.

The Spring Handbook of Auditory Research series last focused on speech perception in 2004 with volume 18 that was entitled *Speech Perception in the Auditory System* (Eds. Steven Greenberg, Arthur N. Popper, and Richard R. Fay). Like the present volume, this previous review examined how the brain proceeds from sound to meaning in speech communication and provided a snapshot of the field at the time. From a historical perspective, it will be informative for readers to undertake a casual examination of the two volumes, separated by nearly 20 years. What becomes clear is that whereas many of the core theoretical questions remain unchanged, methodological advances have radically shaped how the field engages with these questions. Overall, the current volume demonstrates that evolving techniques now provide unprecedented access to neural data from human listeners, and theoretical perspectives of speech perception are making more and more contact with auditory neuroscience to draw upon whole-brain explanations and constructs of attention, learning, and cognitive processing that were less common two decades ago. These opportunities challenge researchers to ask questions that continue to further our understanding of speech perception in new and useful ways.

Lori L. Holt, Pittsburgh, PA, USA
Jonathan E. Peelle, St. Louis, MO, USA
Allison B. Coffin, Vancouver, WA, USA
Arthur N. Popper, College Park, MD, USA
Richard R. Fay, Chicago, IL, USA

Contents

Contributors

Milene Bonte Department of Cognitive Neuroscience, Faculty of Psychology and Neuroscience, Maastricht University, Maastricht, The Netherlands

Maastricht Brain Imaging Center, Maastricht University, Maastricht, The Netherlands

Bharath Chandrasekaran Department of Communication Science and Disorders, The University of Pittsburgh, Pittsburgh, PA, USA

Edward F. Chang Department of Neurological Surgery, University of California, San Francisco, San Francisco, CA, USA

Matthew H. Davis MRC Cognition and Brain Sciences Unit, Cambridge University, Cambridge, UK

Elia Formisano Department of Cognitive Neuroscience, Faculty of Psychology and Neuroscience, Maastricht University, Maastricht, The Netherlands

Maastricht Brain Imaging Center, Maastricht University, Maastricht, The Netherlands

Neal P. Fox Department of Neurological Surgery, University of California, San Francisco, San Francisco, CA, USA

Judit Gervain DPSS, Università degli Studi di Padova, Padova, Italy

INCC, CNRS and Université de Paris, Paris, France

G. Nike Gnanateja Department of Communication Science and Disorders, The University of Pittsburgh, Pittsburgh, PA, USA

Laura Gwilliams Department of Neurosurgery, University of California, San Francisco, San Francisco, CA, USA

Lori L. Holt Department of Psychology, Carnegie Mellon University, Pittsburgh, PA, USA

Ingrid S. Johnsrude Department of Psychology & Brain and Mind Institute, University of Western Ontario, London, ON, Canada

National Centre for Audiology & School of Communication Sciences and Disorders, University of Western Ontario, London, ON, Canada

Jonas Obleser Department of Psychology I, University of Lübeck, Lübeck, Germany

Yulia Oganian Department of Neurological Surgery, University of California, San Francisco, San Francisco, CA, USA

Jonathan E. Peelle Department of Otolaryngology, Washington University in St. Louis, St. Louis, MO, USA

Chad S. Rogers Department of Psychology, Union College, Schenectady, NY, USA

Rachel Tessmer Department of Speech, Language, and Hearing Sciences, The University of Texas at Austin, Austin, TX, USA

Sarah Tune Department of Psychology I, University of Lübeck, Lübeck, Germany

Shruti Ullas Department of Cognitive Neuroscience, Faculty of Psychology and Neuroscience, Maastricht University, Maastricht, The Netherlands

Maastricht Brain Imaging Center, Maastricht University, Maastricht, The Netherlands

Stephen C. Van Hedger Department of Psychology, Huron University College, London, ON, Canada

Department of Psychology & Brain and Mind Institute, University of Western Ontario, London, ON, Canada

Jean Vroomen Department of Cognitive Neuropsychology, Tilburg University, Tilburg, The Netherlands

Chapter 1
The Auditory Cognitive Neuroscience of Speech Perception in Context

Lori L. Holt and Jonathan E. Peelle

Abstract Speech is undeniably significant as a conspecific human communication signal, and it is also perhaps the most ubiquitous class of acoustic signals encountered by the human auditory system. However, historically there was little integration between speech research and the field of auditory neuroscience. Much of this divide can be traced back to the Motor Theory of speech perception, which framed speech not as an auditory process but as one grounded in motor gestures. Recent decades have seen a marked shift in perspective, with mutual interest from researchers in understanding both how neuroscientific principles can be used to study speech perception and, conversely, how speech as a complex acoustic stimulus can advance auditory neuroscience. This introductory chapter reviews this historical context for the modern field of auditory cognitive neuroscience before placing the remaining chapters of the book in context. A number of important themes emerge: methodological improvements, particularly in human brain imaging; the ability to study more natural speech (stories and conversations, rather than isolated stimuli); an appreciation for ways in which different listeners (e.g., of different ages or hearing levels) perceive speech; and incorporation of regions outside traditional auditory and language networks into our neuroanatomical frameworks for speech perception. Evolving techniques, theories, and approaches have provided unprecedented progress in understanding speech perception. These opportunities challenge researchers to ask new questions and to fully integrate speech perception into auditory neuroscience.

Keywords Hearing · Brain · Cognition · Language · Neuroimaging · Neurosciences · Motor Theory

L. L. Holt (✉)
Department of Psychology, Carnegie Mellon University, Pittsburgh, PA, USA
e-mail: loriholt@cmu.edu

J. E. Peelle
Department of Otolaryngology, Washington University in St. Louis, St. Louis, MO, USA
e-mail: jpeelle@wustl.edu

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022

L. L. Holt et al. (eds.), *Speech Perception*, Springer Handbook of Auditory Research 74, https://doi.org/10.1007/978-3-030-81542-4_1

1.1 Speech Perception Research: A Historical Perspective

In many circumstances, speech is crucial to guiding human behavior. The focus of the current book is on auditory speech: acoustically complex changes in sound waves uttered by a talker with the intention of conveying information to a listener. Whether catching up on a favorite television series, taking in the news on public radio, chatting with a friend at a café, or listening to a colleague describe her latest idea, we consume a daily perceptual diet of acoustically complex speech that originates from diverse talkers and blends with distinct acoustic backgrounds. A classic and oft-cited analysis attributes 70–80% of the workday to communication, with about 55% of this time devoted to speech listening (Klemmer and Snyder 1972). Even our own voice provides us with rich input; systematic recordings of natural conversations indicate that we utter an average of about 16,000 words each day (Mehl et al. 2007). Although changes in technology and culture over the years may affect the specifics of these estimates, speech is undeniably significant as a conspecific human communication signal, and it is also perhaps the most ubiquitous class of acoustic signals encountered by the human auditory system. It may seem surprising through a modern perspective, then, that interdisciplinary efforts linking auditory neuroscience and speech perception were not always appreciated. At least some of the reasons for this historical divide can be traced back to the progression of early theoretical ideas about speech perception.

When researchers began investigating speech perception in earnest in the 1950s and 1960s (reviewed in Diehl et al. 2004; Samuel 2011), their landmark research resulted in the discovery of a list of perceptual phenomena that appeared to be present for speech perception, but not for perception of other auditory signals (Cooper et al. 1951; Liberman 1957). This work provided the foundation for what is known about how acoustic cues map to linguistic units like phonemes and revealed the complexity of this relationship (Peterson and Barney 1952; Delattre et al. 1955). Evidence emerged that acoustic information relevant to perceiving phonemes like those that differentiate *bear* from *pear* was categorical and context-dependent – not invariant – and further that it smeared across adjacent phonemes; speech was not as simple as an acoustic alphabet (Fowler 2001). A theory took shape from these observations that had an incredibly strong influence on the course of research spanning many decades.

1.1.1 Motor Theory

Alvin Liberman and his colleagues at the Haskins Laboratories became convinced that perceived phonemes and features have a more nearly one-to-one relationship to articulation than to speech acoustics, and this gave rise to the *Motor Theory* of speech perception (Liberman 1957; Liberman et al. 1967). This Motor Theory took as a first principle that speech signals, by virtue of being human vocalizations providing entry to language, engage human-specific processing entirely distinct from processing other sounds (Liberman et al. 1967; Liberman and Mattingly 1985). In the strongest form of Motor Theory, speech was posited to be perceived as a motor object, not an acoustic one. Specifically, the objects of speech perception were proposed to be the intended phonetic articulatory gestures of the speaker represented as the invariant motor commands that called upon articulator movements to speak. Thus, the Motor Theory imagined the invariant motor commands to be a common currency linking speaking and listening. Crucially, the theory argued that this perceptual-motor relationship did not emerge as a learned association by virtue of having been both a speaker and a listener. Instead, the link was posited to be innately specified as a human-specific mode of perception as part of a larger specialization for language with an adaptive advantage provisioned by the "common currency" to automatically translate from sound to articulatory gesture. Of course, from this perspective, it made very little sense to study the auditory system to understand speech perception, or to study speech perception to understand auditory perception of complex signals. The two were simply distinct systems.

Although Motor Theory was extremely influential in early speech perception research, it was not without controversy. Intellectual debates raged in the 1980s and 1990s. By the early 2000s, weaker versions of Motor Theory were proposed (Galantucci et al. 2006) to accommodate empirical observations that systematically ticked off the list of phenomena purported to differentiate perception of speech from perception of other sounds by demonstrating that under the right conditions, speech and nonspeech perceptual phenomena align (Diehl et al. 2004). In the end, considering nonspeech perception in richer contexts that drew upon attention, learning, and cognitive control demonstrated that the hallmarks of speech perception could often be replicated in nonspeech signals when listeners were afforded the right expertise or listening context (Holt and Lotto 2010; Heald and Nusbaum 2014). Categorical perception provides an example (Harnad 1987). Perhaps the best-known pattern of speech perception, categorical perception refers to the observation that speech sounds gradually changing in their acoustics tend to be perceived categorically, with a sharp boundary in how they are labeled rather than a gradual, graded change in perception that mirrors the acoustics. Further, when listeners discriminate pairs of stimuli drawn from a series of speech sounds, the resulting discrimination function is discontinuous. It is nearly perfect for stimuli that lie on opposite sides of the sharp identification boundary, whereas it is very poor for pairs of stimuli that are equally acoustically distinct but fall on the same side of the identification boundary. Categorical perception was thought to be a peculiarity of speech perception, not evident for nonspeech sounds (Liberman et al. 1957). However, later research demonstrated that categorical perception could emerge for nonspeech sounds when listeners trained to apply category labels to them (Mirman et al. 2004).

Further in contrast to the predictions of Motor Theory, research demonstrated that speech and nonspeech acoustics interacted strongly in perception providing more evidence for a shared substrate (Lotto and Kluender 1998; Holt 2005). Moreover, nonhuman animal listeners (who lack a human speech motor system) were found to exhibit some of the very speech perception behaviors that were thought to differentiate speech from nonspeech perception, including categorical perception (Kuhl and Miller 1978; Kluender et al. 1987), and context effects in perception of speech (Lotto et al. 1997). Finally, damage to the motor speech areas (e.g., in Broca's aphasia) did not produce the speech perception deficits that would be predicted by Motor Theory (Moineau et al. 2005; Hickok 2009). The overall weight of the empirical evidence did not side with the elegant, parsimonious predictions of the Motor Theory.

As evidence contrasting with predictions of the Motor Theory accumulated, the lively – often impassioned – debates regarding the objects of speech perception ultimately moved the field forward. But, there were casualties. The field lost decades of opportunity for realizing the reciprocal benefits of studying the human auditory system in alignment with human speech perception and aligning it with interpretative frameworks from nonhuman animal auditory research. More, it was denied the broader enterprise of understanding the human auditory system using one of the richest, most complex perceptual challenges: speech.

1.1.2 Speech Perception from an Auditory Perspective

Like most pervasive aspects of our lives, it is easy to take speech for granted. We are so adept at speech perception that it hardly seems a major accomplishment. However, the ease with which we perceive speech belies the complexity of the perceptual, cognitive, and neural mechanisms involved and the rich opportunities for advancing understanding of *general* human auditory perceptual abilities by studying the specific perceptual challenges introduced by speech. The fundamental units of speech that carry information may exist for mere moments. These units are complex and may be signaled by a dozen or more variable acoustic dimensions even for simple distinctions that change meaning, like *bear* from *pear.*

Complicating matters further, acoustic speech is often mixed with considerable noise, and even overlapping speech from other talkers. Yet, from this fleeting and complex acoustic signal, we are able to apprehend the linguistic message of the speaker as well as information about her gender, age, region of origin, identity, and emotional state (Kraus et al. 2019). Speech thus provides a rich testbed for understanding general principles of auditory processing, and for observations of auditory processing directed at other nonspeech acoustic signals to inform how we understand the mechanisms available to speech perception. As a complex, ecologically significant acoustic signal, speech presents challenging perceptual dilemmas spanning sensory encoding, prediction, attention, learning, memory, and integration with multimodal sensory inputs as well as other important sensory, perceptual, and cognitive issues. There is much to be gained by investigating the human auditory system through the lens of speech perception.

1.1.3 *Speech Perception Today*

Contemporary research is realizing this promise. The field of speech perception has radically shifted to embrace these reciprocal benefits, with a methodological toolbox equipped to support the endeavor. With the advent of noninvasive functional neuroimaging using hemodynamic (Evans and McGettigan 2017; Peelle 2017) and electrophysiological (Wöstmann et al. 2017) approaches, and the application of invasive neurosurgical approaches to speech perception (Leonard and Chang 2016), there is unprecedented opportunity to examine the human brain's response to speech. Accelerating benefits, auditory science more generally has developed a nascent appreciation for the cognitive aspects of auditory processing, the field of auditory cognitive neuroscience has begun to develop traction, and general cognitive and perceptual mechanisms are increasingly understood to play a role in speech communication (Pichora-Fuller et al. 2016; Peelle 2018).

At the same time, theoretical models of speech originating from cognitive science have greatly informed neurobiological approaches to understanding speech perception. Early cognitive models of speech and the human behavioral research that tested them provided evidence of hierarchically organized levels of representation whereby speech signals activate acquired representations for lower-level phonetic features, categories, and words (McClelland and Elman 1986; Norris 1999), and there is interactive processing across levels (Elman and McClelland 1988) that is modulated by attention (Mirman et al. 2008), the history of experienced that shaped the acquired representations (Kronrod et al. 2016), and online adaptation to short-term input regularities (Norris et al. 2003; Kraljic et al. 2008).

Yet, there remains much to be learned from cognitive science and behavioral approaches; indeed, the very nature of speech representations is actively under debate (Samuel 2020). Nonetheless, at this point in time, general auditory mechanisms, whether described at the cognitive or neurobiological level, are so systematically integrated into accounts of speech perception that early career researchers will likely find it most unusual to learn that the literature raged for decades about whether this was appropriate. In this volume of the *Springer Handbook of Auditory Research*, we showcase these advances at a truly exciting time for research.

1.2 The Auditory Cognitive Neuroscience of Speech Perception

This book is organized such that interested readers can dip into individual chapters of interest, or read the book cover to cover. Although it would be impossible to review the auditory cognitive neuroscience of speech perception in its entirety in a single volume, the chapters included here survey a broad range of theoretical perspectives, methodological approaches, and listening contexts that highlight current successes, challenges, and controversies in the field.

In Chap. 2, Bharath Chandrasekaran, Rachel Tessmer, and G. Nike Gnanateja provide an overview of the subcortical processing of speech sounds. This perspective is important, in part, because it is possible to develop a “cortical bias” in understanding how the brain processes speech, particularly in the context of its role in language. Nonetheless, as Chandrasekaran and colleagues review, there are important subcortical contributions to speech perception. Rather than simply relaying acoustic information to higher-order centers of the auditory system, contemporary research reveals substantial cortical-subcortical interactions in speech processing. There is significant bottom-up as well as top-down processing, a theme that recurs across this book’s chapters. Chandrasekaran, Tessmer, and Gnanateja guide readers through a thorough review of cortical and subcortical anatomy and physiology to situate discussion of the role of subcortical processing in extraction, encoding, and experience-dependent modulation of incoming speech.

In Chap. 3, Yulia Oganian, Neal P. Fox, and Edward F. Chang review contributions of human intracranial recordings to our understanding of speech perception, focusing on the superior temporal gyrus. Although electrophysiology using nonhuman animal models has long played a role in understanding speech perception (Palmer and Shamma 2004; Quam et al. 2017), there are inherent limitations in how much we can learn from species that do not, themselves, use speech to communicate. Oganian, Fox, and Chang provide readers with an overview of empirical findings and the computational tools that have been essential in revealing speech perception in human auditory cortex. Supported in equal parts by the availability of human intracranial data collected in the context of human neurosurgery and advanced computational approaches to analyzing these data, the past two decades have seen an incredibly rapid expansion of our understanding of how auditory regions of the superior temporal gyrus represent speech information. In harmony with empirical literature and theoretical models reviewed by other chapters in this book, these discoveries include demonstrations that the neural representation of speech is nonlinear, and not a faithful representation of the input. Rather, it enhances behaviorally relevant information and is influenced strongly by top-down processing.

In Chap. 4, Sarah Tune and Jonas Obleser explore the role of neural oscillations, the rhythmic or repetitive patterns of neural activity in the central nervous system that generally arise from feedback connections among neurons that result in synchronization of firing patterns, in speech perception. Tune and Obleser provide an introduction to the key characteristics of neural oscillations, as well as their origins and the functions they are thought to support. The authors argue that neural oscillations, studied extensively across sensory and cognitive domains, provide a parsimonious connection of speech perception to broader strategies for sensory, perceptual, and cognitive processing by the brain. Whereas the authors caution against the allure of ascribing distinct oscillations to specific functions, they also present a case for why understanding neural oscillations more generally will allow researchers to relate the complex dynamics of speech perception to neural dynamics. Finally, in linking to other book chapters, Chap. 4 critically examines evidence for the role neural oscillations may play in the perceptual analysis of continuous speech, from analysis of the sounds of speech to sentence-level comprehension.

In Chap. 5, Laura Gwilliams and Matthew H. Davis introduce an information-based approach to speech communication, grounded in statistical properties of speech content and the linguistic information conveyed by speech. Information-based frameworks for spoken communication have a long history in the field, and have recently found new utility in cognitive neuroscience. The chapter provides an overview of the evidence that the neural processing of speech is influenced by linguistic structure of a language – the morpheme and word-level statistical properties of the information conveyed by the acoustic speech signal. The authors situate these findings in information theoretic measures entropy and surprisal, demonstrating their value in understanding neural responses to speech. The authors argue that modeling the *information* content of the speech signal helps to explain the interface between sensory information conveyed by speech and how that interacts with listeners' sensitivity to the statistically structured patterns of linguistic input learned through years of experience. Importantly, information-based approaches can be applied at different levels of analysis (phonemes, words, sentences, and so on), providing a common currency for comparing responses at each of these levels.

In Chap. 6, Stephen C. Van Hedger and Ingrid S. Johnsrude explore how listeners understand speech in adverse listening conditions. They cover a range of challenges listeners might encounter, including background noise, competing talkers, an unfamiliar talker, and more. They provide a systematic review of behavioral and neurobiological evidence demonstrating that even minor challenges to listening demand interactions across perceptual, cognitive, and linguistic processes. The authors make the case that abstract knowledge and context are particularly important when the acoustic speech input is degraded and that although listeners likely draw upon multiple mechanisms to cope with the diversity of adverse listening conditions, the processes generally appear to be attentionally demanding. They describe evidence for the involvement of the cingulo-opercular network – especially anterior insula – in directing the cognitive effort involved in speech perception under adverse listening conditions. Finally, the chapter highlights the importance of the interaction of various listening contexts with individual differences in the cognitive resources available to speech perception, a theme that appears also in Chap. 9 (Rogers and Peelle).

In Chap. 7, Shruti Ullas, Milene Bonte, Elia Formisano, and Jean Vroomen review evidence that the mappings from acoustics to phonetic categories representing the speech sounds of a native language are flexible, rather than fixed. It has long been observed that context can resolve ambiguous speech acoustics (see Chap. 6, Van Hedger and Johnsrude). The movement of the speaker's lips, the context of the sound in a familiar word, and adjacent speech sounds each can provide contextual support to resolve ambiguity in the mapping from acoustics to phonetic categories. The chapter reviews studies that demonstrate that when listeners experience repeated instances of this contextual resolution, longer-lasting perceptual learning or recalibration can occur such that perception of the ambiguous speech acoustics is shifted even when contextual support is no longer available. The chapter also reviews a rich literature that has developed to investigate this adaptive plasticity in speech perception and relates these investigations to theories of speech perception and neuroimaging data that inform its neural underpinnings.

In Chap. 8, Judit Gervain reviews the development of speech perception. Before we are native speakers, we are native listeners – infants begin learning about the patterns of speech in their native language even before birth and, by their first birthday, exhibit substantial experience-dependent reorganization of auditory processing of speech that accommodates the sound patterns of the native language(s). In this way, examination of speech perception across early development provides a window into experience-dependent auditory processing. The chapter reviews the major milestones of the development of speech perception, beginning prenatally and continuing through the first year of life and into the toddler years when word learning and bootstrapping of grammar by the prosodic properties of speech become apparent. The review makes clear that the developing brain orchestrates acquisition of spoken language in parallel across multiple levels of representation that ultimately support speech perception in the native language(s).

Finally, in Chap. 9, Chad S. Rogers and Jonathan E. Peelle discuss interactions between audition and cognition in hearing loss and aging. Earlier chapters (Chap. 5, Gwilliams and Davis, and Chap. 6, in particular, Van Hedger and Johnsrude) make the case that speech perception involves a distributed network of processes, including cognitive processes that vary rather substantially across individuals. Rogers and Peelle highlight the central role of cognitive processes in speech perception among older adults with hearing loss. The chapter reviews age-related changes in both hearing and cognition and describes converging evidence demonstrating their interplay – the evidence indicates that when confronted with acoustically challenging speech, cognitive effort is required. Individual differences in hearing and cognitive abilities determine the cognitive demand of a listener in a particular listening context, and therefore the cognitive and neural resources that contribute to speech perception.

1.3 Common Threads and Future Directions

Although each chapter covers its own specific topic, there are also a number of important themes that cut across chapters that are important to highlight.

A clear shift in the field has occurred with the widespread availability of functional neuroimaging and electrophysiological measurements to study how the brain processes speech. Every chapter in this volume reviews neural data collected from human listeners that were simply unavailable during earlier eras of speech perception research. In fact, reading this volume alongside an earlier volume on speech perception in the *Springer Handbook of Auditory Research* series provides an excellent bird's-eye view of how the field has evolved as new approaches to examining the human brain became ubiquitous (Greenberg and Ainsworth 2004). Keeping pace with advances in data collection methods, there have been increasingly sophisticated approaches to modeling data that incorporate acoustic or linguistic features and permit extraction of neural signatures of specific aspects of the speech signal (Chap. 3, Oganian, Fox, and Chang; Chap. 4, Tune and Obleser; Chap. 5, Gwilliams and Davis). A challenge introduced by this wealth of approaches is to keep sight of the value of integrating what we learn across different methods, levels of analysis, time domains, and populations to advance deeper understanding. Just as important, it will be crucial for the field to not only address "old" questions with these new techniques but also to reconceptualize speech in the context of distributed processing across an interactive brain, and to start asking questions from this new perspective.

Supported by methodological advancements, there is now also an increasing use of more "natural" speech signals, including movies and short stories, to study speech perception (Chap. 3, Oganian, Fox, and Chang; Chap. 4, Tune and Obleser; Chap. 5, Gwilliams and Davis). Of course, most of our everyday communication does not happen listening to isolated phonemes, words, or sentences over headphones while lying in an MRI scanner, and the move toward ever-more natural speech is a positive one. At the same time, almost by definition, these natural stimuli are not well controlled for various acoustic or linguistic features of interest. Thus, the strongest claims will likely need to be backed by converging evidence from both "traditional" experimental paradigms (offering tight control over experimental conditions) and naturalistic listening (verifying real-world applicability).

Another dimension along which our understanding of speech perception is broadening relates to the people doing the listening. There is increasing realization that the challenges (and, hopefully, successes) of speech perception depend not only on the acoustic properties of the speech signal but on the auditory, linguistic, and cognitive abilities of individual listeners (highlighted in Chap. 6, Van Hedger and Johnsrude; Chap. 8, Gervain; and Chap. 9, Rogers and Peelle). The ways in which different listeners perceive speech are important not only to ensure generalizability of our theoretical approaches but also to test specific hypotheses. For example, if we have a hypothesis about how acoustic clarity affects speech perception, then studying speech perception in hearing-impaired listeners is one way to empirically test our claim. Considering speech perception across the lifespan, listeners with different abilities, and a variety of listening environments will ensure that the field converges on robust mechanistic accounts that accommodate the true demands on speech perception.

Neuroanatomically, there is still much focus on core auditory regions including the hindbrain and midbrain (Chap. 2, Chandrasekaran, Tessmer, and Gnanateja) and superior temporal gyrus (Chap. 3, Oganian, Fox, and Chang). However, there is also an increasing appreciation for speech as a whole-brain activity. For example, the fact that regions outside traditional speech and language networks are engaged during adverse listening situations (Chap. 6, Van Hedger and Johnsrude; Chap. 9, Rogers and Peelle) highlights the systems-level interactions required for speech perception (at least under some circumstances). Recognition of these " extra-auditory" brain regions as crucial to speech perception goes hand in hand with the developing appreciation that learning, attention, and cognitive control are crucial components to any full theoretical account of speech perception.

In this regard, speech perception offers a rich testbed for cognitive science and cognitive neuroscience, more broadly. For example, although the Motor Theory did not hold up to empirical scrutiny, there remains important work to be done in understanding the nuanced interactions between speech perception and speech production. Future work also will be needed to blur the arbitrary lines that have traditionally been drawn between perception, learning, attention, and cognition – even outside of speech perception. Speech presents a model case for making progress in this regard; even "online" speech *perception* engages learning (Chap. 7, Ullas, Bonte, Formisano, and Vroomen), and attention (Chap. 6, Van Hedger and Johnsrude), and cognitive processing (Chap. 9, Rogers and Peelle). Similarly, given the intimate connection of speech input with distinct levels of language processing (phonemes, words, etc.), speech provides an ideal model for advancing general understanding of the interplay of hierarchical levels of representation and of predictive models in neural processing (Chap. 5, Gwilliams and Davis).

1.4 Summary

In summary, evolving techniques have provided unprecedented access to neural data, and theoretical perspectives of speech perception are making more and more contact with auditory neuroscience. These opportunities challenge researchers to ask questions that continue to further our understanding of speech perception in new and useful ways. It is an exciting time to be studying speech perception.

Acknowledgments This work was supported in part by grants R01 DC014281, R21 DC016086, R21 DC015884, and R56 AG059265 from the US National Institutes of Health to JEP as well as R01DC017734, R03HD099382, and R21DC019217 from the US National Institutes of Health and BCS1950054 and BCS1655126 from the US National Science Foundation to LLH.

Compliance with Ethics Requirements Lori L. Holt declares that she has no conflict of interest.

Jonathan E. Peelle declares that he has no conflict of interest.

References

Cooper FS, Liberman AM, Borst JM (1951) The interconversion of audible and visible patterns as a basis for research in the perception of speech. Proc Natl Acad Sci U S A 37:318–325

Delattre PC, Liberman AM, Cooper FS (1955) Acoustic loci and transitional cues for consonants. J Acoust Soc Am 27:769–773

Diehl RL, Lotto AJ, Holt LL (2004) Speech perception. Annu Rev Psychol 55:149–179

Elman JL, McClelland JL (1988) Cognitive penetration of the mechanisms of perception: compensation for coarticulation of lexically restored phonemes. J Mem Lang 27:143–165

Evans S, McGettigan C (2017) Comprehending auditory speech: previous and potential contributions of functional MRI. Lang Cogn Neurosci 32:829–846

Fowler CA (2001) Obituary: Alvin M. Liberman (1917-2000). Am Psychol 56:1164–1165

Galantucci B, Fowler CA, Turvey MT (2006) The motor theory of speech perception reviewed. Psychon Bull Rev 13:361–377

Greenberg S, Ainsworth WA (2004) Speech processing in the auditory system: an overview. Springer, New York

Harnad S (1987) Categorical perception: The groundwork of cognition. Cambridge University Press, Cambridge

Heald S, Nusbaum HC (2014) Speech perception as an active cognitive process. Front Syst Neurosci 8:35

Hickok G (2009) Eight problems for the mirror neuron theory of action understanding in monkeys and humans. J Cogn Neurosci 21:1229–1243

Holt LL (2005) Temporally nonadjacent nonlinguistic sounds affect speech categorization. Psychol Sci 16:305–312

Holt LL, Lotto AJ (2010) Speech perception as categorization. Atten Percept Psychophys 72:1218–1227

Klemmer ET, Snyder FW (1972) Measurement of time spent communicating. J Commun 22:142–158

Kluender KR, Diehl RL, Killeen PR (1987) Japanese quail can learn phonetic categories. Science 237:1195–1197

Kraljic T, Samuel AG, Brennan SE (2008) First impressions and last resorts: how listeners adjust to speaker variability. Psychol Sci 19:332–338

Kraus MJ, Torrez B, Park JW, Ghayebi F (2019) Evidence for the reproduction of social class in brief speech. Proc Natl Acad Sci USA 116:22998–23003

Kronrod Y, Coppess E, Feldman NH (2016) A unified account of categorical effects in phonetic perception. Psychon Bull Rev 23:1681–1712

Kuhl PK, Miller JD (1978) Speech perception by the chinchilla: identification function for synthetic VOT stimuli. J Acoust Soc Am 63:905–917

Leonard MK, Chang EF (2016) Direct cortical neurophysiology of speech perception. In: Hickok G, Small SL (eds) Neurobiology of language. Academic Press, London, pp 479–489

Liberman AM (1957) Some results of research on speech perception. J Acoust Soc Am 29:117–123

Liberman AM, Mattingly IG (1985) The motor theory of speech perception revised. Cognition 21:1–36

Liberman AM, Harris KS, Hoffman HS, Griffith BC (1957) The discrimination of speech sounds within and across phoneme boundaries. J Exp Psychol 54:358–368

Liberman AM, Cooper FS, Shankweiler DP, Studdert-Kennedy M (1967) Perception of the speech code. Psychol Rev 74:431–461

Lotto AJ, Kluender KR (1998) General contrast effects in speech perception: effect of preceding liquid on stop consonant identification. Percept Psychophys 60:602–619

Lotto AJ, Kluender KR, Holt LL (1997) Perceptual compensation for coarticulation by Japanese quail (*Coturnix coturnix japonica*). J Acoust Soc Am 102:1135–1140

McClelland JL, Elman JL (1986) The TRACE model of speech perception. Cogn Psychol 18:1–86

Mehl MR, Vazire S, Ramírez-Esparza N et al (2007) Are women really more talkative than men? Science 317:82

Mirman D, Holt LL, McClelland JL (2004) Categorization and discrimination of nonspeech sounds: differences between steady-state and rapidly-changing acoustic cues. J Acoust Soc Am 116:1198–1207

Mirman D, McClelland JL, Holt LL, Magnuson JS (2008) Effects of attention on the strength of lexical influences on speech perception: behavioral experiments and computational mechanisms. Cogn Sci 32:398–417

Moineau S, Dronkers NF, Bates E (2005) Exploring the processing continuum of single-word comprehension in aphasia. J Speech Lang Hear Res 48:884–896

Norris D (1999) The merge model: speech perception is bottom-up. J Acoust Soc Am 106:2295–2295

Norris D, McQueen JM, Cutler A (2003) Perceptual learning in speech. Cogn Psychol 47:204–238

Palmer A, Shamma S (2004) Physiological representations of speech. In: Greenberg S, Ainsworth WA (eds) Speech processing in the auditory system: an overview. Springer, New York

Peelle JE (2017) Optical neuroimaging of spoken language. Lang Cogn Neurosci 32:847–854

Peelle JE (2018) Listening effort: how the cognitive consequences of acoustic challenge are reflected in brain and behavior. Ear Hear 39:204–214

Peterson GE, Barney HL (1952) Control methods used in a study of the vowels. J Acoust Soc Am 24:175–184

Pichora-Fuller MK, Kramer SE, Eckert MA, Edwards B, Hornsby BW, Humes LE, Lemke U, Lunner T, Matthen M, Mackersie CL, Naylor G, Phillips NA, Richter M, Rudner M, Sommers MS, Tremblay KL, Wingfield A (2016) Hearing impairment and cognitive energy: the framework for understanding effortful listening (FUEL). Ear Hear 37:5S–27S

Quam RM, Ramsier MA, Fay RR, Popper AN (2017) Primate hearing and communication. Springer, Cham

Samuel AG (2011) Speech perception. Annu Rev Psychol 62:49–72

Samuel AG (2020) Psycholinguists should resist the allure of linguistic units as perceptual units. J Mem Lang 111:104070

Wöstmann M, Fiedler L, Obleser J (2017) Tracking the signal, cracking the code: speech and speech comprehension in non-invasive human electrophysiology. Lang Cogn Neurosci 32:855–869

Chapter 2
Subcortical Processing of Speech Sounds

Bharath Chandrasekaran, Rachel Tessmer, and G. Nike Gnanateja

Abstract Extant literature identifies the subcortical auditory system as critical to the encoding of key acoustic features relevant to speech. In this chapter, rather than view the subcortex as "lower-level" passive relay stations exclusively involved in speech encoding, a systems neuroscience approach is adopted that argues for active subcortical-cortical interactions during speech processing, subserved by afferent (bottom-up) as well as efferent (top-down) connectivity. These interactions are not only relevant to speech encoding but are critical to the process of mapping highly variable, temporally ephemeral signals to meaningful, behaviorally relevant units. An overview of subcortical and relevant cortical anatomy and physiology is provided as well as a discussion of contemporary neuroscience methodology used to study speech processing. Subcortical plasticity as a function of positive and negative individual experience is discussed, highlighting an emerging understanding of subcortical processes in the extraction, encoding, and experience-dependent modulation of speech signals.

Keywords Auditory cognitive neuroscience · Representation · Auditory nerve · Inferior colliculus · Basal ganglia · Neuroplasticity · Frequency-following responses

B. Chandrasekaran (✉) · G. N. Gnanateja
Department of Communication Science and Disorders, The University of Pittsburgh, Pittsburgh, PA, USA
e-mail: b.chandra@pitt.edu; nikegnanateja@pitt.edu

R. Tessmer
Department of Speech, Language, and Hearing Sciences, The University of Texas at Austin, Austin, TX, USA
e-mail: rachel.tessmer@austin.utexas.edu

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022

L. L. Holt et al. (eds.), *Speech Perception*, Springer Handbook of Auditory Research 74, https://doi.org/10.1007/978-3-030-81542-4_2

2.1 Introduction

The subcortex is loosely defined as central nervous system structures located below the cerebral cortex. Relative to the cerebral cortex, subcortical structures are evolutionarily older brain structures and ontogenetically earlier to mature (Rakic 2009). The subcortex plays a critical role in filtering and orienting to sensory stimuli; coordinating stereotypical motoric responses controlling arousal; and mediating emotion, learning, and memory (Murdoch and Whelan 2009). This chapter discusses the subcortical processing of speech signals. A systems neuroscience approach will be used to describe subcortical systems as an integral part of cortical-subcortical circuits working dynamically to reconstruct, extract, and map speech to linguistically meaningful constructs (Yeo and Eickhoff 2016). This approach moves away from the traditional cortico-centric descriptions of the subcortex as "lower" centers and the cerebral cortex as "higher" centers (Parvizi 2009). When viewed through the systems neuroscience lens, characterizing subcortical function *within* the extensive cortical-subcortical circuitry for speech processing becomes a critical imperative.

2.2 Overview of Subcortical Anatomy

The subcortex includes the deeper parts of the forebrain, the midbrain, and the hindbrain. Primary subcortical forebrain structures include the basal ganglia, the extended limbic system that encompasses the amygdala, the hippocampus, the thalamus, and the hypothalamus. The basal ganglia are a group of subcortical nuclei with complex circuitry that are interconnected with the cerebral cortex (Alexander et al. 1986; Lim et al. 2014). These nuclei include the putamen, caudate nucleus, globus pallidus, subthalamic nucleus, and substantia nigra (Lim et al. 2014). The caudate nucleus and putamen are the main basal ganglia structures that receive input from cortical areas, while the globus pallidus is primarily an output structure (Murdoch and Whelan 2009). Output signals from the basal ganglia project through the thalamus back to the cerebral cortex via open and closed loops (Alexander et al. 1986; Hélie et al. 2015). Open loops target cortical sites that do not directly connect with the basal ganglia, while closed loops target cortical sites that project to the basal ganglia (Alexander et al. 1986). The temporal lobe is one of many output targets of the basal ganglia (Middleton and Strick 1996). Several functionally distinct loops have been identified that regulate important motoric, motivational, executive, and learning-related functions (Seger 2006). The amygdala and the hippocampus are part of an extended limbic network that plays a critical role in emotion processing, learning, and memory function. The thalamus includes the medial geniculate body (MGB; see Table 2.1 for abbreviations), an auditory nucleus that projects to the primary auditory cortex.

Midbrain structures include the tectum, composed of the superior colliculus and inferior colliculus (IC), and the tegmentum, which consists of the reticular network

Table 2.1 Table of abbreviations

Abbreviation	Full name
ABR	Auditory brainstem response
AN	Auditory nerve
cABR	Complex auditory brainstem response
CF	Characteristic frequency
DLS	Dual-learning systems
EEG	Electroencephalography
FFR	Frequency-following response
fMRI	Functional magnetic resonance imaging
IC	Inferior colliculus
MEG	Magnetoencephalography
MGB	Medial geniculate body

and the substantia nigra. The hindbrain includes the cerebellum, the pons, and the medulla oblongata. While the cerebellum is critically involved in motor learning, movement coordination, and maintaining equilibrium, there is considerable cerebellar involvement in regulating cognitive functions, including speech and language (Jueptner et al. 1997; Doya 2000). Similar to the basal ganglia, the cortex is connected to the cerebellum via cerebrocerebellar loops (Murdoch and Whelan 2009). Anatomically and functionally relevant bidirectional connectivity exists between the cerebellum and the basal ganglia (Hoshi et al. 2005; Bostan et al. 2010). Taken together, there is an emerging systems-level view of the cerebello-basal ganglia-thalamo-cortical network that subserves cognitive, motor, and sensory functioning.

The midbrain and the hindbrain structures are also called the brainstem and form the most vital transmission and processing route for auditory processing from the inner ear to the auditory cortex. Figure 2.1 illustrates auditory neural pathways in the brainstem and their ascending and descending connections. Macroanatomically, the auditory cortex is divided into the core, belt, and parabelt regions. The core is most caudally located and is surrounded by the belt, which, in turn, is surrounded by the parabelt. Microanatomically, the auditory cortex is arranged in a columnar pattern with five layers that show different patterns of subcortical connections with the ascending and descending auditory networks (Linden and Schreiner 2003; Kral and Eggermont 2007). The thalamus and the brainstem (composed of the midbrain, pons, and medulla oblongata) encompass several nuclei that are specialized for auditory processing. These include the MGB of the thalamus, which relays information to the primary and secondary auditory cortices located within the temporal lobe of the cerebral cortex. Ascending lemniscal pathways from the ventral and dorsal MGB terminate in layers III and IV of the auditory cortex. In contrast, those from the medial MGB terminate in all the layers (Bartlett 2013). Non-lemniscal pathways from the deep and caudal dorsal nuclei and the medial nuclei of the MGB terminate in layers I, III, and V of the auditory cortex. Beyond its traditionally identified role in sensory gatekeeping, the MGB is also involved in task-dependent modulation of

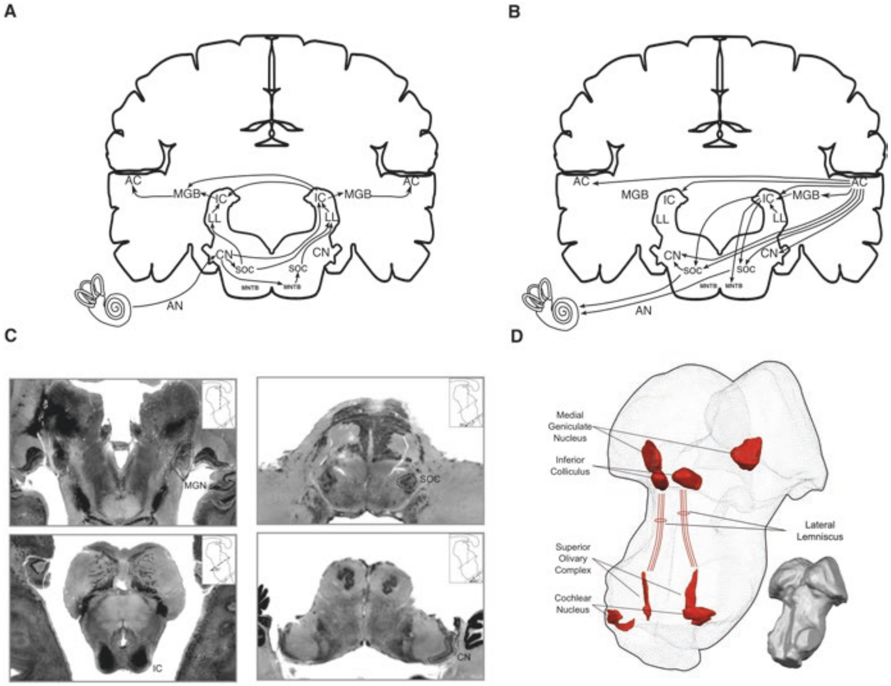

Fig. 2.1 Illustration of the ascending (**a**) and descending (**b**) pathways in the central auditory system. *AC* auditory cortex, *MGB* medial geniculate body, *IC* inferior colliculus, *LL* lateral lemniscus, *SOC* superior olivary complex, *CN* cochlear nucleus, *AN* auditory nerve. (Based on Kral and Eggermont (2007)). (**c**) The cross section of the brainstem with the major auditory nuclei marked, while (**d**) shows the three-dimensional MRI reconstruction of the brainstem with major auditory nuclei highlighted. Inset in (**d**) shows the MRI reconstruction of the brainstem. ((**c**, **d**) Adapted from Sitek et al. (2019) with permission)

speech signals (von Kriegstein et al. 2008). This modulation is mediated by descending projections to the MGB from layers V and VI of the auditory cortex.

The MGB receives ascending projections from the IC. The IC is considered a major hub within the auditory system, to the extent that some researchers consider the IC a computational and functional equivalent of the primary visual cortex (Nelken 2008). The IC has three subdivisions: a central nucleus, a lateral nucleus, and a dorsal nucleus. The central nucleus receives ascending projections from various brainstem nuclei and is organized on the basis of tonotopic and periodotopic mapping (Schreiner and Langner 1997; Baumann et al. 2011). The IC also receives descending projections from layer V of the auditory cortex.

The dorsal nucleus receives descending connections from the cerebral cortex, including direct projections from the primary and secondary auditory cortices (Winer 2005). Inferior to the IC, the superior olivary complex is located in the pons, while the cochlear nucleus is located in the medulla oblongata. The superior olivary complex is the first point of convergence (decussation) for input from the left and right ears (Rasmussen 1946). The cochlear nucleus receives ipsilateral input from the auditory nerve (AN) and connects to the IC via the lateral lemniscus. The representation of speech across these early auditory pathways is discussed in Sect. 2.3.1. The superior olivary complex and the cochlear nucleus also receive direct descending connections from layer V of the auditory cortex.

2.3 Subcortical Speech Representation

2.3.1 Insights from Animal Models

A primary goal of auditory cognitive neuroscience is to understand the neural mechanisms underlying how sound information is transformed into behaviorally relevant constructs. Invasive studies on animal models, including birds, rodents, and primates, have provided rich information on the subcortical processing of conspecific vocalizations as well as conditioned sounds that have been rendered behaviorally relevant (Portfors et al. 2009; Romanski and Averbeck 2009). There is extensive literature on subcortical processing of human speech from these animal models using a comparative approach. Insights gained from this approach have to be tempered by the fact that the speech signal may not be behaviorally relevant intrinsically to animals and may not yield information regarding processes that are specific to humans or the human experience.

Prior work has argued for the validity of a comparative approach on the basis that the anatomical and physiological neural infrastructure in the early auditory pathway of several animal models is largely consistent with the organization found in humans. There is remarkable neural consistency in subcortical speech representation, evidenced by thorough comparisons of scalp-recorded potentials across human and animal models (Ayala et al. 2017). Concerning the “special” status of speech to human processing, examination of speech representation in animal models has yielded rich information regarding phenomena previously thought to be exclusive to human speech processing (Kuhl 1981; Lotto et al. 1997). For example, definitive work in chinchillas (*Chinchilla lanigera*) demonstrated that categorical perception of speech is a phenomenon that is not specific to humans (Kuhl and Miller 1978). A general conclusion is that animal models provide crucial information, at least at the level of phonemic representation and contextual modulation. Studying the representation of speech in animal models is limiting insofar as to what insights can be revealed regarding behavioral ramifications. To counter this limitation, several studies have trained animals to distinguish human speech sounds, providing abundant information regarding brain-behavior correspondences (Engineer et al. 2008; Ranasinghe et al. 2012).

Previous research has utilized animal models to examine the role of the subcortex in reconstructing speech signals using synthesized speech, speech analogs, and a combination of these stimuli. These studies have leveraged pioneering work on

AN representation of speech. Systematic studies on major phonetic classes, including stop consonants, fricatives, and vowels, have revealed the basic neural properties used to code linguistically relevant units at the AN (Delgutte and Kiang 1984a, b). AN representation of speech is highly sensitive to sound onset and demonstrates frequency-specific adaptation. Both onset responsivity and adaptation phenomena are preserved across the auditory subcortical network. Neural adaptation can be leveraged to robustly define onset information as well as encode successive, acoustically distinct elements critical to consonant perception (Delgutte 1990). For example, neurons with high characteristic frequency (CF) robustly encode the consonant burst, while neurons with low CF do not adapt to the high-frequency bursts and robustly capture the onset of voicing. Together, the two tonic components, along with neural adaptation, can provide information regarding voice onset time, a phonologically relevant cue for stop consonant perception (Delgutte 1997).

Concerning critical spectral components that define properties related to vowels (e.g., formants) and consonants (e.g., burst spectra), several coding schemes have been identified. In a rate-place scheme, the discharge rate is greatest for neurons with CFs closest to critical spectral peaks (e.g., first, second, and third formants) (Young and Sachs 1979). The rate-place scheme provides an excellent representation of spectral features at low stimulus intensity levels but not at high levels: a well-described phenomenon referred to as the dynamic range problem in auditory neuroscience (Evans 1981). More complex rate-place models incorporate weighted combinations of neurons with different spontaneous discharge rates (Delgutte 1990). For example, low spontaneous firing neurons show a higher dynamic range but are fewer in number. Therefore, a weighted combination may capture a greater dynamic range that is closer to the broad dynamic range evidenced in human perception. Temporal schemes involving phase-locking have also been suggested as a putative mechanism underlying spectral processing at the level of the AN, which can phase-lock up to 5 kHz (Delgutte 1997). Interspike interval distributions from the AN provide rich information regarding pitch coding (Delgutte 1990). Rate-place codes also contribute to pitch percept, although there is evidence that they are not as robust as temporal processing schemes.

While the current understanding of AN responses to speech signals in animal models is significant, fewer studies have probed the speech representation schema in subcortical structures. In general, the response properties of neurons in the cochlear nucleus and the IC are significantly more abstract relative to the AN. This complexity is driven by a large number of converging inputs from various neural centers as well as the presence of several combination-sensitive neurons at each ascending level in the central auditory system (Nelken 2008). There is also a great amount of diversity in neuronal cell types and their response properties (Kim et al. 1986; Ranasinghe et al. 2013). Despite this complexity, there is some commonality in representation and processing schemes in the auditory brainstem relative to the AN. The diversity in neuronal subtypes ensures complementary response patterns. For instance, primary-like cells in the cochlear nucleus, similar to the AN, have high phase-locking capability yet show a reduced dynamic range (Kim et al. 1986;

Delgutte 1997). In contrast, chopper neurons show a broader and more stable dynamic range but poorer phase-locking at higher frequencies.

In the auditory brainstem, the complexity of neuronal responses makes it challenging to relate response properties to specific acoustic properties. One approach to characterizing response preference is to examine the spectro-temporal receptive fields (preferred spectral and temporal properties) of neurons (Fritz et al. 2003; David et al. 2007). In the IC, this approach demonstrates a close correspondence between derived receptive fields and acoustic properties that are behaviorally relevant in the signal and may be a useful approach to quantify subcortical speech representation (Pasley et al. 2012). This approach has yielded rich information in the understanding of cortical speech representation. When comparing single unit and population responses to speech sounds across the IC, the MGB, and the auditory cortex, a key finding is that responses in the auditory cortex and the MGB are much more abstract relative to the IC (Ranasinghe et al. 2013). There is also a large amount of information redundancy in IC neurons that is lost at the level of the auditory cortex, where the code is sparser and more efficient (Chechik et al. 2006). Thus, the animal models used to study subcortical encoding of speech show specialized responses to different parameters of speech, which are transformed at multiple levels to retain and decode relevant acoustic information.

2.3.2 Insights from Human Studies

Human studies examining subcortical speech representation using non-invasive methods provide complementary information to the animal studies reviewed in Sect. 2.3.1. The behavioral relevance, application to clinical disorders, and easy trainability of humans are critical advantages relative to animal models. The drawback is that the non-invasive nature of human neuroimaging methods makes it challenging to specify the underlying neural sources and mechanisms/processing schemes. Despite this constraint, human studies have corroborated prior work on animal models and advanced auditory cognitive neuroscience in new and exciting directions. Over the last decade, there have been a plethora of studies examining experience-dependent plasticity in the subcortical representation of speech signals (Chandrasekaran and Kraus 2010; Chandrasekaran et al. 2014b). The general finding across these studies is that speech cues are represented with remarkable fidelity at the subcortical level, and these representations are highly shaped by long-term and short-term auditory experiences (Xie et al. 2017; Reetzke et al. 2018). Further, this subcortical speech representation is shown to be highly specific to individual experiences (Skoe and Chandrasekaran 2014).

It is important to note that the rich information from the auditory brainstem is distributed across multiple cortical regions that are fine-tuned to extract specific features in speech and aid in robust speech perception. It is at the cortex that the lower-level features of the sound are transformed into higher-level linguistic and acoustic features, which are essential for speech perception. The core region in the auditory cortex maintains the rich representation of the acoustic features in speech which are transmitted to the non-core auditory regions in the lateral parts of the superior temporal gyrus for mapping onto phoneme representations (Nourski 2017). Mapping the spectro-temporal features to phoneme features is also mediated by inputs from various regions in the cortex such as the inferior frontal and medial frontal gyri. The following sections review non-invasive neuroimaging of subcortical auditory processing (Sect. 2.3.2.1), subcortical representation of segmental and suprasegmental speech features (Sect. 2.3.2.2), neuroplasticity in subcortical speech representation (Sect. 2.4), subcortical systems in learning novel speech categories (Sect. 2.5), and subcortical processing in individuals with clinical disorders (Sect. 2.6).

2.3.2.1 Non-invasive Neuroimaging of Subcortical Auditory Processing

Examining the subcortical representation of speech features in humans is challenging with non-invasive neuroimaging methods. It can be challenging to image deep subcortical structures using some of these neuroimaging techniques (Crosson et al. 2010). Subcortical nuclei are typically small, heterogeneous, and located deep in the brain near neural centers that regulate breathing and cardiac rhythm. Non-invasive functional neuroimaging methods that assay hemodynamic properties (e.g., functional magnetic resonance imaging, fMRI) are affected by all of these subcortical characteristics. The timescale of the blood oxygenation level-dependent signal and the negative impact of scanner noise make it difficult to study the representation of speech sounds in subcortical structures using fMRI (Guimaraes et al. 1998; Chandrasekaran et al. 2014b). Numerous methodological advances have attempted to overcome these obstacles. These include high-resolution imaging and depth-based analyses aimed to tackle issues related to size and depth, and the use of cardiac gating and retrospective physiological noise correction to enhance subcortical signal-to-noise (Glover et al. 2000; Ress and Chandrasekaran 2013). Despite these available methods, few studies have examined speech processing across the entire subcortical auditory system.

Magnetoencephalography (MEG) has high temporal and spatial resolution and is relatively less affected by the skull and the volume-conducting media (Cuffin and Cohen 1979). Unlike fMRI, MEG does not produce a lot of noise that can energetically mask the auditory subcortical responses. MEG is highly sensitive to the tangential electrical dipoles generated by the primary neuroelectric activity at the brain surface. However, MEG is not very sensitive to the radial dipoles that are a characteristic of deeply located subcortical brain regions. Due to the IC being deeply seated in the brain and its low response magnitude, MEG is not the most efficient method to record subcortical activity.

Given the limitations of fMRI and MEG, the majority of studies in humans have utilized electroencephalography (EEG) to assess the subcortical encoding of speech signals. EEG has excellent temporal resolution, and, due to its sensitivity to radially oriented dipoles, it can be used to reliably track the subcortical encoding of rapidly varying spectro-temporal cues in speech. Further, EEG is relatively inexpensive, and a single electrode channel on the scalp (when appropriately referenced and grounded) is often sufficient to record the subcortical responses to speech with high fidelity. However, EEG suffers from poor spatial resolution, due to which the differential subcortical encoding of speech cues at different levels in the brainstem cannot be reliably measured. However, multichannel recordings and advanced source localization methods may yield insightful information regarding specific subcortical nuclei (Bidelman 2015). Nevertheless, by leveraging its excellent temporal precision, EEG studies have contributed to a rich understanding of subcortical speech processing.

Two subcortical EEG components have been used to study the neural encoding of the speech signal: the auditory brainstem response (ABR) and the frequency-following response (FFR). The ABR measures obligatory onset-related auditory evoked responses from neurons in the brainstem that are phase-locked to the stimulus onset. The FFR is a neurophonic response that reflects activity from subcortical neural ensembles that are phase-locked to the periodicities in the stimulus (Hecox and Galambos 1974). The ABR and the FFR to speech signals likely reflect different neural brainstem ensembles specialized in processing onset and sustained (phase-locked) information in speech signals (Bidelman 2015). Both components can also be elicited in animal models (Ayala et al. 2017). Work on invasive models suggests that these components are both primarily driven by subcortical activity from the neural ensembles in the cochlear nucleus to the IC (Marsh et al. 1970; Smith et al. 1975). The ABR and the FFR evoked to periodic, complex stimuli have been referred to as a complex auditory brainstem response (cABR) in the literature (Skoe and Kraus 2010). It has to be mentioned that these terms can be misleading, where ABR leads one to think that the response exclusively arises from the brainstem. A consensus suggests the preferred use of the term “FFR” as it does not presuppose the anatomical source of activity (Kraus et al. 2017a; Coffey et al. 2019). Though the cABRs mentioned above are technically onset/transient responses, they are considered to fit under the umbrella term FFR to differentiate them from the ABRs that are conventionally recorded to clicks and tone-bursts.

Typically, a consonant-vowel stimulus can evoke both components (see Fig. 2.2). An onset component of the FFR is seen at a delay of approximately 5–10 ms, which is consistent with the neural delay from the ear to the brainstem nuclei. This response delay is considered a critical signature of subcortical processing, as onset responses from cortical sources show delays in the order of approximately 50 ms (Coffey et al. 2016). The sustained portion of the FFR lasts the duration of the periodic stimulus and has been extensively used as a measure of the integrity of encoding of periodic properties within the speech signal (Chandrasekaran et al. 2014b). The FFRs are typically much smaller (in the nanovolts scale) than cortical responses and typically require a large number of trials to obtain desirable signal-to-noise ratios. The FFRs are also considered pre-attentive and can be robustly elicited even when the participant is not attending to the repetitively presented stimulus. In fact, studies examining FFRs to speech stimuli are frequently recorded either when the participant is asleep or while the participant watches a subtitled movie without audio.

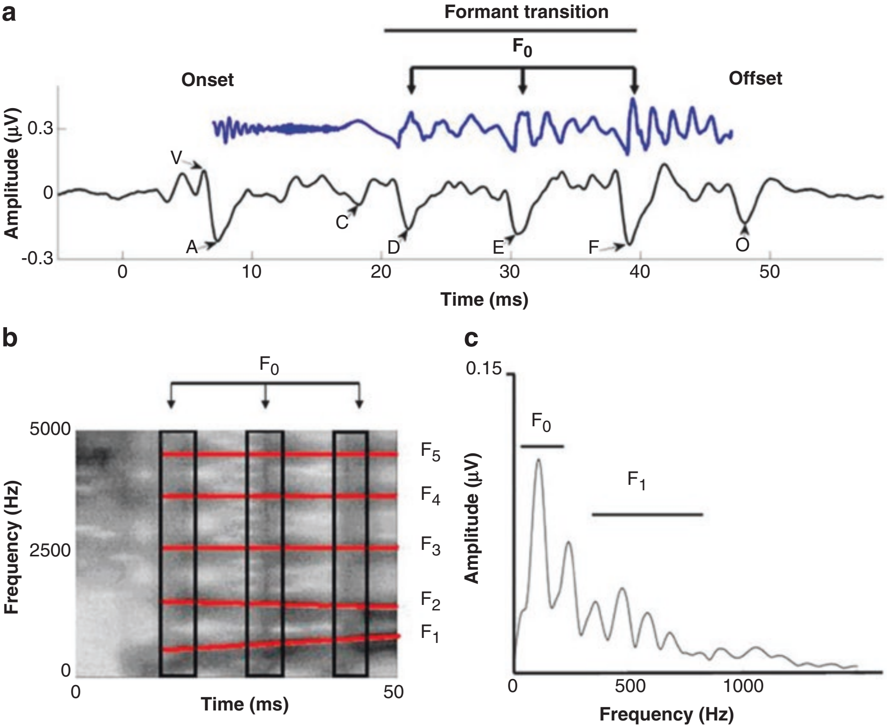

Fig. 2.2 (**a**) The time-amplitude waveform of a 40 ms synthesized /da/ (blue) time-shifted to align with the time-locked brainstem response (black). *A* marks the onset peak (~6–10 ms) that follows the well-characterized wave V of ABR; *C* captures the formant transition from the consonant to the vowel; *D*, *E*, and *F* mark responses to the dominant periodic elements of the vowel stimuli, including the fundamental frequency; and *O* represents the offset of the stimuli. (**b**) The broadband spectrogram of the /da/ stimulus, with darker areas indicating greater energy. The relative spacing between the frequencies of the first formant (*F1*) and the second formant (*F2*) relates to vowel identity. (**c**) The fast Fourier transform analysis of the brainstem response to /da/, showing representation of the fundamental frequency and its harmonics. (From Chandrasekaran and Kraus (2010))

Along with the onset and sustained portions of the FFR, an offset response has also been elicited in some studies, but the extent to which this component can be disassociated from the sustained portion of FFR is unclear. There are several comprehensive reviews on these responses that speak to the neural origin and characteristics of the onset, sustained, and offset components in the FFR. Skoe and Kraus (2010) offer a detailed guide on how to record and analyze these scalp-recorded components in humans. Reviews also cover the mechanisms underlying neuroplasticity, the clinical and biological relevance of onset and FFR components, and what these components may reveal about the organization of the auditory system (Chandrasekaran et al. 2014b; Kraus and White-Schwoch 2015). Section 2.4 focuses on what these components reveal about the representation of speech signals: how experiences can shape subcortical speech representation and conceptual challenges that arise when interpreting the large body of literature on using these components as markers of subcortical speech representation.

Due to the relatively high temporal and spatial resolution, and the possibility of noise-free scanning, MEG has been well utilized to assess cortical processing of auditory signals, including speech (Hämäläinen et al. 1993). Advances in MEG have been leveraged to record reliable FFRs in humans with high temporal and spatial resolution (Coffey et al. 2016, 2019). With advanced source localization algorithms and recordings with high signal-to-noise ratio, MEG has been utilized to record radial dipolar activity from several deep brain structures, including subcortical auditory nuclei (Coffey et al. 2016, 2017). For example, evoked activity in the cochlear nucleus and IC can be recorded in response to repetitively presented speech sounds (Coffey et al. 2016). Additionally, with source analysis, it has been shown that the FFRs consist of phase-locked components generated both in the brainstem and the auditory cortices. The cortical contribution to the FFR suggests that the response does not purely reflect subcortical processes. Rather, consistent with a systems neuroscience viewpoint, the FFRs to speech signals reflect an integrated response from subcortical and cortical circuitry. Multichannel EEG studies show that subcortical responses dominate the FFRs, and the cortical contribution rolls off beyond periodicities above 150 Hz (Bidelman 2015, 2018).

2.3.2.2 Subcortical Representation of Segmental and Suprasegmental Speech Features

Onset and FFR components have been elicited in response to segmental, suprasegmental, and paralinguistic information in speech (Kraus et al. 2017a). Stimuli are typically repetitively presented syllables, although more advanced analysis methods in studies have enabled the use of continuous, naturalistic speech stimuli (Forte et al. 2017; Maddox and Lee 2018). Steady-state vowel stimuli across a range of intensities can also evoke the FFR. FFRs have been shown to reliably encode formant feature patterns critical for vowel identity. This was demonstrated using back vowel stimuli so that the first two formants were within the FFR phase-locking limits (Krishnan 2002). The harmonics corresponding to the formant frequency peaks of the vowels were enhanced while the non-formant harmonics were suppressed. This is consistent with data from population models of AN activity incorporating rate-place algorithms, suggesting that the FFR, at least in part, inherits properties of the AN. Using a data-driven machine learning approach, a study examined the extent to which the vowels /æ/ and /u/ could be decoded from single-trial FFRs using novel machine learning–based techniques (Yi et al. 2017). This type of single-trial vowel decoding was based on the formant spectral patterns that best differentiated the vowels from a large database of natural speech utterances (Hillenbrand et al. 1995). The potential to extract phonemic information from single-trial FFR based on interpretable spectral features sets the stage to examine the online modulation of speech signals.

FFRs have also been measured in response to dynamic speech stimuli. For example, a 40 ms segment of the syllable /da/ has been extensively used to evoke onset and FFR components. The evoking stimulus is characterized by a sharp onset burst, a short transition to the vowel, and a steady-state period that reflects the low back vowel (Skoe and Kraus 2010). Figure 2.2 also shows the typical neural response to the syllable /da/. In the response, A marks the onset peak (approximately 6–10 ms) that follows the well-characterized wave V of the ABR; C captures the transition from the consonant to the vowel; D, E, and F mark responses to the dominant periodic elements of the vowel stimuli that correspond to the fundamental frequency; and O represents the offset of the stimuli. Spectral analyses of the responses reveal broad peaks in the fundamental frequency and the first formant regions. The characteristic neural response to /da/ can be evoked without the need to attend to the stimuli and demonstrates high test-retest reliability (Hornickel et al. 2012). The characteristic neural response to /da/ has also been elicited from animal models (White-Schwoch et al. 2016; Ayala et al. 2017).

Figure 2.2 also reveals a constraint in the use of the onset and sustained components of FFR to index speech representation. While the FFRs can reveal reliable phase-locking up to about 1200 Hz (Bidelman and Powers 2018), speech signals contain critical information beyond this frequency. For example, some phonological contrasts are differentiated by differences in the third formant (/l/ vs. /r/) that is well outside the phase-locking limit observed using the FFR. This limitation at least partially reflects physiology. The phase-locking limit systematically reduces across each ascending nucleus in the auditory system as the response becomes increasingly abstract (Chandrasekaran and Kraus 2010). Nevertheless, differences that are above the subcortical phase-locking limitations can still be represented in the neural response. For instance, stop consonants differing by the spectral content within the burst in the second and third formant range evoke consistent shifts in the latency of the onset response that could be used as a cue to differentiate phonemic contrasts (Hornickel et al. 2009). For example, /ga/ evokes an earlier response relative to /da/ or /ba/ (Warrier et al. 2011). Such phase differences are not epiphenomenal and are likely generated at the level of the IC. While subtle, the latencies of these responses can differentiate between phonological contrasts cued by spectral properties outside phase-locking limits.

In addition to segmental information, the FFRs can also assay suprasegmental information in speech (Krishnan et al. 2004; Xie et al. 2017). In tonal languages such as Mandarin, pitch changes within syllables can distinguish word meanings (Yip 2002). FFRs can be reliably elicited in response to Mandarin pitch contours in native as well as non-native listeners (see Fig. 2.3). Various metrics have been used to establish the robustness of pitch encoding as well as the fidelity of the response relative to the stimulus. Typically, interpeak intervals are calculated from the FFRs and the eliciting stimuli using autocorrelation (Skoe and Kraus 2010). Metrics related to the robustness and accuracy of phase-locking are used to quantify the FFRs. Advanced machine learning algorithms have advanced the decoding of pitch contours from single-trial FFRs. Such advances in evaluating the neural encoding of pitch contours from spectro-temporally rich stimuli have paved the way for the

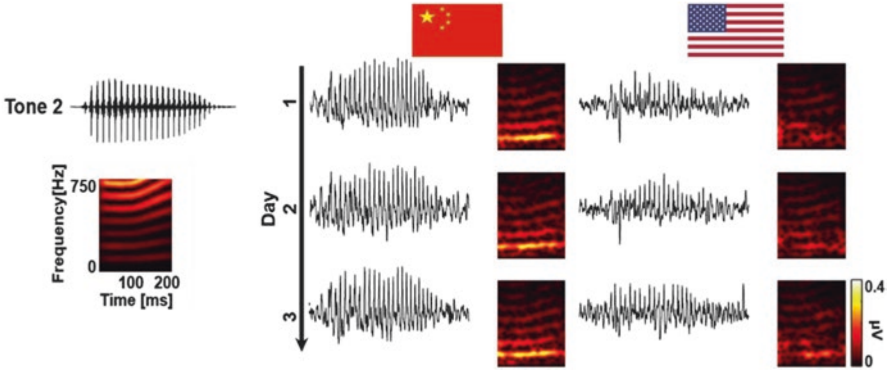

Fig. 2.3 (**a**) Waveforms and spectrograms of rising Mandarin Tone 2 that were used to elicit the frequency-following response (FFR). (**b**) Waveforms and spectrograms of FFRs elicited by Tone 2 across 3 days in an example native Mandarin-speaking participant. (**c**) Waveforms and spectrograms of FFRs elicited by Tone 2 across 3 days in an example native English-speaking participant. The FFRs are highly stable across days within participants. (Data from Xie et al. (2017))

application of FFRs in paradigms that involve behavioral measures of pitch encoding and categorization of linguistic pitch contours (Xie et al. 2018, 2019). Such metrics can be further used to trace neural plasticity associated with online learning in categorization tasks at different timescales.

2.4 Neuroplasticity in Subcortical Speech Representation

In a seminal study, Krishnan et al. (2005) demonstrated cross-language differences in FFRs evoked to Mandarin tones. Prior to this study, the dominant view was that language-specific encoding occurred at cortical stages of processing (Davis and Johnsrude 2003). Multidimensional scaling studies on the judgment of tone similarity have yielded significant cross-language differences in the perception of dimensions underlying tones, with native speakers of contour-tone languages weighing pitch direction more than height, reflecting perceptual warping to enhance dimensions critical to disambiguating phonological contrasts (Gandour 1983). Krishnan et al.'s (2005) study also showed that Chinese listeners had enhanced phase-locking and more faithful pitch encoding of Mandarin tone contours relative to English listeners. These cross-language differences persist for non-speech homologs of the tone patterns but are eliminated when the pitch patterns are not ecologically valid (Xu et al. 2006; Swaminathan et al. 2008). Follow-up studies have revealed that cross-language differences reflected by the FFR to tones are not an effect of a better signal-to-noise ratio or an overall gain effect; rather, such plasticity is specific to critical regions of the signal that are highly dynamic (Krishnan and Gandour 2009).

Figure 2.3 shows the FFRs to a rising Mandarin tone contour elicited from a representative native speaker of Mandarin and a representative native speaker of

English across 3 days of recording. The FFRs represent the average of 1000 trials. For both participants, the FFRs reflect periodic content in the stimuli including the fundamental frequency and the lower harmonics. The FFRs also reveal consistency in the cross-language differences across recording days. The FFR from the native Mandarin participant is more robust in regard to phase-locking and yields greater fidelity to the stimulus pitch pattern relative to the native English participant. Thus, the cross-language differences seen in the FFR are stable.

Long-term language experience can shape the subcortical representation of speech. Such plasticity likely reflects the long-term reorganization of the subcortical structures to better accommodate linguistically relevant features in the signal (Krishnan and Gandour 2009). As a follow-up to studies examining long-term, experience-dependent plasticity in subcortical speech representation, researchers have asked questions related to the specificity and length of training required for subcortical auditory plasticity. Reetzke et al. (2018) showed that plasticity-related changes in FFRs with systematic everyday training on a tone categorization task emerge much after behavioral performance plateaus, which takes about 3 weeks of continuous training. Jeng et al. (2011) compared FFRs elicited in response to lexical tones from adults and neonates whose native languages were either Mandarin or English. They only found significant differences in neural pitch tracking for adults, suggesting that language-dependent differences require some amount of exposure to the sound properties of the native language.

Concerning the specificity of experience, long-term music experience has been found to enhance the representation of specific linguistic tone contours (Wong et al. 2007). This suggests that there may be a cross-domain effect wherein long-term music training enhances pitch processing, which spills over to the encoding of pitch patterns in speech. Musicians also demonstrate faster and more accurate learning of tone patterns, suggesting that sensory advantages may have behavioral relevance (Smayda et al. 2015). Native speakers of a tone language demonstrate an advantage in the subcortical representation of non-native tones, relative to speakers of non-tonal languages (Krishnan et al. 2010). Beyond tones, long-term music training and bilingualism can alter the subcortical representation of speech signals (Krizman et al. 2012; Skoe and Kraus 2012). In particular, the fidelity of critical timing and pitch-related cues is enhanced in musicians and bilinguals, relative to non-musicians and monolinguals, respectively. The outcomes of subcortical plasticity are more evident in challenging listening environments, wherein tracking the dominant cue may be more difficult.

The studies reviewed so far unambiguously demonstrate that long-term auditory experiences can shape subcortical speech processing. The question of whether the mature subcortical auditory system is malleable enough to reorganize in response to newly learned speech information remains unanswered. This question directly ties into a debate in animal models regarding the role of subcortical plasticity (Weinberger 2004). Models of auditory plasticity have largely differed on the time-course and relevance of subcortical plasticity (Suga and Ma 2003; Weinberger 2004). One influential model suggests that IC plasticity is rapid and directly contributes to enhancing auditory cortical plasticity in response to behaviorally relevant signals

(Gao and Suga 2000). This model also suggests that subcortical plasticity is required for auditory learning. Other models suggest that cortical plasticity is not "inherited" and that it may work independently from subcortical plasticity (Weinberger 2004).

Studies with humans have investigated the extent to which short-term learning can enhance subcortical representation (Song et al. 2008; Chandrasekaran et al. 2012). Using a paradigm typically involving multiple talkers, trial-by-trial feedback, and natural stimuli, it has been shown that the FFRs can be modulated by short-term training (Chandrasekaran et al. 2012). While subcortical plasticity is not evidenced on the day of training (although behavioral gains are evident), more robust representations of newly learned categories emerge after overnight consolidation (Xie et al. 2017). A study demonstrated that systematic short-term training for non-native linguistic pitch contour categorization resulted in enhanced pitch representation in the FFRs, suggesting that the subcortical pathway is malleable and plastic in adults (Reetzke et al. 2018).

Two different models have been proposed as mechanistic explanations for subcortical plasticity in the representation of speech signals. Top-down, corticofugal modulation is thought to be an important mechanism for guiding subcortical neural plasticity (Skoe and Kraus 2012; Chandrasekaran et al. 2014b). This modulation is thought to drive subcortical enhancements for behaviorally relevant signals, such as language-relevant stimuli, and bolster transfer effects between speech and music. In contrast, experience-dependent effects in FFRs may also be driven by local reorganization of synaptic plasticity within the subcortical network that enhances frequently encountered features in the speech signals (Krishnan et al. 2005, 2010). Per this latter model, neural ensembles in the brainstem may recalibrate over time to preferentially encode signals that are frequent within one's auditory environment (Chandrasekaran et al. 2014b). These models are not mutually exclusive. It is possible that a developmental approach could bridge these models. Kral and Eggermont (2007) propose that local reorganization within the auditory system drives plasticity early in development. By maturity, top-down effects may dominate via corticofugal pathways. This is largely consistent with principles underlying learning as well. While early speech learning can be primarily driven by unsupervised, Hebbian learning processes, speech learning in the mature system requires supervision and task-related attention (Chandrasekaran et al. 2014a).

Disambiguating between bottom-up and top-down effects is a challenge with EEG-based approaches. Due to the superior spatial resolution relative to EEG, MEG has been used to assess the sources underlying the FFR to speech syllables. In addition to activity localized to the cochlear nucleus and the IC, MEG has also revealed a dominant auditory cortical source that contributes to the FFR (Coffey et al. 2016). This raises interesting questions, as at least some portion of the FFR may be driven by cortical neurons that can phase-lock to frequencies in the range of the male fundamental frequency. Evidence from MEG studies that show cortical contributions to the FFRs provide a new perspective into neural plasticity observed in the above studies. It is not clear if the observed training or experience-related effects in the FFRs emerge from the plastic changes in the auditory cortex or the subcortex. It is also possible that the cortical component may play a substantial role in mediating subcortical plasticity. Future studies using methods to disambiguate the cortical and subcortical components of the FFR have the potential to contribute to the understanding of the mechanisms underlying subcortical neuroplasticity to speech.

The work reviewed thus far has largely examined the subcortical representation of repetitively presented single syllables produced by a single talker. However, speech signals rarely occur in isolation. In connected contexts, speech is highly co-articulated. Robust speech perception also requires that humans adjust their auditory processing to accommodate large inter-speaker and intra-speaker differences in speech production. There is now considerable behavioral evidence that the higher-order auditory system keeps track of long-term spectral statistics that influence the perception of the incoming speech stream (Holt and Lotto 2002; Holt 2005). Such contextual modulations are critical for the various normalization processes (e.g., talker normalization) operating during speech processing and likely reflect general auditory processes that are not specific to humans (Ladefoged and Broadbent 1957; Lotto et al. 1997). There is evidence that adaptation to the mean statistics of the input occurs in several subcortical nuclei, suggesting a putative basis for context effects in speech (Willmore et al. 2016). Non-invasive human neuroimaging studies have revealed that subcortical speech encoding is more robust in predictive and patterned contexts, relative to contexts in which the incoming speech sounds are less predictable (Chandrasekaran et al. 2009; Lau et al. 2017). Further repetition of the same speech sound has been shown to alter perception over time and is known as the verbal transformation effect (Warren 1961). This verbal transformation effect is also influenced by several external and internal factors. FFRs have also been shown to be affected by the verbal transformation effect (Galbraith et al. 1997). While stimulus repetition is vital for obtaining good signal-to-noise ratios, the extraneous effects associated with stimulus repetition often obscure the effects that researchers intend to observe.

Another limitation of stimulus repetition is that, due to the time taken for recording, only a handful of conditions can be tested in a single session, which may or may not generalize to speech perception in a real-world environment. Advances in recording and analysis techniques that have enabled researchers to study FFR responses using single trials (Yi et al. 2017) and natural speech (Forte et al. 2017) could be leveraged to obtain FFRs to various conditions, contexts, speakers, etc., circumventing these problems associated with stimulus repetition. Future studies can examine the neural mechanisms underlying online contextual influences that impact speech perception in various listening situations.

The subcortical auditory systems are not just important in encoding the lower-level features of the sound through the ascending and descending neural pathways, but are also involved in higher-level cognitive processes that shape auditory perception. The following section will focus on the role of the subcortical structures in facilitating the learning of novel speech categories.

2.5 Subcortical Systems in Learning Novel Speech Categories

So far in this chapter, the focus has been on the neural processing of native speech signals. There is considerable evidence that also shows that novel speech categories can be acquired in adulthood with training. What is the role of the subcortical systems in the acquisition of novel speech categories? In the cognitive neuroscience literature, at least three forms of learning have been identified. During *unsupervised learning*, representations are constructed solely on the basis of the statistical properties of the input. During *reinforcement learning*, representations are constructed by the learned output that maximizes the possibility of reward. The third form of learning is *supervised learning*, wherein representations are constructed on the basis of output, with the goal of reducing the input-output mapping error. The cerebral cortex, basal ganglia, and cerebellum have been viewed as key substrates for unsupervised, reinforcement, and supervised learning, respectively (Doya 2000). Thus, the complementary profiles of these neural structures may allow different forms of learning.

Neuroimaging and computational modeling approaches suggest that native speech categories are acquired via unsupervised learning, subserved by the auditory associative cortex (Vallabha et al. 2007; Feng et al. 2018). While unsupervised learning is the likely candidate for shaping neural representations to speech categories in infancy, such learning is less labile as humans age. Indeed, learning with at least some amount of reinforcement enhances non-native speech learning relative to unsupervised learning (Vallabha and McClelland 2007; Vallabha et al. 2007). Laboratory-based training paradigms have been successful in training adults to acquire even difficult-to-learn non-native speech categories (Jamieson and Morosan 1986; Lively et al. 1993). Such paradigms typically provide trial-by-trial feedback, with an eye towards error monitoring. Feedback could drive learning via supervision or reinforcement. Several lines of evidence suggest that reinforcement learning may be the optimal candidate driving speech category acquisition in adulthood. First, with respect to the information content in feedback, rich, informative feedback is less effective in enhancing learning relative to minimal feedback that only informs on correctness. Second, studies have employed video-game-based associative learning procedures wherein no explicit feedback is provided (Lim and Holt 2011). Such approaches have resulted in robust and rapid learning. Third, neuroimaging studies examining speech learning demonstrate significant activation of the caudate and putamen. This activation relates to individual differences in learning success (Feng et al. 2019; Lim et al. 2019). Figure 2.4 shows the differential brain activation patterns in native English listeners acquiring Mandarin tone categories while processing correct feedback versus incorrect feedback.

A dual-learning systems (DLS) model has been proposed to account for the different learning systems involved in auditory and speech categorization (Chandrasekaran et al. 2014a). In line with a proposal that speech perception can be viewed as a categorization process involving many-to-one mapping, the DLS model

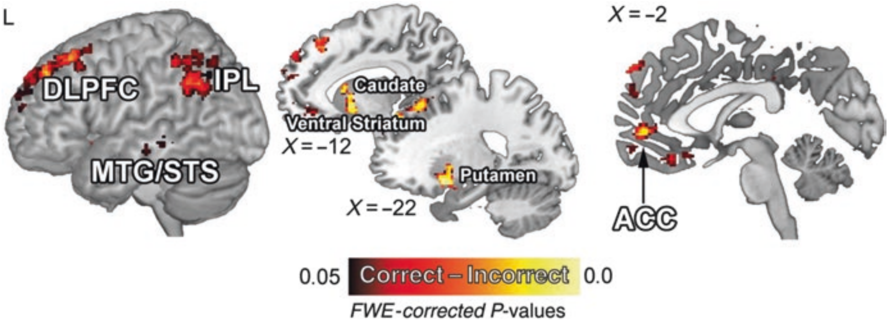

Fig. 2.4 Activation while processing correct versus incorrect feedback during a tone category learning task. During feedback processing, activation is observed in the anterior cingulate cortex (ACC), left caudate nucleus, bilateral putamen, ventral striatum, left dorsolateral prefrontal cortex (DLPFC), ventral striatum, left inferior parietal lobule (IPL), and left middle temporal gyrus/superior temporal sulcus (MTG/STS). (From Yi et al. (2014))

outlines the computational processes involved in learning to categorize with feedback. The DLS model is based on contemporary neurobiology and emphasizes two competing corticostriatal learning systems involved in feedback processing: a reflective learning system, where processing is under conscious control, and a reflexive learning system that is not under conscious control (Ashby and Ell 2001; Ashby and Maddox 2011).

The reflective system uses working memory and executive attention to develop and test rules based on feedback for explicit classification (DeCaro et al. 2008; Yi et al. 2016). Processing in this system is verbalizable, available to conscious awareness, and is mediated primarily by a circuit involving the prefrontal cortex, anterior cingulate, and the head of the caudate (Ashby and Ell 2001; Ashby and Ennis 2006). In contrast, the reflexive learning system is not consciously penetrable, is nonverbalizable, and operates by associating perception with actions that lead to reward (Nomura and Reber 2008; Chandrasekaran et al. 2015). Learning in the reflexive system is mediated primarily by the posterior caudate nucleus and the putamen (Nomura et al. 2007). Rather than relying on working memory processes to construct verbalizable rules, the reflexive system uses dopaminergic reward learning to associate regions within the stimulus space with a response (Ashby et al. 1998).

The DLS model predicts that speech categorization is reflexive-optimal due to the multi-dimensional nature of speech sounds that make generating optimal rules particularly challenging (Chandrasekaran et al. 2014a). The DLS model also predicts that early learning is subserved by the reflective system, and eventually control is released to the more optimal reflexive system. This model has been systematically tested using computational modeling, behavioral designs, and neuroimaging and can account for the large individual differences in speech category learning success (Smayda et al. 2015; Llanos et al. 2020).

2.6 Subcortical Speech Processing in Individuals with Clinical Disorders

While in previous studies, focus was placed on experience-dependent plasticity in subcortical speech representation, it is important to note that experiences do not always have a positive impact on subcortical function. In this section, studies that demonstrate subcortical *dysfunction* in individuals with neurological disorders and differences that impact communication will be discussed. Dysfunction in the subcortical representation of speech can be informative for clinical diagnoses, as specific response patterns can reflect different types of hearing loss, brainstem dysfunction, or auditory neuropathy. Individuals with various communication impairments have somewhat characteristic neural signatures as a result of impaired auditory function (Kraus and Anderson 2015; Kraus et al. 2017a). Among individuals with distinct clinical diagnoses, the FFR can show reduced and inefficient tracking for various speech features. Prior work has utilized subcortical metrics to assess differences in neural timing and response variability in various populations (Hornickel et al. 2012; Reetzke et al. 2017). Key findings from studies examining subcortical speech representation in individuals with various disorders that impact communication are highlighted below.

Developmental dyslexia is a common neurological disorder that impairs reading and spelling skills. While theories regarding core deficits abound in the literature, it is highly likely that developmental dyslexia is multifactorial (Bishop 2015). Several theories have focused on abnormalities in subcortical function, including the cerebellar deficit theory and the magnocellular deficit theory (Stein and Walsh 1997). At the core of the cerebellar deficit theory is the finding that cerebellar activation is abnormal in individuals with dyslexia who are performing sequencing tasks and that some individuals with developmental dyslexia demonstrate deficits on cerebellar tests (Nicolson et al. 2001). On the other hand, the magnocellular theory posits a selective morphological deficit that impacts the magnocellular layers within the lateral geniculate nucleus. This impact, in turn, affects the sensory processing of fast, temporal information. Originally developed to account for visual processing deficits, the magnocellular theory has been extended to speech processing.

Neuroimaging studies have found abnormal neural activation in the MGB during the processing of phonemes, but this has not been seen for other speech dimensions (Díaz et al. 2012). These results are consistent with an emerging view that the thalamus, mediated by thalamocortical ascending and corticothalamic descending pathways, plays an active role as a sensory gatekeeper and that thalamic dysfunction can lead to sensory processing deficits. The reasons for the phonological deficits are much debated, and this debate is reflected in the literature on subcortical function. One possibility is that speech representations are intact, but access to the representations is dysfunctional (Boets et al. 2013). Another possibility is that disruption in bottom-up auditory processing may lead to “fuzzy” speech representations.

Individuals with dyslexia have been found to have impaired responses to harmonics, formant frequency timing, and onset timing as well as less stable neural responses to speech signals (Banai et al. 2009; Kraus and Nicol 2014). There is also evidence that these individuals may have impaired context-dependent modulation of speech signals (Chandrasekaran et al. 2009). In line with work showing poorer speech perception in challenging listening environments, individuals with dyslexia also exhibit a larger neural delay in subcortical responses to speech sounds (Anderson et al. 2010; White-Schwoch et al. 2015). Children with impaired reading ability have been found to have more variable responses to speech sounds than children with typical reading ability, particularly for formant transitions (Hornickel and Kraus 2013). Studies have also shown that children with dyslexia have difficulty both differentiating and identifying stop consonants, perhaps due to the more quickly changing formant transitions relative to vowel sounds (Hornickel et al. 2009).

While the FFRs of individuals with dyslexia differ on several speech-relevant features, other characteristics of their responses do not differ from individuals without dyslexia (Kraus and Nicol 2014). For instance, the subcortical representation of the fundamental frequency does not differ between individuals with and without dyslexia (Banai et al. 2009). A key finding is that for non-speech stimuli, such as click-evoked ABRs, children with dyslexia have similar responses to children without dyslexia, lending further support to the notion that the observed deficits in dyslexia are related to challenges underlying speech-specific subcortical processing (Billiet and Bellis 2011; Kumar and Singh 2015). It is difficult to reach a singular conclusion about the nature of the subcortical deficit from these diverse findings. Indeed, dysfunctional subcortical speech processing seen in developmental dyslexia appears to fit several different theories. Dynamic elements in the signal appear to be particularly challenging to encode, in line with theories that argue for a bottom-up, fast-temporal processing deficit (Banai et al. 2009). On the other hand, in line with a noise-exclusion deficit in developmental dyslexia, speech representation in noisy conditions indexed by the FFR in preschoolers is highly predictive of emerging literacy (Sperling et al. 2005; White-Schwoch et al. 2015). Thus, while the FFR may be a sensitive marker of subcortical dysfunction, it is unlikely to provide information about the causative mechanism underlying dyslexia.

Relative to dyslexia, "auditory processing disorder," a term encompassing a group of disorders that impact auditory information processing, is a more nebulous diagnosis (Moore 2006). While the exact underlying cause(s) of the manifestation of auditory processing disorder is/are still unknown, underlying brainstem dysfunction is suspected (Medwetsky 2011). Children with auditory processing disorder have both temporal auditory processing deficits and abnormal timing in their FFR, including difficulty with harmonics, formant frequency timing, and onset timing (Rocha-Muniz et al. 2012; Kraus and Nicol 2014). A study of auditory steady-state responses in individuals with auditory processing disorders found that phase coherence was worse in individuals with disproportionately low speech understanding scores (Ali and Jerger 1992). Simões (2009) also looked at auditory steady-state responses and found significantly increased thresholds in children with auditory processing disorder, possibly reflecting underlying deficits in temporal processing as a result of poor phase-locking. Rocha-Muniz et al. (2016) found that 85% of children with abnormal FFRs showed deficits on tests for auditory processing.

However, despite these differences, individuals with auditory processing disorder have been shown to process simple acoustic information, such as clicks, similar to typically developing individuals (Filippini and Schochat 2009). The FFR deficits in individuals with auditory processing disorder may be linked to dysfunction in effectively connecting sound to meaning, a key skill for language learning (Hornickel and Kraus 2013). It may also be the case that poor synchrony of neural firing leads to poorer representations of speech sounds (Schochat et al. 2017).

Prior work has also examined subcortical speech processing in individuals with autism spectrum disorder. Relative to age-matched neurotypicals, individuals with autism spectrum disorder have been found to have impaired pitch tracking and reduced responses for onset timing (Russo et al. 2009; Kraus and Nicol 2014). The ability to track changes in fundamental frequency over time relates to the ability to perceive pitch changes over time, an important cue for processing communicative intent (Kraus and Anderson 2015). Impaired pitch tracking evidenced by the FFR may be one reason why individuals with autism spectrum disorder have difficulty understanding emotional content and intention, such as sarcasm, in speech (Hornickel and Kraus 2013). High-functioning children with autism spectrum disorder have less stable FFRs relative to neurotypical children to a variety of speech stimuli, including clicks, stop consonants, and glides with rising and falling pitch (Otto-Meyer et al. 2018). However, baseline noise levels do not differ between neurotypical children and those on the autism spectrum. This suggests an encoding difference that manifests in the inability to reliably differentiate signal from noise in the brain (Baron-Cohen and Belmonte 2005).

Subcortical function has been assessed in individuals across the lifespan in both typical and atypical aging. Older adults have been found to have delays across nearly all FFR metrics (Anderson et al. 2012). With advanced age, impairments in tracking pitch, harmonics, onset timing, and formant frequency are seen (Kraus and Nicol 2014). Older adults also have less consistent responses, greater neural noise in their FFRs, and exhibit delayed neural timing to aspects of speech sounds, suggesting deficits in central processing (Vander Werff and Burns 2011; Anderson et al. 2012). Age-related deficits in temporal precision are another explanation posited for older adults' auditory processing deficits that result in suboptimal speech perception (Anderson 2017). Presacco et al. (2015) compared FFRs to /a/ and /da/ and found older adults showed a reduction in sustained phase-locking and, unlike younger adults, did not show differences in peak latencies. Anderson et al. (2011) also examined FFRs to /da/ and found smaller amplitudes and reduced phase-locking for transitions and steady-state regions in both temporal and frequency domains in older adults. Response amplitudes to speech syllables, particularly onset and offset regions, are also impacted by aging (Vander Werff and Burns 2011; Clinard and Tremblay 2013).

Older adults with normal hearing may have difficulty understanding speech in noisy and reverberant conditions, which may be due to degraded neural representations of acoustic signals, decreased neural inhibition, and temporal jitter (Pichora-Fuller et al. 2007; Fujihira et al. 2017). Fujihira et al. (2017) found that older adults listening to /da/ showed decreased amplitudes for both the fundamental frequency and the first formant frequency in reverberant conditions. Behavioral performance on speech recognition tasks also decreased in reverberant conditions. A study by Anderson et al. (2011) found that the FFRs of older adults with low speech perception in noise performance had smaller response amplitudes to the fundamental frequency and lower correlations between responses in quiet and noisy conditions relative to high-performing older adults. Reduced context-dependent modulation of the FFR has also been shown in older adults with normal hearing (Maruthy et al. 2017). Older adults and younger adults differed in the relationship between contextual modulation of the FFR and speech perception in noise. Taken together, these findings suggest that older adults' difficulty understanding speech in challenging listening environments may be at least partially driven by poorer subcortical encoding of critical speech features.

The impact of hearing loss on subcortical speech processing in older adults is also well studied. Hearing loss can have various effects on the FFR, including degraded representation of the temporal fine structure and the envelope for lower frequency steady stimuli (Anderson 2017). At the same time, individuals with hearing loss may exhibit enhanced envelope encoding and intact encoding of higher frequencies of the temporal fine structure (Anderson et al. 2013). In animal models, consistent findings are shown in chinchillas with noise-induced hearing loss that exhibit amplified envelope coding at higher frequencies (Zhong et al. 2014). Imbalance in the representation of the envelope relative to temporal fine structure in older adults with hearing loss may distract from salient cues, which could underlie difficulty understanding speech-in-noise (Anderson et al. 2013). Older adults with hearing loss may also have a disturbed balance between envelope representation and fine structure representation relative to individuals with normal hearing, which may be the key to their underlying difficulty understanding speech in noise.

Poor encoding of both fundamental frequency and higher frequencies in FFRs have also been found in individuals with hearing loss (Plyler and Ananthanarayan 2001; Ananthakrishnan et al. 2016). These differences have been linked to differential manifestations of hearing loss on auditory frequency selectivity and the frequency and intensity characteristics of the stimuli used. Further, changes in sound intensity neither improved intelligibility for individuals with hearing loss nor affected the magnitude of the first formant in the FFRs (Ananthakrishnan et al. 2016). This may be linked to altered basilar membrane input-output functions as a result of broadened auditory filters. The manifestation of hearing loss on the FFRs supports the notion that perceptual deficits for temporal fine structure cues associated with hearing loss contribute to deficits in speech perception (Lorenzi et al. 2009; King et al. 2014). Older adults can show enhanced encoding of the envelope and more sensitive thresholds for gap detection (Füllgrabe et al. 2003; Horwitz et al. 2011). The downside of this enhancement is a possible reduction in the salience of temporal fine structure cues important for hearing in noise (Ananthakrishnan et al. 2016). These findings suggest a disruption in the balance of coding the envelope and temporal fine structure in individuals with hearing loss.

Subcortical encoding of speech information has also been studied in individuals with neurodegeneration and brain trauma. Mild cognitive impairment, a cognitive transition phase that can progress to more severe forms of dementia or follow the progression of typical cognitive aging, is associated with the dysfunctional encoding of speech signals (Bidelman et al. 2017). Bidelman et al. (2017) recorded FFRs and cortical onset responses in older adults with and without mild cognitive impairment as they listened to a 5-step vowel continuum from /u/ to /a/. Individuals with mild cognitive impairment had hypersensitive encoding at both brainstem and cortical levels, perhaps due to an exacerbation of reduced neural inhibition found in healthy aging. Interestingly, brainstem responses were a better predictor of the severity of cognitive impairment relative to cortical activity. Although there were neural differences, behavioral differences between older adults with and without mild cognitive impairment were not found, suggesting that neurophysiological changes in the auditory system may precede deficits in behavior and communication skills (Johnson and Lin 2014; Bidelman et al. 2017).

Head trauma can lead to pervasive neurological damage (Kraus et al. 2017b). Kraus and colleagues found that individuals who had experienced a concussion had smaller fundamental frequency responses to speech stimuli, suggesting that brain injuries can result in lasting damage to the midbrain and affect the fine granular subcortical processing of sounds. Fundamental frequency is a key cue for identifying sounds and assists sound processing in complex communication environments, such as a noisy restaurant, that can be challenging for individuals with brain injuries (Gallun et al. 2012). These patterns are also found in children who have sustained a concussion, suggesting that concussions leave lasting traces on auditory processing (Kraus et al. 2016, 2017b). Taken together, neural signatures reflected in FFRs have been shown to be sensitive to differences in subcortical speech processing in individuals with a variety of disorders and differences affecting communication.

2.7 Future Directions

While FFRs have provided many insights into the subcortical processing of speech, future work can address several important questions. One such area that requires further exploration is the nature and extent of cortical contributions to the FFR. Better characterization of the cortical contributions to the FFR may further elucidate both the role of subcortical structures in speech processing and important subcortical-cortical connections. Though advances have been made in regard to the number of trials required for stable FFR recordings, the development of more fine-tuned techniques to record FFRs with a small number of trials could enable researchers to obtain fast estimates of FFRs across different conditions and different stimuli in a single, short recording session. Another key future direction for understanding subcortical speech processing is assessing FFRs to ecologically valid, naturalistic speech stimuli. By developing tools to achieve this, researchers may better understand the subcortical contributions to speech processing in real-world communicative environments.

The development of normative databases for FFR metrics across the life span from a multi-lab study could provide important insight into subcortical speech processing and how it may change over time. This should emerge from a consensus on recording protocols and stimuli which can be easily applied across different populations. Relatedly, more work on the use of portable EEG systems to record FFRs in different settings with different populations should be explored, evaluating validity, consistency, and diagnostic suitability. Future work should also leverage FFRs to evaluate individual differences in speech processing across a wide range of populations. By systematically evaluating the influences of demographic variables such as age, gender, race, ethnicity, socio-economic background, as well as communication differences and reading abilities, researchers could not only identify possible neural signatures of these variables; they could incorporate appropriate corrections in normative databases. Beyond the FFR, advances in non-invasive neuroimaging that maintain the high temporal precision of EEG but allow greater spatial precision at deeper brain levels beyond current MEG capabilities could reveal further insights into subcortical activity during speech processing.

2.8 Summary

Contemporary understanding of the role of the subcortical system in speech processing is undergoing a massive revision. Both animal and human models have contributed to the understanding of the subcortical neural schema involved in processing critical information in speech. Subcortical auditory processing of speech information is not hard-wired and can continue to change throughout the life span as a function of positive and negative experiences. Such plasticity reflects integrative processing between the cortex and the subcortex, subserved by ascending as well as descending feedback loops. Subcortical systems, including the basal ganglia and cerebellum, are also involved in learning novel speech categories. At this juncture, there is a critical need to use a systems neuroscience approach to go beyond the traditional characterization of the subcortex as “lower” sensory/perceptual structures. Understanding subcortical function within a larger cortical-subcortical circuit is critical for a holistic understanding of the neurobiology of speech perception. Such an approach will also allow the interpretation and mechanistic characterization of subcortical dysfunction in clinical disorders that affect learning and communication to be better understood.

Acknowledgments The authors wish to acknowledge the support of the National Institute on Deafness and Other Communication Disorders of the National Institutes of Health under Award Numbers R01DC015504 (BC) and R01DC013315 (BC). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Compliance with Ethics Requirements Bharath Chandrasekaran declares that he has no conflict of interest.

Rachel Tessmer declares that she has no conflict of interest.
G. Nike Gnanateja declares that he has no conflict of interest.

References

Alexander GE, DeLong MR, Strick PL (1986) Parallel organization of functionally segregated circuits linking basal ganglia and cortex. Annu Rev Neurosci 9(1):357–381

Ali AA, Jerger J (1992) Phase coherence of the middle-latency response in the elderly. Scand Audiol 21(3):187–194

Ananthakrishnan S, Krishnan A, Bartlett E (2016) Human frequency-following response: neural representation of envelope and temporal fine structure in listeners with normal hearing and sensorineural hearing loss. Ear Hear 37(2):e91–e103

Anderson S (2017) Clinical translation: aging, hearing loss, and amplification. In: Kraus N, Anderson S, White-Schwoch T et al (eds) The frequency-following response. Springer handbook of auditory research, vol 6. Springer, Cham, pp 267–294

Anderson S, Skoe E, Chandrasekaran B et al (2010) Brainstem correlates of speech-in-noise perception in children. Hear Res 270(1–2):151–157

Anderson S, Parbery-Clark A, Yi HG et al (2011) A neural basis of speech-in-noise perception in older adults. Ear Hear 32(6):750–757

Anderson S, Parbery-Clark A, White-Schwoch T et al (2012) Aging affects neural precision of speech encoding. J Neurosci 32(41):14156–14164

Anderson S, Parbery-Clark A, White-Schwoch T et al (2013) Effects of hearing loss on the subcortical representation of speech cues. J Acoust Soc Am 133(5):3030–3038

Ashby FG, Ell SW (2001) The neurobiology of human category learning. Trends Cogn Sci 5(5):204–210

Ashby FG, Ennis JM (2006) The role of the basal ganglia in category learning. Psychol Learn Motiv 46:1–36

Ashby FG, Maddox WT (2011) Human category learning 2.0. Ann N Y Acad Sci 1224:147–161

Ashby FG, Alfonso-Reese LA, Waldron EM (1998) A neuropsychological theory of multiple systems in category learning. Psychol Rev 105(3):442–481

Ayala YA, Lehmann A, Merchant H (2017) Monkeys share the neurophysiological basis for encoding sound periodicities captured by the frequency-following response with humans. Sci Rep 7(1):1–11

Banai K, Hornickel J, Skoe E et al (2009) Reading and subcortical auditory function. Cereb Cortex 19(11):2699–2707

Baron-Cohen S, Belmonte MK (2005) Autism: a window onto the development of the social and the analytic brain. Annu Rev Neurosci 28:109–126

Bartlett EL (2013) The organization and physiology of the auditory thalamus and its role in processing acoustic features important for speech perception. Brain Lang 126(1):29–48

Baumann S, Griffiths TD, Sun L et al (2011) Orthogonal representation of sound dimensions in the primate midbrain. Nat Neurosci 14(4):423–425

Bidelman GM (2015) Multichannel recordings of the human brainstem frequency-following response: scalp topography, source generators, and distinctions from the transient ABR. Hear Res 323:68–80

Bidelman GM (2018) Subcortical sources dominate the neuroelectric auditory frequency-following response to speech. NeuroImage 175:56–69

Bidelman GM, Powers L (2018) Response properties of the human frequency-following response (FFR) to speech and non-speech sounds: level dependence, adaptation and phase-locking limits. Int J Audiol 57(9):665–672

Bidelman GM, Lowther JE, Tak SH et al (2017) Mild cognitive impairment is characterized by deficient hierarchical speech coding between auditory brainstem and cortex. J Neurosci 37(13):3610–3620

Billiet CR, Bellis TJ (2011) The relationship between brainstem temporal processing and performance on tests of central auditory function in children with reading disorders. J Speech Lang Hear Res 54(1):228–242

Bishop DV (2015) The interface between genetics and psychology: lessons from developmental dyslexia. Proc Biol Sci 282:1–8

Boets B, de Beeck HPO, Vandermosten M et al (2013) Intact but less accessible phonetic representations in adults with dyslexia. Science 342(6163):1251–1254

Bostan AC, Dum RP, Strick PL (2010) The basal ganglia communicate with the cerebellum. Proc Natl Acad Sci U S A 107(18):8452–8456

Chandrasekaran B, Kraus N (2010) The scalp-recorded brainstem response to speech: neural origins and plasticity. Psychophysiology 47(2):236–246

Chandrasekaran B, Hornickel J, Skoe E et al (2009) Context-dependent encoding in the human auditory brainstem relates to hearing speech in noise: implications for developmental dyslexia. Neuron 64(3):311–319

Chandrasekaran B, Kraus N, Wong PC (2012) Human inferior colliculus activity relates to individual differences in spoken language learning. J Neurophysiol 107(5):1325–1336

Chandrasekaran B, Koslov SR, Maddox WT (2014a) Toward a dual-learning systems model of speech category learning. Front Psychol 5:1–17

Chandrasekaran B, Skoe E, Kraus N (2014b) An integrative model of subcortical auditory plasticity. Brain Topogr 27(4):539–552

Chandrasekaran B, Yi HG, Blanco NJ, McGeary JE, Maddox WT (2015) Enhanced procedural learning of speech sound categories in a genetic variant of FOXP2. J Neurosci 35(20):7808–7812

Chechik G, Anderson MJ, Bar-Yosef O et al (2006) Reduction of information redundancy in the ascending auditory pathway. Neuron 51(3):359–368

Clinard CG, Tremblay KL (2013) Aging degrades the neural encoding of simple and complex sounds in the human brainstem. J Am Acad Audiol 24(7):590–599

Coffey EB, Herholz SC, Chepesiuk AM et al (2016) Cortical contributions to the auditory frequency-following response revealed by MEG. Nat Commun 7(1):1–11

Coffey EB, Musacchia G, Zatorre RJ (2017) Cortical correlates of the auditory frequency-following and onset responses: EEG and fMRI evidence. J Neurosci 37(4):830–838

Coffey EB, Nicol T, White-Schwoch T et al (2019) Evolving perspectives on the sources of the frequency-following response. Nat Commun 10(1):1–10

Crosson B, Ford A, McGregor KM et al (2010) Functional imaging and related techniques: an introduction for rehabilitation researchers. J Rehabil Res Dev 47(2):7–34

Cuffin BN, Cohen D (1979) Comparison of the magnetoencephalogram and electroencephalogram. Electroencephalogr Clin Neurophysiol 47(2):132–146

David SV, Mesgarani N, Shamma SA (2007) Estimating sparse spectro-temporal receptive fields with natural stimuli. Network 18(3):191–212

Davis MH, Johnsrude IS (2003) Hierarchical processing in spoken language comprehension. J Neurosci 23(8):3423–3431

Decaro MS, Thomas RD, Beilock SL (2008) Individual differences in category learning: sometimes less working memory capacity is better than more. Cognition 107(1):284–294

Delgutte B (1990) Two-tone rate suppression in auditory-nerve fibers: dependence on suppressor frequency and level. Hear Res 49(1–3):225–246

Delgutte B (1997) Auditory neural processing of speech. In: Hardcastle WJ, Laver J, Gibbon FE (eds) The handbook of phonetic sciences. Blackwell, Oxford, pp 507–538

Delgutte B, Kiang NY (1984a) Speech coding in the auditory nerve: I. Vowel-like sounds. J Acoust Soc Am 75(3):866–878

Delgutte B, Kiang NY (1984b) Speech coding in the auditory nerve: IV. Sounds with consonant-like dynamic characteristics. J Acoust Soc Am 75(3):897–907

Díaz B, Hintz F, Kiebel SJ et al (2012) Dysfunction of the auditory thalamus in developmental dyslexia. Proc Natl Acad Sci U S A 109(34):13841–13846

Doya K (2000) Complementary roles of basal ganglia and cerebellum in learning and motor control. Curr Opin Neurobiol 10(6):732–739

Engineer CT, Perez CA, Chen YH et al (2008) Cortical activity patterns predict speech discrimination ability. Nat Neurosci 11(5):603–608

Evans EF (1981) The dynamic range problem: place and time coding at the level of cochlear nerve and nucleus. In: Syka J, Aitkin L (eds) Neuronal mechanisms of hearing. Springer, Boston, pp 69–85

Feng G, Gan Z, Wang S et al (2018) Task-general and acoustic-invariant neural representation of speech categories in the human brain. Cereb Cortex 28(9):3241–3254

Feng G, Yi HG, Chandrasekaran B (2019) The role of the human auditory corticostriatal network in speech learning. Cereb Cortex 29(10):4077–4089

Filippini R, Schochat E (2009) Brainstem evoked auditory potentials with speech stimulus in the auditory processing disorder. Braz J Otorhinolaryngol 75(3):449–455

Forte AE, Etard O, Reichenbach T (2017) The human auditory brainstem response to running speech reveals a subcortical mechanism for selective attention. elife 6:e27203

Fritz J, Shamma S, Elhilali M et al (2003) Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex. Nat Neurosci 6(11):1216–1223

Fujihira H, Shiraishi K, Remijn GB (2017) Elderly listeners with low intelligibility scores under reverberation show degraded subcortical representation of reverberant speech. Neurosci Lett 637:102–107

Füllgrabe C, Meyer B, Lorenzi C (2003) Effect of cochlear damage on the detection of complex temporal envelopes. Hear Res 178(1–2):35–43

Galbraith GC, Jhaveri SP, Kuo J (1997) Speech-evoked brainstem frequency-following responses during verbal transformations due to word repetition. Electroencephalogr Clin Neurophysiol 102(1):46–53

Gallun FJ, Diedesch AC, Kubli LR et al (2012) Performance on tests of central auditory processing by individuals exposed to high-intensity blasts. J Rehabil Res Dev 49(7):1005–1025

Gandour J (1983) Tone perception in Far Eastern languages. J Phon 11(2):149–175

Gao E, Suga N (2000) Experience-dependent plasticity in the auditory cortex and the inferior colliculus of bats: role of the corticofugal system. Proc Natl Acad Sci U S A 97(14):8081–8086

Glover GH, Li TQ, Ress D (2000) Image-based method for retrospective correction of physiological motion effects in fMRI: RETROICOR. Magn Reson Med 44(1):162–167

Guimaraes AR, Melcher JR, Talavage TM et al (1998) Imaging subcortical auditory activity in humans. Hum Brain Mapp 6(1):33–41

Hämäläinen M, Hari R, Ilmoniemi RJ et al (1993) Magnetoencephalography-theory, instrumentation, and applications to noninvasive studies of the working human brain. Rev Mod Phys 65(2):413–497

Hecox K, Galambos R (1974) Brain stem auditory evoked responses in human infants and adults. Arch Otolaryngol 99(1):30–33

Hélie S, Ell SW, Ashby FG (2015) Learning robust cortico-cortical associations with the basal ganglia: an integrative review. Cortex 64:123–135

Hillenbrand J, Getty LA, Clark MJ et al (1995) Acoustic characteristics of American English vowels. J Acoust Soc Am 97(5):3099–3111

Holt LL (2005) Temporally nonadjacent nonlinguistic sounds affect speech categorization. Psychol Sci 16(4):305–312

Holt LL, Lotto AJ (2002) Behavioral examinations of the level of auditory processing of speech context effects. Hear Res 167(1–2):156–169

Hornickel J, Kraus N (2013) Unstable representation of sound: a biological marker of dyslexia. J Neurosci 33(8):3500–3504

Hornickel J, Skoe E, Nicol T et al (2009) Subcortical differentiation of stop consonants relates to reading and speech-in-noise perception. Proc Natl Acad Sci U S A 106(31):13022–13027

Hornickel J, Knowles E, Kraus N (2012) Test-retest consistency of speech-evoked auditory brainstem responses in typically-developing children. Hear Res 284(1–2):52–58

Horwitz AR, Ahlstrom JB, Dubno JR (2011) Level-dependent changes in detection of temporal gaps in noise markers by adults with normal and impaired hearing. J Acoust Soc Am 130(5):2928–2938

Hoshi E, Tremblay L, Féger J et al (2005) The cerebellum communicates with the basal ganglia. Nat Neurosci 8(11):1491–1493

Jamieson DG, Morosan DE (1986) Training non-native speech contrasts in adults: acquisition of the English /ð/–/θ/ contrast by francophones. Percept Psychophys 40(4):205–215

Jeng FC, Hu J, Dickman B et al (2011) Cross-linguistic comparison of frequency-following responses to voice pitch in American and Chinese neonates and adults. Ear Hear 32(6):699–707

Johnson M, Lin F (2014) Communication difficulty and relevant interventions in mild cognitive impairment: implications for neuroplasticity. Top Geriatr Rehabil 30(1):18–34

Jueptner M, Frith CD, Brooks DJ et al (1997) Anatomy of motor learning. II. Subcortical structures and learning by trial and error. J Neurophysiol 77(3):1325–1337

Kim DO, Rhode WS, Greenberg SR (1986) Responses of cochlear nucleus neurons to speech signals: neural encoding of pitch, intensity and other parameters. In: Moore BCJ, Patterson RD (eds) Auditory frequency selectivity. Nato ASI series, vol 119. Springer, Boston, pp 281–288

King A, Hopkins K, Plack CJ (2014) The effects of age and hearing loss on interaural phase difference discrimination. J Acoust Soc Am 135(1):342–351

Kral A, Eggermont JJ (2007) What's to lose and what's to learn: development under auditory deprivation, cochlear implants and limits of cortical plasticity. Brain Res Rev 56(1):259–269

Kraus N, Anderson S (2015) Low socioeconomic status linked to impaired auditory processing. Hear J 68(5):38–40

Kraus N, Nicol T (2014) The cognitive auditory system: the role of learning in shaping the biology of the auditory system. In: Popper A, Fay R (eds) Perspectives on auditory research. Springer handbook of auditory research, vol 50. Springer, New York, pp 299–319

Kraus N, White-Schwoch T (2015) Unraveling the biology of auditory learning: a cognitive–sensorimotor–reward framework. Trends Cogn Sci 19(11):642–654

Kraus N, Thompson EC, Krizman J et al (2016) Auditory biological marker of concussion in children. Sci Rep 6:1–10

Kraus N, Anderson S, White-Schwoch T (2017a) The frequency-following response: a window into human communication. In: Kraus N, Anderson S, White-Schwoch T et al (eds) The frequency-following response. Springer handbook of auditory research, vol 61. Springer, Cham, pp 1–15

Kraus N, Lindley T, Colegrove D et al (2017b) The neural legacy of a single concussion. Neurosci Lett 646:21–23

Krishnan A (2002) Human frequency-following responses: representation of steady-state synthetic vowels. Hear Res 166(1–2):192–201

Krishnan A, Gandour JT (2009) The role of the auditory brainstem in processing linguistically-relevant pitch patterns. Brain Lang 110(3):135–148

Krishnan A, Xu Y, Gandour JT et al (2004) Human frequency-following response: representation of pitch contours in Chinese tones. Hear Res 189(1–2):1–12

Krishnan A, Xu Y, Gandour JT et al (2005) Encoding of pitch in the human brainstem is sensitive to language experience. Cognitive Brain Res 25(1):161–168

Krishnan A, Gandour JT, Bidelman GM (2010) The effects of tone language experience on pitch processing in the brainstem. J Neurolinguistics 23(1):81–95

Krizman J, Marian V, Shook A et al (2012) Subcortical encoding of sound is enhanced in bilinguals and relates to executive function advantages. Proc Natl Acad Sci U S A 109(20):7877–7881

Kuhl PK (1981) Discrimination of speech by nonhuman animals: basic auditory sensitivities conducive to the perception of speech-sound categories. J Acoust Soc Am 70(2):340–349

Kuhl PK, Miller JD (1978) Speech perception by the chinchilla: identification functions for synthetic VOT stimuli. J Acoust Soc Am 63(3):905–917

Kumar P, Singh NK (2015) BioMARK as electrophysiological tool for assessing children at risk for (central) auditory processing disorders without reading deficits. Hear Res 324:54–58

Ladefoged P, Broadbent DE (1957) Information conveyed by vowels. J Acoust Soc Am 29(1):98–104

Lau JC, Wong PC, Chandrasekaran B (2017) Context-dependent plasticity in the subcortical encoding of linguistic pitch patterns. J Neurophysiol 117(2):594–603

Lim SJ, Holt LL (2011) Learning foreign sounds in an alien world: videogame training improves non-native speech categorization. Cogn Sci 35(7):1390–1405

Lim SJ, Fiez JA, Holt LL (2014) How may the basal ganglia contribute to auditory categorization and speech perception? Front Neurosci 8:230

Lim SJ, Fiez JA, Holt LL (2019) Role of the striatum in incidental learning of sound categories. Proc Natl Acad Sci U S A 116(110):4671–4680

Linden JF, Schreiner CE (2003) Columnar transformations in auditory cortex? A comparison to visual and somatosensory cortices. Cereb Cortex 13(1):83–89

Lively SE, Logan JS, Pisoni DB (1993) Training Japanese listeners to identify English /r/ and /l/ II: the role of phonetic environment and talker variability in learning new perceptual categories. J Acoust Soc Am 94(3):1242–1255

Llanos F, McHaney JR, Schuerman WL, Han GY, Leonard MK, Chandrasekaran B (2020) Non-invasive peripheral nerve stimulation selectively enhances speech category learning in adults. NPJ Sci Learn 5(1):1–11

Lorenzi C, Debruille L, Garnier S et al (2009) Abnormal processing of temporal fine structure in speech for frequencies where absolute thresholds are normal. J Acoust Soc Am 125(1):27–30

Lotto AJ, Kluender KR, Holt LL (1997) Perceptual compensation for coarticulation by Japanese quail (*Coturnix coturnix japonica*). J Acoust Soc Am 102(2):1134–1140

Maddox RK, Lee AKC (2018) Auditory brainstem responses to continuous natural speech in human listeners. eNeuro 5(1):1–13

Marsh JT, Worden FG, Smith JC (1970) Auditory frequency-following response: neural or artifact? Science 169(3951):1222–1223

Maruthy S, Kumar UA, Gnanateja GN (2017) Functional interplay between the putative measures of rostral and caudal efferent regulation of speech perception in noise. J Assoc Res Otolaryngol 18(4):635–648

Medwetsky L (2011) Spoken language processing model: bridging auditory and language processing to guide assessment and intervention. Lang Speech Hear Serv Sch 42(3):286–296

Middleton FA, Strick PL (1996) The temporal lobe is a target of output from the basal ganglia. Proc Natl Acad Sci U S A 93(16):8683–8687

Moore DR (2006) Auditory processing disorder (APD): definition, diagnosis, neural basis, and intervention. Audiol Med 4(1):4–11

Murdoch BE, Whelan BM (2009) Speech and language disorders associated with subcortical pathology. Wiley & Sons Ltd, West Sussex

Nelken I (2008) Processing of complex sounds in the auditory system. Curr Opin Neurobiol 18(4):413–417

Nicolson RI, Fawcett AJ, Dean P (2001) Developmental dyslexia: the cerebellar deficit hypothesis. Trends Neurosci 24(9):508–511

Nomura EM, Reber PJ (2008) A review of medial temporal lobe and caudate contributions to visual category learning. Neurosci Biobehav Rev 32(2):279–291

Nomura EM, Maddox WT, Filoteo JV et al (2007) Neural correlates of rule-based and information-integration visual category learning. Cereb Cortex 17(1):37–43

Nourski KV (2017) Auditory processing in the human cortex: an intracranial electrophysiology perspective. Laryngoscope Investig Otolaryngol 2(4):147–156

Otto-Meyer S, Krizman J, White-Schwoch T et al (2018) Children with autism spectrum disorder have unstable neural responses to sound. Exp Brain Res 236(3):733–743

Parvizi J (2009) Corticocentric myopia: old bias in new cognitive sciences. Trends Cogn Sci 13(8):354–359

Pasley BN, David SV, Mesgarani N et al (2012) Reconstructing speech from human auditory cortex. PLoS Biol 10(1):1–13

Pichora-Fuller MK, Schneider BA, MacDonald E et al (2007) Temporal jitter disrupts speech intelligibility: a simulation of auditory aging. Hear Res 223(1–2):114–121

Plyler PN, Ananthanarayan AK (2001) Human frequency-following responses: representation of second formant transitions in normal-hearing and hearing-impaired listeners. J Am Acad Audiol 12(10):523–533

Portfors CV, Roberts PD, Jonson K (2009) Over-representation of species-specific vocalizations in the awake mouse inferior colliculus. Neuroscience 162(2):486–500

Presacco A, Jenkins K, Lieberman R et al (2015) Effects of aging on the encoding of dynamic and static components of speech. Ear Hear 36(6):e352–e363

Rakic P (2009) Evolution of the neocortex: a perspective from developmental biology. Nat Rev Neurosci 10(10):724–735

Ranasinghe KG, Vrana WA, Matney CJ et al (2012) Neural mechanisms supporting robust discrimination of spectrally and temporally degraded speech. J Assoc Res Otolaryngol 13(4):527–542

Ranasinghe KG, Vrana WA, Matney CJ et al (2013) Increasing diversity of neural responses to speech sounds across the central auditory pathway. Neuroscience 252:80–97

Rasmussen GL (1946) The olivary peduncle and other fiber projections of the superior olivary complex. J Comp Neurol 84(2):141–219

Reetzke R, Xie Z, Chandrasekaran B (2017) Neurobiology of literacy and reading disorders. In: Kraus N, Anderson S, White-Schwoch T et al (eds) The frequency-following response. Springer handbook of auditory research, vol 6. Springer, Cham, pp 251–266

Reetzke R, Xie Z, Llanos F et al (2018) Tracing the trajectory of sensory plasticity across different stages of speech learning in adulthood. Curr Biol 28(9):1419–1427

Ress D, Chandrasekaran B (2013) Tonotopic organization in the depth of human inferior colliculus. Front Hum Neurosci 7:1–10

Rocha-Muniz CN, Befi-Lopes DM, Schochat E (2012) Investigation of auditory processing disorder and language impairment using the speech-evoked auditory brainstem response. Hear Res 294(1–2):143–152

Rocha-Muniz CN, Filippini R, Neves-Lobo IF et al (2016) Can speech-evoked Auditory Brainstem Response become a useful tool in clinical practice? CoDAS 28(1):77–80

Romanski LM, Averbeck BB (2009) The primate cortical auditory system and neural representation of conspecific vocalizations. Annu Rev Neurosci 32:315–346

Russo N, Nicol T, Trommer B et al (2009) Brainstem transcription of speech is disrupted in children with autism spectrum disorders. Dev Sci 12(4):557–567

Schochat E, Rocha-Muniz CN, Filippini R (2017) Understanding auditory processing disorder through the FFR. In: Kraus N, Anderson S, White-Schwoch T et al (eds) The frequency-following response. Springer handbook of auditory research, vol 6. Springer, Cham, pp 225–250

Schreiner CE, Langner G (1997) Laminar fine structure of frequency organization in auditory midbrain. Nature 388(6640):383–386

Seger CA (2006) The basal ganglia in human learning. Neuroscientist 12(4):285–290

Simões MB (2009) Auditory steady state response in children with dyslexia and with (central) auditory processing disorders. Master's dissertation, University of São Paulo, Brazil

Sitek KR, Gulban OF, Calabrese E et al (2019) Mapping the human subcortical auditory system using histology, postmortem MRI and in vivo MRI at 7T. elife 8:1–36

Skoe E, Chandrasekaran B (2014) The layering of auditory experiences in driving experience-dependent subcortical plasticity. Hear Res 311:36–48

Skoe E, Kraus N (2010) Auditory brainstem response to complex sounds: a tutorial. Ear Hear 31(3):302–324

Skoe E, Kraus N (2012) A little goes a long way: how the adult brain is shaped by musical training in childhood. J Neurosci 32(34):11507–11510

Smayda KE, Chandrasekaran B, Maddox WT (2015) Enhanced cognitive and perceptual processing: a computational basis for the musician advantage in speech learning. Front Psychol 6:1–14

Smith JC, Marsh JT, Brown WS (1975) Far-field recorded frequency-following responses: evidence for the locus of brainstem sources. Clin Neurophysiol 39(5):465–472

Song JH, Skoe E, Wong PC et al (2008) Plasticity in the adult human auditory brainstem following short-term linguistic training. J Cogn Neurosci 20(10):1892–1902

Sperling AJ, Lu ZL, Manis FR et al (2005) Deficits in perceptual noise exclusion in developmental dyslexia. Nat Neurosci 8(7):862–863

Stein J, Walsh V (1997) To see but not to read: the magnocellular theory of dyslexia. Trends Neurosci 20(4):147–152

Suga N, Ma X (2003) Multiparametric corticofugal modulation and plasticity in the auditory system. Nat Rev Neurosci 4(10):783–794

Swaminathan J, Krishnan A, Gandour JT (2008) Pitch encoding in speech and nonspeech contexts in the human auditory brainstem. Neuroreport 19(11):1163–1167

Vallabha GK, McClelland JL (2007) Success and failure of new speech category learning in adulthood: consequences of learned Hebbian attractors in topographic maps. Cogn Affect Behav Neurosci 7(1):53–73

Vallabha GK, McClelland JL, Pons F et al (2007) Unsupervised learning of vowel categories from infant-directed speech. Proc Natl Acad Sci U S A 104(33):13273–13278

Vander Werff KR, Burns KS (2011) Brain stem responses to speech in younger and older adults. Ear Hear 32(2):168–180

von Kriegstein K, Patterson RD, Griffiths TD (2008) Task-dependent modulation of medial geniculate body is behaviorally relevant for speech recognition. Curr Biol 18(23):1855–1859

Warren RM (1961) Illusory changes of distinct speech upon repetition—the verbal transformation effect. Br J Psychol 52(3):249–258

Warrier CM, Abrams DA, Nicol TG et al (2011) Inferior colliculus contributions to phase encoding of stop consonants in an animal model. Hear Res 282(1–2):108–118

Weinberger NM (2004) Specific long-term memory traces in primary auditory cortex. Nat Rev Neurosci 5(4):279–290

White-Schwoch T, Carr KW, Thompson EC et al (2015) Auditory processing in noise: a preschool biomarker for literacy. PLoS Biol 13(7):e1002196

White-Schwoch T, Nicol T, Warrier CM et al (2016) Individual differences in human auditory processing: insights from single-trial auditory midbrain activity in an animal model. Cereb Cortex 27(11):5095–5115

Willmore BD, Schoppe O, King AJ et al (2016) Incorporating midbrain adaptation to mean sound level improves models of auditory cortical processing. J Neurosci 36(2):280–289

Winer JA (2005) Decoding the auditory corticofugal systems. Hear Res 207(1–2):1–9

Wong PCM, Skoe E, Russo NM et al (2007) Musical experience shapes human brainstem encoding of linguistic pitch patterns. Nat Neurosci 10(4):420–422

Xie Z, Reetzke R, Chandrasekaran B (2017) Stability and plasticity in neural encoding of linguistically relevant pitch patterns. J Neurophysiol 117(3):1409–1424

Xie Z, Reetzke R, Chandrasekaran B (2018) Taking attention away from the auditory modality: context-dependent effects on early sensory encoding of speech. Neuroscience 384:64–75

Xie Z, Reetzke R, Chandrasekaran B (2019) Machine learning approaches to analyze speech-evoked neurophysiological responses. J Speech Lang Hear Res 62(3):587–601

Xu Y, Krishnan A, Gandour JT (2006) Specificity of experience-dependent pitch representation in the brainstem. Neuroreport 17(15):1601–1605

Yeo BT, Eickhoff SB (2016) Systems neuroscience: a modern map of the human cerebral cortex. Nature 536(7615):152–154

Yi HG, Smiljanic R, Chandrasekaran B (2014) The neural processing of foreign-accented speech and its relationship to listener bias. Front Hum Neurosci 8: 768

Yi HG, Maddox WT, Mumford JA et al (2016) The role of corticostriatal systems in speech category learning. Cereb Cortex 26(4):1409–1420

Yi HG, Xie Z, Reetzke R et al (2017) Vowel decoding from single-trial speech-evoked electrophysiological responses: a feature-based machine learning approach. Brain Behav 7(6):e00665

Yip M (2002) Tone. Cambridge University Press, New York

Young ED, Sachs MB (1979) Representation of steady-state vowels in the temporal aspects of the discharge patterns of populations of auditory-nerve fibers. J Acoust Soc Am 66(5):1381–1403

Zhong Z, Henry KS, Heinz MG (2014) Sensorineural hearing loss amplifies neural coding of envelope information in the central auditory system of chinchillas. Hear Res 309:55–62

Chapter 3
Cortical Representation of Speech Sounds: Insights from Intracranial Electrophysiology

Intracranial Electrophysiology of Speech Sound Processing

Yulia Oganian, Neal P. Fox, and Edward F. Chang

Abstract The superior temporal gyrus (STG) has long been recognized as crucial to the human ability to perceive and comprehend spoken language. However, the nature of the neuronal computations and cortical representations responsible for this sensory and cognitive feat remain a mystery. The recent advance of methodologies for intracranial electrophysiology (iEEG) recordings, together with the emergence of novel computational approaches, have heralded progress toward understanding how neural processing in auditory cortex gives rise to the perceptual experience of speech. This chapter describes a collection of intracranial neurophysiology studies that illustrate two fundamental properties of STG encoding of speech sounds. First, this neural representation of speech is firmly rooted in the analysis of high-order acoustic features in the sensory stimulus. Second, the neural representation also differs dramatically from a linear representation of sound acoustics. The STG encodes an imperfect spectrotemporal representation of speech, sacrificing faithfulness to the sensory signal where it enhances the robust encoding of linguistically and behaviorally relevant information. Besides being *insensitive* to behaviorally *irrelevant* information carried by the speech signal, STG is also *sensitive* to behaviorally *relevant* information *not* contained within the speech signal (i.e., top-down cues). Overall, mounting evidence suggests that STG is a sensory-perceptual hub for the human speech perception system, functionally characterized by the behaviorally relevant cortical representation of speech that emerges therein.

Yulia Oganian and Neal P. Fox contributed equally with all other contributors.

Y. Oganian · N. P. Fox · E. F. Chang (✉)
Department of Neurological Surgery, University of California, San Francisco,
San Francisco, CA, USA
e-mail: Yulia.Oganian@ucsf.edu; nealpfox@gmail.com; Edward.Chang@ucsf.edu

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022

L. L. Holt et al. (eds.), *Speech Perception*, Springer Handbook of Auditory
Research 74, https://doi.org/10.1007/978-3-030-81542-4_3

Keywords Human speech perception · Electrocorticography (ECoG) · Intracranial EEG (iEEG) · Superior temporal gyrus (STG) · Phonetic features · Invariance · Top-down processing · Categorical perception · Cocktail party · Intonation · Relative pitch · High-gamma band activity · Stimulus reconstruction · Spectrotemporal receptive field (STRF)

3.1 Introduction

A fundamental goal of sensory neuroscience is to understand how neural representations of sensory inputs generate a perceptual experience of the world (Field 1994; Holdgraf et al. 2017). In auditory neuroscience, one of the most well-studied sensory inputs is human speech. Several decades of multidisciplinary research have aimed to elucidate the neurobiological mechanisms responsible for listeners' remarkable capacity to effortlessly extract meaning from the speech signal. Much is known about the earliest stages of auditory processing, such as the mapping of any acoustic signal (speech included) onto temporally precise, frequency-specific neural firing patterns in the human cochlea (Schnupp et al. 2011). However, far less is known about how auditory signals in general, and human speech sounds in particular, are represented in human cortex.

Of particular consequence for the study of human speech perception is how speech is represented within the superior temporal gyrus (STG; see Table 3.1 for abbreviations), a cortical region that has long been recognized for its crucial role in speech perception and auditory language comprehension (Wernicke 1874; Howard et al. 2000). Despite how prominently the STG figures in leading functional-neurobiological models of speech perception (Hickok and Poeppel 2007; DeWitt and Rauschecker 2012), a detailed understanding of how different speech sounds

Table 3.1 Table of abbreviations

Abbreviation	Full name
BOLD	Blood oxygenation levels
ECoG	Electrocorticography
EEG	Electroencephalography
ERP	Event-related potential
fMRI	Functional magnetic resonance imaging
iEEG	Intracranial electrophysiology
LFPs	Local field potentials
MEG	Magnetoencephalography
MMN	Mismatch negativity
STG	Superior temporal gyrus
STRF	Spectrotemporal receptive field
STS	Superior temporal sulcus

are encoded by patterns of neural activity in STG has been elusive. However, the last decade has witnessed unprecedented progress in understanding the mapping between auditory stimuli, neural response patterns, and perceptual behavior. Many of these advances have been possible due to the advent of innovative human electrophysiological recording methods (Mukamel and Fried 2012; Chang 2015), in combination with modern analytical techniques that leverage computational advances and state-of-the-art machine learning paradigms. The goal of this chapter is to review some of this recent research characterizing the neurophysiological response of human auditory cortex to speech and linking these neural response properties to the perceptual experience of speech.

This chapter describes recent evidence from intracranial electrophysiology (iEEG), showing that the cortical representation of speech is grounded in acoustic features of the sensory stimulus, but also that it differs dramatically from a veridical analog representation of the acoustic signal in several ways. Importantly, a parallel argument is fundamental to prevailing cognitive theories of speech perception: Although listeners' perception of speech is undoubtedly primarily driven by the acoustic signal, it is clear from decades of behavioral psychophysical and psycholinguistic research that the perceptual experience of speech differs in key ways from the sensory input (Davis and Johnsrude 2007; Samuel 2011). Striking parallels between the emerging picture of how human STG encodes speech and how listeners perceive speech suggest that the STG is a sensory-perceptual interface for the human speech perception system, functionally characterized by the behaviorally relevant cortical representation of speech that emerges therein. For instance, the same word will sound differently when produced by different speakers (e.g., due to accents, or voice height). Yet, listeners are able to separate the invariant aspects of speech sounds from such speaker-dependent variability, to arrive at a robust percept. We will discuss how the STG arrives at such invariant representations of speech sounds later in this chapter.

3.2 Spectrotemporal Encoding of Speech Sounds in Human Auditory Cortex

3.2.1 Non-invasive Approaches to the Study of Auditory Cortical Pathways

As described in Sect. 3.1, one general class of questions that is central to human neuroscience involves characterizing the mapping from sensory inputs to the neural responses they evoke. More specifically, in the context of speech, the questions are as follows: (1) What regions in human brain encode speech sounds? (2) What information about auditory sensory inputs is encoded in neuronal firing patterns in these brain regions? Our ability to answer both questions is limited by available methodological approaches for recording and analyzing neural activity.

In the past, the first question has been addressed using methods that can discriminate between neural activity originating in different parts of the brain, that is methods with a high spatial resolution, such as functional magnetic resonance imaging (fMRI) (see also Chap. 7, Ullas, Bonte, Formisano, and Vroomen). fMRI has yielded clear evidence that neural activity in the bilateral superior temporal lobe – including STG, superior temporal sulcus (STS), primary auditory cortex (A1) on Heschl's gyrus, planum temporale, and planum polare – increases in responses to sounds. In particular, neural activity in bilateral STG is increased in response to speech as opposed to non-speech sounds, cementing the view of the STG as the central auditory sensory area that is specialized for the processing of speech sounds (Liebenthal et al. 2003; Zevin and McCandliss 2005).

However, fMRI is not well suited to address the second question, as it measures neural responses indirectly, by tracking changes in blood oxygenation levels (BOLD). Changes in BOLD occur much slower (on the scale of seconds) than changes in neural activity and in the speech signal, which fluctuate on the scale of milliseconds. Thus, to study the neural dynamics of cortical speech sound representations, cognitive neuroscience has long relied on electroencephalography (EEG) and magnetoencephalography (MEG): electrophysiological methods that capture rapid fluctuations in cortical neural dynamics. Using MEG and EEG, it is possible to track neural responses in "real-time" with fluctuations in the speech signal. However, as MEG and EEG capture neural activity from outside the scalp, these methods reflect neural activity summed across a large number of cortical sources. To describe how neural dynamics of spatially confined local neural populations represent speech sounds, it is necessary to record neural activity with both high temporal and high spatial resolution. To achieve such a high spatiotemporal resolution, it is necessary to measure neural activity directly on the cortical surface or with invasive probes from inside the cortex. This can be achieved with invasive electrophysiological recordings, which we introduce in the next section.

3.2.2 Invasive Electrophysiological Recordings in Animals and Humans

Although direct recordings of neural firing are rare in human neuroscience, they are common in research with animal models (Brugge 1992; Theunissen and Shaevitz 2006). Because many basic components and computations in the auditory system appear to be largely evolutionarily conserved, this work has shaped current neurophysiological models of auditory processing in humans (Rauschecker and Scott 2009; Steinschneider et al. 2013).

Invasive neurophysiology research has several advantages over the prevailing non-invasive neural recording methods (e.g., MEG, EEG, fMRI). Whether an experimenter employs intracellular recordings, which monitor spiking activity of a single neuron (unit), or extracellular recordings of local field potentials (LFPs) summed across a number of neurons located in proximity of the probe, which can be used to infer single- and multi-unit activity (Buzsáki et al. 2012; Einevoll et al. 2013; Pesaran et al. 2018), neurophysiological recordings provide direct access to the dependent variable of greatest interest: neuronal firing. Moreover, the spatial localization and temporal precision of these recording methods are unparalleled; millisecond-resolution data about individual spikes can be collected concurrently from many individual neurons. Additionally, direct neurophysiological recordings are typically characterized by very high signal-to-noise ratios, which allow experimenters to robustly estimate detailed coding properties of individual neurons by presenting relatively few tokens sampled from the stimulus space of interest.

Although there is evidence that many aspects of speech perception in humans rely on more basic auditory mechanisms that are shared with other species (Kuhl 1986; Kluender et al. 2005), the degree of similarity between human and non-human cortical representations of speech is not known. As such, direct neurophysiological recordings from humans represent a unique source of information about the cortical encoding of speech sounds. However, because of the invasive nature of direct cortical recordings, it is not possible to ethically collect these data from humans except in rare cases where neurosurgical procedures that expose auditory cortex are clinically necessary (Crone et al. 2006; Parvizi and Kastner 2018).

For example, for some patients with medically refractory (i.e., resistant to pharmacological interventions) temporal lobe epilepsy, treatment may involve surgical resection of the neural tissue implicated in generating the patient's seizures (Ojemann 1987). Frequently, this procedure requires the subdural implantation of non-penetrating electrophysiological recording arrays (ECoG arrays), in direct contact with the cortical surface, or penetrating depth electrodes for access to subcortical structures (e.g., the amygdala, stereoEEG), for 1–2 weeks prior to surgical resection in order to allow for seizure localization and/or functional neuroanatomical mapping (e.g., of sensorimotor and language areas of cortex). Occasionally, patients implanted with such intracranial EEG (iEEG) contacts also volunteer to participate in research over the course of their treatment (Fig. 3.1). Although electrode placement for each patient is determined solely on the basis of clinical necessity, temporal lobe coverage is often included, providing rare and invaluable access to direct cortical recordings from STG (and sometimes of primary auditory cortex) in awake, behaving humans with typical hearing and language abilities.

Most frequently, iEEG records fluctuations in LFPs. From this signal, various components of the LFP can be extracted that correspond to changes occurring over different time-scales (or frequency bands; see Chap. 4, Tune and Obleser). Of particular interest is neural activity in the high-gamma band range (~70–200 Hz), which correlates with local neuronal population activity (Ray and Maunsell 2011; Leszczyński et al. 2019). Research has consistently found that changes in high-gamma amplitude are temporally resolved (Crone et al. 1998, 2001), spatially focal (Menon et al. 1996; Łęski et al. 2013), and reliably evoked by sensory stimuli, even during single trials (Flinker et al. 2010), providing a combination of spatiotemporal resolution and a signal-to-noise ratio that is unmatched by non-invasive recording technologies for human neuroscience.

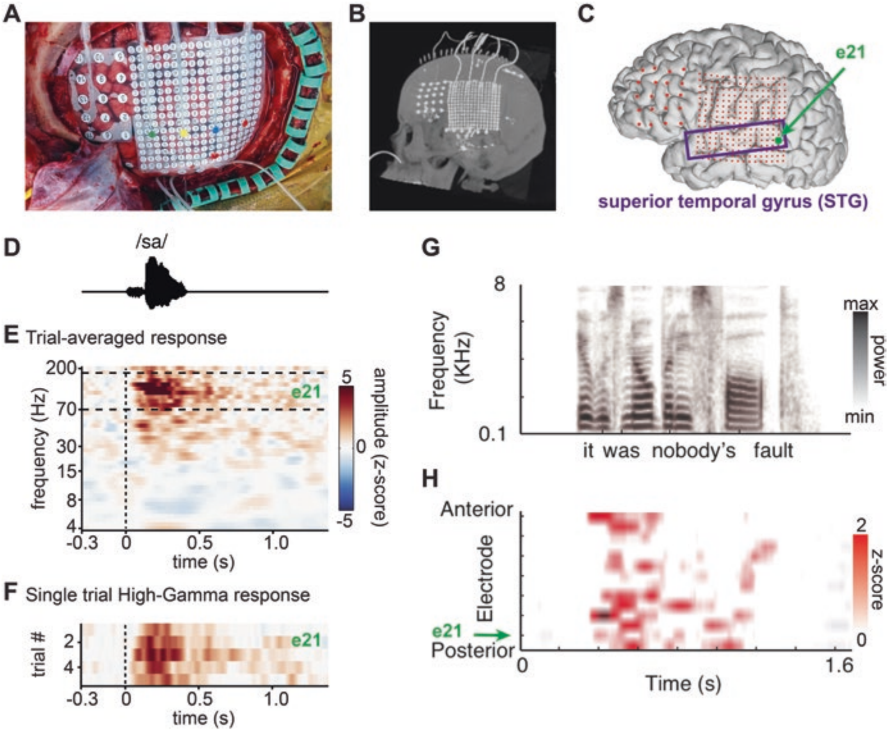

Fig. 3.1 Intracranial recordings of neural activity from human superior temporal gyrus (STG). (**A**) Placement of electrode grids during implantation surgery. (**B**) Localization of grids with postoperative CT. (**C**) Reconstruction of single electrode location on participant's structural MRI scan. Green: Example electrode 21. Purple box: Electrodes located over the STG. (**D–E**) Neural responses on electrode 21, temporally aligned to onset of syllable /sa/. (Adapted from Chang (2015)). (**D**) Stimulus waveform. (**E**) Time-frequency representation of neural response power, averaged across five presentations of the stimulus token in D. Horizontal lines mark boundaries of high gamma (HG) frequency range (70–200 Hz) (**F**) Single trial HG response amplitude in response to /sa/, averaged across single frequency bands in HG range. (**G**, **H**) Spectrotemporal representation of a single sentence (**G**) and HG response amplitude time-aligned to sentence onset on a set of STG electrodes (**H**). (Adapted from Mesgarani et al. (2014))

The high spatiotemporal resolution and signal-to-noise ratio of iEEG come at the cost of clinical limitation of electrode placement and patient availability. First, as the location of contacts is dictated by clinical needs, some brain regions may receive more coverage than others, thus dictating the neuroanatomical research focus. Second, any single patient will only require contacts to be placed in a small number of brain regions, limiting the experimenter's ability to concurrently record neural activity across wide parts of cortex, as is possible, e.g., with MEG. Finally, the demographics of patients participating in an iEEG study are determined by the patients' clinical needs, rather than the experimenter's scientific endeavor. Thus, iEEG research is not as well suited to study special populations (e.g., developmental, aging, learning-disabled, or multilingual populations). For all these reasons, iEEG must be supplemented and combined with non-invasive approaches to human cognition (Johnson et al. 2020).

Of course, any measure of neural activity will have little utility for understanding the cortical representation of speech unless it varies reliably depending on properties of the speech stimulus. Next, we present the recent methodological advances in the application of machine learning methods to speech neuroscience that make it possible to study this link. We will then exemplify these methods on two studies: In short, different spoken words and sounds elicit spatially and temporally specific, highly discriminable patterns of neural activity within human STG. This work provides evidence for the functional relevance of high-frequency responses in auditory cortex for speech sound encoding. More importantly, though, this work illustrates what is perhaps the most fundamental principle of how speech sounds are represented in human STG: sensitivity to rapid, fine-grained spectrotemporal modulations present in the speech signal that are critical to distinguishing among different sounds, and, therefore, among different meanings.

3.2.3 Experimental Design and Data Analytic Approaches in Intracranial Electrophysiology

A straightforward approach to the quest of linking systematic changes in neural activity to changes in the speech stimulus is to present different speech stimuli to a listener and to compare the neural responses evoked by distinct stimuli (Donders 1969). One illustration of this classical "subtraction" approach to cognitive neuroscience relies on the mismatch negativity (MMN) response recorded with MEG and EEG (Näätänen et al. 2007) MMN is an electrophysiological index of an infrequent change in a sequence of repeating stimuli, as measured by a deflection in the event-related potential (ERP), time-locked to the onset of the "mismatching" stimulus (Näätänen and Picton 1987). For example, the classic auditory oddball paradigm exposes a listener to repeated presentations of a frequent ("standard") stimulus, interspersed with presentations of a rare ("deviant") stimulus. Countless studies have shown that deviant spoken syllables, similar to deviant tones or other non-speech sounds, elicit a more negative scalp potential than standard stimuli, with the MMN difference peaking within ~100 ms of stimulus onset (Näätänen 2001). These results provide evidence that neural processing in human cortex *is sensitive* to acoustic differences among speech sounds. Yet, it does not tell us *how* different speech sounds are encoded by different neural firing patterns. To see why this is the case, note that embedding deviant */pa/* syllables among standard */ba/* syllables will induce an MMN response that is indistinguishable from the MMN evoked when */ba/* is the deviant and */pa/* is the standard. As such, the MMN alone cannot directly explain how different speech sounds are encoded by auditory cortex. Moreover, the MMN response is contingent on artificial experimental paradigms that are not well suited for investigating the cortical representations that support humans' ability to perceive and distinguish among a multitude of speech sounds embedded in natural speech (Assmann and Summerfield 2004; Mattys et al. 2012).

To directly assess *how* different speech sounds are encoded by different neural firing patterns, it is necessary to use a technique that explicitly models how measured neural activity co-varies with acoustic (or other; see Sects. 3.3 and 3.4) dimensions of speech stimuli. If a reliable mapping between stimulus properties and neural response properties can be established, then it should ultimately be possible to decode speech from the neural activity alone (Herff and Schultz 2016; Moses et al. 2016). Modern machine learning techniques have allowed neurophysiologists to develop computationally sophisticated methods for examining how auditory stimuli are encoded by neural activity, both locally (e.g., by a single neuron) and at the population level (Depireux et al. 2001; Theunissen et al. 2001). A prominent class of data modeling approaches relies on regularized linear models with explicit modeling of the temporal structure in the stimulus. In the next section, we exemplify this method on two studies. Throughout the remaining of this chapter, we focus on the advances in speech neuroscience that were made possible by the combination of intracranial recordings and this statistical modeling framework. However, it is notable that more and more recent work with fMRI (Mitchell et al. 2008; Huth et al. 2016) and MEG/EEG (Di Liberto et al. 2015; Khalighinejad et al. 2017) has taken a parallel complimentary approach.

3.2.4 Linear Spectrotemporal Encoding of Speech: Distributed Neural Responses in Human Superior Temporal Gyrus

A way to establish a proof-of-concept demonstration that human STG does, indeed, encode the spectrotemporal properties of speech (Fig. 3.2) is the following: If neural firing patterns in human STG robustly encode the spectrotemporal properties of speech sounds, then it should be possible to reconstruct the speech signal that a listener heard from STG neural activity alone. In the past, this approach was used to demonstrate that auditory areas encode sound dynamics, e.g., that neurons in ferret primary auditory cortex A1 encode the spectrotemporal dynamics of speech sounds (Mesgarani et al. 2009) This approach, however, requires the resolution of neural recordings to match the rapid spectrotemporal dynamics of natural speech. This became possible with the advent of human intracranial recordings, as was demonstrated in a seminal study by Pasley and colleagues (2012). They recorded ECoG activity while epilepsy patients listened to spoken words that varied in their acoustic properties (e.g., *deep*, *jazz*, *cause*). For each trial (i.e., word), two corresponding measurements were defined: (1) the spectrogram of the stimulus (relative power in different frequency bands as a function of time), and (2) the stimulus-driven spatiotemporal neural response pattern (high-gamma activity at different electrodes over time, see Sect. 3.2.2). Under the hypothesis that the high-gamma activity evoked on

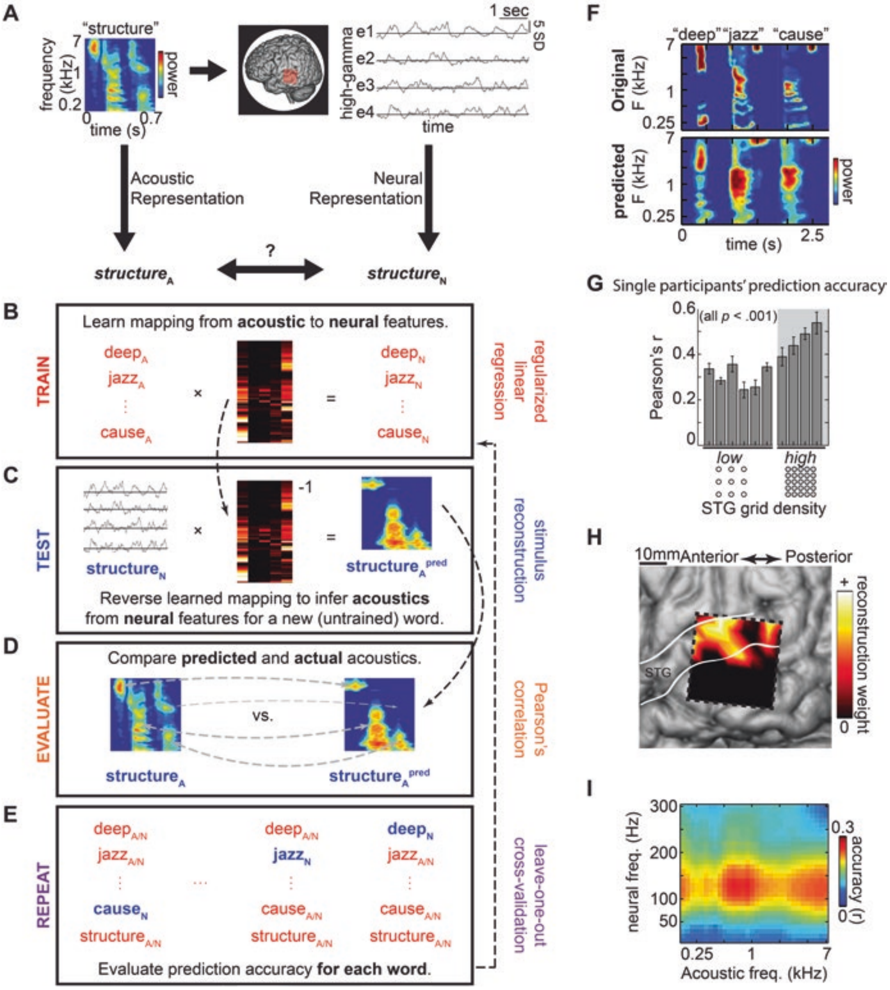

Fig. 3.2 Speech reconstruction from STG ECoG recordings. (**A**) Schematic of experimental setup as employed by Pasley et al. (2012). Participants listen to single words, while high-gamma activity was recorded over their lateral temporal cortex using ECoG grids. (**B–E**) Schematic of model training and testing for stimulus spectrogram reconstruction from HG amplitude time series, for the example test stimulus word "structure." A: acoustic; N: neural. (**B**) Training: Regularized linear regression is used to find the best linear weights, mapping stimulus spectrograms to population neural activity on the electrocorticography (ECoG) grid. (**C**) Testing: These weights are inversed to predict the spectrogram of a single word stimulus that was not used during training. (**D**) Prediction quality is evaluated as the correlation between the true and predicted stimulus spectrograms. (**E**) This procedure is repeated for each single stimulus word serving as test token in a leave-one-out cross-validation setup. (**F**) Original and predicted spectrograms for three example words. (**G**) Average prediction accuracy in a subset of participants. Prediction accuracy was higher in participants with high density grids. (**H**) Electrodes over STG contributed most to stimulus reconstruction. (**I**) Comparing reconstruction accuracy across different acoustic frequency bands and neural frequency bands. For all acoustic frequency bands, best reconstruction was achieved based on the neural high-gamma band. (Adapted from Pasley et al. (2012))

any given trial is directly related to the word's acoustic spectrogram, the authors implemented a linear regression-based model-fitting procedure that finds the optimal mapping between spectrotemporal features of a stimulus and spatiotemporal features of the neural response. If the stimulus-response mapping was stable, a model trained in this way would be capable of predicting the expected pattern of high-gamma activity over time and across multiple cortical recording sites for any arbitrary sound (Bialek et al. 1991; Ramirez et al. 2011).

Complementarily, if the neural activity encodes the acoustic spectrogram in this way, it should also be possible to decode which word was presented to a participant based only on the spatiotemporal pattern of high-gamma activity measured by ECoG. Following this logic, Pasley and colleagues (2012) evaluated how well a linear spectrotemporal encoding model could reconstruct the spectrogram of novel words. For each word that was presented to a given participant, the authors generated a model-predicted spectrogram (reconstructed stimulus) based on the optimal stimulus-response mapping (fit using all other trials) and the actual neural response for that trial. This prediction was then compared to the true spectrogram that was presented to the subject. Overall, they found that the reconstructed speech closely matched the original speech. This was true across all patients, with especially high performance in subjects with high-density electrode coverage of the STG. In other words, single-trial measurements of high-gamma activity in human STG do, indeed, accurately encode the spectrotemporal details of spoken words. Moreover, a speech recognition classifier could often predict which word had been presented based on the spectrogram that had been reconstructed from brain activity. In fact, even when the classifier's prediction was wrong, the authors found that incorrectly guessed words tended to be acoustically similar to the correct word, further corroborating the tight link between stimulus acoustics and STG response patterns.

This study exemplifies the power of human ECoG recordings to elucidate how temporally dynamic, spatially distributed neural activity encodes key properties of the sensory signal. Unlike studies using indirect indices of neural activity, like the MMN response, Pasley and colleagues (2012) demonstrated that the pattern of neural activity measured in human STG in response to a word was linearly related to how that word sounded. Different words evoked different patterns of high-gamma activity, but similar-sounding words evoked similar patterns of activity. Additionally, these data provide compelling support for high-gamma activity as a reliable neurophysiological measure of STG encoding of speech. Indeed, additional analyses showed that spectrogram reconstruction was only possible using the high-gamma band (as opposed to other frequency bands; see also Chap. 4, Tune and Obleser, on possible role of low-frequency bands in speech processing).

It is important to recognize that the key conclusion of this study – that the STG's response to speech is grounded in the spectrotemporal features present in the raw acoustic signal – may be best construed as a proof-of-concept. The fact that a linear model can relate a spectrotemporal representation of the stimulus to a spatiotemporal pattern of the neural response leaves open many theoretical questions about the details of this stimulus-response mapping. Since Pasley and colleagues (2012) modeled cortical activity that was distributed across many spatially discrete neural populations, it is not trivial to understand how individual, local neuronal populations contribute to the overall ability to decode stimulus properties. For a more complete understanding of cortical speech encoding, it is necessary to explicitly characterize how this distributed spectrotemporal representation emerges from the stimulus-dependent activity of individual, distinct local neuronal populations.

3.2.5 Linear Spectrotemporal Encoding of Speech: Local Neural Responses in Human Superior Temporal Gyrus

Across sensory systems and species, cortical neurons very often exhibit highly specific tuning for particular stimulus features and dimensions, such as orientation tuning in vision (Hubel and Wiesel 1962), frequency tuning in vibrotactile somato-sensation (Bolanowski et al. 1988), and frequency tuning in audition (Merzenich and Brugge 1973; Merzenich et al. 1975). In all of these cases, a neuron's tuning curve can be determined by measuring its responsiveness to a range of stimuli. Typically, a given neuron will tend to respond preferentially to a certain stimulus or sub-range of stimuli, while other neurons will tend to respond to different stimuli. This diversity in the local tuning for stimulus features underlies the cortical encoding of sensory stimuli. Similarly, the cortical encoding of speech observed by Pasley et al. (2012) may be subserved by diversely tuned neural populations in STG.

One way to explore this hypothesis is to determine which stimulus features tend to evoke neural responses at each cortical site, which can be achieved by applying the same basic encoding model framework used by Pasley et al. (2012) to data from each cortical site separately. While the stimulus features remain the same (the speech spectrogram), the stimulus-driven neural response is confined to the high-gamma time series for a single electrode. Applying a similar regression-fitting procedure yields a linear filter of weights that relates specific spectral and temporal stimulus properties to changes in amplitude (excitation or inhibition) in the electrode's neural (high-gamma) response.

This approach is not new. It has long been a standard technique in auditory neurophysiology, where the filter of weights relating the spectrotemporal description of the speech signal to neural response amplitude (or spike probability when estimated for single neurons) is known as a *spectrotemporal receptive field*, or STRF (Theunissen et al. 2001). In order to estimate a linear STRF for the neural population beneath a given ECoG electrode, neural activity at each point in time (t) is modeled as a linear combination of stimulus features in a time window preceding t. That is, it is assumed that acoustic energy in any of the speech spectrogram's frequency bands could drive increases (or decreases) in neural firing, and that the latencies of these effects on neural firing could vary. A variety of linear model-fitting techniques allow for the estimation of a set (or filter) of regression weights (the STRF) that best explains a neural population's spectrotemporal tuning and can best predict its response to a novel stimulus spectrogram.

Linear STRF modeling is a dominant paradigm in the field of auditory neurophysiology, where the technique has been applied to study spectrotemporal coding properties throughout the ascending auditory pathway in animal models (deCharms et al. 1998), and, more recently, in human auditory cortex (Ding and Simon 2012; Hullett et al. 2016). However, the underlying approach exemplified by STRF modeling (Klein et al. 2000) is quite general. For instance, instead of training a STRF, which predicts the neural response based on a spectrotemporal representation of the stimulus, the same method can estimate a filter that predicts neural responses from other features (e.g., phonetic, semantic, or pitch features in a stimulus), yielding *feature temporal receptive fields,* or F-TRFs. For example, a speech stimulus can be described by the timing and duration of occurrence of single phonemes. In this case, the complete feature set would contain a separate binary feature for each phoneme of a language, which would take the value 1 whenever that phoneme is present in the stimulus, and 0 otherwise. By comparing the predictive power of models encompassing different sets of features, it is possible to determine what level of stimulus description captures neural responses the best. For example, one might compare a phoneme TRF model with spectrotemporal model, to test whether a neural population is tuned to single phonemes (e.g., [f]) or to certain spectrotemporal patterns, such as energy in high-frequency bands (Mesgarani et al. 2014).

For temporally continuous stimulus features, such as the spectrogram representation, temporal receptive fields can be conceptualized as the optimal stimulus that induces the largest neural response. For discrete features, such as phonetic features, the estimated model weights are best conceptualized as the average neural response to that feature.

Crucially, STRF-fitting estimates the weights for different features and feature combinations even if that feature sometimes overlapped with other features in the stimulus. Consequently, this robust approach is capable of identifying stimulus features that drive a neural response based on data collected during presentation of large, naturalistic stimulus sets (David et al. 2007). Thus, spectrotemporal- and feature-TRF-based analyses are well suited to analyze continuous neural data that are not neatly separated into carefully controlled discrete trials (Hamilton and Huth 2018).

TRF-based analyses offer a window through which we can view average encoding properties (i.e., tuning) of single neurons or, in the case of ECoG, of local ensembles of neurons. For instance, STRFs in human A1 bear functional similarities with those of neurons found in the primary auditory cortex of non-human mammals (Nelken et al. 2003; Wang et al. 2005). In particular, individual neurons tend to be tuned for relatively simple stimulus features, such as the occurrence of a neuron's characteristic frequency or frequencies (Bitterman et al. 2008; Griffiths et al. 2010). However, human neurophysiological evidence suggests that STG neurons and local ensembles of neurons are not as narrowly tuned to single frequencies as in A1, but rather it is clear that STG must encode spectrotemporal properties of the stimulus (Hullett et al. 2016; Berezutskaya et al. 2017). These observations raise an obvious question: What are local STG neural populations tuned to detect?

3.2.6 Phonetic Feature Representation of Spectrotemporal Properties in Speech

One possibility is that neural populations in human STG are tuned to detect complex spectrotemporal patterns in speech known as *phonetic features*. In phonological theory, phonetic features describe acoustic dimensions of human speech that tend to co-vary due to the physical and dynamical properties of the articulatory system (Stevens 2002). Producing a given speech sound involves specific manipulations of the lips, tongue, glottis, and larynx, with different configurations resulting in an assemblage of acoustic consequences that tend to pattern together in complex ways. For instance, when a speaker produces a [d], she must place the tip of her tongue beneath the alveolar ridge just behind the teeth, completely occlude airflow through the vocal tract temporarily, and then rapidly release the blockage while simultaneously initiating vibration of her vocal cords. In terms of phonetic features, a [d] can be described as a *voiced alveolar plosive consonant*. Acoustically, the rapid release of a complete occlusion (characteristic of plosive consonants) is accompanied by a brief broadband burst of noise, while the onset of voicing is signaled by harmonic formant structure caused by resonances of the vocal tract, the details of which depend on the identity of the subsequent vowel. Importantly, phonetic features are common to all individual realizations of speech sounds, even though a lot of variability exists between how different speakers produce the same [d] sound, depending on individual differences in vocal tracts and surrounding speech sounds (coarticulation). Thus, a neural representation based on phonetic features would require separating systematic differences between phonetic features from within-feature variability; in other words, it would be *invariant* to such subphonetic acoustic variability.

Importantly, phonetic features are grounded in the articulatory capabilities of a speech motor system that is shared by all humans. Thus, the complete inventory of human speech sounds can be captured by relatively small space of vocal tract configurations and articulatory maneuvers, even though most human languages use only a subset of phonetic features and their combinations (e.g., in tonal languages, unlike in English, relative pitch is a feature of vowel sounds) (Jakobson et al. 1951; Chomsky and Halle 1968). Thus, a neural system tuned to detect phonetic features would constitute a computationally efficient, theoretically motivated, language-independent representation of speech sounds. Under such a model, successful spectrogram reconstruction from distributed cortical activity (Holdgraf et al. 2017) would be possible because individual cortical sites encode stereotyped spectrotemporal events that arise from particular articulatory operations.

Moreover, and critically, from a behavioral standpoint, such a model would also allow listeners to reliably discriminate speech sounds that distinguish among meanings because phonetic features are the building blocks that comprise *phonemes*, the smallest contrastive units in the sound system of a spoken language (de Saussure 1916; Sapir 1925). In the context of spoken word production, changing one or more phonetic features will change not only the acoustic realization of a speech sound, but also the meaning of a word.

For instance, slightly delaying the initiation of voicing (thereby changing the *voicing* feature from *voiced* to *voiceless*) will change *dill* to *till*, which is acoustically cued by a temporal lag between the burst and onset of harmonic vowel structure. On the other hand, occluding airflow with the lips instead of the tongue (thereby changing the *place of articulation* feature from *alveolar* to *labial*) will change *dill* to *bill*, which is acoustically realized by a change in the spectral content of the burst and transition into the vowel (Stevens and Blumstein 1978), factors that depend on the precise physical arrangement of articulators when a stop is released. Alternatively, maintaining the oral occlusion (instead of releasing it) while allowing air to escape through the nose (thereby changing the *manner of articulation* feature from *plosive* to *nasal*) will change *dill* to *nil*, a difference that listeners easily detect because of the increase in low-frequency energy caused by the expansion of the resonant cavity for the nasal. Finally, implementing the latter two changes simultaneously will change *dill* to *mill*. In short, encoding speech sound identity by detecting acoustic-phonetic features provides a low-dimensional, highly generalizable sensory representation of the spectrotemporal cues that are relevant to meaning, and along which speech sounds universally tend to vary.

3.2.7 Phonetic Feature Encoding by Local Neural Responses in Superior Temporal Gyrus

Two important studies (Chan et al. 2014; Mesgarani et al. 2014) found direct evidence to support the hypothesis that local neural populations in the STG selectively encode the spectrotemporal patterns corresponding to phonetic features. Mesgarani and colleagues (2014) (Fig. 3.3) acquired ECoG recordings directly from STG while patients passively listened to hundreds of naturally spoken sentences from the TIMIT corpus (Garofolo et al. 1993), which contains many instances of all English phonemes. For each electrode positioned over the STG, the average time-locked response to each phoneme across all of its occurrences was extracted in order to determine how phoneme identity modulated high-gamma activity. As discussed earlier, this technique yields a phonetic feature receptive field for each electrode. Results showed that neural populations contributing to neural responses on single electrodes did not respond to all phonemes equally, nor were their responses limited to single phonemes. Rather, neural populations responded with different amplitudes to different phonemes, with any given cortical site in STG tending to respond to a

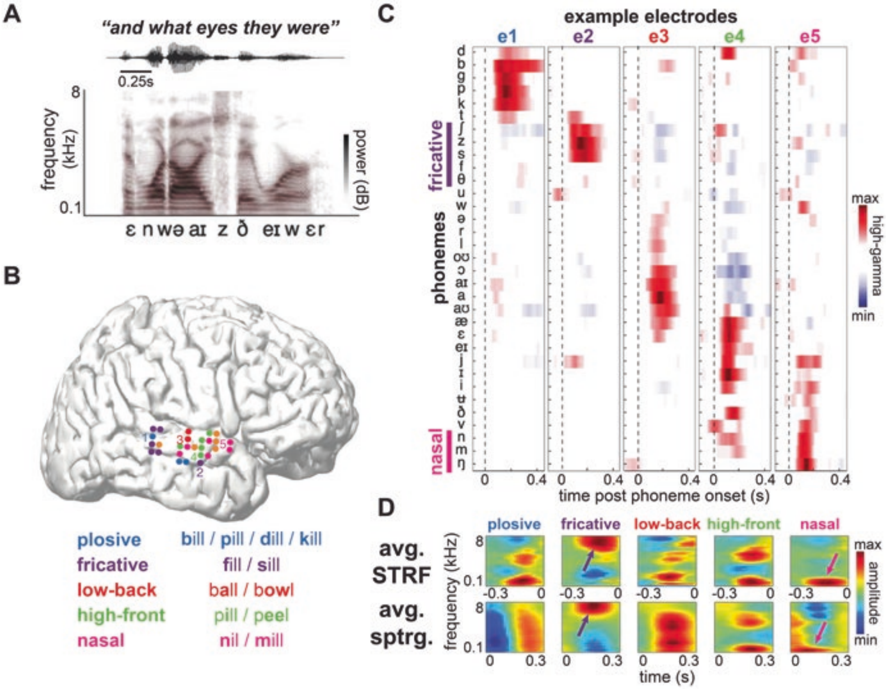

Fig. 3.3 Local neural populations on STG represent phonetic features. (**a**) Example stimulus sentence waveform and spectrogram, as used by Mesgarani et al. (2014). (**b**) STG electrodes in an example participant, colored by phonetic features represented on each location. (**c**) Average HG response magnitudes on single example electrodes, time-aligned to onset of single phonemes. Phonemes are sorted by manner of articulation. (**d**) Average spectrotemporal receptive fields (STRF, top row) across all electrodes that represented distinct manner of articulation, and spectrograms averaged across all phonemes in each feature class (bottom row). (Adapted from Mesgarani et al. (2014))

few phonemes. Importantly, though, the subset of phonemes that elicited activity on a single electrode was not random; rather, local responses reflected phonetic feature theory. For example, electrodes that responded to the phoneme /*t*/, also tended to respond to other plosives (e.g., /*d*/, /*k*/, /*p*/), but not to phonetically (and acoustically) dissimilar phonemes, such as fricatives (e.g., /*s*/, /*z*/, /*ʃ*/). Meanwhile, other spatially distinct neural populations exhibited tuning for other phonetic features, selectively responding to fricatives, nasals (e.g., /*m*/, /*n*/), or vowels with specific spectral properties (e.g., /*a*/ and /*ae*/ vs. /*i*/ and /*I*/).

A secondary STRF-based analysis confirmed the relationship between the apparent phonetic feature selectivity and the observed spectrotemporal sensitivities of these cortical sites. STRFs were fit for each electrode to the continuous high-gamma response throughout all phonemes in order to identify which spectrotemporal stimulus properties were associated with increased local neural activity. These STRFs were averaged across all electrodes that exhibited the same phonetic feature preference (plosive-selective, fricative-selective, etc.). Then, Mesgarani and colleagues computed average spectrograms of all phonemes belonging to a given phonetic feature class (plosives, fricatives, etc.). They found that the average spectrograms for phonemes of a given phonetic feature class and the average STRFs for electrodes tuned to that feature were well correlated. In other words, local neural populations exhibited complex STRF tuning that recapitulated complex spectrotemporal patterns that tend to co-occur in human speech and mark certain phonetic features. This finding was further corroborated by Chan et al. (2014) study of single neurons' responses to speech sounds in anterior STG, showing that these complex spectrotemporal response patterns reflect the tuning of single units. Overall, these results suggest that the ability to decode spectrotemporal details of the speech input from distributed patterns of high-gamma activity in STG arises from the local, selective encoding of spectrotemporally coherent phonetic features at spatially discrete cortical sites in STG.

3.2.8 Superior Temporal Gyrus Representation of Speech: Beyond Linear Spectrotemporal Encoding Models

So far, evidence has been presented to show that distributed and local patterns of neural activity in human STG represent the spectrotemporal properties of the speech signal. However, it is also widely recognized that the response in the STG is not a direct, linear transformation of the acoustic signal (Obleser and Eisner 2009; David 2018). That is, the assumption of a linear stimulus-response relationship that is common to many locally derived STRF models and to the population-level reconstruction approach of Pasley and colleagues (2012) is an oversimplification. Indeed, evidence for this fact is reported in both of the studies described above. For instance, Pasley and colleagues pointed out that stimulus spectrogram reconstruction was not uniformly accurate across all acoustic features, but, rather, that performance was particularly strong along spectral and temporal dimensions known to be critical to speech intelligibility (Chi et al. 1999; Elliott and Theunissen 2009). That is, the STG's encoding of speech is most reliable and detailed for acoustic features that are most important for speech comprehension. This observation suggests a provocative hypothesis: that the STG encodes an imperfect spectrotemporal representation of speech, sacrificing faithfulness to the sensory signal where it enhances the robust encoding of linguistically and/or behaviorally relevant information.

If true, this hypothesis would be consistent with at least two complementary predictions. First, the neural representation may remove some information present in the acoustic signal that is *not* relevant to understanding the meaning of speech, ultimately leading to the emergence of robust, acoustically invariant representations of behaviorally relevant properties in the bottom-up stimulus. Second, the neural representation may reflect the integration of information that is not present in the acoustic signal, but *is* relevant to decoding the meaning of speech, thereby incorporating reliable top-down cues to achieve a more robust perceptual representation. Understanding how the neural response of auditory cortex diverges from a straightforward linear encoding of the acoustic signal is likely to reveal novel insights about speech perception and its neurobiological bases. In Sect. 3.3, we describe a collection of studies that illustrate multiple domains in which cortical speech representations exhibit abstraction and invariance.

3.3 Invariance in Speech Encoding by Human Auditory Cortex

3.3.1 The Representation of An Attended Speech Stream Emerges in the Superior Temporal Gyrus

One of the most basic problems in speech neuroscience concerns the remarkable robustness of speech perception under noisy listening conditions. In order to perceive speech with a high degree of accuracy, listeners need to be able to rapidly "tune-out" background noise, even when the sound level of the signal is dwarfed by the noise (Luce and Pisoni 1998; Sarampalis et al. 2009). One popularly studied paradigm that showcases this faculty is referred to as the "cocktail party problem," so named because a listener's ability to hold a conversation at a raucous cocktail party exemplifies the noise-robustness of the speech perception system (Cherry 1953; McDermott 2009). In the classic experimental task that simulates this problem, a listener is simultaneously presented with two overlapping voices and told to selectively attend to one speaker's utterances while ignoring the ongoing "background" speech by the other speaker. From a sensory neuroscience perspective, this represents a daunting challenge, but healthy listeners rarely experience this "problem" as such and are typically unable to report any details about the content of the unattended speech stream.

An important question is whether the auditory cortical responses during "cocktail party" scenarios reflected the sensory signal (mixed speakers, as predicted by a linear spectrotemporal encoding model) or the listeners' noise-invariant perceptual experience (only the attended speaker, somehow filtering out the unattended distractor speech). In particular, Mesgarani and Chang (2012) recorded cortical activity with ECoG from the STG of patients while they listened to stimuli from each of three conditions (Fig. 3.4): a sentence spoken by Speaker 1 ("SP1"), a sentence spoken by Speaker 2 ("SP2"), or two distinct sentences spoken simultaneously by Speakers 1 and 2 ("MIX"). All of the sentences followed the same structure (e.g., *Ready RINGO, go to BLUE TWO now*), but with different combinations of call-sign (e.g., *RINGO* vs. *TIGER*), color (e.g., *BLUE* vs. *RED*), and number (e.g., *TWO* vs. *THREE*). On a given trial from the SP1 or SP2 condition (i.e., conditions with a single speaker), participants listened to a sentence and could easily identify which color and number the lone speaker had uttered. However, during MIX trials,

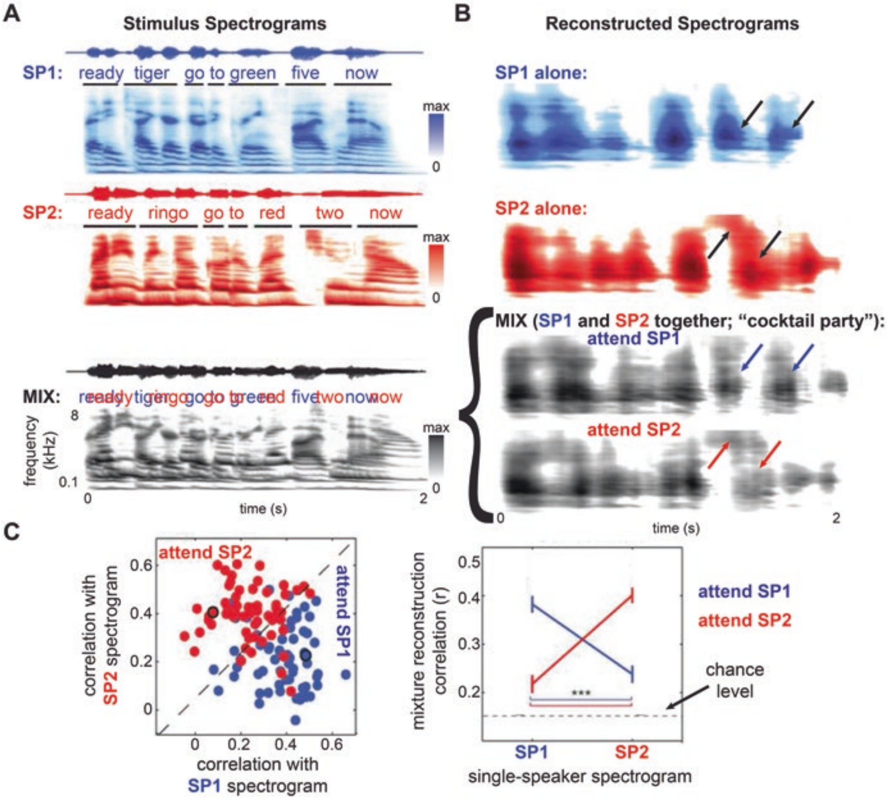

Fig. 3.4 Attention modulates STG representation of speech. (**a**) Spectrograms of examples stimulus sentences used by Mesgarani and Chang (2012), separately for speaker 1 (SP1), speaker 2 (SP2), and mixture stimulus, with both speakers overlayed. (**b**) Spectrograms reconstructed from STG neural activity in an example patient, for the same example sentences in single speaker (top) and MIX (bottom) conditions. In MIX condition, reconstructed spectrogram more closely reflected the spectrogram of the attended speaker. (**c**) Correlation between original and reconstructed spectrograms for single speaker 1 and speaker 2 sentences. Correlations are consistently higher for the attended than for the unattended speaker

listeners were faced with the more difficult "cocktail party" task, during which a visually presented target call-sign (e.g., *RINGO* or *TIGER*) indicated which voice the listener should tune into. For a given MIX trial, the speaker who uttered this target call-sign was the *attended* speaker, and the other was the *unattended* speaker. Despite the rapid attention-shifting demands, listeners were able to accurately report the color and number uttered by the attended speaker on 75% of MIX trials.

In order to evaluate the extent to which human auditory cortex encodes the spectrotemporal properties of each speaker (the attended/unattended speakers), or whether the encoding was faithful to the acoustic complexity of the signal in the MIX condition, Mesgarani and Chang employed a stimulus reconstruction approach (see Sect. 3.2.2). Their primary analysis revealed three key results. First, unsurprisingly, spectrogram reconstructions based on neural response recordings during the single-speaker conditions were close matches to the actual spectrograms on a trial-by-trial basis. For instance, the reconstruction of a SP1 (alone) spectrogram from the neural response to that stimulus was well correlated with the true SP1 (alone) stimulus spectrogram. Second, entirely different spectrogram reconstructions were produced when comparing the patterns of neural activity evoked in matched pairs of identical MIX trials that had the same sensory signal but differed as to which speaker was attended. This result offers clear evidence that the STG representation does not adhere to a strict spectrotemporal encoding model, which would have no means of predicting differential responses to the same acoustic input as a function of attention. Finally, and most strikingly, reconstruction of the SP1 stimulus spectrogram from the neural response to a MIX condition stimulus where the participant attended to Speaker 1 was just as good as the reconstruction from the neural response to SP1 only. In other words, the neural encoding of the attended speech was as robust when there was an unattended background speaker as when the attended speech was heard alone. This result provides direct evidence for noise-invariance in the STG representation of attended speech, a result which diverges sharply from a pure spectrotemporal encoding of the speech input's spectrogram. Overall, this and other related invasive and non-invasive electrophysiological work on the cocktail party effect in humans indicate that the STG's speech representation, although rooted in the sensory signal, also reflects the listener's perceptual experience of the signal (see also (Zion Golumbic et al. 2013).

A central question in speech neuroscience concerns the extent to which encoding of speech in STG differs from the encoding in other auditory cortical areas, most prominently in primary auditory cortex. To address this, O'Sullivan and colleagues (O'Sullivan et al. 2019) compared the effects of attention on speech representation in STG and A1, using the same cocktail party paradigm. To record neural responses from both areas, they employed a combination of ECoG grids on STG and depth electrodes along Heschl's gyrus. They found a clear dissociation between A1 and STG: Neural populations in STG represented the attended speaker and contained little information about the unattended speaker. In contrast, responses on individual electrodes in A1 preferred single speakers, independently of the attention manipulation. In other words, while STG representations are re-tuned by attentional demands, A1 neural populations are tuned to certain spectral content, e.g., low frequencies, and encode information about speakers that fit their receptive fields, e.g., speakers with a low voice, independently of attention. This study exemplifies how the representation of speech sound shifts from a veridical representation of sound content to a focus onto perceptually relevant aspects of the input along the auditory hierarchy (Nourski et al. 2019).

Ongoing work, including in humans (Holdgraf et al. 2016) and in animal models (Moore et al. 2013; Rabinowitz et al. 2013), aims to identify the precise neurophysiological mechanisms that could give rise to the flexible selection of auditory representations in human STG and the difference between processing of and adaptation to different types of background noise (Khalighinejad et al. 2019).

3.3.2 Invariant Representation of Phonetic Categories

Based on the evidence presented so far, it is clear that cortical responses in STG encode acoustic differences among different speech sounds. At a local level, this means that activity within discrete neural populations signals the presence of specific spectrotemporal patterns (phonetic features) at specific times (Mesgarani et al. 2014). At the population level, this distributed, feature-based encoding scheme makes it possible to decode which phonemes and words are present in the speech signal (Pasley et al. 2012). The ability to distinguish among phonemic categories is, of course, critical to the perceptual system's ability to extract meaning from the signal.

However, it is not clear from these findings to what extent auditory cortex also represents spectrotemporal differences *within* a speech sound category. One of the most foundational observations about human speech perception is that there is no one-to-one mapping from spectrograms to phonemes; any given phoneme can be acoustically realized in countless ways (Perkell and Klatt 1986; Diehl et al. 2004). For example, the specific spectral content of a vowel depends on its surrounding consonants, a phenomenon known as co-articulation.

Of course, the particular details that distinguish one exemplar of a category from another exemplar of that category will generally matter little to a listener whose goal is to understand the meaning of the incoming speech. In fact, listeners sometimes behave as if they do not perceive differences among distinct realizations of a single phoneme, as long as they all are perceived as belonging to the same category.

For instance, in a classic study by Liberman and colleagues (1957), a set of synthetic spectrograms was created that parametrically manipulated a single acoustic variable (initial second formant frequency), yielding a continuum of equally spaced syllables ranging perceptually from */ba/* to */da/* to */ga/*. However, when subjects labeled tokens along the continuum, their identification behavior revealed sharp nonlinearities in their perception of linearly spaced tokens along the continuum, suggesting that the perceptual representation of speech is not a direct linear mapping of spectrotemporal space. Moreover, when listeners were presented with a pair of syllables and asked only to decide whether they were the same or different, their discrimination performance was far more accurate if the two tokens corresponded to different phonetic labels (e.g., a */ba/* and a */da/*) than when they were members of the same category (e.g., two exemplars of */ba/*), even if both pairs were equally distant acoustically (i.e., had equal distances in initial second formant frequency). This nonlinear perceptual warping of acoustic space, a phenomenon known as *categorical perception* (Liberman et al. 1967), suggests that the auditory cortical representation of speech reflects not only a purely spectrotemporal encoding of the signal, but also its behavioral or linguistic relevance.

An ECoG study by Chang and colleagues (2010) aimed to explicitly test whether the auditory cortical representation of speech was categorical, consistent with the observed perceptual behavior, or sensitive to gradual acoustic variation, consistent with a linear spectrotemporal model (Fig. 3.5). Neural responses (LFPs) to syllables

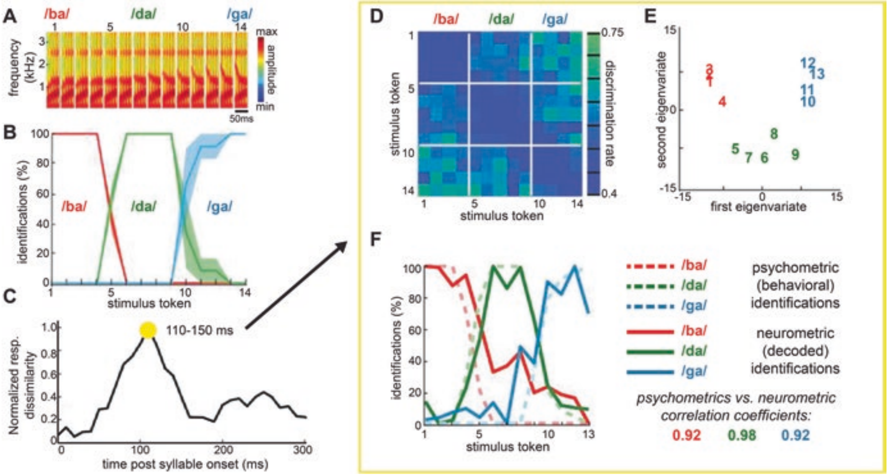

Fig. 3.5 Categorical phoneme representation on STG. (**A**) Stimuli on the /ba/-/da/-/ga/ continuum, employed by Chang et al. (2010). (**B**) Psychometric behavioral identification curves in study participants showed categorical perception of task stimuli. (**C**) Dissimilarity between neural responses to differently categorized stimulus tokens peaked at 110–150 ms post stimulus onset. Time series of the total normalized neural pattern dissimilarity derived from classifier performance aggregated across all pair-wise stimulus comparisons are shown. (**D**) Neural decoding confusion matrix in this time window. (**E**) Projection of neural dissimilarity into a two-dimensional neural space using multi-dimensional scaling (MDS) analysis of neural responses. (**F**) Psychometric and neurometric token identification curves overlapped, showing correlation between perceptual and neural representation of the phonemic continuum. Neurometric identification functions were determined using the MDS distance between each stimulus position and the three cluster means. (Adapted from Chang et al. (2010))

along a comparable linearly spaced acoustic continuum were recorded directly from STG, and pattern classification analyses were used to assess the discriminability of neural responses to different pairs of syllables. Results showed that the neural response pattern differed consistently between syllables belonging to different categories, with discriminability peaking ~110–150 ms after stimulus onset. However, discriminability among within-category differences was far poorer. In fact, neurally derived identification and discrimination functions closely matched the categorical psychometric functions derived from behavioral responses to the same stimuli. These results offer direct evidence for categorical representations of linguistically relevant phonetic information in human STG, and are consistent with converging evidence from non-invasive electrophysiological recordings (Dehaene-Lambertz 1997; Sharma and Dorman 1999) as well as intracranial studies of other phonetic cues, such as voice onset time, discriminating between voiced (/b/) and unvoiced (/p/) plosives (Fox et al. 2020).

3.3.3 Joint Representation of Phonetic and Subphonetic, Indexical Information

As evidence pointing to invariant representations of speech sound categories has grown, further research with non-invasive neuroimaging (Myers 2007; Toscano et al. 2018) and electrophysiology (Sharma et al. 1993; Frye et al. 2007) has shown that the STG's representation of speech is not purely categorical. Similarly, mounting behavioral evidence has made clear that listeners are, in fact, sensitive to fine-grained subphonetic differences in speech stimuli (Pisoni and Tash 1974). Although these results may at first seem to conflict with data supporting invariant speech encoding, it is important to note that categorical phonetic representations and graded subphonetic representations need not be considered mutually exclusive. In fact, modern theories of speech perception recognize the functional relevance of both representations, with categorical perception serving to promote access to stable phoneme categories that distinguish meanings (McClelland and Elman 1986; Grossberg 2003), while within-category sensitivity is a critical prerequisite for perceptual learning and robust cue integration, especially in noisy, unstable listening environments (Clayards et al. 2008; Norris and McQueen 2008).

One example of concurrent representation of phonetic and subphonetic information in speech concerns the neural encoding of so-called indexical information in speech (Abercrombie 1967; Pisoni 1997), or sources of variability that affect the acoustic realization of a phoneme but that are incidental to meaning. For instance, cues to speaker identity comprise a quintessential example of indexical information in the speech signal, because there is enormous variation in how different individuals tend to pronounce any given phoneme (Peterson and Barney 1952; Allen et al. 2003). For instance, vowels produced by taller speakers (with longer vocal tracts) tend to have lower first formants (first vocal tract resonance; F1) than vowels produced by shorter speakers. This sort of indexical (talker-dependent) variability poses a challenge to the speech perception system because the F1 property of vowels is also the main acoustic cue that distinguishes between certain vowel categories, such as /u/ (low F1) vs. /o/ (high F1) (Ladefoged and Johnson 2014). That is, acoustic cues to phoneme identity and to speaker identity are confounded in the speech signal; it may be impossible to know from the spectrogram alone whether a particular speech token was a tall person's /o/ or a short person's /u/ (Ladefoged 1989). Behavioral evidence shows that listeners solve this problem by leveraging contextual cues to compute "normalized" speech representations (Nearey 1989; Johnson 2005).

Although the neurophysiological mechanisms supporting speech sound normalization processes remain unclear, a study by Moses and colleagues (2016) suggests that the STG encoding of phonetic information is speaker invariant. The main goal of this study was to design a "neural speech recognition" system that could decode phonemes and sequences of phonemes from spatiotemporal patterns of high-gamma activity recorded with ECoG from the STG of patients listening to continuous speech (Pasley et al. 2012). One striking result reported by Moses et al. was that a phoneme-decoding model that was trained on neural responses to speech produced by one speaker (for instance, a woman) generalizes to speech produced by a different person (for instance, a man). That is, it was possible to accurately decode the phonemes in continuous speech uttered by a female speaker from neural activity, even when the stimulus-response mapping between phonemes and neural activity had originally been trained using neural data collected while the patient was hearing a male speaker, and vice versa. Since decoding performance was just as high whether the speaker at test matched the speaker during training or was new, this result suggests that the patterns of neural activity capable of robustly encoding phoneme identity are also speaker invariant. Given that Mesgarani et al. (2014) discovered phonetically selective cortical sites by comparing phoneme-evoked responses that were averaged across instances of a phoneme spoken by hundreds of different voices, Moses et al.'s result was, perhaps, foreseeable. Nonetheless, this result provides another example of invariance in the phonetic representation of speech that cannot be accounted for by a purely spectrotemporal cortical encoding of speech (Sjerps et al. 2019).

One key conclusion that can be drawn from the data that have been presented so far is that neural activity in human STG robustly encodes behaviorally relevant spectrotemporal information in the speech signal. The tuning properties of local neural populations seem optimized for detecting linguistically universal phonetic features that are critical for distinguishing among different meanings. This neural representation of the stimulus exhibits noise invariance, acoustic invariance, and speaker invariance – all of which are desirable properties for a system whose goal is to map a highly variable speech signal onto stable representations of meaning. Critically, though, the existence of a cortical representation that has these properties does not preclude the existence of cortical representations that are sensitive to subphonetic information like within-category acoustic detail or indexical speaker information, as discussed in the next section.

3.3.4 Intonational Prosody in Speech: Linguistic Meaning Beyond Phonetic Categories

Besides cues to phonetic category identity, another type of acoustic information in the speech signal that is relevant for understanding meaning is carried by vocal pitch (Cutler et al. 1997; Ladd 2008). In English, patterns of rising and falling pitch over the course of a sentence – or *intonational prosody* – can indicate whether a speaker is making a statement (flat or falling pitch contour) or asking a question (rising pitch at the end of the sentence), and emphasis on different words in a sentence, as cued by pitch accents, can change the meaning of a sentence without changing the phonemes or words (e.g., *Anna* likes oranges vs. Anna likes *oranges*). In many languages, such as Mandarin Chinese, pitch can also carry lexical information, distinguishing between meanings (words) which are phonetically identical (Howie

1976). Thus, listeners extract both phonetic cues and pitch cues from the signal to understand the meaning of some perceived speech (Shattuck-Hufnagel and Turk 1996).

However, just as acoustic cues that carry linguistically relevant phonetic information (e.g., vowel F1 frequency) also carry linguistically irrelevant information (e.g., about the length of the speaker's vocal tract), vocal pitch carries both linguistically relevant intonational information and linguistically irrelevant information (Van Dommelen 1990). In particular, the average pitch (fundamental frequency; F0) of a female speaker's voice is higher than that of a male speaker (Titze 1989). In order to extract the linguistically relevant information carried by changes in pitch, then, auditory cortex must track the *relative pitch* across a particular sentence rather than just the *absolute pitch* across the sentence. In this way, encoding the intonational contour of a sentence may require normalization at two temporal levels: Within single syllabic segments to extract speaker-invariant phonetic representations of speech sounds, and across syllabic segments, to extract prosodic information. Although behavioral evidence indicates that listeners map absolute pitch values in the signal onto relative pitch percepts (Gussenhoven et al. 1997; Wong and Diehl 2003), it has previously been unknown how auditory cortical regions implicated in vocal pitch perception encode this abstract component of linguistic meaning (Patterson et al. 2002; Bendor and Wang 2005).

3.3.5 Invariant Representation of Intonational Prosody

Tang and colleagues (2017) used ECoG to examine the neural encoding of intonational information in spoken sentences (Fig. 3.6). Neurophysiological responses were recorded while patients listened to synthesized spoken English sentences from a tightly controlled stimulus set in which a given sentence (e.g., *movies demand minimal energy*) was heard with each of four linguistically distinct intonational contours: an unaccented falling pitch contour (typical of declarative statements), rising pitch throughout *energy* (typical of a question), or containing a pitch accent emphasizing either the word *movies* or *minimal*. To compare the encoding of relative versus absolute pitch, each sentence was synthesized twice in each intonation condition, with the only difference being a shift in the absolute pitch, effectively simulating a male voice (low absolute pitch) and a female voice (high absolute pitch). Tang et al. repeated this process for several different sentences, thereby allowing for the examination of how linguistic information from intonational versus phonetic cues in speech are simultaneously represented by auditory cortex. In this way, the stimuli varied orthogonally in their relative pitch (intonational cues), absolute pitch (a speaker-dependent, nonlinguistic cue), and phonetic cues. By comparing neural responses to the sentences across stimulus conditions, Tang et al. systematically assessed whether and how the STG simultaneously represents these three different types of acoustic information in speech. The results revealed three key findings,

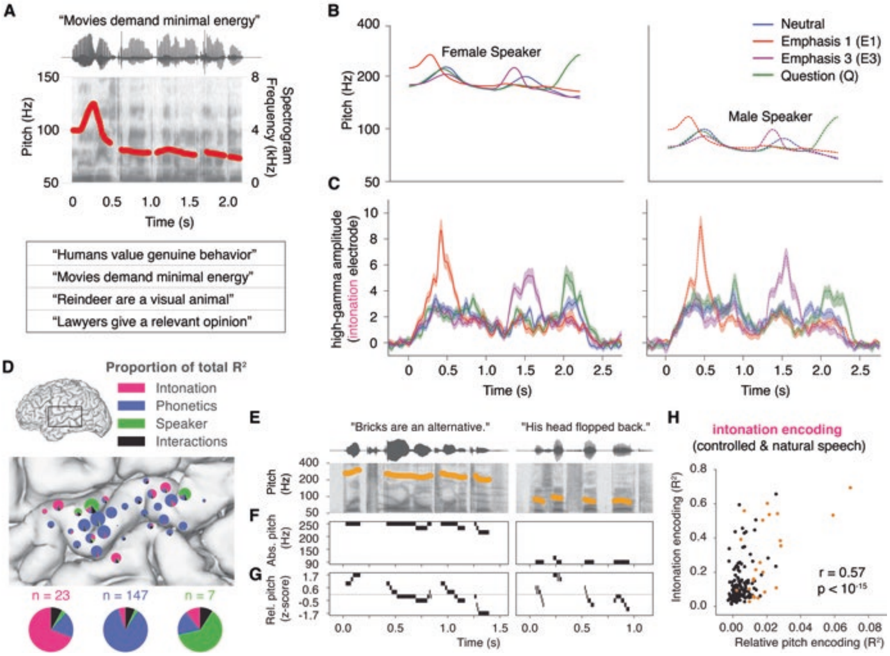

Fig. 3.6 Phonetic information, absolute pitch, and relative pitch are independently encoded on STG. (**A**) Waveform, spectrogram, and pitch contour for an example stimulus sentence used by Tang et al. (2017). (**B**) Pitch contours for all four conditions and different speakers in this experiment. (**C**) Neural responses on an example electrode show peaks in HG amplitude following deflections in stimulus pitch. (**D**) Top: Map of intonation, sentence, and speaker encoding for one participant, pie charts indicate relative percent of total variance explained by each type of predictor. Bottom: Proportion of variance explained by main effects and interactions across time points when the full model was significant, for all significant electrodes across all 10 participants. Each electrode was classified as either intonation, sentence, or speaker electrode, according to which stimulus dimension was maximally encoded. n: number of electrodes across patients. (**E**) Example tokens from the TIMIT speech corpus used in this study. (**F**) Absolute-pitch (in ln Hz) feature representation. Bins represent different values of absolute pitch. (**G**) Relative-pitch (z-scored ln Hz within speaker) feature representation. The gray line indicates a relative-pitch value of 0. (**H**) Encoding of relative pitch in TIMIT corpus and neural discriminability of intonation contours were correlated across electrodes. Colored markers show electrodes with significant (permutation $p < 0.05$) encoding of relative pitch. (Adapted from Tang et al. (2017))

relating to the cortical encoding of intonational, phonetic, and speaker-dependent information in speech, respectively.

First, they discovered neural populations in STG that track relative pitch (intonation contour), but not absolute pitch. That is, evoked high-gamma responses at these intonation-encoding cortical sites discriminated questions from statements and distinguished among sentences with different emphasis, but these were identical across male (low pitch) and female (high pitch) voices. Additional analyses employing *pitch temporal receptive fields* (cf. STRFs, Sect. 3.2.4) confirmed that these same intonation-encoding sites also tracked relative pitch during naturally spoken sentences. Moreover, the responses at intonation-encoding cortical sites were also invariant across sentences that had the same prosodic information but differed in their phonetic information. In sum, linguistically meaningful cues in vocal pitch are extracted from incoming speech and represented by neural activity in human STG in an abstract way (i.e., relative to a given speaker's pitch range).

Second, and concordant with the results of Mesgarani et al. (2014), Tang et al. (2017) identified another (larger) set of electrodes that responded differently to different phonemes in a feature-selective manner. Critically, the responses measured at these phonetically tuned electrodes did not vary based on speaker identity, nor did they vary based on intonation. That is, neural responses to unique instances of the same phoneme at these cortical sites remained stable, even though the acoustic realization of that phoneme was variable. Taken together with the first result, Tang et al. found direct evidence that abstract (i.e., speaker invariant) phoneme identity and intonational contour information are each represented independently in human STG by high-gamma responses at spatially distinct neural populations.

Finally, in addition to finding cortical sites that exhibited tuning preferences for abstract linguistic properties of speech, Tang et al. (2017) identified a third spatially distinct set of electrodes whose responses discriminated speaker identity. These electrodes' responses did not robustly encode phonetic or intonational differences among sentences. Together, these results point to a three-way dissociation between the encoding of phonetic, intonational, and nonlinguistic (speaker-dependent) information in the speech stream.

3.3.6 Distinct Parallel Processing Streams in Superior Temporal Gyrus Encode Invariant Speech Representations Relevant for Meaning

Overall, the studies discussed in the previous sections provide persuasive evidence for abstraction from a detailed spectrotemporal representations toward a representation tuned to perceptually relevant sound dimensions as an emerging principle of the STG representation of speech. Parallel results emerged for two qualitatively different classes of meaningful spectral cues in speech – phonetic and intonational cues, operating at segmental and suprasegmental scales, respectively. In both cases, evoked neural activity in human STG seems to encode linguistically relevant features of the speech input. That is, there exist spatially discrete neural populations that robustly represent acoustic differences among speech stimuli *only* when those differences were meaningful. This result is conceptually linked with results supporting categorical neural representations of phoneme identity (Chang et al. 2010; Steinschneider et al. 2011) and speaker-invariant phonetic encoding (Mesgarani et al. 2014; Moses et al. 2016), which suggest that acoustic differences among speech sounds are only represented when they are linguistically meaningful (i.e., when they are exemplars of different phonemes). Thus, the current evidence shows that neural activity in STG reflects linguistic abstraction in its phonetically invariant and speaker-invariant encoding of speech stimuli.

Finally, as exemplified by the final (speaker encoding) result of Tang et al. (2017), evidence for abstraction or invariance of cortical representation does not preclude simultaneous, parallel encoding that is non-invariant. Indeed, the observation that these representations are spatially segregated may be a key insight into the neurophysiological mechanisms by which the auditory pathway maps spectrotemporal modulations in speech onto behaviorally relevant abstract categories that allow listeners to access meaning, while still retaining more detailed information that has been shown to affect behavior. Invariant and detailed representations of speech are not mutually exclusive, either in theory or according to the empirical data, but the emergence of invariance is critical for understanding the role of STG in speech perception and comprehension.

3.4 The Integration of Top-Down and Bottom-Up Information During Speech Perception in Human Superior Temporal Gyrus

Although the bottom-up mapping acoustic cues in speech onto linguistically relevant representations is at the core of speech perception, the sensory signal is not the only source of information that is relevant to understanding meaning. Because listeners have extensive experience with the structure and statistics of language, their decoding of the speech they hear is informed by their expectations. It is easy to see how integrating so-called "top-down" information about what words or sounds may be more likely to facilitate comprehension of incoming speech when the input is noisy or ambiguous. Indeed, behavioral evidence suggests that listeners are sensitive to such information not only when faced with perceptual ambiguity (e.g., lexical effects on phoneme identification; Ganong 1980; McClelland and Elman 1986), but even when the acoustic signal is clear (e.g., lexical frequency effects on recognition accuracy and speed; Marslen-Wilson 1987).

It is, however, largely unknown how or to what extent prior knowledge and expectations constrain or alter the cortical representations of incoming speech sounds. Recent work with ECoG has begun to close this gap by directly quantifying how cortical activity in STG is affected by information from other sources during the perception of spoken words. However, additional ECoG evidence shows that neurophysiological responses in human STG are modulated by the statistics of speech sound sequences (phonotactics: (Leonard et al. 2015)), speech sound similarity between words (lexical cohort size: (Cibelli et al. 2015)), prior expectations about whether an auditory input contains speech sounds (recognition of degraded speech: (Holdgraf et al. 2016; Khoshkhoo et al. 2018), or visual speech (Micheli et al. 2018; Karas et al. 2019). All of these information sources provide crucial support for robust recognition of speech under adverse listening conditions, when recognition based purely on the incoming acoustic stimulus may not be possible.

In this chapter, we focused on the superior temporal lobe, whose contribution to speech comprehension has been studied most extensively with iEEG thus far, not least due to clinical limitations on electrode placement. However, other brain regions also contribute to speech comprehension, and a growing number of studies is investigating their role. For example, intracranial recordings found that speech sound representations emerge with comprehension in the inferior frontal gyrus (Leonard et al. 2016; Khoshkhoo et al. 2018) and are evident in sensorimotor cortex (Cheung et al. 2016). Additional studies are required to understand the nature of the interaction between these regions and the auditory cortex during speech comprehension. Complimentary, while our focus here was on speech sound representations in the STG, a growing body of evidence suggests that the STG contributes to additional processes related to speech and language, such as error monitoring during speech production (Chang et al. 2013) and morphological processing (Lee et al. 2018).

3.5 Concluding Remarks

In summary, the STG emerges as a central computational hub, critical to the extraction of linguistically and perceptually relevant information from continuous sounds. The comparison of representations between the STG and A1 shows that speaker-independent representation of speech-relevant spectrotemporal sound patterns is unique to the STG, and not present at earlier stages of auditory processing. Intracranial studies to date were mostly focused on the representation of single phonetic units. One important focus for future studies will be to address the question of how the STG integrates across single phonemes, toward the representation of sequences of phonetic and syllabic units. Finally, we want to mention two central questions that have been discussed in the past.

First, none of the above studies found evidence for a meso-anatomical organization of neural populations tuned to different types of spectral information in speech. This is markedly different from the tonotopic, retinotopic, and somatotopic organization of primary sensory cortices. Rather, neural populations representing different speech features are intermixed in mid-posterior STG. Several recent studies, however, discovered a mesoscopic spatial organization that emerges with respect to the representation of temporal landmarks in speech: while populations in posterior STG respond to speech onsets at the phrasal level (Hamilton et al. 2018; Forseth et al. 2020), populations in mid-STG encode the timing of syllabic nucleus onsets, so-called acoustic edges in speech (Oganian and Chang 2019). Future studies will need to elucidate how temporal and phonetic information is integrated during speech perception, and why only the former is organized anatomically along the posterior-to-anterior axis of the STG.

Second, we chose to focus on aspects of speech processing in STG that appear largely shared between the two hemispheres. In fact, none of the studies described in this chapter found hemispheric differences between representations. This stands in contrast to findings from lesion studies and cortical stimulation, where only damage to the dominant (typically left) hemisphere impairs speech comprehension. This discrepancy suggests that while bilateral temporal lobes contain comparable representations of speech, only the left hemispheric representations are necessary for comprehension in the healthy brain, possibly reflected in hemispheric differences in sensitivity to spectral and temporal modulations (Flinker et al. 2019). Future studies will be required to elucidate possible hemispheric differences in more detail, as well as to study the contribution of non-dominant representations to speech comprehension.

Compliance with Ethics Requirements Neal P. Fox declares that he has no conflicts of interest.

Yulia Oganian declares that she has no conflicts of interest.
Edward F. Chang declares that he has no conflicts of interest.

References

Abercrombie D (1967) Elements of general phonetics. Aldine, Chicago

Allen JS, Miller JL, DeSteno D (2003) Individual talker differences in voice-onset-time. J Acoust Soc Am 113(1):544. https://doi.org/10.1121/1.1528172

Assmann P, Summerfield Q (2004) The perception of speech under adverse conditions. In: Greenberg S et al (eds) Speech processing in the auditory system. Springer, New York, pp 231–308. https://doi.org/10.1007/0-387-21575-1_5

Bendor D, Wang X (2005) The neuronal representation of pitch in primate auditory cortex. Nature 436(7054):1161–1165. https://doi.org/10.1038/nature03867

Berezutskaya J, Freudenburg ZV, Güçlü U et al (2017) Neural tuning to low-level features of speech throughout the perisylvian cortex. J Neurosci 37(33):7906–7920. https://doi.org/10.1523/JNEUROSCI.0238-17.2017

Bialek W, Rieke F, de Ruyter van Steveninck RR et al (1991) Reading a neural code. Science 252(5014):1854–1857. https://doi.org/10.1126/SCIENCE.2063199

Bitterman Y, Mukamel R, Malach R et al (2008) Ultra-fine frequency tuning revealed in single neurons of human auditory cortex. Nature 451(7175):197–201. https://doi.org/10.1038/nature06476

Bolanowski SJ, Gescheider GA, Verrillo RT et al (1988) Four channels mediate the mechanical aspects of touch. J Acoust Soc Am 84(5):1680–1694. https://doi.org/10.1121/1.397184

Brugge JF (1992) An overview of central auditory processing. In: The mammalian auditory pathway: neurophysiology, vol 2. Springer, New York, pp 1–33. https://doi.org/10.1007/978-1-4612-2838-7_1

Buzsáki G, Anastassiou CA, Koch C (2012) The origin of extracellular fields and currents — EEG, ECoG, LFP and spikes. Nat Rev Neurosci 13(6):407–420. https://doi.org/10.1038/nrn3241

Chan AM, Dykstra AR, Jayaram V et al (2014) Speech-specific tuning of neurons in human superior temporal gyrus. Cereb Cortex 24(10):2679–2693. https://doi.org/10.1093/cercor/bht127

Chang EF (2015) Towards large-scale, human-based, mesoscopic neurotechnologies. Neuron 86(1):68–78

Chang EF, Rieger JW, Johnson K et al (2010) Categorical speech representation in human superior temporal gyrus. Nat Neurosci 13(11):1428–1432. https://doi.org/10.1038/nn.2641

Chang EF, Niziolek CA, Knight RT et al (2013) Human cortical sensorimotor network underlying feedback control of vocal pitch. PNAS 110(7):2653–2658. https://doi.org/10.1073/pnas.1216827110

Cherry EC (1953) Some experiments on the recognition of speech, with one and with two ears. J Acoust Soc Am 25(5):975–979. https://doi.org/10.1121/1.1907229

Cheung C, Hamilton LS, Johnson K et al (2016) The auditory representation of speech sounds in human motor cortex. elife 5:1–19. https://doi.org/10.7554/eLife.12577

Chi T, Gao Y, Guyton MC et al (1999) Spectro-temporal modulation transfer functions and speech intelligibility. J Acoust Soc Am 106(5):2719–2732. https://doi.org/10.1121/1.428100

Chomsky N, Halle M (1968) The sound pattern of English. Harper & Row, New York

Cibelli ES, Leonard MK, Johnson K et al (2015) The influence of lexical statistics on temporal lobe cortical dynamics during spoken word listening. Brain Lang 147:66–75. https://doi.org/10.1016/j.bandl.2015.05.005

Clayards MA, Tanenhaus MK, Aslin RN et al (2008) Perception of speech reflects optimal use of probabilistic speech cues. Cognition 108(3):804–809. https://doi.org/10.1016/j.cognition.2008.04.004

Crone NE, Miglioretti DL, Gordon B et al (1998) Functional mapping of human sensorimotor cortex with electrocorticographic spectral analysis. II. Event-related synchronization in the gamma band. Brain 121(12):2301–2315. https://doi.org/10.1093/brain/121.12.2301

Crone NE, Hao L, Hart J et al (2001) Electrocorticographic gamma activity during word production in spoken and sign language. Neurology 57(11):2045–2053

Crone NE, Sinai A, Korzeniewska A (2006) High-frequency gamma oscillations and human brain mapping with electrocorticography. Prog Brain Res 159:275–295. https://doi.org/10.1016/S0079-6123(06)59019-3

Cutler A, Dahan D, van Donselaar W (1997) Prosody in the comprehension of spoken language: a literature review. Lang Speech 40(2):141–201. https://doi.org/10.1177/002383099704000203

David SV (2018) Incorporating behavioral and sensory context into spectro-temporal models of auditory encoding. Hear Res 360:107–123. https://doi.org/10.1016/J.HEARES.2017.12.021

David SV, Mesgarani N, Shamma SA (2007) Estimating sparse spectro-temporal receptive fields with natural stimuli. Netw Comput Neural Syst 18(3):191–212. https://doi.org/10.1080/09548980701609235

Davis MH, Johnsrude IS (2007) Hearing speech sounds: top-down influences on the interface between audition and speech perception. Hear Res 229(1–2):132–147

de Saussure F (1916) Nature of the linguistic sign. In: Bally C, Sechehaye A (eds) Cours de linguistique générale. McGraw Hill Education

deCharms RC, Blake DT, Merzenich MM (1998) Optimizing sound features for cortical neurons. Science 280(5368):1439–1443. https://doi.org/10.1126/SCIENCE.280.5368.1439

Dehaene-Lambertz G (1997) Electrophysiological correlates of categorical phoneme perception in adults. Neuroreport 8(4):919–924. https://doi.org/10.1097/00001756-199703030-00021

Depireux DA, Simon JZ, Klein DJ, Shamma SA (2001) Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex. J Neurophysiol 85(3):1220–1234. https://doi.org/10.1152/jn.2001.85.3.1220

DeWitt I, Rauschecker JP (2012) Phoneme and word recognition in the auditory ventral stream. PNAS 109(8):E505–E514. https://doi.org/10.1073/pnas.1113427109

Di Liberto GM, O'Sullivan JA, Lalor EC (2015) Low-frequency cortical entrainment to speech reflects phoneme-level processing. Curr Biol 25(19):2457–2465. https://doi.org/10.1016/j.cub.2015.08.030

Diehl RL, Lotto AJ, Holt LL (2004) Speech perception. Annu Rev Psychol 55(1):149–179. https://doi.org/10.1146/annurev.psych.55.090902.142028

Ding N, Simon JZ (2012) Emergence of neural encoding of auditory objects while listening to competing speakers. PNAS 109(29):11854–11859. https://doi.org/10.1073/pnas.1205381109

Donders FC (1969) On the speed of mental processes. Acta Psychol 30:412–431. https://doi.org/10.1016/0001-6918(69)90065-1

Einevoll GT, Kayser C, Logothetis NK, Panzeri S (2013) Modelling and analysis of local field potentials for studying the function of cortical circuits. Nat Rev Neurosci 14(11):770–785. https://doi.org/10.1038/nrn3599

Elliott TM, Theunissen FE (2009) The modulation transfer function for speech intelligibility. PLoS Comput Biol 5(3):e1000302. https://doi.org/10.1371/journal.pcbi.1000302

Field DJ (1994) What is the goal of sensory coding? Neural Comput 6(4):559–601. https://doi.org/10.1162/neco.1994.6.4.559

Flinker A, Chang EF, Kirsch HE et al (2010) Single-trial speech suppression of auditory cortex activity in humans. J Neurosci 30(49):16643–16650. https://doi.org/10.1523/JNEUROSCI.1809-10.2010

Flinker A, Doyle WK, Mehta AD et al (2019) Spectrotemporal modulation provides a unifying framework for auditory cortical asymmetries. Nat Hum Behav 3(April):393–405. https://doi.org/10.1038/s41562-019-0548-z

Forseth KJ, Hickok G, Rollo PS, Tandon N (2020) Language prediction mechanisms in human auditory cortex. Nat Commun 11(1):1–14. https://doi.org/10.1038/s41467-020-19010-6

Fox NP, Leonard MK, Sjerps MJ, Chang EF (2020) Transformation of a temporal speech cue to a spatial neural code in human auditory cortex. elife 9:1–43. https://doi.org/10.7554/ELIFE.53051

Frye RE, Fisher JM, Coty A et al (2007) Linear coding of voice onset time. J Cogn Neurosci 19(9):1476–1487. https://doi.org/10.1162/jocn.2007.19.9.1476

Ganong WF (1980) Phonetic categorization in auditory word perception. J Exp Psychol Hum Percept Perform 6(1):110–125. https://doi.org/10.1037/0096-1523.6.1.110

Garofolo JS, Lamel LF, Fisher WM et al (1993) TIMIT acoustic-phonetic continuous speech corpus LDC93S1. Linguistic Data Consortium, Philadelphia

Griffiths TD, Kumar S, Sedley W et al (2010) Direct recordings of pitch responses from human auditory cortex. Curr Biol 20(12):1128–1132. https://doi.org/10.1016/J.CUB.2010.04.044

Grossberg S (2003) Resonant neural dynamics of speech perception. J Phon 31(3–4):423–445. https://doi.org/10.1016/S0095-4470(03)00051-2

Gussenhoven C, Repp BH, Rietveld A, Rump HH, Terken J (1997) The perceptual prominence of fundamental frequency peaks. J Acoust Soc Am 102(5):3009–3022. https://doi.org/10.1121/1.420355

Hamilton LS, Huth AG (2018) The revolution will not be controlled: natural stimuli in speech neuroscience. Lang Cogn Neurosci 35(5):573–582. https://doi.org/10.1080/23273798.2018.1499946

Hamilton LS, Edwards E, Chang EF (2018) A spatial map of onset and sustained responses to speech in the human superior temporal gyrus. Curr Biol 28(12):1860–1871.e4. https://doi.org/10.1016/j.cub.2018.04.033

Herff C, Schultz T (2016) Automatic speech recognition from neural signals: a focused review. Front Neurosci 10:429. https://doi.org/10.3389/fnins.2016.00429

Hickok G, Poeppel D (2007) The cortical organization of speech processing. Nat Rev Neurosci 8(5):393–402. https://doi.org/10.1038/nrn2113

Holdgraf CR, de Heer W, Pasley BN et al (2016) Rapid tuning shifts in human auditory cortex enhance speech intelligibility. Nat Commun 7(May):13654. https://doi.org/10.1038/ncomms13654

Holdgraf CR, Rieger JW, Micheli C, Martin S, Knight RT, Theunissen FE (2017) Encoding and decoding models in cognitive electrophysiology. Front Syst Neurosci 11(September):61. https://doi.org/10.3389/fnsys.2017.00061

Howard MA, Volkov IO, Mirsky R (2000) Auditory cortex on the human posterior superior temporal gyrus. J Comp Neurol 416(1):79–92

Howie JM (1976) Acoustical studies of Mandarin vowels and tones. Cambridge University Press, New York

Hubel DH, Wiesel TN (1962) Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. J Physiol 160(1):106–154. https://doi.org/10.1113/jphysiol.1962.sp006837

Hullett PW, Hamilton LS, Mesgarani N, Schreiner CE, Chang EF (2016) Human superior temporal gyrus organization of spectrotemporal modulation tuning derived from speech stimuli. J Neurosci 36(6):2014–2026. https://doi.org/10.1523/JNEUROSCI.1779-15.2016

Huth AG, de Heer WA, Griffiths TL, Theunissen FE, Gallant JL (2016) Natural speech reveals the semantic maps that tile human cerebral cortex. Nature 532(7600):453–458. https://doi.org/10.1038/nature17637

Jakobson R, Fant CGM, Halle M (1951) Preliminaries to speech analysis: the distinctive features and their correlates. MIT Press, Cambridge

Johnson K (2005) Speaker normalization in speech perception. In: Handbook of speech perception. Blackwell, pp 363–389

Johnson EL, Kam JWY, Tzovara A, Knight RT (2020) Insights into human cognition from intracranial EEG: a review of audition, memory, internal cognition, and causality. J Neural Eng 17(5):051001. https://doi.org/10.1088/1741-2552/abb7a5

Karas PJ, Magnotti JF, Metzger BA et al (2019) The visual speech head start improves perception and reduces superior temporal cortex responses to auditory speech. elife 8:1–19. https://doi.org/10.7554/eLife.48116

Khalighinejad B, da Silva GC, Mesgarani N (2017) Dynamic encoding of acoustic features in neural responses to continuous speech. J Neurosci 37(8):2176–2185. https://doi.org/10.1523/JNEUROSCI.2383-16.2017

Khalighinejad B, Herrero JL, Mehta AD, Mesgarani N (2019) Adaptation of the human auditory cortex to changing background noise. Nat Commun 10(1):1–11. https://doi.org/10.1038/s41467-019-10611-4

Khoshkhoo S, Leonard MK, Mesgarani N, Chang EF (2018) Neural correlates of sine-wave speech intelligibility in human frontal and temporal cortex. Brain Lang 187:83–91. https://doi.org/10.1016/j.bandl.2018.01.007

Klein DJ, Depireux DA, Simon JZ, Shamma SA (2000) Robust spectrotemporal reverse correlation for the auditory system: optimizing stimulus design. J Comput Neurosci 9(1):85–111. https://doi.org/10.1023/A:1008990412183

Kluender KR, Lotto AJ, Holt LL (2005) Contributions of nonhuman animal models to understanding human speech perception. In: Greenberg S, Ainsworth W (eds) Listening to speech: an auditory perspective. Oxford University Press, New York, pp 203–220

Kuhl PK (1986) Theoretical contributions of tests on animals to the special-mechanisms debate in speech. Exp Biol 45(3):233–265

Ladd DR (2008) Intonational phonology. Cambridge University Press, New York

Ladefoged P (1989) A note on "Information conveyed by vowels". J Acoust Soc Am 85:2223–2224

Ladefoged P, Johnson K (2014) A course in phonetics. Nelson Education

Lee DK, Fedorenko E, Simon MV et al (2018) Neural encoding and production of functional morphemes in the posterior temporal lobe. Nat Commun 9(1):1–12. https://doi.org/10.1038/s41467-018-04235-3

Leonard MK, Bouchard KE, Tang C, Chang EF (2015) Dynamic encoding of speech sequence probability in human temporal cortex. J Neurosci 35(18):7203–7214. https://doi.org/10.1523/JNEUROSCI.4100-14.2015

Leonard MK, Baud MO, Sjerps MJ, Chang EF (2016) Perceptual restoration of masked speech in human cortex. Nat Commun 7:13619. https://doi.org/10.1038/ncomms13619

Łęski S, Lindén H, Tetzlaff T, Pettersen KH, Einevoll GT (2013) Frequency dependence of signal power and spatial reach of the local field potential. PLoS Comput Biol 9(7):e1003137. https://doi.org/10.1371/journal.pcbi.1003137

Leszczyński M, Barczak A, Kajikawa Y et al (2019) Dissociation of broadband high-frequency activity and neuronal firing in the neocortex. BioRxiv (August):1–13. https://doi.org/10.1101/531368

Liberman AM, Harris KS, Hoffman HS, Griffith BC (1957) The discrimination of speech sounds within and across phoneme boundaries. J Exp Psychol 54(5):358–368. https://doi.org/10.1037/h0044417

Liberman AM, Cooper FS, Shankweiler DP, Studdert-Kennedy M (1967) Perception of the speech code. Psychol Rev 74(6):431–461. https://doi.org/10.1037/h0020279

Liebenthal E, Ellingson ML, Spanaki MV, Prieto TE, Ropella KM, Binder JR (2003) Simultaneous ERP and fMRI of the auditory cortex in a passive oddball paradigm. NeuroImage 19(4):1395–1404. https://doi.org/10.1016/S1053-8119(03)00228-3

Luce PA, Pisoni DB (1998) Recognizing spoken words: the Neighborhood Activation Model. Ear Hear 19(1):1–36

Marslen-Wilson WD (1987) Functional parallelism in spoken word-recognition. Cognition 25(1–2):71–102. https://doi.org/10.1016/0010-0277(87)90005-9

Mattys SL, Davis MH, Bradlow AR, Scott SK (2012) Speech recognition in adverse conditions: a review. Lang Cogn Process 27(7–8):953–978. https://doi.org/10.1080/01690965.2012.705006

McClelland JL, Elman JL (1986) The TRACE model of speech perception. Cogn Psychol 18(1):1–86. https://doi.org/10.1016/0010-0285(86)90015-0

McDermott JH (2009) The cocktail party problem. Curr Biol 19(22):R1024–R1027. https://doi.org/10.1016/j.cub.2009.09.005

Menon V, Freeman WJ, Cutillo BA et al (1996) Spatio-temporal correlations in human gamma band electrocorticograms. Electroencephalogr Clin Neurophysiol 98(2):89–102. https://doi.org/10.1016/0013-4694(95)00206-5

Merzenich MM, Brugge JF (1973) Representation of the cochlear partition on the superior temporal plane of the macaque monkey. Brain Res 50(2):275–296. https://doi.org/10.1016/0006-8993(73)90731-2

Merzenich MM, Knight PL, Roth GL (1975) Representation of cochlea within primary auditory cortex in the cat. J Neurophysiol 38(2):231–249. https://doi.org/10.1152/jn.1975.38.2.231

Mesgarani N, Chang EF (2012) Selective cortical representation of attended speaker in multi-talker speech perception. Nature 485(7397):233–236. https://doi.org/10.1038/nature11020

Mesgarani N, David SV, Fritz JB, Shamma SA (2009) Influence of context and behavior on stimulus reconstruction from neural activity in primary auditory cortex. J Neurophysiol 102(6):3329–3339. https://doi.org/10.1152/jn.91128.2008

Mesgarani N, Cheung C, Johnson K, Chang EF (2014) Phonetic feature encoding in human superior temporal gyrus. Science 343(6174):1006–1010. https://doi.org/10.1126/science.1245994

Micheli C, Schepers IM, Ozker M, Yoshor D, Beauchamp MS, Rieger JW (2018) Electrocorticography reveals continuous auditory and visual speech tracking in temporal and occipital cortex. Eur J Neurosci 51(5):1364–1376. https://doi.org/10.1111/ejn.13992

Mitchell TM, Shinkareva SV, Carlson A et al (2008) Predicting human brain activity associated with the meanings of nouns. Science 320(5880):1191–1195. https://doi.org/10.1126/science.1152876

Moore RC, Lee T, Theunissen FE (2013) Noise-invariant neurons in the avian auditory cortex: hearing the song in noise. PLoS Comput Biol 9(3):e1002942. https://doi.org/10.1371/journal.pcbi.1002942

Moses DA, Mesgarani N, Leonard MK, Chang EF (2016) Neural speech recognition: continuous phoneme decoding using spatiotemporal representations of human cortical activity. J Neural Eng 13(5):056004. https://doi.org/10.1088/1741-2560/13/5/056004

Mukamel R, Fried I (2012) Human intracranial recordings and cognitive neuroscience. Annu Rev Psychol 63(1):511–537. https://doi.org/10.1146/annurev-psych-120709-145401

Myers EB (2007) Dissociable effects of phonetic competition and category typicality in a phonetic categorization task: an fMRI investigation. Neuropsychologia 45(7):1463–1473

Näätänen R (2001) The perception of speech sounds by the human brain as reflected by the mismatch negativity (MMN) and its magnetic equivalent (MMNm). Psychophysiology 38(1):1–21. https://doi.org/10.1111/1469-8986.3810001

Näätänen R, Picton T (1987) The N1 wave of the human electric and magnetic response to sound: a review and an analysis of the component structure. Psychophysiology 24(4):375–425. https://doi.org/10.1111/j.1469-8986.1987.tb00311.x

Näätänen R, Paavilainen P, Rinne T, Alho K (2007) The mismatch negativity (MMN) in basic research of central auditory processing: a review. Clin Neurophysiol 118(12):2544–2590. https://doi.org/10.1016/j.clinph.2007.04.026

Nearey TM (1989) Static, dynamic, and relational properties in vowel perception. J Acoust Soc Am 85(5):2088. https://doi.org/10.1121/1.397861

Nelken I, Fishbach A, Las L, Ulanovsky N, Farkas D (2003) Primary auditory cortex of cats: feature detection or something else? Biol Cybern 89(5):397–406. https://doi.org/10.1007/s00422-003-0445-3

Norris D, McQueen JM (2008) Shortlist B: a Bayesian model of continuous speech recognition. Psychol Rev 115(2):357–395. https://doi.org/10.1037/0033-295X.115.2.357

Nourski KV, Steinschneider M, Rhone AE, Kovach CK, Kawasaki H, Howard MA (2019) Differential responses to spectrally degraded speech within human auditory cortex: an intracranial electrophysiology study. Hear Res 371:53–65. https://doi.org/10.1016/j.heares.2018.11.009

O'Sullivan JA, Herrero J, Smith E et al (2019) Hierarchical encoding of attended auditory objects in multi-talker speech perception. Neuron 104(6):1195–1209.e3. https://doi.org/10.1016/j.neuron.2019.09.007

Obleser J, Eisner F (2009) Pre-lexical abstraction of speech in the auditory cortex. Trends Cogn Sci 13(1):14–19. https://doi.org/10.1016/J.TICS.2008.09.005

Oganian Y, Chang EF (2019) A speech envelope landmark for syllable encoding in human superior temporal gyrus. Sci Adv 5(11):eaay6279. https://doi.org/10.1126/sciadv.aay6279

Ojemann GA (1987) Surgical therapy for medically intractable epilepsy. J Neurosurg 66(4):489–499. https://doi.org/10.3171/jns.1987.66.4.0489

Parvizi J, Kastner S (2018) Promises and limitations of human intracranial electroencephalography. Nat Neurosci 21:474–483. https://doi.org/10.1038/s41593-018-0108-2

Pasley BN, David SV, Mesgarani N et al (2012) Reconstructing speech from human auditory cortex. PLoS Biol 10(1):e1001251. https://doi.org/10.1371/journal.pbio.1001251

Patterson RD, Uppenkamp S, Johnsrude IS, Griffiths TD (2002) The processing of temporal pitch and melody information in auditory cortex. Neuron 36(4):767–776. https://doi.org/10.1016/S0896-6273(02)01060-7

Perkell JS, Klatt DH (1986) Invariance and variability in speech processes. Lawrence Erlbaum, Hillsdale

Pesaran B, Vinck M, Einevoll GT (2018) Investigating large-scale brain dynamics using field potential recordings: analysis and interpretation. Nat Neurosci 21(7):903–919. https://doi.org/10.1038/s41593-018-0171-8

Peterson GE, Barney HL (1952) Control methods used in a study of the vowels. J Acoust Soc Am 24(2):175–184. https://doi.org/10.1121/1.1906875

Pisoni DB (1997) Some thoughts on "normalization" in speech perception. In: Johnson K, Mullennix JW (eds) Talker variability in speech processing. Academic Press, San Diego, pp 9–32

Pisoni DB, Tash J (1974) Reaction times to comparisons within and across phonetic categories. Percept Psychophys 15(2):285–290

Rabinowitz NC, Willmore BDB, King AJ, Schnupp JWH (2013) Constructing noise-invariant representations of sound in the auditory pathway. PLoS Biol 11(11):e1001710. https://doi.org/10.1371/journal.pbio.1001710

Ramirez AD, Ahmadian Y, Schumacher J (2011) Incorporating naturalistic correlation structure improves spectrogram reconstruction from neuronal activity in the songbird auditory midbrain. J Neurosci 31(10):3828–3842. https://doi.org/10.1523/JNEUROSCI.3256-10.2011

Rauschecker JP, Scott SK (2009) Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. Nat Neurosci 12(6):718–724. https://doi.org/10.1038/nn.2331

Ray S, Maunsell JHR (2011) Different origins of gamma rhythm and high-gamma activity in macaque visual cortex. PLoS Biol 9(4):e1000610. https://doi.org/10.1371/journal.pbio.1000610

Samuel AG (2011) Speech perception. Annu Rev Psychol 62:49–72. https://doi.org/10.1146/annurev.psych.121208.131643

Sapir E (1925) Sound patterns in language. Language 1(2):37–51. https://doi.org/10.2307/409004

Sarampalis A, Kalluri S, Edwards B, Hafter E (2009) Objective measures of listening effort: effects of background noise and noise reduction. J Speech Lang Hear Res 52(5):1230. https://doi.org/10.1044/1092-4388(2009/08-0111)

Schnupp J, Nelken I, King AJ (2011) Auditory neuroscience: making sense of sound. MIT Press

Sharma A, Dorman M (1999) Cortical auditory evoked potential correlates of categorical perception of voice-onset time. J Acoust Soc Am 106(2):1078–1083

Sharma A, Kraus N, McGee TJ, Carrell T, Nicol T (1993) Acoustic versus phonetic representation of speech as reflected by the mismatch negativity event-related potential. Electroencephalogr Clin Neurophysiol 88(1):64–71. https://doi.org/10.1016/0168-5597(93)90029-O

Shattuck-Hufnagel S, Turk AE (1996) A prosody tutorial for investigators of auditory sentence processing. J Psycholinguist Res 25(2):193–247. https://doi.org/10.1007/BF01708572

Sjerps MJ, Fox NP, Johnson K, Chang EF (2019) Speaker-normalized sound representations in the human auditory cortex. Nat Commun 10(1):1–9. https://doi.org/10.1038/s41467-019-10365-z

Steinschneider M, Nourski KV, Kawasaki HOH, Brugge JF, Howard MA (2011) Intracranial study of speech-elicited activity on the human posterolateral superior temporal gyrus. Cereb Cortex 21(Cv):2332–2347. https://doi.org/10.1093/cercor/bhr014

Steinschneider M, Nourski KV, Fishman YI (2013) Representation of speech in human auditory cortex: is it special? Hear Res 305:57–73

Stevens KN (2002) Toward a model for lexical access based on acoustic landmarks and distinctive features. J Acoust Soc Am 111(4):1872–1891. https://doi.org/10.1121/1.1458026

Stevens KN, Blumstein SE (1978) Invariant cues for place of articulation in stop consonants. J Acoust Soc Am 64(5):1358–1368. https://doi.org/10.1121/1.382102

Tang C, Hamilton LS, Chang EF (2017) Intonational speech prosody encoding in the human auditory cortex. Science 357(6353):797–801. https://doi.org/10.1126/science.aam8577

Theunissen FE, Shaevitz SS (2006) Auditory processing of vocal sounds in birds. Curr Opin Neurobiol 16(4):400–407. https://doi.org/10.1016/J.CONB.2006.07.003

Theunissen FE, David SV, Singh NC, Hsu A, Vinje WE, Gallant JL (2001) Estimating spatio-temporal receptive fields of auditory and visual neurons from their responses to natural stimuli. Netw Comput Neural Syst 12(3):289–316. https://doi.org/10.1080/net.12.3.289.316

Titze IR (1989) On the relation between subglottal pressure and fundamental frequency in phonation. J Acoust Soc Am 85(2):901–906. https://doi.org/10.1121/1.397562

Toscano JC, Anderson ND, Fabiani M, Gratton G, Garnsey SM (2018) The time-course of cortical responses to speech revealed by fast optical imaging. Brain Lang 184:32–42. https://doi.org/10.1016/J.BANDL.2018.06.006

Van Dommelen WA (1990) Acoustic parameters in human speaker recognition. Lang Speech 33(3):259–272. https://doi.org/10.1177/002383099003300302

Wang X, Lu T, Snider RK, Liang L (2005) Sustained firing in auditory cortex evoked by preferred stimuli. Nature 435(7040):341–346. https://doi.org/10.1038/nature03565

Wernicke C (1874) Der aphasische Symptomencomplex: eine psychologische Studie auf anatomischer Basis. M. Cohn und Weigert

Wong PCM, Diehl RL (2003) Perceptual normalization for inter- and intratalker variation in cantonese level tones. J Speech Lang Hear Res 46(2):413. https://doi.org/10.1044/1092-4388(2003/034)

Zevin JD, McCandliss BD (2005) Dishabituation of the BOLD response to speech sounds. Behav Brain Funct 1:4. https://doi.org/10.1186/1744-9081-1-4

Zion Golumbic EM, Ding N et al (2013) Mechanisms underlying selective neuronal tracking of attended speech at a “cocktail party”. Neuron 77(5):980–991. https://doi.org/10.1016/j.neuron.2012.12.037

Chapter 4
A Parsimonious Look at Neural Oscillations in Speech Perception

Sarah Tune and Jonas Obleser

Abstract Neural oscillations are a prominent feature of the brain's electrophysiological signal, observable at different temporal and spatial scales across many species. This chapter asks how neural oscillations support speech perception. It provides the reader with a synergistic yet critical overview on what has been learned about the functional relationship of neural oscillations to speech perception. To help understand the role that neural oscillations play in speech communication, the chapter offers a concise survey of the origins of neural oscillations, their key features, the operational mechanisms they engage in, as well as core functions they are thought to support. In linking speech perception to the domain of neural oscillations, this chapter focuses on a set of cognitive core functions such as timing, binding, memory, and prediction.

Keywords Speech comprehension · Timing · Memory · Prediction · Entrainment · Speech tracking

4.1 Introduction

Speech perception, the acoustic analysis and categorization of speech sounds, and their subsequent combination and interpretation as part of speech comprehension are important prerequisites for successful communication. Both depend on the coordinated interaction of different cognitive processes that operate at nested timescales and levels of complexity. From a theoretical standpoint, human speech can be neatly decomposed into discrete units or features of varying size. Phonemes are then combined to syllables, syllables to words, and words to phrases and sentences.

From a mechanistic standpoint, which is one that asks how the human brain implements language processing, the task of speech perception and comprehension is a much more complex matter. Processes such as speech segmentation,

S. Tune (✉) · J. Obleser
Department of Psychology I, University of Lübeck, Lübeck, Germany
e-mail: sarah.tune@uni-luebeck.de; jonas.obleser@uni-luebeck.de

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022

L. L. Holt et al. (eds.), *Speech Perception*, Springer Handbook of Auditory Research 74, https://doi.org/10.1007/978-3-030-81542-4_4

categorization, and identification are challenged by the true nature of spoken language. Human speech does not present itself as a sequence of discrete elements but as a continuous stream of rapidly changing sounds with only limited acoustic boundary cues and a large degree of contextual variation. The fact that we are able to process speech quickly and effortlessly suggests the involvement of complex, yet highly flexible, neural dynamics. The goal of understanding the precise nature and function of these neural dynamics is shared across many disciplines (Buzsáki 2006), and in these investigations, the role of neural oscillations has received particular attention.

Neural oscillations, characterized by rhythmic fluctuation in brain activity at different frequencies, are a prominent feature of the electrophysiological signal. These are observable at different temporal and spatial scales across many species. These oscillations are thought to reflect the orchestrated communication of large neural ensembles (Fries 2015). In humans, there has been a growing interest in understanding the role of neural oscillations and their core functions not only in sensory perception but also in higher-order cognition including speech and language processing (Ward 2003; Giraud and Poeppel 2012).

In light of the growing body of evidence, this chapter provides an overview on the functional relationship of neural oscillations to speech perception. The focus is on the emergent, stable patterns of insight gleaned across studies, and will thus largely abstract from the details and results of individual studies. Consequently, this chapter does not intend to provide a comprehensive and detailed review of the existing literature but will point the reader to key articles.

4.1.1 Core Ideas

Overall, this chapter puts forward a perspective that is guided by three fundamental assumptions on the nature and function of neural oscillations, and on how their contributions to speech perception may be best understood and studied. These core ideas, briefly delineated below, will resonate throughout the chapter.

First, studying and understanding how neural oscillations are used in the service of speech perception is most promising under a maximally parsimonious approach: The field of speech and language should strive to integrate its results on the involvement of neural oscillations in speech perception with algorithms and functions that have been ascribed to neural oscillations more generally. Neural oscillations have been extensively studied across different sensory and cognitive domains, and important insights into their origins and overarching functions have been derived entirely irrespective of speech (Kopell et al. 2010; Wang 2010). The ability to produce and comprehend speech is commonly perceived as unique to humans, and one may thus be tempted to assume it is supported by language-specific oscillatory processes. However, as argued in this chapter, a speech researcher might do well in treating the observable oscillatory dynamics during speech first and foremost as reflections of domain-general mechanistic principles instead of language-specific computational operations.

Second, there is no one-to-one correspondence between distinct neural oscillations and speech processing steps: Characterizing the relationship between specific oscillatory dynamics and distinct levels of speech analysis entails a mapping not only between fundamentally different levels of observation but also between two highly complex systems. For this reason, it is generally difficult if not impossible to associate distinct neural oscillations, as defined by their particular frequency band, with single cognitive operations (Cohen 2017). This chapter thus aims to strike a balance between the extraction of well-established and reliable patterns in the relationship of cortical oscillations and processing of speech, and the critical discussion of aspects of language processing for which the involvement of oscillations remains poorly understood.

Third, understanding the functional role of neural oscillations in speech will, in turn, provide insights into how neural dynamics relate to complex behavior more generally. The question whether the observed patterns suggest causal or merely correlational links between brain and behavior is inherent in all investigations of neural oscillations in sensory perception and cognitive processing but rarely explicitly addressed (Fries 2009; Thut et al. 2012). In other words, does the involvement of particular oscillatory signatures indicate that they constitute necessary and sufficient determinants of speech processing or are they themselves epiphenomenal to underlying non-oscillatory neural dynamics (Ding and Simon 2014; Kösem and van Wassenhove 2017)? Indirect evidence of a functional relationship would be provided by reliable correlational patterns of neural dynamics supporting a particular cognitive operation and behavioral variability. More direct evidence could be gleaned from experimental protocols that aim to alter or disrupt ongoing oscillations in order to study the behavioral consequences of such interventions.

This chapter focuses on the available evidence on how neural oscillations may support speech perception and comprehension and will draw particular attention to experimental findings that speak to their functional importance. As will become obvious, despite an ever-growing body of oscillation-centered electrophysiological studies on speech processing, evidence pointing to a causal link between neural oscillations and behavior in the domain of speech processing is still relatively sparse.

The chapter is organized as follows. Section 4.2 will provide the reader with a brief introduction to the key features of neural oscillation, as well as their core functions. Section 4.3 will then link these core functions to speech processing, starting at the level of auditory perception and finally arriving at more complex comprehension processes. Finally, Sect. 4.4 concludes the chapter by outlining how future studies can help to better understand the role neural oscillations serve in speech processing and also highlights the aspects of language processing that may be difficult to examine through the looking glass of neural oscillations.

4.2 Overview on the Features and Functions of Neural Oscillations

Complex behavior such as human speech processing relies on the orchestrated action of multiple processes that involve spatially distributed brain networks, either in cascades or in parallel. To bring together the computational outcomes of these different processes and networks, functional mechanisms are needed to flexibly control the flow and integration of neural information in response to external input and internal goals (Meyer et al. 2019). To help understand the role neural oscillations play in such functional mechanisms, this section provides a concise overview of the origins of neural oscillations, their key features, and the functional mechanisms they engage in, as well as the domain-general core functions they are thought to support.

4.2.1 Key Features of Neural Oscillations

The human brain is never silent. Spontaneous and stimulus-driven electrical activity generated by large ensembles of neurons can be recorded from different parts of the brain using noninvasive methods such as magneto- (MEG) and electroencephalography (EEG; Hansen et al. 2010; da Silva 2013), or invasively by intracranial recording methods using surface or depth electrodes (electrocorticography, ECoG; Crone et al. 2006). While it is technically possible to measure the electrophysiological dynamics of only a small number of neurons, due to their highly invasive nature, these approaches are typically restricted to animal research (see Chap. 3, Oganian, Fox, and Chang).

A prominent feature of the EEG across species is the co-occurrence and superposition of brain rhythms oscillating at nested frequencies and across different spatial scales (Niedermeyer and Lopes da Silva 1999; Nunez and Srinivasan 2006). Neural oscillations arise from the dynamics of structurally organized neural ensembles in which two different types of neurons, excitatory pyramidal cells and inhibitory interneurons, mutually influence one another (Whittington et al. 2000). More precisely, neural oscillations reflect periodic fluctuations in postsynaptic potentials summed across space and time (Buzsáki et al. 2012). The amplitude of an oscillation describes the strength (or energy) of these voltage fluctuations (see Fig. 4.1a).

Across different species, neural oscillations occur across a wide range of frequencies from around 0.1 up to 600 cycles per second. An important characteristic is their hierarchical organization into several distinct frequency bands (Buzsáki and Draguhn 2004). Despite some disagreement over the defining boundaries, electrophysiological evidence collected in countless studies supports a differentiation in humans into (at least) five canonical frequency bands. In 1929, Hans Berger was first to describe the most dominant, large-amplitude rhythm of around 10 Hz. This rhythm is best observed over occipital cortex in awake subjects who have their eyes

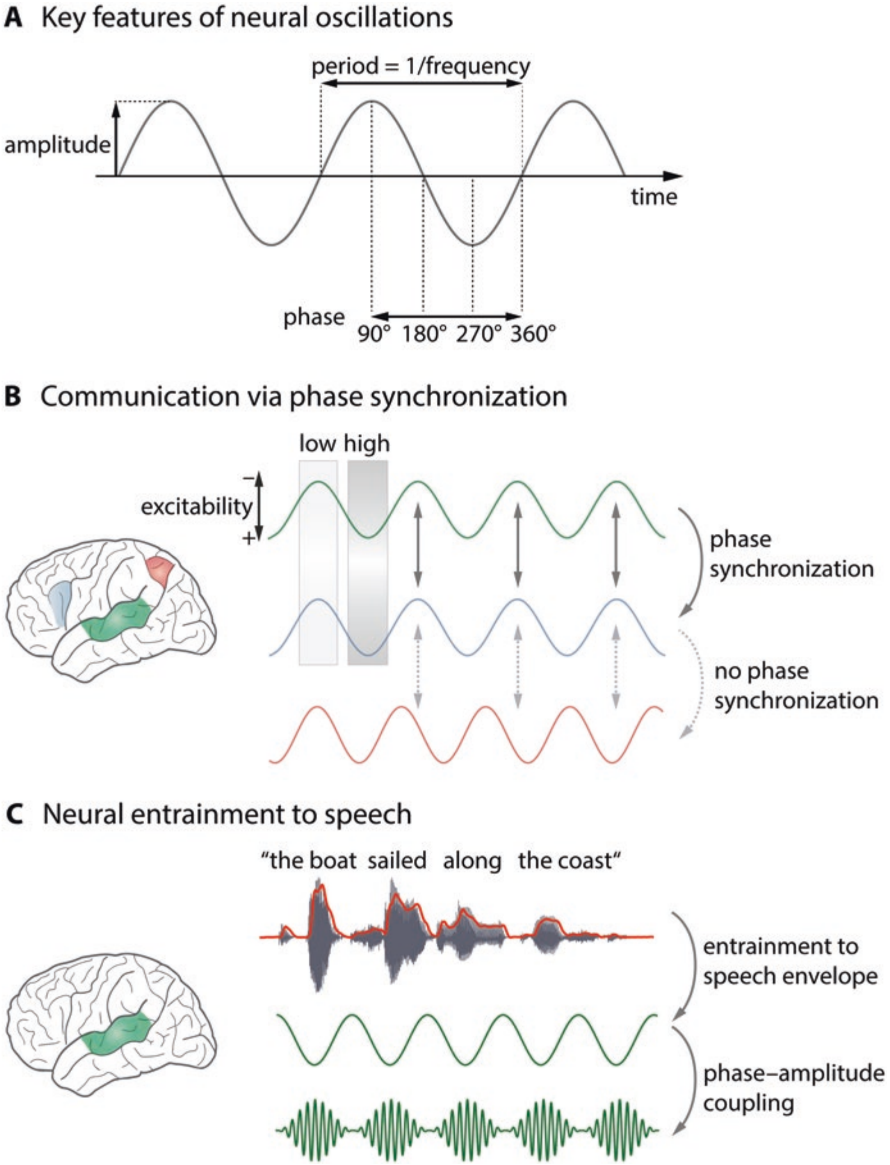
A Key features of neural oscillations
period = 1/frequency
amplitude
time
phase
90° 180° 270° 360°
B Communication via phase synchronization
low high
excitability
−
+
phase synchronization
no phase synchronization
C Neural entrainment to speech
"the boat sailed along the coast"
entrainment to speech envelope
phase–amplitude coupling

Fig. 4.1 (**a**) Schematic illustration of a sinusoidal oscillation and its defining features: amplitude, frequency, and phase. (**b**) Neural oscillations establish recurring phases of high and low excitability. The three waveforms represent oscillatory activity generated by spatially separated neuronal populations. Communication between neuronal populations is enabled by means of phase synchronization that ensures a stable alignment of periods of high or low excitability (as shown for the green and blue waveforms). Communication is blocked when the activity in two neuronal ensembles oscillates out of phase (as shown for the blue and orange waveforms) as inputs from one population arrive during the non-excitable phase of the target population. (**c**) Ongoing oscillatory activity in auditory cortex entrains to syllable-rate fluctuations in the temporal amplitude envelope of speech. The speech waveform is shown for a short sentence together with the corresponding speech envelope (red line). Low-frequency oscillations in the theta range (top waveform) align their high-excitable phases with periods of high energy in the speech signal. The additional coupling of high-frequency amplitude (bottom waveform) to low-frequency phase allows for the analysis of speech at different temporal scales.

closed (Berger 1929). Berger termed this rhythm the "alpha" rhythm and a less pronounced but faster rhythm he observed when subjects had their eyes open "beta" waves. Neural rhythms discovered in the following decades were analogously labeled with Greek letters and resulted in the following human EEG nomenclature of frequency bands: delta (1–4 Hz), theta (4–8 Hz), alpha (8–12 Hz), beta (12–30 Hz), and gamma (>30 Hz) range (Steriade et al. 1990; Buzsáki 2006).

Neural oscillations have a number of additional key features that endow them with a rich repertoire of complex dynamics. These dynamics can be exploited for flexible communication within and across distributed neural networks. The most important effect of neural oscillations is the establishment of alternating periods of high and low excitability (Fries et al. 2007; see Fig. 4.1b). Because inhibitory interneurons play a vital role in the emergence of neural oscillations (Wang 2010), one can also think about this cycle as periods of minimal and maximal inhibition (Klimesch et al. 2007).

Accordingly, the fate of incoming signals depends on their timing relative to these different phases: Inputs arriving close to the peak in excitability are more likely to trigger action potentials that will propagate the signal further downstream than inputs arriving during the opposite phase of the excitation-inhibition cycle. In this way, neural oscillations organize the neural response of oscillating microcircuits into short pulses during which continuous fluctuations in postsynaptic potentials are translated into discrete spiking rates. The resulting synchronization of rhythmic spike trains is a highly effective way to increase the impact on their common projection targets (Masquelier et al. 2009; Fell and Axmacher 2011).

The frequency of a sinusoidal oscillation also determines how long the duty cycle, that is, the high-excitability phase, lasts. This, in turn, affects not only the level of temporal precision with which receiving neurons respond to and encode neural activity but also the spatial scope of an oscillation. Fast oscillations, for example, in the gamma band (>30 Hz), integrate neural activity from nearby local neurons within a very narrow time window spanning tens of milliseconds. By contrast, slower oscillations with their considerably longer time windows of integration are able to recruit larger and more distant neuronal populations. However, they do so at the expense of temporal precision. In sum, amplitude, phase, and frequency of an oscillation define its fundamental functional principles.

4.2.2 Mechanistic Functions Supported by Neural Oscillations

Neural oscillations are thought to implement a number of distinct mechanisms, all of which depend on the dynamic modulation of at least one of their defining features: amplitude, frequency, or phase. Changes in these key features serve as dependent measures in electrophysiological studies using M/EEG or ECoG, including those on speech perception. Furthermore, they are important building blocks for a number of assumed domain-general core functions that are the backbone of many cognitive operations. Because these neural dynamics have been observed in studies covering different domains and are therefore implicated in various functions such as feature binding, memory formation, and motor coordination, a detailed review of the evidence would be beyond the scope of this chapter. Instead, this chapter focuses on a mechanistic level of description.

Mechanisms that exploit the versatility of neural oscillations can be broadly grouped into two different categories. On the one hand, there are stimulus-driven modulations, that is, transient changes in the amplitude or phase of an oscillation in response to the presence of a salient external or internal event. On the other hand, there is a family of mechanisms that enable the communication of neural ensembles across different frequencies and brain regions.

In time-frequency analyses of EEG data, two of the most commonly studied modulations evoked by sensory stimuli or task demands are changes in amplitude or phase. Changes in amplitude (or power, representing squared amplitude) reflect changes of the degree to which neurons oscillate in synchrony. Consequently, increases and decreases in amplitude are referred to as event-related synchronization and desynchronization, respectively (Klimesch et al. 2007).

In the case of phase modulation, stimulus-driven changes in the consistency of phase are observed across trials, in particular for temporally unordered or isolated events. Phase angles of an ongoing oscillation will be uniformly distributed before stimulus onset but highly consistent after stimulus onset (Voloh and Womelsdorf 2016). This phenomenon is referred to as phase resetting, and it is thought to contribute to the generation of event-related potentials observed in the time domain (Sauseng et al. 2007).

Interesting dynamics arise from the interaction of ongoing neural oscillations with one another or in response to rhythmic inputs. Neural oscillations tend to fluctuate at a preferred frequency that is determined by intrinsic frequency properties of the involved neurons. However, in the interaction with other periodically changing signals, they can shift away from their preferred frequency. This form of synchronization between two oscillators that also yields a stable phase relationship is referred to as entrainment. For entrainment to take place, the difference in the preferred frequencies of the two oscillators has to be small. As the difference increases, more force, in the form of a higher degree of coupling strength, is needed (Glass and Sun 1994; Pikovsky et al. 2003). If the difference surpasses a certain limit, stable synchronization can no longer be achieved. Figure 4.1c illustrates the entrainment of ongoing oscillations in auditory cortex to rhythmic fluctuations in the external speech signal.

Finally, there are additional ways for oscillations to synchronize, jointly referred to as cross-frequency coupling, which allow for the functional coordination of neural circuits oscillating at frequencies spaced further apart (Canolty and Knight 2010; for a recent review, see Hyafil et al. 2015b). To enable cooperation of brain rhythms oscillating at different frequencies, some form of synchrony or stable relationship needs to be established. Depending on the involved features, different kinds of cross-frequency coupling signatures are possible. In coupling schemes that involve the synchronization of two different features, as in phase-amplitude coupling (PAC) or phase-frequency coupling (PFC), it is the amplitude or frequency of the faster oscillation that changes relative to the phase of the slower oscillation. In the case of PAC, this is also referred to as *nesting*, and the coupling of gamma band activity relative to phases of excitability in the theta band is a prominent example (Canolty et al. 2006; see Fig. 4.1c).

4.2.3 Emerging Core Functions and Computational Principles

Based on the reviewed fundamental properties of neural oscillations and the mechanisms that operate around them, what kinds of overarching core functions might neural oscillations support in cognitive operations? The understanding of the functional importance of neural oscillations is still undergoing constant refinements. It has become clear, however, that their specific roles are co-determined by the neural dynamics and the connectivity profiles of the underlying generators (Fries 2009). To date, a number of testable working hypotheses on their general computational functions have been put forward (Sejnowski 2006; Singer 2018).

The most fundamental overarching function of hierarchically related (i.e., nested) frequencies is to provide the neural system with internal clocks. Here, different frequencies provide different levels of temporal precision. This timing function establishes basic temporal reference frames that serve as the backbone to more complex computations.

In addition to timing, further domain-general functions have been ascribed to neural oscillations: the encoding and binding of sensory information, the controlled routing of information across different areas of the brain, and the encoding, storage, and retrieval of information in processes of memory and learning. Yet another function that builds on all of the thus far outlined computations is the implementation of internal models that continuously generate and update predictions about incoming signals (Engel et al. 2001). The proposed functions are not mutually exclusive but work in concert to allow for successful behavior (Bonnefond et al. 2017). Importantly, the perception and comprehension of speech, along with other forms of higher-order cognition, crucially rely on the implementation of these core functions as will be shown in detail in Sect. 4.3.

The proposed role of neural oscillations in the representation and binding of sensory information is based on the idea that information becomes encoded via the coordinated spiking behavior of excitatory neuron information. If sensory information is encoded in spike rates (defined as the number of spikes within a given time window), then neural oscillations provide a way to discretize time and concentrate neural spiking within the established time window. Even more dynamic encoding of information is possible when the precise temporal grouping of spikes occurs in reference to the phase of an ongoing oscillation, for example, in the theta band (Jensen and Lisman 2000; Masquelier et al. 2009). The periodicity of an oscillation naturally establishes a temporal reference frame relative to spiking behavior. In addition, the ability to control the spike patterns of neural populations, for example, via synchronization in the gamma band, enables the binding of features across hierarchically organized brain regions (Gray et al. 1989; Singer and Gray 1995).

Another perspective put forward focuses on the role of neural oscillations in the controlled routing of information that includes the selective sampling, amplification, or inhibition of information. The two most prominent accounts, the communication through coherence (CTC; Fries 2005, 2015) and gating by inhibition (GBI; Jensen and Mazaheri 2010) theories, assume that oscillations in the gamma or alpha band serve to implement spatiotemporal filters in line with attentional demands. These filters enable the selective sampling of information and its controlled relay to downstream targets. According to the CTC theory, the flow of information is coordinated via selective entrainment of different neural ensembles operating in the gamma frequency band. By contrast, the GBI theory associates modulation in local alpha power with the active inhibition of task-irrelevant brain areas. In each case, the ability to flexibly control communication between local or more distant nodes in a network consequently allows for the dynamic integration and segregation of functional networks across the brain.

Lastly, it is important to consider how the synchronization of neural activity via oscillatory rhythms plays a vital role in synaptic plasticity and is therefore crucial for memory and learning processes (see Fell and Axmacher 2011 for a comprehensive review). As reviewed in Sect. 4.2.2, the synchronization of two oscillations via phase-phase or PFC allows for a precise encoding of temporal relations in spiking behavior. The sustained phase synchronization between distributed brain regions within the theta and gamma frequency bands as well as their cross-frequency coupling has been implicated in long- and short-term memory processes (Kahana 2006). In line with the concept of Hebbian learning (Hebb 1949), optimized communication between neuronal groups due to gamma phase synchronization leads to the strengthening of their synaptic connections (Caporale and Dan 2008).

To summarize the evidence reviewed up to this point, it has become obvious that neural oscillations afford a wide range of neural computations that may support complex behavior in general. The ability to tightly control the spiking activity within and across neural ensembles allows for the temporally precise and selective integration of neural signals within distributed neural networks. The integration of local and global computations via the coupling of fast and slow oscillations enables the joint influence from bottom-up and top-down signals onto perception and cognition. In sum, and viewed most parsimoniously, neural oscillations establish a functional framework in which internal models can continuously generate and update predictions based on external input and internal states. Figure 4.2 delineates how the set of cognitive core functions discussed thus far may help to bridge the gap between specific neural frequency bands on the one hand and linguistic processing steps on the other.

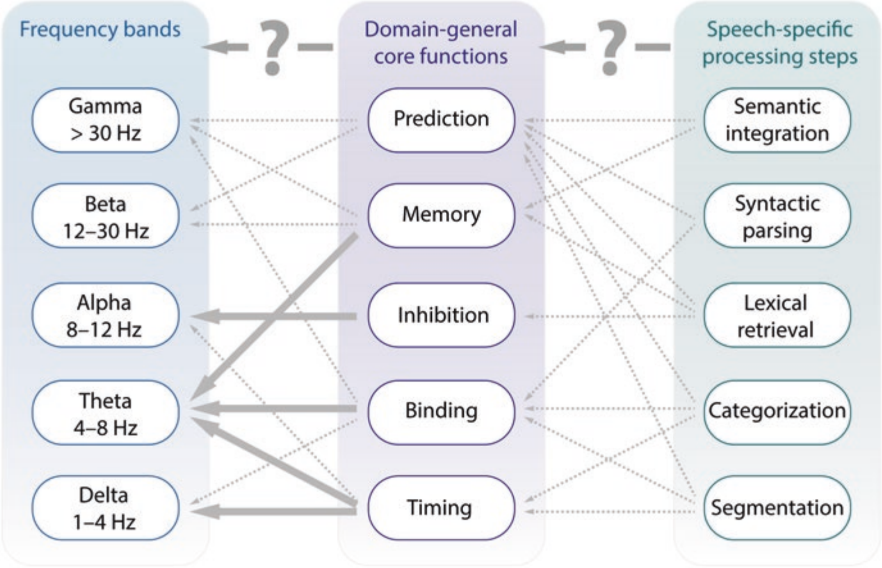

Fig. 4.2 The challenge of mapping between linguistic processing steps, cognitive core functions supported by neural oscillations, and distinct frequency bands. Solid thick arrows denote the links between core function and frequency bands that have been reliably attested by empirical evidence. Thin dashed arrows indicate connections between linguistic analysis step and core function or between core functions and specific frequency bands that have been previously proposed but are not yet generally accepted. Importantly, the link between linguistic processes and specific frequency band is most promisingly established via the set of domain-general core functions.

4.3 Neural Oscillations in Speech Perception and Comprehension

Building on the mechanistic principles and core functions discussed above, this section reviews the empirical evidence that links neural oscillations to processes involved in speech processing. The review starts at the level of auditory perception and ends with more complex comprehension processes. Not incidentally, the role of oscillatory dynamics is empirically more strongly supported and therefore better understood for the lower-level processing steps. Therefore, this section ends with a discussion of the more complex and abstract linguistic units and operations, for which the connection to oscillatory dynamics is less obvious.

In line with the fundamental assumptions put forward in Sect. 4.1, the goal here is not to establish a one-to-one mapping between specific frequency bands and linguistic units or operations. Instead, this section highlights to which extent the computational mechanisms operating around neural oscillations may provide solutions to the challenges posed by speech processing. Lastly, it reviews any available evidence that speaks to the functional relevance of speech-related oscillatory patterns.

4.3.1 Perceptual Analysis of Continuous Speech

In order to analyze how neural oscillations may support the perceptual analysis of speech, one may begin by breaking down the process into its basic computations and necessary prerequisites. Simply speaking, to extract perceptual units of a particular length from an ongoing acoustic signal, an adequate temporal reference frame is needed. However, the establishment of such a temporal grid alone will not suffice – it also needs to be aligned to the continuous signal in a sensible way. Therefore, the signal has to provide acoustic cues that mark the boundaries of the relevant chunks so that meaningful rather than random segments can be extracted. Lastly, it is important to keep track of where a particular segment came from within the continuous signal so that it can be combined with adjacent segments into more complex units.

A number of implications result from linking this computational dissection back to speech perception and neural oscillations: First of all, the decomposition of continuous speech into smaller units critically depends on their physical representation in the auditory signal. Second, for neural oscillations to play a role in their tracking, the perceptual units would have to occur in a manner that is reasonably temporally regular. Third, ongoing neural oscillations in the auditory system would need to operate at the timescales of perceptual units in order to function as temporal reference frames for speech segmentation. In addition, the intrinsic oscillations must be able to flexibly align their phase and period to the quasi-regular patterns in the external speech signal. Lastly, for the combination of low-level features into more complex units, computations at different processing levels and timescales need to interact with one another.

Reconciling the outlined hypotheses with the available evidence, we find that many of these underlying assumptions have been empirically supported. The temporal features of speech, in particular its amplitude envelope and temporal fine structure, provide acoustic cues for the encoding of linguistic elements at different temporal scales. Slow amplitude modulations in the theta range (4–8 Hz) coincide with the syllable rate, while faster fluctuations (~30–80 Hz) linked to the temporal fine structure of speech are associated with the representation of phonemic information. Even slower envelope fluctuations in the lower delta range (about 1–2 Hz) correlate with occurrence of intonational phrases (Rosen 1992). Even though these acoustic variations do not qualify as periodic fluctuations in the strict sense, they are rhythmic enough for neural oscillations to synchronize to them (Obleser et al. 2017).

Electrophysiology in mammals has demonstrated the presence of a hierarchical regime of intrinsic delta, theta, and gamma oscillations in auditory cortex that synchronize to rhythmic acoustic sequences (Lakatos et al. 2005; Kayser et al. 2015). Furthermore, this neural entrainment is enhanced for behaviorally relevant stimuli, emphasizing its role as a tool for the selective sampling of sensory input (Lakatos et al. 2008; Schroeder et al. 2010). Importantly, there is ample evidence that similar principles are involved in human speech processing. The results from numerous studies using M/EEG or ECoG provide evidence for the neural tracking of speech via phase-locked oscillatory activity in auditory cortex. Ongoing neural activity in auditory cortex was found to synchronize most strongly to syllable-rate fluctuations in the temporal envelope of speech (Ghitza 2011; Giraud and Poeppel 2012). However, the precise mechanistic principles and functional roles of this neural tracking in speech perception are a matter of ongoing debate (for review see Kösem and van Wassenhove 2017; Obleser and Kayser 2019).

On the one hand, there are accounts suggesting that entrainment of intrinsic nested oscillations in auditory cortex primarily reflects the bottom-up perceptual analysis of speech (Ghitza 2011; Giraud and Poeppel 2012). According to this perspective, ongoing cortical oscillations at distinct frequencies enable the discretization of continuous speech into linguistic units that are represented on different timescales. The close correspondence of the frequencies at which intrinsic oscillations operate to the rates at which distinct linguistic units occur allows for the parallel tracking of phonemes, syllables, and intonational phrases. Salient events in the speech envelope are assumed to trigger the phase reset of spontaneous theta oscillations (Luo and Poeppel 2007). The phase-aligned theta oscillations then instantiate a temporal reference frame for the dynamics of coupled gamma oscillations that analyze the speech signal at the phonemic scale (Ghitza 2011). In this way, coupled theta and gamma oscillations work together to sample speech at the rate of syllables while also optimizing the decoding of speech at the underlying phonemic level.

On the other hand, there are accounts that argue for a much more domain-general role of entrainment in the active, selective sampling of sensory input (Schroeder and Lakatos 2009; Zion Golumbic et al. 2013). According to this perspective, the degree to which intrinsic oscillations in the delta and theta band exhibit phase locking in response to rhythmic auditory input does not only depend on acoustic cues in the external signal but is also modulated by top-down cognitive factors such as selective attention. Here, the modulation of the strength of entrainment acts as a filter that selectively enhances and attenuates the representation of the behaviorally relevant and irrelevant input, respectively (Lakatos et al. 2013).

Both accounts, as well as further related hypotheses, build on a large body of electrophysiological and behavioral evidence on nonhuman sensory as well as human speech processing. To gain insight into the functional relevance of neural entrainment to speech comprehension, it is crucial to understand which low-level sensory or high-level cognitive factors influence the strength of cortical speech tracking and whether changes in the neural dynamics correlate with changes in behavior.

Different approaches have thus been taken to examine the link between envelope entrainment and speech comprehension. A common strategy is to manipulate speech intelligibility. Here, the core question is whether acoustic characteristics crucially determine the strength of neural entrainment and thereby influence speech intelligibility or whether a certain level of speech intelligibility and thus the involvement of high-level linguistic processes is necessary for neural entrainment to occur. The employed acoustic manipulations that decrease the speech signal's intelligibility include its temporal reversal (Howard and Poeppel 2010; Gross et al. 2013) or compression (Ahissar et al. 2001; Nourski et al. 2009), the degradation of spectral content via noise vocoding (Peelle et al. 2013; Ding et al. 2014), and the addition of background noise (Ding and Simon 2013; Zoefel and VanRullen 2016).

However, the overall pattern of results obtained from these studies does not provide a clear-cut answer to the question of whether acoustic cues or speech intelligibility modulates neural entrainment to speech. This is due to several reasons. First of all, the findings that speak to the relationship of neural entrainment and speech intelligibility have been mixed. While many studies find that decreased speech entrainment correlates with a decrease in speech intelligibility (Ahissar et al. 2001; Luo and Poeppel 2007), others have failed to find such a relationship (Peña and Melloni 2012; Zoefel and VanRullen 2016). Furthermore, in the majority of these studies, changes in acoustic properties are confounded with changes in speech intelligibility which makes it difficult if not impossible to tease apart their contribution to the neural tracking of speech (Peelle and Davis 2012; Ding and Simon 2014). Nevertheless, these investigations still had a substantial impact on our understanding of how entrainment to the rhythm of speech may be utilized in speech perception. The heterogeneous nature of these results challenges the assumption that neural entrainment to speech may be driven by only a narrow range of acoustic cues such as low-frequency modulation in the speech amplitude envelope (Herrmann and Henry 2012). On the contrary, the results emphasize the intricate nature of human speech perception and speech comprehension that depend on the interplay of bottom-up sensory cues and top-down linguistic knowledge.

An excellent example for the mutual influence of sensory cues and linguistic information on phase locking to speech was provided by Peelle et al. (2013). The authors manipulated the spectral complexity of speech using noise vocoding and spectral rotation to tease apart the relative impact of acoustic properties versus speech intelligibility on phase-locking strength.

One of the key findings in this study, shown in Fig. 4.3a, was the observation of phase-locked responses in the range of 4–7 Hz in the bilateral auditory cortices (and surrounding areas) to one-channel vocoded speech that is completely unintelligible. However, they also found that strength of phase locking increased along with an increase in spectral detail and thus in intelligibility. Crucially, the comparison of the original and rotated four-channel vocoded conditions revealed that the enhanced phase locking was at least partially driven by the availability of linguistic content and not just by an increase in spectral fidelity. Interestingly, this difference was most pronounced in the left middle temporal gyrus rather than auditory cortex proper. Taken together, the results demonstrate the significance of low-level sensory cues beyond syllable-rate fluctuation in the speech envelope as well as the impact of high-level linguistic processes involved in the comprehension of speech. The results also speak to the underlying neural substrate that generates and maintains oscillatory dynamics, and to their importance in determining the precise functional roles of these neural dynamics.

Importantly, not only linguistic information impacts the degree to which ongoing oscillations align their phase to temporal regularities in speech. Also, and arguably closely intertwined, such neural tracking of speech is under attentional control (Ding and Simon 2012a; Henry et al. 2017). Using a cocktail party paradigm in

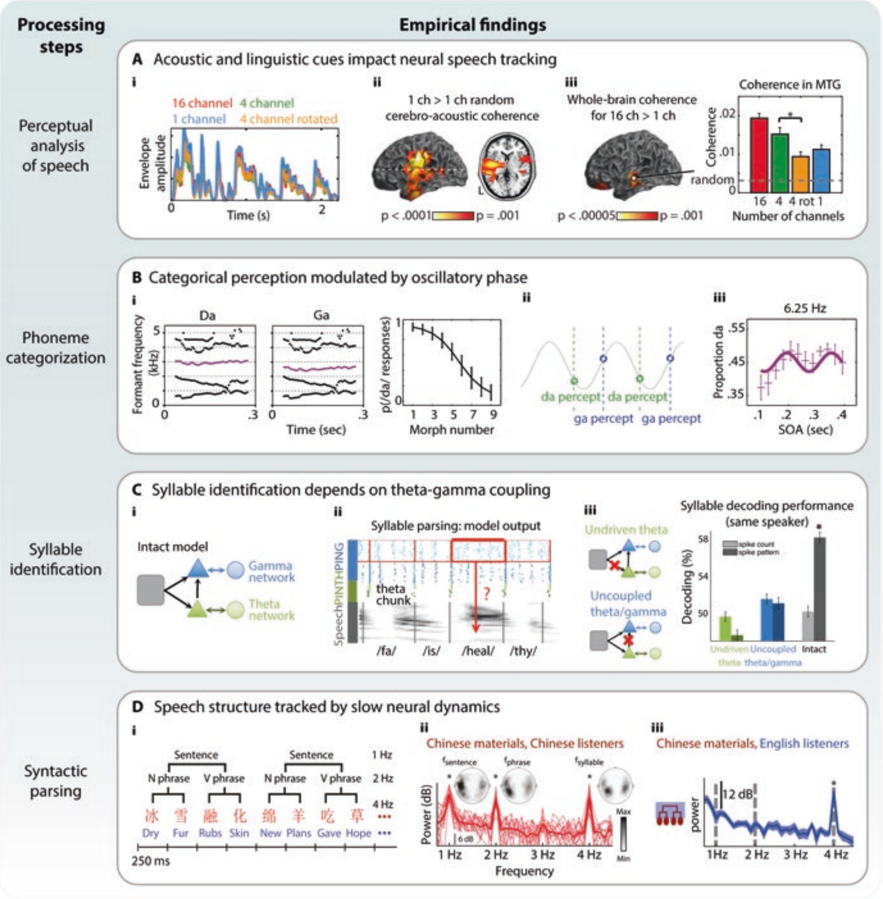

Fig. 4.3 Empirical findings that speak to the role of neural oscillations and neural dynamics in human speech processing. Left panel describes the investigated linguistic processes, and right panel presents the key results from selective studies. (**a**) Top-down and bottom-up cues modulate the neural tracking of speech. (i) Noise vocoding alters temporal fine structure but not the temporal envelope of speech. (ii) Source localization of phase-locked responses to unintelligible one-channel vocoded speech compared to permutation-based null baseline. (iii) The availability of linguistic content influences the phase coupling of brain responses to the speech envelope. Left, increased coherence for intelligible (16-channel vocoded) compared to unintelligible (1-channel vocoded) speech. Right, increased phase locking to moderately intelligible four-channel speech compared to unintelligible four-channel spectrally rotated speech stimuli. (Adapted with permission from Figs. 1, 3, and 4 in Peelle et al. (2013)). (**b**) Phoneme perception is influenced by oscillatory phase. (i) Stimuli consisted of synthesized ambiguous syllables identified as either /da/ or / ga/. (ii) Hypothesized relationship between the phase of a speech-entrained oscillation and the categorical perception of an ambiguous /daga/ morphed stimulus. (iii) Empirical results show an oscillatory behavioral pattern depending on the stimulus presentation relative to the phase of the entrained oscillation at 6.25 Hz. (Adapted with permission from Figs. 1, 3, and 4 in ten Oever and Sack (2015)). (**c**) Syllable decoding in a microcircuit model of coupled theta and gamma oscillations. (i) Schematic representation of the full model architecture that includes the coupling of

(continued)

which participants had to attend to one of the two concurrent speakers, Zion Golumbic et al. (2013) recorded brain activity intracranially from electrode grids covering large portions of the left or right lateral cortex. The authors were able to show that the phase of slow oscillations operating across the delta and theta range (1–7 Hz), as well as high-gamma power (75–150 Hz), tracked the temporal envelope of the target speaker (see also Mesgarani and Chang 2012). In line with known differences in the spatial scope of slow and fast oscillations, the tracking of the to-be-attended speech envelope via phase alignment of low-frequency oscillations was spatially more widespread than the tracking by modulation of high-gamma power. However, while both low-frequency phase and high-gamma power showed generally more reliable tracking of the to-be-attended compared to the to-be-ignored speaker, cortical sites differed with respect to their response profiles. Electrodes situated close to low-level auditory cortices preferentially tracked the speech envelope of the attended speaker but still encoded, albeit to a lesser degree, the speech envelope of the to-be-ignored speaker. By contrast, recording sites distributed in higher-order areas such as the prefrontal cortex, inferior parietal lobule, and association cortices in the temporal lobes showed a much more exclusive tracking of the to-be-attended speaker (Zion Golumbic et al. 2013).

In sum, these findings speak in favor of the neural tracking of speech as a tool to regulate auditory perception in line with attentional demands. Taken together with the results on neural entrainment and speech intelligibility, they strongly support the assumption that the entrainment to rhythmic features of speech does not represent a singular phenomenon that serves one particular purpose for the perceptual analysis of speech. Instead, it is much more likely that depending on the involved neural circuits and their connectivity profiles, neural entrainment of ongoing oscillations may be used for different computational purposes.

Despite the growing body of work investigating the functional role of neural entrainment in speech perception and comprehension, evidence supporting a causal relationship between the fidelity of neural entrainment and speech comprehension is sparse. There are at least two reasons for this paucity of evidence. On the one hand, not all of the studies concerned with the role of neural entrainment in speech perception have included behavioral measures that could speak to its functional relevance (Ding and Simon 2012b; Zion Golumbic et al. 2013). On the other hand, even if arguably coarse behavioral measures of speech comprehension were included, they did not always vary systematically with changes in the observed

Fig. 4.3 (continued) gamma- and theta-generating networks. (ii) Syllable decoding based on the spike pattern generated by gamma neurons within a theta cycle. (iii) Decoding accuracy for the full model outperformed the output of two alternative model architectures. (Adapted with permission from Figs. 3 and 4 in Hyafil et al. (2015a)). (**d**) Hierarchical structures of speech are tracked by slow neural dynamics. (i) Sequences of monosyllabic Chinese words were presented at a fixed rate of 4 Hz. Words could be grouped into phrases and phrases into sentences. (ii) The brain responses of Chinese listeners show pronounced peaks coinciding with the presentation rates of syllables, phrases, and sentences. (iii) The brain responses of English listeners show a neural tracking at the acoustic level only. (Adapted with permission from Figs. 1 and 2 in Ding et al. (2016))

neural dynamics, and their relationship would still be correlational in nature (O'Sullivan et al. 2014).

A promising alternative approach to probing the causal role of oscillatory dynamics in the multi-scale analysis of speech is the use of computational modeling. A clear advantage of such an in silico approach over in vivo work in the human is that it allows the researcher to directly test the influence of different neural circuit configurations on the model-generated "behavior." Hyafil et al. (2015a) implemented a biophysically plausible model of coupled theta- and gamma-generating modules that simulate the dynamics of intrinsic oscillations in auditory cortex (see Fig. 4.3c). Using this model, the authors were able to test several key assumptions about the functional relevance of speech-entrained theta oscillations and theta-gamma cross-frequency coupling for the decomposition, parsing, and encoding of running speech (Giraud and Poeppel 2012; Kayser et al. 2012). In short, they found that the neural output from this model operating in a purely bottom-up fashion closely resembled empirical observations from electrophysiological recordings from the human auditory cortex (Luo and Poeppel 2007; Nourski et al. 2009). Crucially, the authors were able to directly examine the functional relevance of the coupling of theta and gamma oscillation by comparing the outcome of network configurations that possess or lack such a mechanistic feature (see also Hovsepyan et al. 2020). Thus, even though these simulations provide only small-scale examples of the actual underlying architecture that do not yet incorporate any top-down influences, they are promising complementary approaches to understanding the impact of neural dynamics and network connectivity.

Finally, the use of electrical brain stimulation to directly alter ongoing neural oscillations offers a fruitful approach to studying their functional relevance for communication behavior. Two insightful studies probed the impact of neural entrainment to the speech envelope for speech intelligibility by applying transcranial electric currents in the shape of the speech envelope (Riecke et al. 2018; Wilsch et al. 2018). Importantly, modulating neural entrainment to the speech envelope in this way did indeed lead to a systematic modulation of speech intelligibility. These results thus provide important evidence for a causal role of neural speech entrainment in speech comprehension.

4.3.2 Categorical Perception

Following the fine-grained spectro-temporal analysis of speech in core auditory regions, the neural output of these operations is relayed to higher-order brain areas along the superior temporal lobe that map highly variable speech tokens to invariant linguistic representations such as phonemes or words (DeWitt and Rauschecker 2012). In the case of phoneme categorization, this process operates in a multidimensional space as phonemes can be discriminated based on several acoustic parameters. These parameters encompass different spectral and temporal cues including fundamental frequency, formant frequency, formant transition duration, or voice onset timing (Holt and Lotto 2010). This subsection focuses on how neural oscillations and in particular their entrainment to speech may exploit temporal features for the task of phoneme categorization.

Among the temporal features that can be extracted from the complex speech signal are two phonetic parameters that guide the discrimination of different stop consonants. The voice onset time (VOT) describes the delay between the consonant release and the onset of voicing in consonant-vowel clusters. In English, VOT is close to zero for voiced stop consonants (as in the syllable /ba/) and has an average duration of 40–60 ms for voiceless stop consonants (as in /ta/) and can thus be used to distinguish syllables such as /ba/ and /pa/ or /da/ and /ta/. Another parameter that reflects differences in the manner or place of articulation describes the duration of formant frequency transitions from consonants to vowels or vice versa. The duration of the transition in second and/or third formant frequency would thus help to discriminate between /ba/ and /wa/ syllables that differ in the manner of articulation.

Which computational operations might be involved in processes of categorical perception that are based on temporal cues? First, rapidly changing temporal features such as VOT and formant transition duration would need to be represented in the neural activity of synchronized neuronal populations in auditory cortex so that this information can be exploited by higher-order areas. Intracranial recordings from the human auditory cortex suggest that this is indeed the case (Nourski and Brugge 2011). Second, to encode the temporally specific occurrence of these short-duration cues within a broader context, an additional, coarser grid serving as temporal reference frame is needed. Building on the idea that neural entrainment to speech creates such a syllable-rate temporal grid, one can ask how this mechanism can be utilized in the service of categorical speech perception.

Peelle and Davis (2012) put forward a theoretical account on how categorical perception could be based on a consistent temporal pairing between the phase of an entrained low-frequency oscillation and the onset of voicing in stop-consonant-vowel utterances. They proposed that a change in the relationship of oscillatory phase and voice onset could lead to a change in the categorical perceptions of syllable-initial stop consonants. Specifically, they build on behavioral studies that showed that a given (absolute) VOT is interpreted in reference to the preceding speech rate and can thus lead to either the perception of a voiced or unvoiced stop consonant (Port 1976; Miller et al. 1984). Peelle and Davis reasoned that the entrainment of theta oscillations to the preceding speech rate would allow for a stable alignment of the consonant release to a specific phase of the oscillatory cycle. In their conceptual example, the release of the consonant closure for an unambiguous unvoiced /pa/ would always align with the trough of the entrained low-frequency oscillation, whereas the release in the unambiguous voiced /ba/ stimulus would occur close to the peak of the oscillation. Preserving this relationship consequently ensures that the voice onset, which causes an increase in amplitude, coincides with the phase of high excitability, i.e., the peak, of the entrained oscillation.

An interesting implication that follows from this conceptual account is the speech-rate-dependent perception of ambiguous speech tokens. Following the logic outlined above, for an ambiguous syllable with a VOT halfway between that of a clear /pa/ and a clear /ba/, the rate of the entraining speech rhythm determines during which phase of the ongoing oscillation the consonant release will occur. Depending on this temporal relationship, the ambiguous stimulus would be perceived as either a voiced or unvoiced stop consonant.

The idea that oscillatory phase could directly influence the categorical perception of phonemes was put to the test in an EEG study (Oever ten and Sack 2015). Here, the authors focused on the systematic temporal delay between visual mouth movements and subsequent sound production that can be exploited for the discrimination of voiced stop consonants such as /da/ and /ga/ which differ in their respective visual-to-auditory onset delay by 80 ms. The authors examined the impact of oscillatory phase on syllable identification by presenting a synthesized ambiguous sound (perceived either as /da/ or /ga/) after a period of rhythmic auditory stimulation with 50 ms bursts of white noise presented at 1, 6.25, or 10 Hz. More precisely, they tested the hypothesis that shifting the onset of the presented stimulus relative to the phase of the entrained oscillation should lead to corresponding fluctuation in categorical perception.

Indeed, following the perturbation of ongoing neural activity by auditory stimulation at either 6.25 or 10 Hz, they observed a systematic relationship of syllable identification as either /da/ or /ga/ and stimulus onset relative to the phase of the entrained oscillation (see Fig. 4.3b). More precisely, the measured phase difference in which perception was biased toward one of the two syllables agreed with their difference in visual-to-auditory temporal delay of 80 ms. These results provide compelling evidence for the impact of phase coding on categorical perception (see also Oever ten et al. 2020). However, the results did not speak to the specific neural networks involved in the representation of abstract linguistic representations.

Lastly, an EEG study by di Liberto et al. (2015) showed that even low-frequency (i.e., 1–4 and 4–8 Hz) EEG responses to continuous speech contain enough temporal information to distinguish between different phonetic features and phoneme categories. Using linear regression to estimate the relationship between recorded EEG activity and various features of natural speech, they found that the EEG responses were best predicted by a model that included low-level acoustic as well as high-level phonemic information (but see Daube et al. 2019). It is unclear, however, whether these effects are driven by the neural entrainment of intrinsic low-frequency oscillations to the respective low- or high-level features (i.e., "entrainment in a narrow sense") or whether they are reflective of short-lived stimulus-evoked responses that sit on top of ongoing neural activity (i.e., "entrainment in the broad sense"; Obleser and Kayser 2019).

In sum, there is initial evidence that the synchronization of cortical oscillations presumably downstream from primary auditory cortex guides the extraction of abstract, invariant phoneme representation based on the encoding of the temporal fine structure of speech. However, many additional aspects of the highly complex process of categorical perception, such as influence of spectral cues or the context-specific weighting of concurrent acoustic features, remain to be modeled and explained.

4.3.3 Mapping Form to Meaning

Following the identification of meaningful units such as words based on the results of the acoustic-phonetic analysis of continuous speech, these units have to be mapped onto their conceptual representations stored in long-term memory. This subsection will abstract from many of the intricate details that are involved in the recognition of spoken words and focus more specifically on the process of lexical retrieval. Despite the ubiquitous use of metaphors such as the “mental lexicon” and the “semantic store” that seem to suggest a unique locus of semantic information in the brain, there is ample evidence that semantic memory relies on the concurrent computations in a distributed network that include modality-specific as well as heteromodal brain areas (Binder and Desai 2011; Fernandino et al. 2016). Neural computations involved in semantic processes must therefore integrate neural information across different brain networks.

The representation and retrieval of semantic information builds on domain-general principles supporting the long-term storage of acquired knowledge. As such, it is important to consider how insights gained into the involvement of neural oscillations in processes supporting encoding and retrieval of lexical memory can be linked back to overarching roles of oscillatory dynamics in various memory operations that are independent of speech processing. In general, the neural underpinnings of lexical retrieval have been extensively studied in aphasic populations, using neuroimaging techniques such as functional magnetic resonance imaging (fMRI) and positron emission tomography (PET) (Davis and Gaskell 2009; DeWitt and Rauschecker 2012) or by focusing on event-related potentials (Kutas and Federmeier 2011). In comparison, studies that have explicitly focused on the contribution of oscillatory mechanisms to long-term memory operations in speech processing are relatively sparse.

In mammals, theta oscillations recorded invasively in the hippocampus are not only one of the most prominent neural oscillations but have also been reliably related to processes supporting memory encoding and retrieval (Lisman 2005; Fell and Axmacher 2011). In this context, theta oscillations have been implicated in the coordination of neural networks and in the adjustment of synaptic weights (Buzsáki 2002). Similarly, in humans, theta oscillations in the neocortex and in hippocampus have been associated with episodic memory, the temporal organization of memory content, and short-term memory processes (Lisman and Jensen 2013; Roux and Uhlhaas 2014) which makes them an obvious candidate for the mediation of memory-related processes in language as well.

Indeed, EEG studies have provided evidence that lexical retrieval as part of speech or language comprehension is associated with changes in the power and phase coherence of cortical theta oscillations. The observed neural signatures include the increase in left-parietal theta power for content words compared to function words (e.g., words with little lexical meaning used to express grammatical relationships; Bastiaansen et al. 2005), the phasic upregulation of theta power over temporal sensors for individual words in a sentence (Bastiaansen et al. 2002), and topographically specific increases in theta power for a lexical decision task on words describing visual or auditory semantic properties (Bastiaansen et al. 2008).

An EEG study on semantic priming by Mellem et al. (2013) tested more directly the idea that long-range communication between distant brain regions in the service of semantic retrieval may be mediated by synchronized theta activity. For unrelated compared to related prime-target pairs, the authors found a theta-specific increase in coherence between anterior and posterior channels that reflected a stronger coupling between distant brain regions presumably involved in lexical retrieval. In line with an interpretation that associates increases in theta power with the difficulty of lexical access, Strauß et al. (2014) report a selective enhancement of frontotemporal theta activity during a challenging lexical decision task (see also Obleser and Weisz 2012 for the influence of acoustic factors).

The overall pattern of results, including those of studies that link changes in theta power and coherence to working-memory processes during sentence comprehension (Dillon et al. 2014; Meyer et al. 2015), is generally in line with the proposed role of theta activity in language-related memory operations. However, the observed neural patterns and proposed functions are severely underspecified with respect to the assumed underlying oscillatory mechanisms and their relevance for behavior. In fact, in many of the discussed studies, it is unclear whether the observed modulations in theta activity are indeed reflective of changes in sustained oscillatory dynamics or whether they rather index short intermittent activity changes (see Jones 2016 for a methodological treatment of this topic). Second, even under the assumption that these studies in fact tap into oscillatory dynamics, the role of theta activity in language-related memory processes is rarely spelled out in terms of the underlying mechanistic principles. Unfortunately, these circumstances do not yet allow for a deeper integration of the observed results into the more general context that links theta waves to memory formation and retrieval (Lisman and Jensen 2013; Staudigl and Hanslmayr 2013).

4.3.4 Syntactic Structure Building

As part of the combinatorial nature of human language, smaller units such as words can be grouped into more complex and abstract linguistic structures such as phrases or sentences. Unlike the integration of phonemes into syllables or syllables into words, processes of structure building encompass operations that are more complex than the linear combination of adjacent elements into larger units, as syntactic structures can also be used to establish long-range dependencies within a sentence. However, the precise nature of syntactic structure is a topic of ongoing debate. On the one hand, there is the question of whether syntactic phrase structure necessarily follows hierarchical principles or whether key observations in language use and acquisition could also be reconciled with a simpler, sequential structure (Frank et al. 2012). On the other hand, there is the question of which cues are most informative for syntactic structure building. Two not necessarily mutually exclusive options are that syntactic analysis is guided by learned grammatical rules or based on statistical regularities that are extracted from speech input (Ding et al. 2017).

In light of these alternative assumptions on the processes that support syntactic structure building, it is difficult to derive a definite set of computational operations that are necessarily involved in the syntactical analysis of speech. Compared to the rich body of empirical work and conceptual accounts on the segmentation and analysis of speech at the timescales of syllables and phonemes, the extent to which neural oscillations may be involved in syntactic processes is relatively poorly understood (but see Meyer 2017 for review). Still, there is an emerging perspective that associates neural responses in the delta band with the tracking of syntactic phrases.

In analogy to the observed syllable-rate neural tracking of speech by oscillations in auditory cortex, it was proposed that similar principles may apply for the neural representation of larger linguistic units such as phrases or sentences (Ding et al. 2016; Ghitza 2017). To date, the most comprehensive investigation into the cortical "tracking" of phrase and sentence structures asks whether the analysis of running speech based on rule-based knowledge alone would elicit separable neural responses for different linguistic structures (Ding et al. 2016; Zhang and Ding 2017). To this end, the authors constructed sequences of monosyllabic Chinese words that were stripped of prosodic cues at the phrasal or sentence level and presented at a constant rate of 4 Hz. Based on syntactic knowledge, two adjacent words could be grouped into noun or verb phrases (at the rate of 2 Hz) which in turn could be combined to form short four-word sentences (at the rate of 1 Hz). The authors presented these internally structured sequences to native speakers of Chinese and English. As illustrated in Fig. 4.3d, for Chinese but not English listeners, they observed prominent peaks in the power spectrum at 1, 2, and 4 Hz which they interpret as a reflection of the concurrent tracking of syllables, phrases, and sentences by neural activity in networks beyond primary auditory cortex.

Going beyond the interpretation of synchronized delta oscillations as a means of tracking linguistic structures at the phrase or sentence level, Meyer and colleagues (Meyer et al. 2017) argue for a causal role of delta oscillations in the application of an internally represented chunking strategy. More precisely, they propose that delta phase serves as a tool to implement an internal bias for establishing phrase boundaries that operates somewhat independent of the acoustic evidence provided. This is an alluring idea but so far is mostly linked to a truly neural oscillatory phenomenon by conjecture. The EEG of listeners showed a significant difference in frontotemporal delta phase for the interpretation of syntactically ambiguous sentences that was either in line with acoustic cues or followed an internal bias for grouping words into phrases (Meyer 2017).

Both interpretations of the role of delta oscillations in syntactic analysis share a number of implicit assumptions. First, they assume that phrasal units, whether they are hierarchically organized or not, occur rhythmically enough to connect their formation or encoding to neural mechanisms of an oscillatory nature – a tenet for neural entrainment as mentioned in Sect. 4.2. While such a strong temporal regularity was clearly given in the paradigm employed by Ding et al. (2016) who presented words at a constant rate of 4 Hz, the same degree of rhythmicity is hardly present in natural, continuous speech. Second, in connecting the observed neural pattern to the phenomenon of neural entrainment, both accounts imply that there are indeed ongoing oscillations operating at adequate frequencies that could rapidly or gradually align their phase and period to an externally or internally generated periodic signal. However, it remains to be empirically shown that this is in fact the case and which neural networks participate in the generation of these spontaneous brain rhythms.

Lastly, it is unclear whether the postulated oscillatory behavior of delta activity would be the cause or the consequence of the encoding of phrasal structure in speech. If the application of internally represented grammatical knowledge leads to cortical tracking of phrase structure via oscillatory delta activity, then the mechanisms instrumental to this internal representation remain unsolved. In turn, if oscillatory delta activity is causing the formation of phrasal structure, then it is equally unclear how the alignment of this oscillatory activity to the speech signal can be mechanistically implemented.

To sum up this section, while there is some evidence of consistent neural signatures in the processing of phrasal structure, either based on acoustic cues or learned generalizations, the translation of the involved computations into oscillatory mechanisms is not always straightforward. Conceptually, the proposed roles of neural oscillations in the generation of phrasal and sentence-level structures need to be spelled out more explicitly so that they can be more rigorously tested in empirical work.

4.3.5 Sentence-Level Speech Comprehension

The establishment of sentence- or discourse-level meaning is the ultimate goal of human speech perception. The integration of meaning across a wider context requires more than a simple combination of word-level meaning. Instead, speech comprehension is guided by a range of contextual cues provided by the acoustic signal, the syntactic structure, global and local semantic associations, as well as pragmatic knowledge. To enable successful communication, speech comprehension has to be highly flexible and adaptive. It thus relies on a complex set of computations that help to integrate information across various sources and to compare the outcome to continuously generated predictions. In this sense, high-level speech comprehension depends, to a varying degree, on all of the thus far discussed core computations and neural networks.

The question of how contextual predictions influence semantic integration at the sentence level has been extensively researched in the context of event-related potentials, especially in reference to the so-called N400 component (Lau et al. 2008; Kutas and Federmeier 2011). The N400 component describes a negative deflection in the time-locked EEG that peaks around 400 ms after stimulus onset and is particularly sensitive to meaningful stimuli including but not limited to auditory or visual words, pictures, and faces (Kutas and Federmeier 2011). However, the role of neural oscillations in the computation of high-level linguistic meaning, and how their involvement relates to other neural signatures of syntactic or semantic processing, remains insufficiently understood. One of the fundamental challenges lies in the functional dissection of the implicated processes into core computations that could be plausibly supported by oscillatory dynamics. The translation between predictive processes at the cognitive level and the implementation of predictive coding at the level of neural microcircuits (Bastos et al. 2012; Womelsdorf et al. 2014) pose a particularly challenging problem.

Early research into the contributions of oscillatory dynamics to sentence-level comprehension has predominantly focused on the role of beta and gamma activity (Lewis et al. 2015). More specifically, the results of numerous M/EEG studies seemed to suggest a differential involvement of beta and gamma band activity in sentence-level comprehension: changes in beta band power or phase coherence were associated with syntactic integration processes (Bastiaansen and Hagoort 2006; Bastiaansen et al. 2010), whereas changes in gamma band activity were thought to impact semantic integration (Hagoort et al. 2004; Hald et al. 2006). More recently, however, it has been argued that the observed dynamic changes in beta and gamma band activity might be better explained in a predictive coding framework (Lewis and Bastiaansen 2015; Lewis et al. 2016).

The proposed interpretation of changes in beta and gamma activity as instances of domain-general predictive processes builds on recent accounts on the principled cortical organization and neural dynamics that enable predictive coding (Engel and Fries 2010; Bastos et al. 2012). Bastos et al. (2012) describe a canonical microcircuit at the level of the cortical column, in which top-down and bottom-up inputs are segregated into distinct cortical layers and frequency bands. They argue for a functional asymmetry in the computational roles of neural activity in superficial compared to deep cortical layers that is mirrored by an asymmetry in frequency content: high-frequency gamma range oscillations generated by cell ensembles in superficial layers are thought to be engaged in the propagation of signals from lower to higher cortical areas in a bottom-up fashion. By contrast, slower oscillations in the beta range that originate from cells in deep cortical layers are associated with top-down feedback projections. This proposal is generally in line with an account by Engel and Fries (2010) who assume a role of beta oscillations in signaling the maintenance or change of a current sensorimotor or cognitive state.

In translating the proposed functional asymmetry of beta and gamma band activity to predictive processes in sentence-level comprehension, Lewis and colleagues (Lewis and Bastiaansen 2015; Lewis et al. 2015) suggest that changes in beta band synchrony serve two distinct purposes. On the one hand, changes relate to the propagation of context-based top-down predictions about upcoming inputs from higher to lower cortical areas. On the other hand, they support the active maintenance of the current cognitive network configuration engaged in the computation and representation of sentence-level meaning. Gamma band oscillations, by contrast, are assumed to play a role in the comparison of predicted to actual neural inputs, and the transmission of the resulting prediction error signals from lower to higher areas in the cortical hierarchy (Sedley et al. 2016).

The notion that predictive processes at the neural as well as at the cognitive level play a crucial role in the highly flexible operations of speech perception and speech comprehension is by now generally accepted (Sohoglu et al. 2012; Peelle and Sommers 2015; Arnal and Giraud 2018). Nevertheless, it is not straightforward to infer the mapping of broad, often scalp-level, modulations in beta and gamma band activity to precise computational mechanisms at the level of cortical microcircuits. As such, the functional role of changes in beta and gamma band activity observed in sentence comprehension remains elusive and will need to be tested much more rigorously.

4.4 Summary, Conclusions, and Future Directions

This chapter has asked how neural oscillations support human speech perception and speech comprehension. In answering this question, it focused on the patterns that emerge from the integration of speech-related neural oscillatory signatures with mechanistic principles and core functions that support perception more generally.

Overall, it has become clear that cognitive core functions of oscillatory dynamics are more readily linked to processes involved in the lower-level perceptual analysis of speech than processes involved in high-level speech comprehension. Does this mean that neural oscillations are not crucially engaged in high-level speech processes? Given the ubiquitous nature of neural oscillations and their proposed role in predictive processes that are highly relevant to complex behavior such as speech comprehension (Park et al. 2015; Chao et al. 2018), this conclusion would seem overly pessimistic. However, understanding how neural oscillations support the perceptual and linguistic analysis of speech will require a deeper understanding of how functional principles studied at the microcircuit level relate to the dynamics of larger distributed neural networks. To push our insights into the contributions of neural oscillations to speech perception means to also recognize their limitations. This entails close scrutiny of whether the neural implementation of a particular linguistic process is indeed best understood from a perspective focused on oscillatory mechanisms.

Nevertheless, the evidence reviewed in this chapter is generally in line with an emerging functional segregation of neural oscillations depending on the core functions they subserve. That is, oscillations in the delta, theta, and gamma band are more closely related to the analysis and propagation of sensory information, whereas beta and alpha oscillations, the latter of which have not been discussed in great detail in this chapter, are engaged in the implementation of internally represented states and goals (van Kerkoerle et al. 2014; Sedley et al. 2016).

Lastly, understanding the degree to which neural oscillations represent necessary, or even sufficient, neural means of speech processing hinges on future studies that will investigate the functional relevance to behavior more directly (Thut et al. 2012). Here, this chapter has highlighted three particularly promising research strategies: studies that (i) bring together changes in neural dynamics with changes in fine-grained behavioral measures, (ii) compare empirical evidence to predictions generated by simplified neurocomputational models, or (iii) that directly modulate ongoing neural oscillations via noninvasive brain stimulation to study the causal link from neural oscillations to communication behavior.

Acknowledgments This work was supported by an ERC grant (ERC-CoG-2014 AUDADAPT) to Jonas Obleser.

Compliance with Ethics Requirements Sarah Tune declares that she has no conflict of interest.

Jonas Obleser declares that he has no conflict of interest.

References

Ahissar E, Nagarajan S, Ahissar M et al (2001) Speech comprehension is correlated with temporal response patterns recorded from auditory cortex. Proc Natl Acad Sci U S A 98:13367–13372. https://doi.org/10.1073/pnas.201400998

Arnal LH, Giraud AL (2018) Cortical oscillations and sensory predictions. Trends Cogn Sci 16:1–9. https://doi.org/10.1016/j.tics.2012.05.003

Bastiaansen M, Hagoort P (2006) Oscillatory neuronal dynamics during language comprehension. Prog Brain Res 159:179–196. https://doi.org/10.1016/S0079-6123(06)59012-0

Bastiaansen MCM, Van Berkum JJA, Hagoort P (2002) Event-related theta power increases in the human EEG during online sentence processing. Neurosci Lett 323:13–16. https://doi.org/10.1016/S0304-3940(01)02535-6

Bastiaansen MCM, Linden MVD, Keurs MT et al (2005) Theta responses are involved in lexical—semantic retrieval during language processing. J Cogn Neurosci 17:530–541. https://doi.org/10.1162/0898929053279469

Bastiaansen MCM, Oostenveld R, Jensen O, Hagoort P (2008) I see what you mean: theta power increases are involved in the retrieval of lexical semantic information. Brain Lang 106:15–28. https://doi.org/10.1016/j.bandl.2007.10.006

Bastiaansen M, Magyari L, Hagoort P (2010) Syntactic unification operations are reflected in oscillatory dynamics during on-line sentence comprehension. J Cogn Neurosci 22:1333–1347. https://doi.org/10.1162/jocn.2009.21283

Bastos AM, Usrey WM, Adams RA et al (2012) Canonical microcircuits for predictive coding. Neuron 76:695–711. https://doi.org/10.1016/j.neuron.2012.10.038

Berger H (1929) Über das Elektrenkephalogramm des Menschen. Eur Arch Psychiatry Clin Neurosci 87:527–570

Binder JR, Desai RH (2011) The neurobiology of semantic memory. Trends Cogn Sci 15:527–536. https://doi.org/10.1016/j.tics.2011.10.001

Bonnefond M, Kastner S, Jensen O (2017) Communication between brain areas based on nested oscillations. eNeuro. https://doi.org/10.1523/ENEURO.0153-16.2017

Buzsáki G (2002) Theta oscillations in the hippocampus. Neuron 33:325–340. https://doi.org/10.1016/S0896-6273(02)00586-X

Buzsáki G (2006) Rhythms of the brain. Oxford University Press

Buzsáki G, Draguhn A (2004) Neuronal oscillations in cortical networks. Science 304:1926–1929. https://doi.org/10.1126/science.1099745

Buzsáki G, Anastassiou CA, Koch C (2012) The origin of extracellular fields and currents — EEG, ECoG, LFP and spikes. Nat Rev Neurosci 13:407–420. https://doi.org/10.1038/nrn3241

Canolty RT, Knight RT (2010) The functional role of cross-frequency coupling. Trends Cogn Sci 14:506–515. https://doi.org/10.1016/j.tics.2010.09.001

Canolty RT, Edwards E, Dalal SS et al (2006) High gamma power is phase-locked to theta oscillations in human neocortex. Science 313:1626–1628. https://doi.org/10.1126/science.1128115

Caporale N, Dan Y (2008) Spike timing–dependent plasticity: a Hebbian learning rule. Annu Rev Neurosci 31:25–46. https://doi.org/10.1146/annurev.neuro.31.060407.125639

Chao ZC, Takaura K, Wang L et al (2018) Large-scale cortical networks for hierarchical prediction and prediction error in the primate brain. Neuron 100:1252–1266.e3. https://doi.org/10.1016/j.neuron.2018.10.004

Cohen MX (2017) Where does EEG come from and what does it mean? Trends Neurosci 40:208–218. https://doi.org/10.1016/j.tins.2017.02.004

Crone NE, Sinai A, Korzeniewska A (2006) High-frequency gamma oscillations and human brain mapping with electrocorticography. In: Neuper C, Klimesch W (eds) Progress in brain research. Elsevier, pp 275–295

da Silva FL (2013) EEG and MEG: relevance to neuroscience. Neuron 80:1112–1128. https://doi.org/10.1016/j.neuron.2013.10.017

Daube C, Ince RAA, Gross J (2019) Simple acoustic features can explain phoneme- based predictions of cortical responses to speech. Curr Biol 29:1924–1937.e9. https://doi.org/10.1016/j.cub.2019.04.067

Davis MH, Gaskell MG (2009) A complementary systems account of word learning: neural and behavioural evidence. Philos Trans R Soc Lond B Biol Sci 364:3773–3800. https://doi.org/10.1523/JNEUROSCI.4587-03.2004

DeWitt I, Rauschecker JP (2012) Phoneme and word recognition in the auditory ventral stream. Proc Natl Acad Sci U S A 109:E505–E514. https://doi.org/10.1073/pnas.1113427109

Di Liberto GM, O'Sullivan JA, Lalor EC (2015) Low-frequency cortical entrainment to speech reflects phoneme-level processing. Curr Biol 25:2457–2465. https://doi.org/10.1016/j.cub.2015.08.030

Dillon B, Chow W-Y, Wagers M et al (2014) The structure-sensitivity of memory access: evidence from Mandarin Chinese. Front Psychol 5:1–16. https://doi.org/10.3389/fpsyg.2014.01025

Ding N, Simon JZ (2012a) Emergence of neural encoding of auditory objects while listening to competing speakers. Proc Natl Acad Sci U S A 109:11854–11859. https://doi.org/10.1073/pnas.1205381109

Ding N, Simon JZ (2012b) Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. J Neurophysiol 107:78–89. https://doi.org/10.1152/jn.00297.2011

Ding N, Simon JZ (2013) Adaptive temporal encoding leads to a background-insensitive cortical representation of speech. J Neurosci 33:5728–5735. https://doi.org/10.1523/JNEUROSCI.5297-12.2013

Ding N, Simon JZ (2014) Cortical entrainment to continuous speech: functional roles and interpretations. Front Hum Neurosci 8:13367. https://doi.org/10.3389/fnhum.2014.00311

Ding N, Chatterjee M, Simon JZ (2014) Robust cortical entrainment to the speech envelope relies on the spectro-temporal fine structure. NeuroImage 88:41–46. https://doi.org/10.1016/j.neuroimage.2013.10.054

Ding N, Zhang H, Tian X, Poeppel D (2016) Cortical tracking of hierarchical linguistic structures in connected speech. Nat Neurosci 19:158–164. https://doi.org/10.1038/nn.4186

Ding N, Melloni L, Tian X, Poeppel D (2017) Rule-based and word-level statistics-based processing of language: insights from neuroscience. Lang Cogn Neurosci 32:570–575. https://doi.org/10.1080/23273798.2016.1215477

Engel AK, Fries P (2010) Beta-band oscillations – signalling the status quo? Curr Opin Neurobiol 20:156–165. https://doi.org/10.1016/j.conb.2010.02.015

Engel AK, Fries P, Singer W (2001) Dynamic predictions: oscillations and synchrony in top–down processing. Nat Rev Neurosci 2:704–716

Fell J, Axmacher N (2011) The role of phase synchronization in memory processes. Nat Rev Neurosci 12:105–118. https://doi.org/10.1038/nrn2979

Fernandino L, Binder JR, Desai RH et al (2016) Concept representation reflects multimodal abstraction: a framework for embodied semantics. Cereb Cortex 26:2018–2034. https://doi.org/10.1093/cercor/bhv020

Frank SL, Bod R, Christiansen MH (2012) How hierarchical is language use? Proc Biol Sci 279:4522–4531. https://doi.org/10.1098/rspb.2012.1741

Fries P (2005) A mechanism for cognitive dynamics: neuronal communication through neuronal coherence. Trends Cogn Sci 9:474–480. https://doi.org/10.1016/j.tics.2005.08.011

Fries P (2009) Neuronal gamma-band synchronization as a fundamental process in cortical computation. Annu Rev Neurosci 32:209–224. https://doi.org/10.1146/annurev.neuro.051508.135603

Fries P (2015) Rhythms for cognition: communication through coherence. Neuron 88:220–235. https://doi.org/10.1016/j.neuron.2015.09.034

Fries P, Nikolić D, Singer W (2007) The gamma cycle. Trends Neurosci 30:309–316. https://doi.org/10.1016/j.tins.2007.05.005

Ghitza O (2011) Linking speech perception and neurophysiology: speech decoding guided by cascaded oscillators locked to the input rhythm. Front Psychol. https://doi.org/10.3389/fpsyg.2011.00130/abstract

Ghitza O (2017) Acoustic-driven delta rhythms as prosodic markers. Lang Cogn Neurosci 32:545–561. https://doi.org/10.1080/23273798.2016.1232419

Giraud AL, Poeppel D (2012) Cortical oscillations and speech processing: emerging computational principles and operations. Nat Neurosci 15:511–517. https://doi.org/10.1038/nn.3063

Glass L, Sun J (1994) Periodic forcing of a limit-cycle oscillator: fixed points, Arnold tongues, and the global organization of bifurcations. Phys Rev E 50:5077–5084. https://doi.org/10.1103/PhysRevE.50.5077

Gray CM, König P, Engel AK, Singer W (1989) Oscillatory responses in cat visual cortex exhibit inter-columnar synchronization which reflects global stimulus properties. Nature 338:334–337. https://doi.org/10.1038/338334a0

Gross J, Hoogenboom N, Thut G et al (2013) Speech rhythms and multiplexed oscillatory sensory coding in the human brain. PLoS Biol 11:e1001752–e1001714. https://doi.org/10.1371/journal.pbio.1001752

Hagoort P, Hald LA, Bastiaansen M, Petersson KM (2004) Integration of word meaning and world knowledge in language comprehension. Science 304:438–441. https://doi.org/10.1126/science.1095455

Hald LA, Bastiaansen MCM, Hagoort P (2006) EEG theta and gamma responses to semantic violations in online sentence processing. Brain Lang 96:90–105. https://doi.org/10.1016/j.bandl.2005.06.007

Hansen P, Kringelbach M, Salmelin R (2010) MEG. Oxford University Press

Hebb DO (1949) The organization of behavior. Wiley, New York

Henry MJ, Herrmann B, Kunke D, Obleser J (2017) Aging affects the balance of neural entrainment and top-down neural modulation in the listening brain. Nat Commun 8:15801. https://doi.org/10.1038/ncomms15801

Herrmann B, Henry MJ (2012) Neural oscillations in speech: don't be enslaved by the envelope. Front Hum Neurosci 6:250. https://doi.org/10.3389/fnhum.2012.00250

Holt LL, Lotto AJ (2010) Speech perception as categorization. Atten Percept Psychophys 72:1218–1227. https://doi.org/10.3758/APP.72.5.1218

Hovsepyan S, Olasagasti I, Giraud AL (2020) Combining predictive coding and neural oscillations enables online syllable recognition in natural speech. Nat Commun 11:1–12. https://doi.org/10.1038/s41467-020-16956-5

Howard MF, Poeppel D (2010) Discrimination of speech stimuli based on neuronal response phase patterns depends on acoustics but not comprehension. J Neurophysiol 104:2500–2511. https://doi.org/10.1152/jn.00251.2010

Hyafil A, Fontolan L, Kabdebon C et al (2015a) Speech encoding by coupled cortical theta and gamma oscillations. eLife Sci 4:e06213–e06245. https://doi.org/10.7554/eLife.06213

Hyafil A, Giraud AL, Fontolan L, Gutkin B (2015b) Neural cross-frequency coupling: connecting architectures, mechanisms, and functions. Trends Neurosci 38:725–740. https://doi.org/10.1016/j.tins.2015.09.001

Jensen O, Lisman JE (2000) Position reconstruction from an ensemble of hippocampal place cells: contribution of theta phase coding. J Neurophysiol 83:2602–2609. https://doi.org/10.1152/jn.2000.83.5.2602

Jensen O, Mazaheri A (2010) Shaping functional architecture by oscillatory alpha activity: gating by inhibition. Front Hum Neurosci 4:186. https://doi.org/10.3389/fnhum.2010.00186

Jones SR (2016) When brain rhythms aren't 'rhythmic': implication for their mechanisms and meaning. Curr Opin Neurobiol 40:72–80. https://doi.org/10.1016/j.conb.2016.06.010

Kahana MJ (2006) The cognitive correlates of human brain oscillations. J Neurosci 26:1669–1672. https://doi.org/10.1523/JNEUROSCI.3737-05c.2006

Kayser C, Ince RAA, Panzeri S (2012) Analysis of slow (theta) oscillations as a potential temporal reference frame for information coding in sensory cortices. PLoS Comput Biol 8:e1002717. https://doi.org/10.1371/journal.pcbi.1002717

Kayser C, Wilson C, Safaai H et al (2015) Rhythmic auditory cortex activity at multiple timescales shapes stimulus-response gain and background firing. J Neurosci 35:7750–7762. https://doi.org/10.1523/JNEUROSCI.0268-15.2015

Klimesch W, Sauseng P, Hanslmayr S (2007) EEG alpha oscillations: the inhibition–timing hypothesis. Brain Res Rev 53:63–88. https://doi.org/10.1016/j.brainresrev.2006.06.003

Kopell N, Kramer MA, Malerba P, Whittington MA (2010) Are different rhythms good for different functions? Front Hum Neurosci 4:1–9. https://doi.org/10.3389/fnhum.2010.00187

Kösem A, van Wassenhove V (2017) Distinct contributions of low- and high-frequency neural oscillations to speech comprehension. Lang Cogn Neurosci 32:536–544. https://doi.org/10.1080/23273798.2016.1238495

Kutas M, Federmeier KD (2011) Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP). Annu Rev Psychol 62:621–647

Lakatos P, Shah AS, Knuth KH et al (2005) An oscillatory hierarchy controlling neuronal excitability and stimulus processing in the auditory cortex. J Neurophysiol 94:1904–1911. https://doi.org/10.1152/jn.00263.2005

Lakatos P, Karmos G, Mehta AD et al (2008) Entrainment of neuronal oscillations as a mechanism of attentional selection. Science 320:110–113. https://doi.org/10.1126/science.1154735

Lakatos P, Musacchia G, O'Connel MN et al (2013) The spectrotemporal filter mechanism of auditory selective attention. Neuron 77:750–761. https://doi.org/10.1016/j.neuron.2012.11.034

Lau EF, Phillips C, Poeppel D (2008) A cortical network for semantics: (de)constructing the N400. Nat Rev Neurosci 9:920–933. https://doi.org/10.1038/nrn2532

Lewis AG, Bastiaansen M (2015) A predictive coding framework for rapid neural dynamics during sentence-level language comprehension. Cortex 68:155–168. https://doi.org/10.1016/j.cortex.2015.02.014

Lewis AG, Wang L, Bastiaansen M (2015) Fast oscillatory dynamics during language comprehension: unification versus maintenance and prediction? Brain Lang 148:51–63. https://doi.org/10.1016/j.bandl.2015.01.003

Lewis AG, Schoffelen J-M, Schriefers H, Bastiaansen M (2016) A predictive coding perspective on beta oscillations during sentence-level language comprehension. Front Hum Neurosci 10:179–176. https://doi.org/10.3389/fnhum.2016.00085

Lisman J (2005) The theta/gamma discrete phase code occurring during the hippocampal phase precession may be a more general brain coding scheme. Hippocampus 15:913–922. https://doi.org/10.1002/hipo.20121

Lisman JE, Jensen O (2013) The theta-gamma neural code. Neuron 77:1002–1016. https://doi.org/10.1016/j.neuron.2013.03.007

Luo H, Poeppel D (2007) Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. Neuron 54:1001–1010. https://doi.org/10.1016/j.neuron.2007.06.004

Masquelier T, Hugues E, Deco G, Thorpe SJ (2009) Oscillations, phase-of-firing coding, and spike timing-dependent plasticity: an efficient learning scheme. J Neurosci 29:13484–13493. https://doi.org/10.1523/JNEUROSCI.2207-09.2009

Mellem MS, Friedman RB, Medvedev AV (2013) Gamma- and theta-band synchronization during semantic priming reflect local and long-range lexical–semantic networks. Brain Lang 127:440–451. https://doi.org/10.1016/j.bandl.2013.09.003

Mesgarani N, Chang EF (2012) Selective cortical representation of attended speaker in multi-talker speech perception. Nature 485:233–236. https://doi.org/10.1038/nature11020

Meyer L (2017) The neural oscillations of speech processing and language comprehension: state of the art and emerging mechanisms. Eur J Neurosci 28:3958. https://doi.org/10.1111/ejn.13748

Meyer L, Grigutsch M, Schmuck N et al (2015) Frontal–posterior theta oscillations reflect memory retrieval during sentence comprehension. Cortex 71:205–218. https://doi.org/10.1016/j.cortex.2015.06.027

Meyer L, Henry MJ, Gaston P et al (2017) Linguistic bias modulates interpretation of speech via neural delta-band oscillations. Cereb Cortex 27:4293–4302. https://doi.org/10.1093/cercor/bhw228

Meyer L, Sun Y, Martin AE (2019) Synchronous, but not entrained: exogenous and endogenous cortical rhythms of speech and language processing. Lang Cogn Neurosci 0:1–11. https://doi.org/10.1080/23273798.2019.1693050

Miller JL, Aibel IL, Green K (1984) On the nature of rate-dependent processing during phonetic perception. Percept Psychophys 35:5–15. https://doi.org/10.3758/BF03205919

Niedermeyer E, Lopes da Silva FH (1999) Electroencephalography. Lippincott Williams & Wilkins

Nourski KV, Brugge JF (2011) Representation of temporal sound features in the human auditory cortex. Rev Neurosci 22:187–203. https://doi.org/10.1515/RNS.2011.016

Nourski KV, Reale RA, Oya H et al (2009) Temporal envelope of time-compressed speech represented in the human auditory cortex. J Neurosci 29:15564–15574. https://doi.org/10.1523/JNEUROSCI.3065-09.2009

Nunez PL, Srinivasan R (2006) Electric fields of the brain. Oxford University Press, USA

O'Sullivan JA, Power AJ, Mesgarani N et al (2014) Attentional selection in a cocktail party environment can be decoded from single-trial EEG. Cereb Cortex 25:1697–1706. https://doi.org/10.1093/cercor/bht355

Obleser J, Kayser C (2019) Neural entrainment and attentional selection in the listening brain. Trends Cogn Sci 23:913–926. https://doi.org/10.1016/j.tics.2019.08.004

Obleser J, Weisz N (2012) Suppressed alpha oscillations predict intelligibility of speech and its acoustic details. Cereb Cortex 22:2466–2477. https://doi.org/10.1093/cercor/bhr325

Obleser J, Henry MJ, Lakatos P (2017) What do we talk about when we talk about rhythm? PLoS Biol 15:e2002794–e2002795. https://doi.org/10.1371/journal.pbio.2002794

Oever ten S, Sack AT (2015) Oscillatory phase shapes syllable perception. Proc Natl Acad Sci U S A 112:15833–15837. https://doi.org/10.1073/pnas.1517519112

Oever ten S, Meierdierks T, Duecker F et al (2020) Phase-coded oscillatory ordering promotes the separation of closely matched representations to optimize perceptual discrimination. iScience 23:101282. https://doi.org/10.1016/j.isci.2020.101282

Park H, Ince RAA, Schyns PG et al (2015) Frontal top-down signals increase coupling of auditory low-frequency oscillations to continuous speech in human listeners. Curr Biol 25:1649–1653. https://doi.org/10.1016/j.cub.2015.04.049

Peelle JE, Davis MH (2012) Neural oscillations carry speech rhythm through to comprehension. Front Psychol 3:320. https://doi.org/10.3389/fpsyg.2012.00320

Peelle JE, Sommers MS (2015) Prediction and constraint in audiovisual speech perception. Cortex 68:1–13. https://doi.org/10.1016/j.cortex.2015.03.006

Peelle JE, Gross J, Davis MH (2013) Phase-locked responses to speech in human auditory cortex are enhanced during comprehension. Cereb Cortex 23:1378–1387. https://doi.org/10.1093/cercor/bhs118

Peña M, Melloni L (2012) Brain oscillations during spoken sentence processing. J Cogn Neurosci 24:1149–1164. https://doi.org/10.1162/jocn_a_00144

Pikovsky A, Rosenblum M, Kurths J (2003) Synchronization. Cambridge University Press

Port R (1976) Influence of tempo on the closure interval cue to the voicing and place of intervocalic stops. J Acoust Soc Am 59:S41–S42. https://doi.org/10.1121/1.2002689

Riecke L, Formisano E, Sorger B et al (2018) Neural entrainment to speech modulates speech intelligibility. Curr Biol 28:1–9. https://doi.org/10.1016/j.cub.2017.11.033

Rosen S (1992) Temporal information in speech: acoustic, auditory and linguistic aspects. Philos Trans R Soc Lond Ser B Biol Sci 336:367–373. https://doi.org/10.1098/rstb.1992.0070

Roux F, Uhlhaas PJ (2014) Working memory and neural oscillations: alpha–gamma versus theta–gamma codes for distinct WM information? Trends Cogn Sci 18:16–25. https://doi.org/10.1016/j.tics.2013.10.010

Sauseng P, Klimesch W, Gruber WR et al (2007) Are event-related potential components generated by phase resetting of brain oscillations? A critical discussion. Neuroscience 146:1435–1444. https://doi.org/10.1016/j.neuroscience.2007.03.014

Schroeder CE, Lakatos P (2009) Low-frequency neuronal oscillations as instruments of sensory selection. Trends Neurosci 32:9–18. https://doi.org/10.1016/j.tins.2008.09.012

Schroeder CE, Wilson DA, Radman T et al (2010) Dynamics of active sensing and perceptual selection. Curr Opin Neurobiol 20:172–176. https://doi.org/10.1016/j.conb.2010.02.010

Sedley W, Gander PE, Kumar S et al (2016) Neural signatures of perceptual inference. eLife Sci. https://doi.org/10.7554/eLife.11476

Sejnowski TJ (2006) Network oscillations: emerging computational principles. J Neurosci 26:1673–1676. https://doi.org/10.1523/JNEUROSCI.3737-05d.2006

Singer W (2018) Neuronal oscillations: unavoidable and useful? Eur J Neurosci 41:403–410. https://doi.org/10.1111/ejn.13796

Singer W, Gray CM (1995) Visual feature integration and the temporal correlation hypothesis. Annu Rev Neurosci 18:555–586. https://doi.org/10.1146/annurev.ne.18.030195.003011

Sohoglu E, Peelle JE, Carlyon RP, Davis MH (2012) Predictive top-down integration of prior knowledge during speech perception. J Neurosci 32:8443–8453. https://doi.org/10.1523/JNEUROSCI.5069-11.2012

Staudigl T, Hanslmayr S (2013) Theta oscillations at encoding mediate the context-dependent nature of human episodic memory. Curr Biol 23:1101–1106. https://doi.org/10.1016/j.cub.2013.04.074

Steriade M, Gloor P, Llinás RR et al (1990) Basic mechanisms of cerebral rhythmic activities. Electroencephalogr Clin Neurophysiol 76:481–508. https://doi.org/10.1016/0013-4694(90)90001-Z

Strauß A, Kotz SA, Scharinger M, Obleser J (2014) Alpha and theta brain oscillations index dissociable processes in spoken word recognition. NeuroImage 97:387–395. https://doi.org/10.1016/j.neuroimage.2014.04.005

Thut G, Miniussi C, Gross J (2012) The functional importance of rhythmic activity in the brain. Curr Biol 22:R658–R663. https://doi.org/10.1016/j.cub.2012.06.061

van Kerkoerle T, Self MW, Dagnino B et al (2014) Alpha and gamma oscillations characterize feedback and feedforward processing in monkey visual cortex. Proc Natl Acad Sci 111:14332–14341. https://doi.org/10.1073/pnas.1402773111

Voloh B, Womelsdorf T (2016) A role of phase-resetting in coordinating large scale neural networks during attention and goal-directed behavior. Front Syst Neurosci 10:308–319. https://doi.org/10.3389/fnsys.2016.00018

Wang X-J (2010) Neurophysiological and computational principles of cortical rhythms in cognition. Physiol Rev 90:1195–1268. https://doi.org/10.1152/physrev.00035.2008

Ward LM (2003) Synchronous neural oscillations and cognitive processes. Trends Cogn Sci 7:553–559. https://doi.org/10.1016/j.tics.2003.10.012

Whittington MA, Traub RD, Kopell N et al (2000) Inhibition-based rhythms: experimental and mathematical observations on network dynamics. Int J Psychophysiol 38:315–336

Wilsch A, Neuling T, Obleser J, Herrmann CS (2018) Transcranial alternating current stimulation with speech envelopes modulates speech comprehension. NeuroImage 172:1–25. https://doi.org/10.1016/j.neuroimage.2018.01.038

Womelsdorf T, Valiante TA, Sahin NT et al (2014) Dynamic circuit motifs underlying rhythmic gain control, gating and integration. Nat Neurosci 17:1031–1039. https://doi.org/10.1038/nn.3764

Zhang W, Ding N (2017) Time-domain analysis of neural tracking of hierarchical linguistic structures. NeuroImage 146:333–340. https://doi.org/10.1016/j.neuroimage.2016.11.016

Zion Golumbic EM, Ding N, Bickel S et al (2013) Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party". Neuron 77:980–991. https://doi.org/10.1016/j.neuron.2012.12.037

Zoefel B, VanRullen R (2016) EEG oscillations entrain their phase to high-level features of speech sound. NeuroImage 124:16–23. https://doi.org/10.1016/j.neuroimage.2015.08.054

Chapter 5
Extracting Language Content from Speech Sounds: The Information Theoretic Approach

Laura Gwilliams and Matthew H. Davis

Abstract Speech comprehension involves recovering a speaker's intended meaning from the speech sounds that they produce. While the sensory-driven components of this process have been widely investigated, the impact of speech content (i.e., linguistic information) on sensory processing is much less understood. Here we summarize the growing body of research demonstrating that neural processing of speech sounds is influenced by morpheme- and word-level statistical properties of the information conveyed. We introduce and review evidence that information theoretic measures such as entropy and surprisal are apparent in neural responses. These findings help uncover fundamental organizational principles of the language system: what units are stored and how they are accessed. Modeling sensitivity to the information content of the speech signal helps explain the interface between (i) auditory processes operating on speech sounds and (ii) the words and meanings that those sounds convey.

Keywords Surprisal · Entropy · Prediction · Brain · Language processing · Information content · Lexical access · Predictive coding

L. Gwilliams (✉)
Department of Neurosurgery, University of California, San Francisco,
San Francisco, CA, USA
e-mail: laura.gwilliams@ucsf.edu

M. H. Davis
MRC Cognition and Brain Sciences Unit, Cambridge University, Cambridge, UK
e-mail: matt.davis@mrc-cbu.cam.ac.uk

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022

L. L. Holt et al. (eds.), *Speech Perception*, Springer Handbook of Auditory
Research 74, https://doi.org/10.1007/978-3-030-81542-4_5

5.1 Introduction

Speech is a means of exchanging information. Through verbal communication, humans have the unique ability to convey a potentially limitless number of thoughts and ideas through their utterances (Chomsky 2000), and infer the thoughts of others from what they say.

Speaker-listener interactions can be formally described as a *communication system* (Shannon 1948) (Fig. 5.1). The role of the speaker is to conceive of a message and encode it in the auditory signals they produce. These signals are decomposed into elemental time-frequency representations by the cochlea (Shamma 1985; Moore 2008), before being passed to auditory cortex. The role of the listener's brain is to decode the intended message from the signals that reach cortex, whereby communication can be considered successful to the extent that the intended conceptual message of the speaker matches the reconstructed conceptual message of the listener.

Although listening to someone talk *feels* like an effortless passive process, speech comprehension involves overcoming some major computational challenges. Not least because the mapping from acoustics to meaning is largely arbitrary (De Saussure 2011), different speakers have vastly different ways of pronouncing words depending on biological, regional, and incidental factors (Stevens and Blumstein 1981), and external noise, such as the voices of surrounding talkers or nonlinguistic noise sources, often masks the signal (Mattys et al. 2012) (see also Chap. 6, Van Hedger and Johnsrude). The extent of this challenge is exemplified by the fact that, despite the vast amounts of money and time invested, current state-of-the-art automatic speech recognition systems do not rival the accuracy, speed, or robustness to

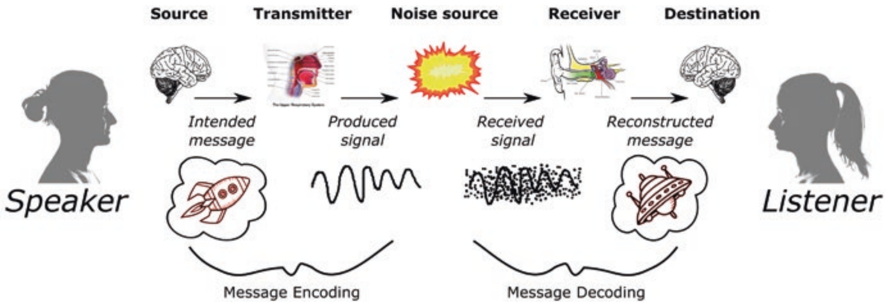

Fig. 5.1 Schematic diagram of the human communication system. Verbal exchanges involve a speaker conceiving of a message and encoding that message into complex temporal-spectral patterns through their vocal articulators. As the signal travels through the air, it may be contaminated with external sources of sound, such as other people speaking or noises from the environment. This contaminated acoustic signal is received by the auditory system of the listener and passed to auditory cortex for processing. The brain of the listener then needs to decode the original message from the auditory signals that were given as input. Here we see that the intended message that was encoded – a red space rocket – closely resembles but is not identical to the decoded message – a red alien spaceship

speaker variability demonstrated by human listeners (O'Shaughnessy 2008; Graves et al. 2013).

The purpose of speech, in sum, is not to exchange auditory signals but to exchange information content. And within the structure of language, content takes the form of linguistic units such as morphemes, words, and phrases (Fig. 5.2). We refer to these information-bearing chunks as "higher-order representations."

Current models posit that the brain transforms the auditory signal of speech into higher-order representations, which can then interface with stored representations in memory. This is achieved by generating increasingly complex and abstract representations of the acoustic input as neural activity propagates through the auditory pathways (Bonte et al. 2006; Hickok and Poeppel 2007; Gwilliams, 2020). This representational hierarchy is naturally supported by the hierarchical organization of auditory cortex: Regions of the auditory core (e.g., Heschl's gyrus) are driven by acoustically simple input features (e.g., frequency, amplitude), and surrounding cortical areas (e.g., superior temporal gyrus (STG), left temporal lobe) are sensitive to more complex spectro-temporal features of the input (Scott et al. 2000; Davis and Johnsrude 2003). Regions further along the anterior and posterior inferior temporal lobe in turn contribute to lexical and semantic processing of speech (Lau et al. 2008; Rauschecker and Scott 2009). Generating a hierarchy of progressively more abstract acoustic and linguistic representations serves to convert the sensory input (i.e., the speech that the listener hears) into meaning (i.e., a reconstruction of the message intended by the speaker). A major goal of the brain during speech comprehension is therefore to access the correct chunks from memory, based on the hierarchy of representations that it generates from the auditory signal.

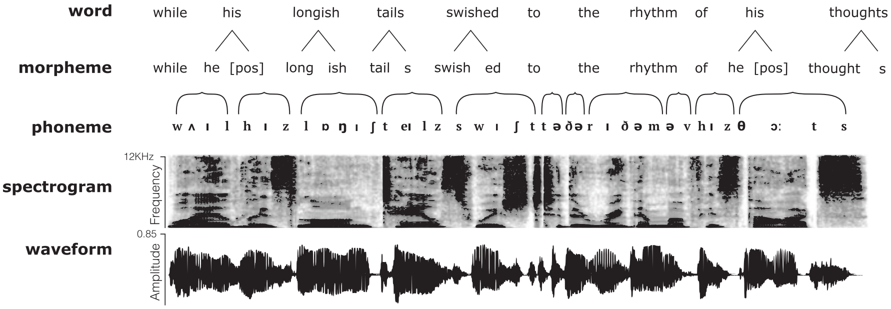

Fig. 5.2 Speech hierarchy. Language is a hierarchically structured stimulus. The acoustic signal can be discretized into a series of phonetic elements, which can be further grouped into morphemes, words, phrases, etc. Here we show an example speech segment, with the raw waveform, derived spectrogram, and corresponding linguistic annotation. Note that while the acoustic representations are continuous, stepping into linguistic features entails discretization of the signal. Also note that the linguistic units are hierarchically structured: e.g., words are comprised of morphemes, which are comprised of phonemes. This hierarchical structure is at the core of what allows us to investigate processes at "higher levels" as a function of neural responses to "lower levels" – because all of the levels are mutually structurally dependent

In this chapter we review evidence that neural processing of speech, even at early auditory processing stages, is fundamentally shaped by the goal of correctly and rapidly accessing higher-order information (e.g., words, meaning). Before reviewing this evidence, we will first consider the processes by which the sounds of speech are converted into discrete representations (e.g., phonemes). We will then introduce quantitative measures of the information content of speech signals; these measures – based on information theory (Shannon 1948) – presuppose that the brain, at some stage during processing, needs to access discrete higher-order representations from memory. These units can be straightforwardly assigned probabilities and therefore information value, which can then be used as a proxy measure of processing higher-order representations. As we discuss in Sect. 5.3, the use of information theoretic measures does not require commitment to a single and specific form of representation. Indeed, one of the strengths of this approach is that it can be equally applied to all levels in the linguistic hierarchy, covering all units between sounds and meanings. The main empirical data reviewed in the sections that follow concentrate on neural responses measured in peri-auditory regions of the STG and linked to specific information-carrying elements (speech segments, morphemes, words, etc.) in single words and in connected speech. We conclude by summarizing the computational and neural mechanisms by which the higher-level content of speech combines with acoustic signals during comprehension.

5.2 Discrete and Binary Representations of Speech Sounds Connect to Linguistic Units

Articulatory gestures of the speaker, and therefore the acoustic signals they produce, are continuous: both over time (any given sound can have variable duration) and in terms of content (spectral power can assume any continuous value). The continuous nature of the speech signal is somewhat at odds with the discrete and binary nature of the higher-order representations that need to be ultimately recognized. For example, the speaker is saying *either* “pit” or “bit” – they cannot be saying both at the same time.

A key challenge in the perception of speech sounds is therefore to convert the continuously varying acoustic input into discrete units that can be used to interface with higher-order representations. While there is some debate as to the specific low-level speech units the brain uses (Daube et al. 2019), they seem to sufficiently resemble phonemes and phonetic features (Chomsky and Halle 1968) for this to be a productive assumption. In order to correctly distinguish between different words, the discrete identity of constituent speech sounds is critical. A spoken consonant such as [p] is defined relative to its manner of articulation (plosive), place of articulation (bilabial), and phonation (voiceless). Each of these distinctive features must be correctly recognized during speech identification – a different speech sound – and hence different words or meanings will be understood if these features are misidentified ("pit" becomes "fit," "kit," or "bit" with changes to the manner, place, or voicing of the initial consonant). Any measurement of the information conveyed by speech sounds must ultimately operate on, and be calculated relative to, discrete representations that are the product of categorizing speech segments.

Yet, identifying the cortical signature of discrete processing of speech sounds has been a challenge for research on speech perception (see Chap. 3, Oganian, Fox, and Chang, for a review). Neural populations in STG around 100 ms after speech onset are sensitive to both the veridical acoustic content of a sound and the discrete phonetic categories to which the sound corresponds (Chang et al. 2010; Mesgarani et al. 2014; Di Liberto et al. 2015). Further, representations of speaker-specific details and other acoustic properties of speech coexist with categorical representations of linguistic content in auditory areas (Formisano et al. 2008; Evans and Davis 2015). Other cortical regions – including motor cortex and frontal regions – are also shown to contribute to coding of the categorical identity of speech sounds (Arsenault and Buchsbaum 2015; Evans and Davis 2015). Further evidence suggests that higher-level, nonauditory representations act top-down to constrain and guide lower-level auditory processing of speech signals (Kilian-Hütten et al. 2011; Sohoglu and Davis 2016; Gwilliams et al. 2017) (see Chap. 7, Ullas, Bonte, Formisano, and Vroomen, for a review). Therefore, the categorical responses to speech shown in the STG likely reflect the outcome of a transformation from the continuous auditory signal into discrete phonemic units, as influenced and constrained by higher-level language processing.

In contrast to the established work on speech sound representations, there is less consensus on the representational units that contribute to higher-level processing of speech (e.g., morphemes, words, and other meaning-carrying units above the level of the phonetic feature or phoneme). This discrepancy can be partly attributed to the fact that it is simpler to investigate features of the representational hierarchy that are closer to the sensory input than more abstract features that are closer to the meaning content, for at least two reasons.

First, whereas the acoustic sensory signal is easily measured and analyzed, higher-order representations only exist within the mind of the listener. One significant unresolved issue shared between cognitive neuroscience and engineering, therefore, is *feature discovery*: determining the feature space that best encodes abstract linguistic information. Modeling these higher-order processes can only be as successful as the suitability of the features selected to define those processes. While engineering approaches to deriving, for example, word meaning representations, have been used to predict neural activity during speech comprehension (Mitchell et al. 2008; Huth et al. 2016), there is currently little evidence in favor of one computational approach over another. Indeed, it has been argued that current engineering approaches to this problem are missing the key ingredients required for sufficient representation of meaning in true comprehension (Bender and Koller 2020).

Second, studying higher-order speech structure comes with analytical challenges. Assuming that the correct features have been identified, it is not always straightforward to relate relevant language features to a particular "moment" in the speech input. For example, if we assume that part of speech (e.g., noun, verb, adjective) is a feature that the brain uses to process words, at what moment in hearing the noun “hippopotamus” can we say that the brain is processing a noun? Does processing begin at the moment that the part of speech can be identified with 100% certainty, for example, when this word is uniquely specified after “hippop-.” or at word offset? Or are multiple part-of-speech hypotheses entertained simultaneously until syntactic class can be established beyond some threshold level of certainty? (See Wurm 1997; Balling and Baayen 2012 for discussions related to this issue.) The situation is further complicated by the fact that the timing of relevant neural processes – for different listeners, and for different utterances – becomes increasingly variable at higher levels of the processing hierarchy (Gwilliams and King 2020). Thus, even if the correct features have been identified, and even if the optimal latency relative to speech input could indeed be established, the time at which corresponding neural representations are activated will also vary. This significantly reduces the average signal strength associated with higher-order processes, making them much more difficult to investigate.

Here we review a set of studies which have investigated the effect of higher-order linguistic structure on processing of phoneme-by-phoneme information content in speech. The approach is to (i) model responses to phonemic units, which produce clear and well-characterized responses in terms of timing and spatial location (Mesgarani et al. 2014) and (ii) contrast segments that differ in the higher-order speech structures they communicate, e.g., whether or not specific speech sounds are predictable given the lexical or semantic context they occur within. The rationale is that by testing responses as a function of the *information* that a discrete speech sound provides about higher-level representations (e.g., syllables, morphemes, words; Fig. 5.2), it is possible to reverse engineer which higher-order representations are relevant to processing, how they are recognized or accessed, and how they interact with other representational units. This approach allows speech research to progress from studying auditory signal processing to information processing. Before we begin to review these studies, we will first provide a brief tutorial on the key quantitative measures of information content that have been employed in the study of speech comprehension.

5.3 Quantifying Information Content in Speech

Recent studies investigating information processing have capitalized on two properties of language. First, while speakers *can* convey a range of information constrained by the vocabulary and grammar of the language, not all expressions are equally likely: Some phoneme sequences, words, and meanings are much more probable than others. For example, English speakers are more likely to describe themselves as “happy” than “exultant” or “jocose”; it is more likely that after hearing /mæ/, you will hear the phoneme sequence /t/ (to create the word “mat”) than /lɪs/ (to create the word “malice”). This difference in likelihood is not encoded in the sensory signal – these likelihoods are reversed after hearing /pæ/ given that “palace” is more frequent than “pat,” for example. This knowledge comes from having an internal model of the language, including the statistical structure (what sound sequences, words, or sentences are more or less likely) as well as linguistic regularity (what sound sequences, words, etc. are permitted). Our view is that listeners employ both of these forms of knowledge during comprehension. We will hence use the term “statistical regularity” to describe these knowledge sources collectively.

This chapter focuses on probabilistic definitions of language knowledge since a wide range of data shows that listeners are exquisitely sensitive to the statistical structure of speech. Variability in the probability of linguistic units leads to differences in behavioral measures of speech comprehension such as response time and accuracy, in addition to the magnitude of neural responses as measured both invasively (electrocorticography (ECoG), stereoelectroencephalography (sEEG)) and noninvasively (electroencephalography (EEG), magnetoencephalography (MEG), functional magnetic resonance imaging (fMRI)) during comprehension. This kind of sensitivity has been demonstrated across many levels of the linguistic hierarchy: at the level of phonemes, morphemes, words, syntactic structures, and semantic content. It has also been demonstrated cross-linguistically, in languages with fundamentally different morphological and syntactic structures. These observations therefore suggest that sensitivity to statistical regularity is not only robust, it is also common across languages and pervasive across linguistic structures. These observations, however, presuppose that statistical regularities can be directly quantified: The quantification procedure is the focus of this section.

5.3.1 Introduction to Information Theory

We begin by detailing key measures of statistical regularity that can be quantified under information theory focusing on the most prevalent in the cognitive neuroscientific literature. For more complete introductions to information theory and derived metrics, see Manning and Schütze (1999) and MacKay (2003).

For demonstration we show how these variables would be computed for the word “mat” in the sentence “the cat sat on the mat,” and the final phoneme “t” in the word “mat.” But, they can be applied to any unit or feature (e.g., phoneme, syllable, part of speech, word identity, word length) within any context. In other words, the metrics we describe are not just applicable to defining the statistical likelihood of phonemes in words, and words in sentences – they can be applied to units of different types, and to contexts of different sizes, to investigate processing at all levels of the linguistic hierarchy.

Tailoring the measures for different levels of representation and context involves estimating x, C, and X:

- x – The linguistic event being modeled. Deciding what unit to use as x determines the units specified as the input. The majority of our examples assume x to be a phonological unit p or a lexical unit w, but this could also be defined as any unit of interest – e.g., phonetic feature, syllable, or morpheme. Our literature review focuses on research where the event of interest is a phonological unit.

- *C* – The relevant preceding events that influence the probability of linguistic event *x*. The below examples assume that the relevant context is the set of preceding phonemes in the current word in the case of *P* (*t ma*) and all preceding words of the sentence in the case of *P* (*mat thecatsatonthe*). Note, however, that *x* and *C* need not be in the same representational format; for example, it is possible to measure the probability of phoneme *p* given all preceding words *w* – indeed, converting between representational formats is necessary if prior words are to constrain processing of word-initial speech sounds. This assumes that the context *C* is relevant to the processing of event *x*.
- *X* – The alternative outcomes for which event *x* is informative. Note that *x* and *X* are necessarily in the same format because one is a single instance of the full set of the other. For example, phoneme *p* (e.g., /t/) is one instance of the cohort of possible phonemes in cohort *P* (e.g., /t/, /p/, /k/); word *w* (e.g., "mat") is one instance of the cohort of possible words in cohort *W* (e.g., "mat," "map," "mac").

How these parameters are estimated involves making theoretical commitments. The quantification of a particular representation can therefore be used to make adjudications between different theoretical alternatives. For example, it is possible to measure whether morphological or lexical context better explains neural responses (Gwilliams and Marantz 2015) by modeling phoneme *x* separately based on morpheme context $C_{morpheme}$ and lexical context $C_{lexical}$.

5.3.1.1 Conditional Probability

The probability of something happening given (i.e., "conditional upon") the other things that have happened. Conditional probability forms the basis of all the metrics we will define here.

For pedagogical purposes, an intuitive nonspeech example follows. In this case, the event being modeled (*x*) is rain, the preceding events (*C*) are clouds, and the alternative outcomes (*X*) are other weather outcomes, such as snow and sunshine. The conditional probability that it is going to rain, given that there are clouds in the sky is

$$P\left(x|C\right)=\frac{\text{freq}\left(C,x\right)}{\text{freq}\left(C\right)}$$

$$P\left(\text{rain}|\text{clouds}\right)=\frac{\text{freq}\left(\textit{rain and clouds}\right)}{\text{freq}\left(\textit{clouds}\right)}$$

This probability is computed by dividing the frequency of event *x* (e.g., rain) occurring in context *C* (e.g., clouds) by the frequency of context *C* alone (e.g., clouds both with and without rain). In these cases, we can say that the probability of event *x* is *conditional* upon the preceding events.

The conditional probability is highest when the two frequencies are similar in magnitude, i.e., if *x* (e.g., rain) almost always follows *C* (e.g., clouds). The probability is lowest when the frequency of *C* is much higher than the probability of *C* followed by *x*, i.e., if *x* (rain) very rarely follows *C* (clouds). Note that by definition the frequency of *C* is always equal to or greater than the frequency of *C* followed by *x*. At the heart of these probability calculations, therefore, are relative frequency counts, which can be derived from large databases of language called corpora. The wildcard "*" is used to denote that all continuations contribute to the frequency count.

The same kind of probabilities can be computed for spoken language. For example, what is the probability that a listener hears the word "mat" given that they just heard the words "the cat sat on the?" Or, what is the probability that a listener hears "*t*" given that they just heard the phoneme sequence /m/, /æ/:

$$P\left(x|C\right)=\frac{\mathrm{freq}\left(C,x\right)}{\mathrm{freq}\left(C\right)}$$

$$P\left(t|\mathrm{ma}\right)=\frac{\mathrm{freq}\left(\mathrm{mat}\right)}{\mathrm{freq}\left(\mathrm{ma}*\right)}$$

$$P\left(\mathrm{mat}|\mathrm{thecatsatonthe}\right)=\frac{\mathrm{freq}\left(\mathrm{thecatsatonthemat}\right)}{\mathrm{freq}\left(\mathrm{thecatsatonthe}*\right)}$$

The same logic we used in the rain example equally applies here. The more similar the frequency of the context and event as compared to the context alone, the higher the probability of the event *conditional upon* the context.

5.3.1.2 Surprisal

With conditional probability in hand, now we move to quantifying the *information gain* of a particular event. This is quantified as *surprisal*, and, as the name intuitively suggests, it scales with how unpredictable event *x* is given the context (Shannon 1948).

This measure is very similar to conditional probability. Computing surprisal involves taking the log transform of the conditional probability of *x* and then negating it to make all the values positive:

$$h\left(x\right)=\log 2\frac{1}{P(x|C)}=-\log 2P\left(x|C\right)$$

$$h(t) = \log 2 \frac{1}{P(t|\mathrm{ma})} = -\log 2P(t|\mathrm{ma})$$

$$h(\mathrm{mat}) = \log 2 \frac{1}{P(\mathrm{mat}|\mathrm{thecatsatonthe})} = -\log 2P(\mathrm{mat}|\mathrm{thecatsatonthe})$$

The resulting value is measured in “bits” – a value which quantifies information content. Surprisal can be construed, therefore, not just as a measure of predictability but also one of information gain. If an outcome was probable, less was “learned” about the state of the world than if it was less probable. If an event outcome has a probability of 1, predictions are entirely confident, and no information is gained from the event occurring because the outcome was already known.

For example, continuing the weather example above, a heat wave would be a high-surprisal event given its small probability in the context of clouds. Conversely, rain would be a low-surprisal event because it is more probable.

Surprisal is bound between zero bits of information gained (100% predictable) and infinity (0% predictable). While such extreme values are rare in natural language, everyday examples are possible. Let’s take the word “trombone,” for instance. If I hear the first syllable /trɒm/, surprisal at hearing the second syllable /bəʊn/ is near zero given that “trombone” (and derived words like “trombonist”) are among the only English words that contain this syllable (see Fig. 5.3b). If a different second syllable is heard – for instance, when one first learns of a type of mushroom called a “trompette” – the conditional probability for the second syllable given the first (i.e., /pɛt/ following /trɒm/) will be near zero, and surprisal will approach infinity.

5.3.1.3 Entropy

The final metric we will define here is *entropy*, which refers to the state of uncertainty about the subsequent event to occur, given the context. High entropy is the result of high uncertainty.

This metric is different from surprisal in that it refers to uncertainty about the upcoming event before the event happens – regardless of what event actually ends up happening. Intuitively speaking, certainty about the outcome of a situation will be lower when each outcome is equally likely. By contrast, certainty will be higher when one outcome is more likely than the alternatives. As an example, when selecting a playing card from a traditional Western deck, there is higher entropy over what suit your card will be (hearts, diamonds, clubs, spades – all 25% likely) than whether it will be a number card or a royalty card (about 23% vs. 77%). Or, referring back to the weather example, a clear blue sky is a lower-entropy context than a cloudy sky, because “sunshine” is one of the only options in the former, whereas there may be several likely outcomes (rain, hail, snow, etc.) in the latter. Thus, entropy expresses how certain you can be about *any* outcome, not the likelihood of *one* outcome.

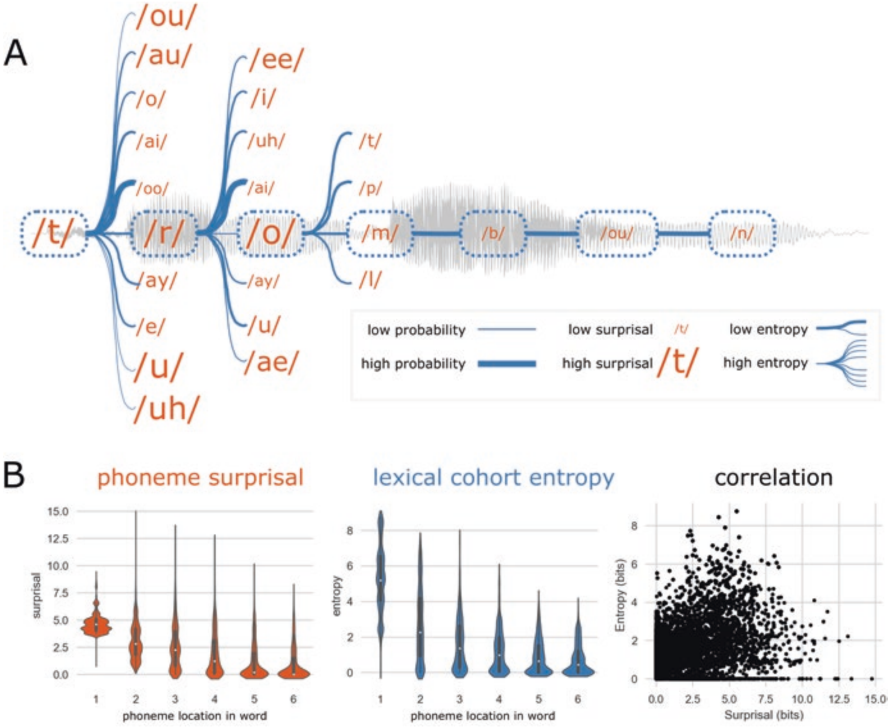

Fig. 5.3 Information theory metrics for phonemes in spoken words. (**a**) Waveform of the spoken word "trombone" superimposed with all the possible phonological alternatives at each phoneme position. Each alternative phoneme is shown with a blue trajectory, for which the thickness of the arrow corresponds to the likelihood of that continuation. Each phoneme continuation is shown in orange, whereby the size of the phoneme also indicates surprisal. As shown in the legend, when there are a number of possible, equally likely, continuations, entropy is higher as compared to when there are fewer, more asymmetrically likely continuations. Note that at the phoneme /m/, the word becomes lexically unique, and all of the subsequent phonemes in the sequence are 100% determined, such that entropy and surprisal are zero at /b/, /ou/, and /n/. Note that technically we depict phoneme-level entropy (uncertainty about the upcoming phoneme) in this sub-panel, which is correlated with – but makes different theoretical commitments from – lexical entropy (uncertainty about the word being said). (**b**) Surprisal and lexical entropy values were computed for around 20,000 phonemes of the audio-book stories used in Gwilliams et al. (2020). Violin plots show quartiles and distributions of surprisal and entropy for different phoneme positions in the word. Correlation strength between the two metrics is $r = 0.57$

In the literature review in Sect. 5.4, we will most often refer to "lexical cohort entropy," that is, uncertainty about what lexical item a person is saying, given the phonological sequence thus far. As discussed above, it is equally possible to compute "phoneme entropy" – uncertainty about what phoneme will be said next – or "word class entropy," uncertainty about whether the person will say a noun or a verb, etc., depending on the hypothesis at hand. Technically, the schematic in Fig. 5.3a depicts entropy over the upcoming phoneme, and Fig. 5.3b depicts entropy over the resulting lexical item. In practice these measures are highly correlated (when it is known what word is to be said, it is also not known what phoneme will come next), but this is not always the case (if many words share a common phoneme sequence, uncertainty about the lexical outcome may be high even if uncertainty about the subsequent phoneme is low).

Mathematically, entropy is equivalent to the expected surprisal of an outcome, i.e., the average surprisal (information gain) of each predicted outcome weighted by the probability of that outcome. In the examples below, *Wo* refers to the entire cohort of possible upcoming words starting with /mae/ and *w* to one instance of the cohort (e.g., "mat"). *Ph* refers to the entire cohort of possible upcoming phonemes (/t/, /p/, /k/ for "mat," "map," or "mac") and *p* to one specific phoneme instance:

$$H(X) = -\sum_{x \epsilon X} P(x|C) \log 2\, P(x|C) = \sum_{x \epsilon X} P(x|C)\, h(x)$$

$$H(\mathrm{Ph}) = -\sum_{p \epsilon \mathrm{Ph}} P(p|C_p) \log 2\, P(p|C_p) = \sum_{p \epsilon \mathrm{Ph}} P(p|C_p)\, h(p)$$

$$H(\mathrm{Wo}) = -\sum_{w \epsilon \mathrm{Wo}} P(w|C_w) \log 2\, P(w|C_w) = \sum_{w \epsilon \mathrm{Wo}} P(w|C_w)\, h(w)$$

Because entropy is a sum over weighted surprisal values, it is also measured in bits. It will assume the highest value when there are a large number of possible outcomes *X* each of which has equal probability, and the lowest value when one outcome is much more probable than its competitors. For the example word "trombone" as shown in Fig. 5.3a, entropy is highest at the first phoneme because there are many possible lexical items that begin with /t/ at onset. However, at the final /m/, entropy reaches zero because that is the phoneme that uniquely identifies that lexical item from all others – "trombone" (and its morphological family) is the only word with those first four phonemes.

As shown in Fig. 5.3b, both phoneme surprisal and lexical entropy tend to reduce for phonemes later on in spoken words. Although these measures are quite strongly correlated, they are not unavoidably so – it is possible to distinguish between neural responses that correlate with one or other, as we will observe in our review of relevant empirical data.

5.3.2 Using Neural Networks to Estimate Surprisal and Entropy

Neural network models are exceptionally good at learning statistical regularities, in particular nonlinear, probabilistic dependencies between representations. Correspondingly, some recent studies have moved from corpus-based estimates to neural network estimates of probability. Under this approach, a network is trained to predict the unit of interest – for example, the next spoken word in a sentence. After sufficient training, the model can then be input with an experimental sentence (e.g., "the cat sat on the …") and queried for its probabilistic prediction of the upcoming word (e.g., mat, floor, sofa). These probabilities are essentially the same as (and can be directly substituted for) conditional probability in the equations provided above. The main difference is that the probability is now "conditional upon" the language that the network was trained on. What language a model has been trained on is critical for interpreting probabilities from such network models.

It is also important to note that neural networks can be used not just to derive probability estimates but also representations of the language input. This is particularly helpful for networks with multiple layers of intervening processing units (i.e., deep neural networks). In these cases, it is possible to derive putative language representations by querying the model for how it represents language in activation values at intermediate processing stages i.e., layers. A prevalent example of using models to derive language representations is word embeddings (Pennington et al. 2014). These features putatively represent the semantic space of lexical items, and have been shown to correlate with neural responses as recorded with fMRI (Huth et al. 2016) or EEG (Broderick et al. 2018). The features are derived from the weights of the one-layer model that predicts a word (e.g., mat) from a context word (e.g., cat). In this chapter, we will not review studies that use models to derive representations of language in this way; we will focus on their ability to produce probability estimates.

In all, whether the basic probability estimates are derived formally from corpora or empirically from trained network models, the same method of surprisal and entropy calculations remains. It is likely that as these models become increasingly sophisticated and accessible, we will see increasing reliance on these networks for estimates of statistical regularity.

5.4 Information Theoretic Measures and Neural Responses

Here we will review recent studies that have employed information theoretic measures to assess the influence of linguistic computations at the level of morphemes, words, and phrases on neural activity during speech perception.

Early studies in this area contrasted fMRI responses to different types of spoken words, e.g., words compared with nonwords, or words with more vs. fewer competitors (e.g., Binder et al. 2000; Bozic et al. 2010), and provide evidence for additional activation of superior temporal and frontal regions for more difficult to identify words. There is some evidence of a dissociation of frontal and temporal regions: More unexpected speech (e.g., high-surprisal segments in nonwords compared to real words) activates superior temporal regions (Davis and Gaskell 2009; Zhuang et al. 2014), whereas words with more competitors (e.g., words like *claim* with an onset-embedded word *clay*) lead to additional activity of inferior frontal regions (Bozic et al. 2010; Zhuang et al. 2011). However, these studies did not directly compare neural activity linked to specific statistical properties (such as surprisal or entropy that co-vary). Furthermore, fMRI lacks the time resolution required to link activation to statistical properties or neural processes for specific speech segments. More consistent evidence that the auditory system is sensitive to the statistical structure of speech input has therefore come from neural measures with higher-temporal resolution (e.g., MEG or EEG). Furthermore, as techniques for estimating language-based statistics from corpora and natural language have improved, so has the application of these computational measures to model neural responses. Here we summarize the main findings demonstrating that phoneme-level metrics of higher-order linguistic structure can influence information processing during speech comprehension.

5.4.1 Phoneme Surprisal and Lexical Entropy in Isolated Words

Gagnepain et al. (2012) conducted one of the first studies that used neural data to differentiate information theoretic measures of lexical processing of speech. Neural responses in left STG were recorded using MEG in response to triples of familiar words (e.g., *formula*), learned novel words (e.g., *formubo*), and untrained novel words (e.g., *formuty*). Neural responses were time-locked to speech before the divergence point (DP) (i.e., *formu*-) and after the divergence point (i.e., -*la* vs. -*bo* vs. -*ty*).

The authors compared responses to item triples containing novel words that were learned and consolidated (and, hence, added to lexical knowledge; cf. Davis and Gaskell 2009), or learned but not consolidated (and hence not lexicalized). They assessed the impact of changes to lexical knowledge on neural activity to adjudicate between two hypotheses (see Fig. 5.3b). First, adding a new word to the lexicon (i.e., once the novel word *formubo* has been consolidated) would lead to an increase in lexical entropy, particularly before the divergence point. Yet, phoneme surprisal would decrease during the same pre-divergence point period due to stronger predictions for shared segments. Second, phoneme surprisal will increase in the post-divergence point window upon hearing phonemes that were not expected (i.e., a phoneme surprisal response will increase for the less expected continuation -*la* once *formubo* had been consolidated).

MEG responses in the pre-divergence point and post-divergence point periods changed in line with phoneme surprisal rather than lexical entropy; hence changes in lexical knowledge can modify phoneme-level responses in such a way that is consistent with computation of lexically generated prediction error. Specifically, responses in left STG from 280 to 350 ms post-divergence point were increased for word neighbors of the consolidated novel word compared to neighbors of learned but not consolidated items. Furthermore, consolidated novel words (but not learned, not consolidated items) showed a reduced response in the same time period. There were no detectable effects of lexical entropy in the pre-DP window as predicted by the lexical inhibition account; rather response reductions were observed in line with stronger segment predictions. These findings suggest that lexical knowledge (one element of the internal language model) generates predictions for upcoming speech segments that are compared with heard speech leading to STG responses that resemble prediction error. Gagnepain and colleagues further suggest that prediction error signals can be used to update lexical probabilities though they do not provide evidence to show these update mechanisms in operation (see Davis and Sohoglu 2020 for discussion).

Building from these results, a collection of studies by Marantz and colleagues capitalized on sensitivity to segmental probability to understand the representation and processing of higher-order linguistic units in single spoken words. Specifically, they tested whether internal (morphological) word structure (e.g., a word like *disappears* is composed of morphemes *dis*, *appear*, *s*) influences segment prediction error or surprisal. Using MEG, a study conducted by Ettinger et al. (2014) revealed a main effect of phoneme surprisal in left STG responses measured 200 ms after segment onset, which was significantly greater for bimorphemic words (*bruis-er*) as compared to phonologically matched monomorphemic words (*bourbon*). There were also later effects of phoneme surprisal toward the end of the word (700+ ms after word onset). Furthermore, they found main effects of lexical cohort entropy from 335 to 377 ms after word onset. The authors conclude that the internal (morphological) structure of words serves to enhance segmental predictions at the phoneme level and that predictions are delayed under conditions of high lexical entropy. This may suggest that segmental predictions are generated, not just at the phoneme-unit level but also at the level of entire morphological units.

To further investigate the influence of morphological units on speech processing, languages with a non-concatenative morphological structure like Arabic and Hebrew are an ideal test case. Whereas in English morphemes are combined one after the other (e.g., *dis-appear-s*), in Arabic, they are interleaved within one another (e.g., the morphemes [k-t-b] and [a-a-a] are combined to form *kataba*). Thus, the linear order with which the auditory signal unfolds is at odds with the nonlinear order that the relevant speech sounds of morphemes are received, allowing the two to be disassociated. Under this rationale Gwilliams and Marantz (2015) assessed neural effects of segmental prediction in order to determine whether spoken Arabic words are processed via their constituent morphemes (k-t-b, a-a-a) or as whole units (e.g., kataba) by opposing two measures of phoneme surprisal. They constructed stimuli that uncorrelated root-based "morpheme" surprisal (probability of a consonant conditioned on the previous consonants in the root morpheme) and word-based "linear" surprisal (probability of a consonant conditioned on all previous phonemes in the word). MEG was recorded while Arabic speakers performed a lexical decision task on spoken isolated words. They analyzed responses to the final consonant of the words (e.g., kata**b**a) as a function of preceding morphological content and preceding whole-word content. Activity in left STG was significantly modulated by morpheme surprisal from 100 to 250 ms. Word-based linear surprisal modulated later responses, from 250 to 300 ms in an overlapping set of sources. Thus, the results suggest that words are processed via morphological units before they are processed as wholes. This research showcases the use of phoneme-level responses to understand the representation and processing of higher-order linguistic structure, by contrasting predictions from different units (e.g., words vs. morphemes). By understanding what information is used to constrain predictions of upcoming information, it allows for inferences about what higher-level information is being accessed, and therefore what information is likely *stored* in lexical memory and deployed in speech perception.

Similar methods have been used to assess the relationship between lexical and semantic processing of spoken words that refer to specific categories of concrete objects. For example, Kocagoncu et al. (2017) show lexical uncertainty (quantified as lexical entropy based on participants' responses in a word-gating task) is encoded in MEG patterns recorded from superior temporal and inferior frontal regions. These responses are earlier than, and partially overlap with, frontal and parietal responses that encode the degree of semantic competition (i.e., lexical uncertainty modulated by semantic dissimilarity). These findings suggest that the brain derives semantic interpretations of spoken words throughout identification and that access to meaning is not delayed until a single lexical item has been identified and settled upon (Zwitserlood 1989).

Along similar lines, Gwilliams et al. (2017) tested whether activation of lexical candidates is weighted by acoustic evidence in favor of one phoneme or another. They recorded MEG responses of subjects listening to words, where the onset phoneme was acoustically manipulated (morphed) along a five-step phonetic continuum from /b/-/p/, /t/-/d/, and /k/-/g/. The authors quantified two measures of surprisal and entropy: First "acoustic weighted" metrics consider both the "b-" and "p-" cohorts of words into the computation of surprisal and entropy, where each cohort is weighted both by word frequency *and* acoustic evidence. The second "switch-based" metrics assume that the brain categorizes phonemes before activating lexical candidates, and so in these surprisal and entropy metrics, *either* the "b-" *or* "p-" onset words will be included in the information theoretic measures. The authors found that when modeling surprisal and entropy from 200 to 250 ms after phoneme onset in left STG, responses to early phoneme locations were better modeled under the "acoustic weighted" account, whereas later phoneme locations were better modeled by the "switch-based" account. The interpretation of these results is that earlier during processing, the brain uses both acoustic detail and lexical frequency equally to activate words, whereas later in processing, the brain favors categorical representations of the input in order to focus more heavily on lexical statistics. These results again showcase the ability to use information theoretic measures at the level of phoneme responses to adjudicate between specific processing hypotheses as they pertain to higher-order structures such as lexical items.

5.4.2 Phoneme Surprisal and Lexical Entropy in Continuous Speech

While these studies tested sensitivity to phoneme predictions within isolated words, in natural speech, expectations can also be generated based on previously heard words: Natural speech provides a *continuous* stream of linguistic information, in which preceding words can serve to constrain the probabilities of upcoming inputs. A study conducted by Gaston and Marantz (2018) asked the critical question of whether, in minimal phrases (e.g., "the clash persisted"), the brain uses preceding words to inform phoneme-level predictions. They tested whether phoneme surprisal and lexical entropy responses could be conditioned *across* word boundaries, based on syntactic constraints provided by the preceding context. In terms of our tutorial above, this would mean contrasting the conditional probabilities that enter into the surprisal and entropy calculations to include either just the prior context within the word or also prior context across multiple words.

MEG was recorded while participants listened to minimal phrases, which were either grammatical (e.g., "the clash persisted") or nongrammatical (e.g., "*the frown darkly") where the first word (the/to) made deterministic predictions about the part of speech of the subsequent word (noun/verb). The results show that both constrained and unconstrained surprisal metrics significantly accounted for neural responses in STG from around 200–400 ms after each phoneme in the noun/verb target, though no significant effects of entropy were observed. Thus, even when prior context has the *potential* to redistribute probabilities on the level of "boundary blind" phoneme sequences, the brain remains sensitive to the context-free word-internal statistics *in parallel to* the context-sensitive statistics. This important observation suggests that lexical and sub-lexical units are activated based on both sources of information, perhaps aggregating over the predictions at a later stage. Similar findings arise from a study that explored the role of semantic constraints in guiding word identification from Klimovich-Gray et al. (2019). For two-word phrases (e.g., "yellow banana"), MEG response patterns in left STG around 150 ms after the start of the second word encode the change in entropy (i.e., surprisal) while lexical interpretation is guided both by prior context and by heard speech sounds. Partial correlation analyses confirm that these effects are independent of entropy and overall semantic similarity of word candidates.

A set of recent studies have also demonstrated the ability to use these same calculations of surprisal and entropy, which assume that the word is presented in isolation ("word-internal metric"), to investigate processing of words in natural, continuous speech such as spoken narratives. Brodbeck et al. (2018) analyzed responses to continuous speech as a function of word-internal information theoretic metrics. They found that phoneme surprisal modulated STG responses peaking around 115 ms after the onset of the relevant speech segments; responses correlated with cohort entropy followed soon after and peaked at around 125 ms. The relative timing of these effects is in line with Gwilliams and Marantz (2015), but earlier than those seen in Gagnepain et al. (2012) and Gaston and Marantz (2018). It might be that neural responses linked to specific speech segments arise at shorter latencies for words in connected speech than for words heard in isolation.

5.4.3 Other Related Metrics that Predict Neural Responses

Speech research has primarily used surprisal and entropy to capture predictive processes – hence our focus on them in this review. These are not exhaustive of all features that can be used for this purpose, however, and different metrics can be used to tap into different putative neural operations.

A good example of this is phonotactics. The phonotactic rules that govern the probability of different phoneme transitions – for example, in English the phoneme sequences /kn/ and /ng/ and /bn/ never occur at the beginning of a word – are correlated with phoneme surprisal, but this kind of linguistic knowledge is contained within the statistics of the phonological sequence itself. It does not depend upon accessing or failing to access lexical items. Di Liberto et al. (2019) analyzed the relative contribution of phonotactic probability and phoneme surprisal when modeling EEG recordings of subjects listening to continuous speech. They replicated the finding that phoneme surprisal modulates responses at around 110 ms. Critically, these phoneme surprisal effects occurred much earlier (~110 ms) than sensitivity to phonotactic transitions of English (~300–400 ms). This therefore suggests that the two metrics tap into two different neural computations. Although phonotactics reflect statistics over phoneme sequences (Jusczyk et al. 1994), the source of the statistical regularity is not related to their higher-order connection to the mental lexicon. Overall, this result indicates that while both surprisal and phonotactic probability relate to statistical processing of phoneme sequences, the statistics that are informative for lexical access produce an earlier and distinct neural response. Speech perception therefore appears to be optimized for, and prioritizes processes that contribute to, identification of spoken words.

Furthermore, lexical entropy quantifies the weighted activation of different lexical candidates, and surprisal quantifies how much the lexical competition needs to be updated. However, other metrics of lexical competition and update can also be derived, which may be indices of independent neural processes. Brodbeck et al. (2018) and Donhauser and Baillet (2020) tested the contribution of regressors that are correlated with lexical entropy and phoneme surprisal, which may otherwise serve as potential confounds: Brodbeck et al. (2018) tested the contribution of lexical cohort size (how many lexical items are possible given the sequence input) and cohort reduction (how many lexical items are no longer consistent with the phonological sequence, given the new phoneme that was just heard). Donhauser and Baillet (2020) also tested the role of cohort reduction. While Brodbeck et al. found no additional contribution of either regressor above the explained variance of the existing analysis factors, Donhauser did find that cohort reduction explained additional variance above the level of phoneme surprisal. This might suggest that previous studies using phoneme surprisal may have actually been tapping into two distinct processes: (i) sensory surprisal, comparing the predicted input to the received input, which is best modeled using phoneme surprisal proper, and (ii) lexical-update surprisal, updates in activated lexical items as a *consequence* of the phoneme input, which is best modeled using a cohort reduction measure.

Another important extension of phoneme surprisal and lexical entropy is to derive metrics which are sensitive to the surrounding sentential context. For example, whether a word is preceded by the article "a" or "an" is highly informative as to the identity of the initial phoneme of that word, and it is likely that the brain is sensitive to such information. Donhauser and Baillet (2020) modeled MEG responses to continuous speech, as a function of context-sensitive phoneme surprisal and phoneme uncertainty. These measures were computed based on a neural network (cf. Elman 1990; Cairns et al. 1997) which was trained to predict upcoming segments in speech sequences, including predictions that cross boundaries between higher-order units. See Sect. 5.3.2 for an explanation of how phoneme surprisal and lexical entropy can be derived from a neural network model. Unlike the "word-internal" metrics discussed so far, this metric also uses information from previous words as part of the prior lexical context. Specifically, their neural network model used a context of the preceding 35 phonemes sufficient to encode several preceding words and (potentially) their meaning and syntactic structure. They found that across-word phoneme surprisal modulated responses from around 80–160 ms and 230–420 ms after phoneme onset, and contextual entropy modulated responses 60–120 ms and 230 ms in primary and association auditory cortex. These results further support the notion that early auditory responses reflect sensitivity to higher-order structure and that surprisal and entropy are metrics that tap into distinct neural computations. When hearing spoken words in isolation, context-sensitive and context-insensitive predictors are perfectly correlated (for instance, in Gagnepain et al. (2012)). Yet, being able to separate these responses in connected speech might support the existence of multiple processes in naturalistic listening, involving both context-specific and more locally computed phoneme probabilities (cf. Gaston and Marantz 2018).

Overall these studies highlight that capturing the complex array of predictive processes involved in speech processing requires a suite of probabilistic regressors, beyond surprisal and entropy that we have proposed contribute most to the extraction of meaning. Some of these regressors should capture lower-level sensory or phonotactic likelihoods, some should capture lexical activation and access, and some should be based on local within-word context and others on the broader sentential context. The results highlighted here indicate that the brain engages in multiple predictive processes in parallel, acting upon linguistic units of different types and as informed by contexts of different sizes. Further research will be needed to establish the unique functional contributions of these putative parallel pathways.

5.5 Predictive Coding and Bayesian Inference

The studies reviewed here show consistent evidence that auditory responses to speech sounds are modulated by the higher-level information content those sounds communicate, within the first 100–400 ms after the onset of speech sounds. These observations demonstrate how responses to low-level units of speech are shaped by the language system's ultimate goal of linking speech sounds to stored linguistic representations in order to reconstruct the higher-level meaning that the speaker intended to communicate.

Two information metrics are most commonly observed to modulate neural responses: phoneme surprisal and lexical entropy. Both metrics modulate neural responses in STG with a similar time range (see Fig. 5.4a for a schematic summary of this literature). Although these measures are highly correlated and show similar spatiotemporal response profiles, these variables have been shown to make independent statistical contributions to neural data (Brodbeck et al. 2018; Donhauser and Baillet 2020) suggesting that these two metrics tap into two distinct neural computations.

How can we place the neural effects of these information theoretical metrics into a computational understanding of speech perception? The goal of speech comprehension can be construed as identifying a sequence of morphemes or words from the auditory signal based on multiple sources of (noisy) information which must be combined with prior knowledge or expectations about the likely words that will occur and their meanings. One popular framework by which to integrate these different sources of information uses Bayesian inference, and other mathematically similar approaches (Mumford 1992; Rao and Ballard 1999; Friston 2005); for a specific treatment of Bayesian inference in speech perception, see Norris and McQueen (2008) and Kleinschmidt and Jaeger (2015).

Under a Bayesian formalization, it is possible to estimate the probability of a specific interpretation (let's say, the identity of the current lexical item) as a function of each incrementally received input (e.g., using the identity of each input phoneme). We focus on the relationship between these lexical and phoneme-level processes as the input and output (the pink and purple nodes in Fig. 5.4b). However, the Bayesian inference process would operate similarly for lower-level processes (e.g., recognizing phonemes given acoustic signals), or higher-level processes (accessing meaning or syntactic information given the words heard).

Identification of words from phoneme sequences would operate as follows: The listener has an internal language model which is formed based on linguistic experience and includes knowledge of the likelihood of different words, and the identity of the speech sounds that make up those words. This statistical knowledge specifies the prior knowledge that the listener uses to make top-down predictions (P) for the sounds of upcoming words. These predictions can incorporate multiple forms of hierarchically structured knowledge; that is, predictions at the phoneme level might be influenced not only by known words and their constituent sounds but also by higher-level contextual knowledge (semantic or syntactic representations of the

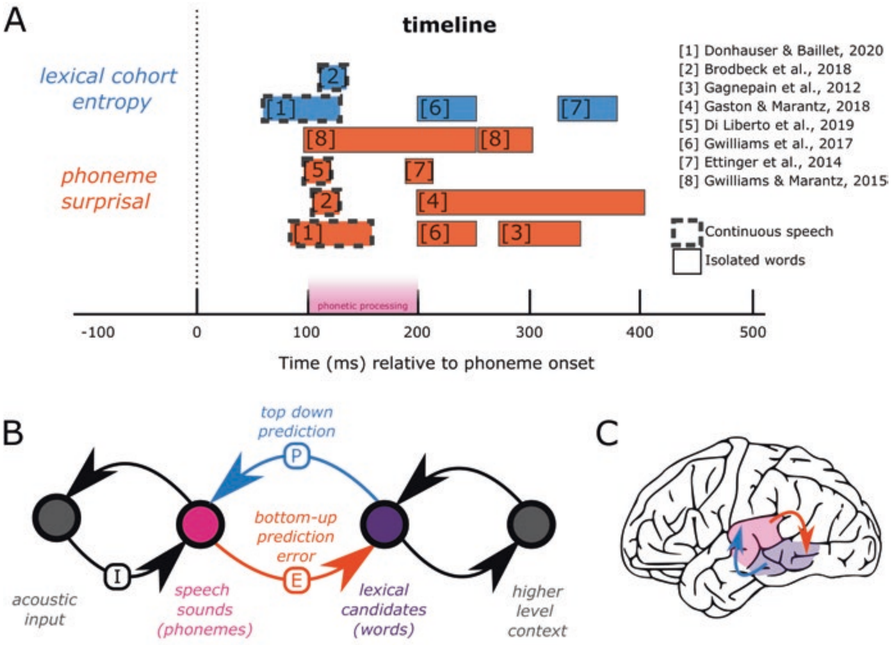
A
timeline
lexical cohort entropy
phoneme surprisal
[1]
[2]
[6]
[7]
[8]
[8]
[5]
[7]
[2]
[4]
[1]
[6]
[3]
[1] Donhauser & Baillet, 2020
[2] Brodbeck et al., 2018
[3] Gagnepain et al., 2012
[4] Gaston & Marantz, 2018
[5] Di Liberto et al., 2019
[6] Gwilliams et al., 2017
[7] Ettinger et al., 2014
[8] Gwilliams & Marantz, 2015
Continuous speech
Isolated words
phonetic processing
-100
0
100
200
300
400
500
Time (ms) relative to phoneme onset
B
top down prediction
P
bottom-up prediction error
E
I
acoustic input
speech sounds (phonemes)
lexical candidates (words)
higher level context
C

Fig. 5.4 Information exchange. (**a**) Summary timeline of when the studies in our review (Gagnepain et al. 2012; Ettinger et al. 2014; Gwilliams et al. 2015, 2017; Brodbeck et al. 2018; Gaston and Marantz 2018; Di Liberto et al. 2019; Donhauser and Baillet 2020) find significant effects of surprisal and entropy. Boxes with a dashed outline refer to studies using continuous speech; boxes with a solid outline refer to the presentation of an isolated word or a minimal phrase. Pink shading between 100 and 200 ms corresponds to approximately when phonetic features are processed, for reference. (**b**) A simple network graph model showing how information is hypothesized to pass between the different of processing during lexical access. (**c**) Putative brain regions involved, and the direction of information flow associated with surprisal and entropy. Here just the left hemisphere is visualized because responses have been mainly tested and validated in the left hemisphere

current utterance) that change the likelihood of different words. Thus, top-down predictions, or priors, provide a probabilistic prediction about the current utterance (i.e., what word is being said), which is computed based on the frequency with which words have been experienced in the past combined with a representation of previous words in the sequence. The extent to which the system has converged on a single prediction of the word is reflected in the *entropy* metric: Recall that entropy is highest when the prior is uniformly distributed across multiple possibilities, and lowest when all predictions are centered around a single outcome (Sect. 5.3.1.3).

The phoneme predictions generated by the prior are then compared to the current input (I), which in our illustration (Fig. 5.4b) would be based on a representation of discrete phonemes. The difference between the phoneme that was predicted, based on the prior of the phoneme sequence of the word, and the phoneme that was heard (P-I) gives the prediction error (E). The magnitude of this prediction error is correlated with phoneme *surprisal*: When hearing an unexpected sound, prediction error and surprisal are higher than when hearing a more strongly predicted sound (Sect. 5.3.1.2). When the prediction error is large, this provides additional information to update the probabilities of possible words (orange arrow to the purple nodes in Fig. 5.4b), because the word that was predicted to be the outcome is no longer the best lexical candidate. This iterative updating process happens at successive speech segments throughout the time course of word processing, until the optimal candidate can be recognized (see Fig. 5.4b, and Blank and Davis (2016) for a simple implementation of this model).

Under this framework, we interpret the surprisal response as a reflection of the extent to which the relative activation of lexical candidates needs updating on receiving each new piece of phonological input. If heard phonemes are strongly expected, there is little information gained, and therefore lexical activations and subsequent predictions will go unaltered. If the phoneme was unexpected, this requires a big shift in which lexical candidates are most likely, which is reflected in the surprisal signal. This surprisal response may therefore reflect the extent to which the internal state of the system needs to be updated (orange arrow, Fig. 5.4b), leading to changes to the predictions generated for subsequent inputs in the phonological sequence (blue arrow, Fig. 5.4b, Gagnepain et al. 2012; Donhauser and Baillet 2020).

Given this iterative updating of predictions based on prediction error, lexical entropy reflects the current state of uncertainty about which lexical candidate will ultimately “win” the recognition process. It is posited that entropy (in domain-general accounts (Feldman and Friston 2010)) serves to boost or dampen the information that is likely to be gained from subsequent sensory signals (blue arrow, Fig. 5.4b). Time points at which lexical entropy is low (i.e., one candidate word is much more likely than the rest) permit easy identification, and subsequent sensory input is not so critical; the likely outcome of word recognition is already known. However, in cases of high entropy (i.e., if multiple candidate words are activated to a similar degree), then resolving which lexical candidate is correct will require a heavier reliance on sensory input. This interpretation is in line with Bayesian accounts of predictive coding (Adams et al. 2013; Davis and Sohoglu 2020): When predictions for upcoming input are uncertain (high entropy), sensory processing plays a more important role in disambiguating the input, and prediction error will tend to be higher to compensate (see Fig. 5.4b).

As can be seen in the summary timeline in Fig. 5.4a, the estimates for *when* these different metrics matter for neural processing are highly varied. Responses have been reported as early at 60 ms and as late at 400 ms, for both phoneme surprisal and lexical entropy, with variability between studies in terms of both the onset latency and duration of neural effects. While some of this variation might be due to differences in statistical power or thresholds in specific studies, other variation may be due to properties of the speech stimuli used. For example, one consistent observation is that latencies are shorter when words are presented in the context of continuous speech rather than in isolation. Studies using naturalistic stimuli find effects of entropy and surprisal at around 120 ms after phoneme onset on average (simultaneous with processing the phonetic features of the speech sound itself), whereas studies using isolated words find sensitivity to the same features around 250 ms after phoneme onset on average. Nonetheless, even allowing for this variation, we observe that in some studies, surprisal effects begin earlier than entropy (Ettinger et al. 2014; Brodbeck et al. 2018), and in other studies the reverse is observed (Donhauser and Baillet 2020). Understanding whether any reliable temporal difference between surprisal and entropy exists, or whether they are better described as simultaneous processes, promises to provide significant insight into the computational operations being applied to the speech signal. In addition, how those responses can be changed with the provision of higher-level context which allows for predictions of upcoming lexical input or with changes to the sensory quality of the speech signal which might permit more rapid or slower speech processing (gray nodes, Fig. 5.4b) is an exciting avenue to explore.

Even though both surprisal and entropy reflect higher-order processes, it is noteworthy that their neural correlates are not located in the cortical areas that are typically associated with lexical access, such as the middle or inferior temporal gyrus or inferior frontal gyrus (Hickok and Poeppel 2007; Davis 2016). Instead, all of the studies we have described broadly localize these responses to auditory brain regions – including the left transverse temporal gyrus and STG – overlapping with where acoustic and phonetic features are known to be processed (Mesgarani et al. 2014; Gwilliams et al. 2018). It is also worth pointing out that all of the reported effects were overwhelmingly left lateralized. Overall this indicates that local, perhaps recurrent, processing of speech sounds in auditory cortex is influenced by higher-order structure, such as sequence statistics, and higher-order computations, such as lexical access. While we and others have assumed functional and anatomical hierarchies in speech processing, this does not imply that higher-level features of speech do not influence lower-level auditory responses. Further investigation, perhaps by taking advantage of the joint spatiotemporal resolution of intracranial recordings, will be required to fully specify the spatial location of sensitivity to phonetic features, surprisal and entropy, and the extent to which they are supported by the same versus neighboring neural populations.

5.6 Conclusion

Overall, the evidence presented here suggests that the brain applies Bayesian-inference (-like) computations in order to decode meaning from the speech signal. The acoustic input (the likelihood) is weighted by probabilities over *what the speaker could say* (the prior) in order to derive *what the speaker is saying* (the posterior).

As we saw in this review, neural responses illustrate how multiple sources of information are potentially computed in parallel, including context-free and context-sensitive measures of prior probability distributions. These context-sensitive measures allow lexical items to be activated based on current syntactic and semantic context (Marslen-Wilson and Welsh 1978), whereas context-free measures rely on within-word phoneme-sequence statistics alone. Aggregating over both measures allows the brain to jointly estimate the best interpretation of upcoming input for cases in which predictions across multiple information sources converge, and to be more skeptical when multiple information sources do not converge. These situations can cue revisions to perceptual interpretations, demand semantic reinterpretation, or allow detection of lexical novelty or speech errors (Davis and Sohoglu 2020).

Any valid account of how the brain achieves speech comprehension needs to explain not just how the acoustic signal is processed but how this signal is used to identify the words being said. Information theoretic measures, such as surprisal and entropy, provide excellent tools for examining such "higher-order" processes. For example, what priors are used to form predictions of upcoming information, and what linguistic units these predictions may comprise, shed light on what representations are accessed and composed online during comprehension. As techniques for estimating such probability measures become increasingly precise, so does our ability to model how the brain uses them for language understanding.

Conflict of Interest The authors declare that they have no conflict of interest.

References

Adams RA, Stephan KE, Brown HR, Frith CD, Friston KJ (2013) The computational anatomy of psychosis. Front Psych 4:47

Arsenault JS, Buchsbaum BR (2015) Distributed neural representations of phonological features during speech perception. J Neurosci 35(2):634–642

Balling LW, Baayen RH (2012) Probability and surprisal in auditory comprehension of morphologically complex words. Cognition 125(1):80–106

Bender EM, Koller A (2020) Climbing towards nlu: on meaning, form, and understanding in the age of data. In: Proceedings of the 58th annual meeting of the Association for Computational Linguistics, pp 5185–5198

Binder JR, Frost JA, Hammeke TA, Bellgowan PS, Springer JA, Kaufman JN, Possing ET (2000) Human temporal lobe activation by speech and nonspeech sounds. Cereb Cortex 10(5):512–528

Blank H, Davis MH (2016) Prediction errors but not sharpened signals simulate multivoxel fmri patterns during speech perception. PLoS Biol 14(11):e1002577

Bonte M, Parviainen T, Hytönen K, Salmelin R (2006) Time course of top-down and bottom-up influences on syllable processing in the auditory cortex. Cereb Cortex 16(1):115–123

Bozic M, Tyler LK, Ives DT, Randall B, Marslen-Wilson WD (2010) Bihemispheric foundations for human speech comprehension. Proc Natl Acad Sci 107(40):17439–17444

Brodbeck C, Hong LE, Simon JZ (2018) Rapid transformation from auditory to linguistic representations of continuous speech. Curr Biol 28(24):3976–3983

Broderick MP, Anderson AJ, Di Liberto GM, Crosse MJ, Lalor EC (2018) Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech. Curr Biol 28(5):803–809

Cairns P, Shillcock R, Chater N, Levy J (1997) Bootstrapping word boundaries: a bottom-up corpus-based approach to speech segmentation. Cogn Psychol 33(2):111–153

Chang EF, Rieger JW, Johnson K, Berger MS, Barbaro NM, Knight RT (2010) Categorical speech representation in human superior temporal gyrus. Nat Neurosci 13(11):1428

Chomsky N (2000) New horizons in the study of language and mind. Cambridge University Press, Cambridge

Chomsky N, Halle M (1968) The sound pattern of English, 1st edn. Harper and Row

Daube C, Ince RA, Gross J (2019) Simple acoustic features can explain phoneme-based predictions of cortical responses to speech. Curr Biol 29(12):1924–1937

Davis MH (2016) The neurobiology of lexical access. In: Hickok G, Small SL (eds) Neurobiology of language. Elsevier, pp 541–555

Davis MH, Gaskell MG (2009) A complementary systems account of word learning: neural and behavioural evidence. Philos Trans R Soc Lond B Biol Sci 364(1536):3773–3800

Davis MH, Johnsrude IS (2003) Hierarchical processing in spoken language comprehension. J Neurosci 23(8):3423–3431

Davis MH, Sohoglu E (2020) Three functions of prediction error for bayesian inference in speech perception. In: Poeppel D, Mangun G, Gazzaniga MS (eds) The cognitive neurosciences, 6th edn. MIT Press, pp 177–189

De Saussure F (2011) Course in general linguistics. Columbia University Press, New York

Di Liberto GM, O'Sullivan JA, Lalor EC (2015) Low-frequency cortical entrainment to speech reflects phoneme-level processing. Curr Biol 25(19):2457–2465

Di Liberto GM, Wong D, Melnik GA, de Cheveigné A (2019) Low-frequency cortical responses to natural speech reflect probabilistic phonotactics. NeuroImage 196:237–247

Donhauser PW, Baillet S (2020) Two distinct neural timescales for predictive speech processing. Neuron 105(2):385–393

Elman JL (1990) Finding structure in time. Cogn Sci 14(2):179–211

Ettinger A, Linzen T, Marantz A (2014) The role of morphology in phoneme prediction: evidence from MEG. Brain Lang 129:14–23

Evans S, Davis MH (2015) Hierarchical organization of auditory and motor representations in speech perception: evidence from searchlight similarity analysis. Cereb Cortex 25(12):4772–4788

Feldman H, Friston K (2010) Attention, uncertainty, and free-energy. Front Hum Neurosci 4:215

Formisano E, De Martino F, Bonte M, Goebel R (2008) "who" is saying "what"? Brain-based decoding of human voice and speech. Science 322(5903):970–973

Friston K (2005) A theory of cortical responses. Philos Trans R Soc Lond B Biol Sci 360(1456):815–836

Gagnepain P, Henson RN, Davis MH (2012) Temporal predictive codes for spoken words in auditory cortex. Curr Biol 22(7):615–621

Gaston P, Marantz A (2018) The time course of contextual cohort effects in auditory processing of category-ambiguous words: Meg evidence for a single "clash" as noun or verb. Lang Cogn Neurosci 33(4):402–423

Graves A, Mohamed A-R, Hinton G (2013) Speech recognition with deep recurrent neural networks. In: 2013 IEEE international conference on acoustics, speech and signal processing. IEEE, pp 6645–6649

Gwilliams L (2020) Hierarchical oscillators in speech comprehension: a commentary on Meyer, Sun, and Martin. Lang Cogn Neurosci 35(9):1–5

Gwilliams L, King J-R (2020) Recurrent processes support a cascade of hierarchical decisions. elife 9:e56603

Gwilliams L, Marantz A (2015) Non-linear processing of a linear speech stream: the influence of morphological structure on the recognition of spoken arabic words. Brain Lang 147:1–13

Gwilliams LE, Monahan PJ, Samuel AG (2015) Sensitivity to morphological composition in spoken word recognition: evidence from grammatical and lexical identification tasks. J Exp Psychol Learn Mem Cogn 41(6):1663

Gwilliams L, Poeppel D, Marantz A, Linzen T (2017) Phonological (un) certainty weights lexical activation. arXiv preprint:1711.06729

Gwilliams L, Linzen T, Poeppel D, Marantz A (2018) In spoken word recognition, the future predicts the past. J Neurosci 38(35):7585–7599

Gwilliams L, King J-R, Marantz A, Poeppel D (2020) Neural dynamics of phoneme sequencing in real speech jointly encode order and invariant content. bioRxiv

Hickok G, Poeppel D (2007) The cortical organization of speech processing. Nat Rev Neurosci 8(5):393–402

Huth AG, De Heer WA, Griffiths TL, Theunissen FE, Gallant JL (2016) Natural speech reveals the semantic maps that tile human cerebral cortex. Nature 532(7600):453–458

Jusczyk PW, Luce PA, Charles-Luce J (1994) Infants' sensitivity to phonotactic patterns in the native language. J Mem Lang 33(5):630

Kilian-Hütten N, Vroomen J, Formisano E (2011) Brain activation during audiovisual exposure anticipates future perception of ambiguous speech. NeuroImage 57(4):1601–1607

Kleinschmidt DF, Jaeger TF (2015) Robust speech perception: recognize the familiar, generalize to the similar, and adapt to the novel. Psychol Rev 122(2):148

Klimovich-Gray A, Tyler LK, Randall B, Kocagoncu E, Devereux B, Marslen-Wilson WD (2019) Balancing prediction and sensory input in speech comprehension: the spatiotemporal dynamics of word recognition in context. J Neurosci 39(3):519–527

Kocagoncu E, Clarke A, Devereux BJ, Tyler LK (2017) Decoding the cortical dynamics of sound-meaning mapping. J Neurosci 37(5):1312–1319

Lau E, Phillips C, Poeppel D (2008) A cortical network for semantics:(de) constructing the N400. Nat Rev Neurosci 9(12):920–933

MacKay DJ (2003) Information theory, inference and learning algorithms. Cambridge university press, Cambridge

Manning CD, Schütze H (1999) Foundations of statistical natural language processing. MIT press, Boston

Marslen-Wilson WD, Welsh A (1978) Processing interactions and lexical access during word recognition in continuous speech. Cogn Psychol 10(1):29–63

Mattys SL, Davis MH, Bradlow AR, Scott SK (2012) Speech recognition in adverse conditions: a review. Lang Cogn Process. 27(7–8):953–978

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

Mitchell TM, Shinkareva SV, Carlson A, Chang KM, Malave VL, Mason RA, Just MA (2008) Predicting human brain activity associated with the meanings of nouns. Science 320(5880):1191–1195

Moore BC (2008) Basic auditory processes involved in the analysis of speech sounds. Philos Trans R Soc Lond B Biol Sci 363(1493):947–963

Mumford D (1992) On the computational architecture of the neocortex. Biol Cybern 66(3):241–251

Norris D, McQueen JM (2008) Shortlist b: a bayesian model of continuous speech recognition. Psychol Rev 115(2):357

O'Shaughnessy D (2008) Automatic speech recognition: history, methods and challenges. Pattern Recogn 41(10):2965–2979

Pennington J, Socher R, Manning CD (2014) Glove: global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp 1532–1543

Rao RP, Ballard DH (1999) Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nat Neurosci 2(1):79–87

Rauschecker JP, Scott SK (2009) Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. Nat Neurosci 12(6):718–724

Scott SK, Blank CC, Rosen S, Wise RJ (2000) Identification of a pathway for intelligible speech in the left temporal lobe. Brain 123(12):2400–2406

Shamma SA (1985) Speech processing in the auditory system: the representation of speech sounds in the responses of the auditory nerve. J Acoust Soc Am 78(5):1612–1621

Shannon CE (1948) A mathematical theory of communication. Bell Syst Tech J 27(3):379–423

Sohoglu E, Davis MH (2016) Perceptual learning of degraded speech by minimizing prediction error. Proc Natl Acad Sci 113(12):E1747–E1756

Stevens KN, Blumstein SE (1981) The search for invariant acoustic correlates of phonetic features. In: Perspectives on the study of speech. Psychology Press, pp 1–38

Wurm LH (1997) Auditory processing of prefixed English words is both continuous and decompositional. J Mem Lang 37(3):438–461

Zhuang J, Randall B, Stamatakis EA, Marslen-Wilson WD, Tyler LK (2011) The interaction of lexical semantics and cohort competition in spoken word recognition: an fmri study. J Cogn Neurosci 23(12):3778–3790

Zhuang J, Tyler LK, Randall B, Stamatakis EA, Marslen-Wilson WD (2014) Optimally efficient neural systems for processing spoken language. Cereb Cortex 24(4):908–918

Zwitserlood, P. (1989). The locus of the effects of sentential-semantic context in spoken-word processing. *Cognition, 32*(1), 25–64

Chapter 6
Speech Perception Under Adverse Listening Conditions

Stephen C. Van Hedger and Ingrid S. Johnsrude

Abstract Perceiving and understanding spoken language is something that most listeners take for granted, at least in favorable listening conditions. Yet, decades of research have demonstrated that speech is variable and ambiguous, meaning listeners must constantly engage in active hypothesis testing of what was said. Within this framework, even relatively minor challenges imposed on speech recognition must be understood as requiring the interaction of perceptual, cognitive, and linguistic factors. This chapter provides a systematic review of the various ways in which listening environments may be considered adverse, with a dual focus on the cognitive and neural systems that are thought to improve speech recognition in these challenging situations. Although a singular mechanism or construct cannot entirely explain how listeners cope with adversity in speech recognition, overcoming listening adversity is an attentionally guided process. Neurally, many adverse listening conditions appear to depend on higher-order (rather than primary) representations of speech in cortex, suggesting that more abstract linguistic knowledge and context become particularly important for comprehension when acoustic input is compromised. Additionally, the involvement of the cinguloopercular (CO) network, particularly the anterior insula, in a myriad of adverse listening situations may indicate that this network reflects a general indication of cognitive effort. In discussing the various challenges faced in the perception and understanding of speech, it is critically important to consider the interaction of the listener's cognitive resources (knowledge and abilities) with the specific challenges imposed by the listening environment.

S. C. Van Hedger (✉)
Department of Psychology, Huron University College, London, ON, Canada

Department of Psychology & Brain and Mind Institute, University of Western Ontario,
London, ON, Canada

I. S. Johnsrude
Department of Psychology & Brain and Mind Institute, University of Western Ontario,
London, ON, Canada
e-mail: svanhedg@uwo.ca

National Centre for Audiology & School of Communication Sciences and Disorders,
University of Western Ontario, London, ON, Canada

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022

L. L. Holt et al. (eds.), *Speech Perception*, Springer Handbook of Auditory
Research 74, https://doi.org/10.1007/978-3-030-81542-4_6

Keywords Speech perception · Attention · Working memory · Cognitive control · Intelligibility · Context · Aging · Listening effort · Cinguloopercular network · Dual-stream auditory processing · Hierarchical representations

6.1 Introduction

Efficient and accurate speech recognition is essential for communication, although people often take this skill for granted. It is difficult to fully appreciate the degree to which individuals rely on speech communication to enrich and provide the essentials of life, and it is similarly difficult to appreciate the processes that support speech recognition across variable listening environments. Diverse listening challenges such as novel voices, speech accents, and a wide range of background noise pose unique perceptual, cognitive, and linguistic demands that often must be solved. This chapter provides an overview of how listeners perceive speech under a variety of adverse listening conditions, with an emphasis on the cognitive and neurobiological foundations that support perception under these conditions. In reviewing the ways in which listeners must overcome listening challenges, this chapter emphasizes that different adverse conditions place different demands on cognitive resources, and so one must consider the specific challenges of a given listening environment to understand how listeners may achieve successful comprehension.

The phrase "adverse listening conditions" might evoke an image of trying to carry on a conversation while sitting on an active airplane runway. What is meant by "adverse" is more varied, more mundane, and more plausible. Imagine, for example, trying to converse with a cashier as they are ringing up items in a crowded grocery store. To successfully perceive the cashier's speech, one must engage in several processes. First, the complex sound wave hitting the ears, which is a mixture of all the audible sounds in the store, must be perceptually organized into discrete sound sources in different locations, based on a variety of cues that enable perceptual grouping and segregation (Darwin and Carlyon 1995). One important cue is harmonicity: the frequency components of the cashier's voice occur at regular harmonic intervals in the spectrum, and one can use this cue to work out which frequency components belong together. This is made more difficult if other sounds, like the "beep" that accompanies the scanning of each grocery item, contain similar frequency components – these can effectively obliterate (energetically mask) the original components from the voice, which would then need to be perceptually restored using knowledge and contextual information. Somehow the noisy and variable speech sounds produced by the cashier are mapped in one's speech/language system onto linguistic representations that are organized (grouped and segregated) into words and phrases, evoking meaning. Several cashiers and customers in the environment may be talking at the same time, making it difficult to determine which words were produced by one's conversational partner and which came from elsewhere, since all of it is processed to some degree by the speech/language system. In other words, masking intelligible speech produces perceptual and cognitive interference (informational masking) (Kidd and Colbourn 2017). If one is deeply familiar with the topic of conversation, and if the linguistic material is very simple and predictable, that will help. If the topic is hard to identify, or if an esoteric word is used, or if the interlocutor has an unfamiliar accent, that adds to the perceptual and cognitive challenge. If one has a hearing impairment, or is an older person, that adds to the challenge as well.

This chapter explores such challenges and what they may involve in more detail. Specifically, the chapter will first explore the cognitive processes underlying successful speech comprehension (Sects. 6.2.1 and 6.2.2), and how this depends on the neurobiology of the human brain (Sect. 6.2.3). From there, the chapter will detail different types of adverse conditions, and the cognitive resources that may be required to overcome them (Sect. 6.3). The role of attention in speech comprehension will then be specifically highlighted, with an emphasis on how the role of attention may differ dramatically depending on listening conditions (Sect. 6.4). The chapter will then introduce the idea of listening effort and explain it as an interaction between the demands imposed by the listening situation, and the unique constellation of cognitive abilities an individual listener brings to bear (Sect. 6.5). Finally, potentially fruitful directions of future research will be identified (Sect. 6.6).

6.2 Important Speech Features for Effective Comprehension

It may be difficult to appreciate the variety of processes required to successfully understand a signal as rich and complex as fluent speech. Just as one is not consciously aware of the complexities of other systems, such as the mechanisms supporting breathing or balance, speech understanding often feels like it occurs effortlessly and automatically. To put the complexity of speech understanding in perspective, briefly consider some of the steps involved in conversing with another individual. One presumably starts with a linguistic thought, which then must be transformed into a physiological code (moving one's lips, tongue, and vocal cords to produce the intended speech), using the distinctive articulations characteristic of a particular individual's accent, idiolect (speech habits peculiar to an individual), and voice. This signal, which now exists as compressions and rarefactions in the air, mixes with other acoustic energy in the environment, creating a complex waveform which impinges on the eardrum of the listener and is transduced into electrical impulses in the auditory nerve in the cochlea. The listener must analyze this complex sound to perceptually organize the auditory scene into discrete sources, segregating the target signal from any background, and mapping sounds onto linguistic representations, eventually resulting in understanding. This *speech chain* (Denes and Pinson 1993) unfolds extremely quickly in naturalistic settings and is aided by listeners' remarkable abilities to segment speech into meaningful units (Sect. 6.2.1), listeners' abilities to hold parts of speech in working memory and use context to improve comprehension (Sect. 6.2.2), and the neurobiology of listeners' auditory systems (Sect. 6.2.3).

6.2.1 Segmentation

One of the most fundamental components of comprehending speech is the parsing of a continuous speech signal into discrete words and phrases. This process is so well rehearsed that many individuals (who do not study speech for a living) are surprised to discover that the boundaries of words are not actually represented by silence or other reliable acoustic markers in the waveform. For example, consider a relatively long single word in English – *unimaginatively* – which contains seven syllables. Even if this word does not appear in everyday conversation, native speakers of English will generally have little trouble grouping these syllables together, easily parsing the word from other words that make up a phrase or sentence, such as *He spoke unimaginatively*. One can experience the issue of speech segmentation firsthand by listening to naturalistic speech from an unfamiliar language. In this exercise, one may get sense of where word boundaries exist, but this will be largely driven by how one segment speech in one's native language. Indeed, this is precisely what Cutler and Norris (1988) demonstrated in a seminal paper. English speakers tend to demarcate lexical items using the rhythmic patterns of their native language, with strong syllables being more likely to correspond to the beginning of a word in English. English listeners in their study were slower to detect a target word embedded in nonsense disyllables when there were two strong syllables (e.g., detecting *mint* in the nonsense disyllable *mintayve*) compared to a strong and a weak syllable (e.g., detecting *mint* in the nonsense disyllable *mintesh*). Thus, listeners must learn the appropriate cues to parse a continuous speech stream into discrete lexical items, but these cues are not universal across languages and are not necessarily reflected in the acoustics of the speech signal.

Further research indicates that a host of other statistical characteristics of a known language, in addition to stress patterns, are used to segment speech into words. Phoneme sequence constraints, or phonotactics, describe the permissible combinations of phonemes in a language at various points in a word, such as onsets and offsets. Listeners have implicit knowledge of the phonotactics of their native language, and word boundaries are inferred when phoneme transitional probabilities are low. For example, the sequence "I'd love lunch" (phonetic notation: /ajdləvləntʃ/) would be heard by English speakers as having a boundary between the /vl/ sequence. This is because the phoneme sequence /vl/ cannot occur at the beginnings of words in English. Rather, /vl/ can only occur in the middle of words (e.g., "unraveling"), or at word offsets (e.g., "unravel") – arguably even then with a schwa (a weak, unstressed vowel, such as the "a" in "about") between the /v/ and /l/. As such, in the example sequence /ajdləvləntʃ/, the only possible perceptual organization that would not leave nonword fragments (e.g., /əntʃ/) would place a word boundary between /ləv/ and /ləntʃ/ (Norris et al. 1997). In addition to acoustic and lexical information, semantic information and context can also drive segmentation – in fact, according to Mattys et al. (2005), knowledge-based lexical and semantic cues are the most important for driving perception, followed by segmental cues such as phonotactics, with stress being perhaps the weakest cue to segmentation. The problem of determining word boundaries is thus complex, requires the balancing of multiple, sometimes conflicting, constraints, and draws on both the acoustics of the signal and prior language-specific linguistic knowledge.

6.2.2 Working Memory and Use of Context

The effective comprehension of speech requires a kind of active hypothesis testing of what was said. The acoustics of speech do not cleanly map onto linguistic categories – a single acoustic event can have multiple phonetic interpretations depending on the speaker and the context of the listening environment, and a single phonetic category can have multiple acoustic realizations. This lack of invariance in speech (Liberman et al. 1967) means that there is a many-to-many mapping between any acoustic event and its linguistic meaning, which poses a computational problem to the listener. As such, working memory – the ability to temporarily store, maintain, and manipulate information in service of complex cognitive tasks (Baddeley and Hitch 1974) – may be particularly important for effectively weighing possible interpretations of incoming speech until the most appropriate interpretation can be selected.

For example, consider the vowels /I/ and /ɛ/, as heard in the words "bit" and "bet." These vowels in American English are highly similar with respect to their formant frequencies, making them particularly confusable. This means that a listener may rely on working memory to understand a spoken sentence in which the intended utterance is not immediately apparent. In the sentence, "The [bill/bell] was so large that it took me by surprise, even though I had previously been to that church," both interpretations of the bracketed words are plausible until the final word, which ultimately provides strong evidence for "bell." Even in this simple example, it should be apparent how working memory is an important component of effective speech comprehension, especially as ambiguity is increased or the strength of the meaningful context in which the ambiguous utterance is decreased, as is often the case in adverse listening conditions.

Ambiguous speech material must be held online in some way until sufficient contextual information is received to disambiguate it. Context, broadly construed, is any information in, or related to, the environment that might constrain interpretation of an ambiguous utterance. Context can include other words in the utterance, or what was previously said, visual cues, or even shared history with the talker. Context influences the perception of speech across multiple levels of analysis, reflecting the inherent ambiguity of how acoustic patterns map onto linguistic categories, how words map onto meaning, and pragmatically how an utterance ought to be interpreted (often beyond its literal meaning).

Robust context effects have been observed at the level of phonemes, even for nonlinguistic context, such as sine waves presented in the frequency range of vowel formants (Holt 2005), which supports the idea that contextual influences on perception of sublexical elements may reflect a more general auditory process. Yet, at the level of the talker, the influence of context depends on a listener's interpretation – that is, whether the listener attributes a particular sound to idiosyncratic variation in articulation (due to a talker's idiolect, or perhaps due to temporary articulatory constraints such as holding a pencil in their teeth) or due to principled changes that are linguistically informative (Kraljic et al. 2008). Such principled changes would include those due to the talker's dialect, coarticulatory effects, the articulators forming the next sound before the previous sound is completely produced, or other kinds of fine phonetic (or subphonemic) detail, signaling, for example, morphological complexity, utterance ending, register, and emotional state (Hawkins 2003). This suggests that the influence of context on phonetic perception is complex, depending both on the level at which the context is operating and the interpretation of the context in service of understanding meaning and talker-specific attributes.

Context is also critical for disambiguating words with multiple meanings and/or syntactic roles (Rodd et al. 2002, 2005). When interpreting an utterance, a listener must use the surrounding words to guide the selection of the appropriate syntactic role and semantic properties of each word. For example, in the phrase "the bank of the river," the initial word "the" indicates that "bank" is being used as a noun and not a verb, while the semantic properties of the word "river" indicate that "bank" is referring to the water's edge and not to, for example, an institution concerned with the borrowing and saving of money. These forms of ambiguity are ubiquitous in language. At least 80% of the common words in a typical English dictionary have more than one definition (Rodd et al. 2002), and many words, such as "run," have dozens of definitions. Each time one of these ambiguous words is encountered, the listener must hold the unfolding utterance in mind until they are able to select the appropriate meaning based on context.

Context is not limited to the auditory modality. In many everyday settings, the recognition of speech occurs in tandem with the processing of visual information, either in the environment or from the talker's face and gestures which establishes a specific context (incorporating a talker's sex, height, and facial attributes) for interpreting the speech signal. Listeners frequently make use of bimodal speech cues that are readily available in conversational settings and that tap existing knowledge. For example, if one is at a busy party and hears the sentence "I wanna eat the Grampa bunny's hearing aids!," knowing that what is shown in Fig. 6.1 is on the table in front of the 10-year-old talker (and that the "hearing aids" on the larger cake are made of marzipan, and that the child loves marzipan) would help enormously.

A wealth of research indicates that auditory and visual information complement each other in speech perception and that the facial gestures available in audiovisual speech make it more intelligible than auditory-alone speech. In one of the earliest publications on the topic, visual speech cues were noted to improve the signal-to-noise ratio (SNR) by up to 15 dB (Sumby and Pollack 1954), dramatically enhancing intelligibility. The use of visual speech information is especially advantageous when speech is semantically and syntactically complex (Reisberg et al. 1987) or when it is impoverished or degraded (Macleod and Summerfield 1990).

Linguistic information (phonotactic, lexical, semantic, syntactic, and facial/gestural), which is used to disambiguate speech, must be stored as long-term, stable

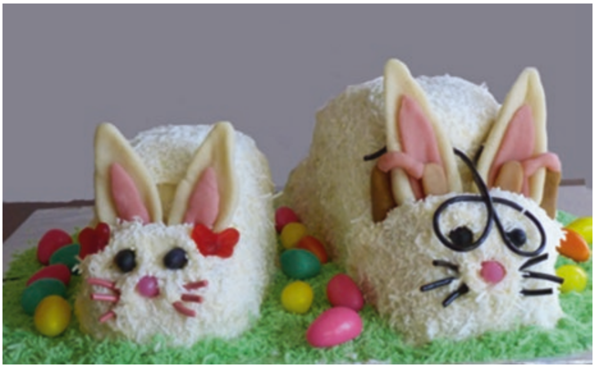

Fig. 6.1 The spoken sentence "I wanna eat the Grampa bunny's hearing aids!" makes a lot more sense when you know that the talker is a 10-year-old who loves marzipan, at a joint Easter birthday party for a 77-year-old man and his 8-year-old granddaughter and that the 10-year-old is looking at these cakes (particularly the one on the right; with "hearing aids" made of marzipan)

representations in the brain. Semantic knowledge and memory are required to comprehend spoken language, such as to interpret an utterance in the context of known facts and events. Information that has been stored about individual talkers can also facilitate intelligibility and comprehension. For example, voices of people that are personally known to a listener are substantially more intelligible than voices of strangers, when heard in a mixture with a competing talker (Johnsrude et al. 2013; Holmes et al. 2018), and better intelligibility also results when listeners are trained with voices in a lab (Nygaard and Pisoni 1998). Thus, long-term knowledge of a talker's articulatory patterns, developed through prior experience, can constrain the interpretation of speech.

6.2.3 Distributed Neurobiology for Effective Comprehension

At this point, it should be apparent that listeners use multiple cues to successfully comprehend fluent speech. This effective understanding of speech must be grounded in the neuroanatomy of the auditory system, as well as a more distributed language network, and so it is worth considering how speech comprehension is supported from a neurobiological perspective. Beginning with general auditory processing, anatomical and neurophysiological findings in nonhuman primates support the idea of multiple parallel streams of processing in the auditory system. Despite 25 million years of divergent evolution, the anatomical organization of cortical auditory system in rhesus macaque monkeys is often taken as a model for human cortical organization (Davis and Johnsrude 2007; Hackett 2011). Processing of auditory information is highly parallel (multiple computations at once) at various levels of the primate auditory system.

Even in the earliest cortical receiving areas (primary, or "core" auditory cortex), multiple representations of the input are available (Jones 2003). The organization of the cortical auditory system is cascaded, with hierarchical connections among auditory core, neighboring secondary or "belt" regions, and adjacent parabelt areas, suggesting at least three discrete levels of processing (Hackett 2011). A distributed, interconnected set of fields, in superior temporal gyrus and sulcus, in the inferior parietal lobule, and in prefrontal cortex, receive inputs from belt and parabelt regions, constituting a potential fourth stage of processing (Hackett 2011; see Fig. 6.2).

Accounts of speech processing in humans emphasize two main processing pathways that radiate out from primary auditory regions on the superior temporal plane (Hickok and Poeppel 2015; but also see Davis and Johnsrude 2007). The "dual-stream" account is based on the observation that temporal, parietal, and frontal connections of macaque auditory cortex are topographically organized. Anterior belt, parabelt, and associated anterior temporal-lobe regions interconnect with anterior and ventral frontal cortical sites (the ventral auditory stream). In contrast, more posterior belt, parabelt, and associated posterior temporal regions interconnect with more posterior and dorsal frontal cortical sites (the dorsal auditory stream) (Hackett 2011). These two routes have been given different putative functional roles. For the ventral stream, these include a role in lexico-semantic comprehension of speech, and in selective retrieval of contextual information associated with words (Hickok and Poeppel 2015). For the dorsal stream, these include motor-articulatory mapping of sound which might be particularly important for understanding when speech is acoustically degraded (Du et al. 2014).

Results of functional neuroimaging studies provide evidence that human speech perception may also be based on multiple hierarchical processing pathways consistent with a comparative neurobiological framework. Early functional magnetic resonance imaging (fMRI) investigations demonstrated that, for listeners hearing nonlinguistic stimuli, more complex sounds (amplitude and frequency-modulated tones, bandpass filtered noise) activated auditory regions beyond the core (belt and parabelt), whereas simpler sounds (pure tones) activated primarily the core (Giraud et al. 2000). Davis and Johnsrude (2003) investigated the hierarchical organization of the speech perception system used a converging-operations approach in which naturalistic sentence-length stimuli were processed three acoustically different ways, each applied parametrically to yield different levels of intelligibility. The

Fig. 6.2 (continued) four levels of processing, including core regions (darkest shading), belt regions (light shading), parabelt regions (hatching), and temporal and frontal regions that interconnect with belt and parabelt (dotted). (Adapted from Hackett et al. (2014).) Dotted lines indicate sulci that have been opened to show auditory regions. (**b**) Schematic of cortical areas in the macaque monkey that are metabolically active during processing of auditory, visual, and audiovisual stimuli. (From Poremba and Mishkin (2007)). (**c**) Model of hierarchical processing of speech summarizing neuroimaging data (see text; Davis and Johnsrude 2003; Okada et al. 2010; after Peelle et al. 2010). *CS* central sulcus, *IPL* inferior parietal lobule, *IPS* intraparietal sulcus, *ITG* inferior temporal gyrus, *MTG* middle temporal gyrus, *PFC* prefrontal cortex, *STG* superior temporal gyrus, *STS* superior temporal sulcus

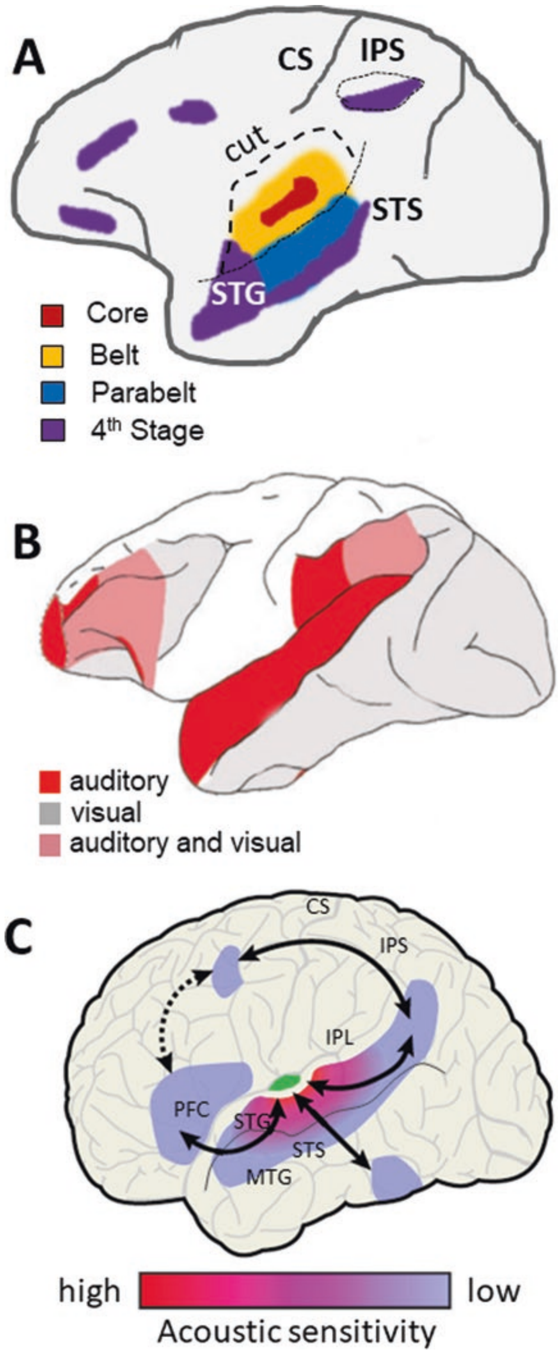

Fig. 6.2 Auditory-responsive cortex in the primate includes many anatomically differentiable regions. All brains show the brain from the side, with the front of the brain (frontal cortex) to the left of the page. (**a**) The anatomical organization of the auditory cortex is consistent with at least (continued)

investigators were able to distinguish three levels of processing. Primary auditory regions were sensitive to any kind of sound, intelligible or not. Activity in more lateral, anterior, and posterior areas in the temporal lobe correlated with intelligibility, but also differed depending on acoustic characteristics (specifically, the type of distortion). More distant intelligibility sensitive regions in the middle and superior temporal lobes and in left inferior frontal gyrus (IFG) were not sensitive to the acoustic form of the stimuli, suggesting that more abstract, nonacoustic processing of speech is performed by these regions. These three levels of processing, reflecting progressive abstraction of the linguistic signal from the acoustic, appear to radiate out from primary auditory cortex in a fashion reminiscent of the anatomical organization of the auditory system in macaques (see Fig. 6.2).

Rodd et al. (2005, 2012) subsequently identified left dorsolateral frontal and posterior inferior temporal regions, even further from primary auditory cortex, which are recruited when listeners hear meaningful, intelligible sentences that contain words with more than one meaning, perhaps consistent with a fourth stage of processing; see Fig. 6.3. Binder et al. (2009) observed imaging results consistent with the idea that linguistic processes at higher processing stages are topographically further away from auditory cortex. In a meta-analytic study of 120 functional imaging reports, they observed that when people had to process the meaning of spoken or read words,

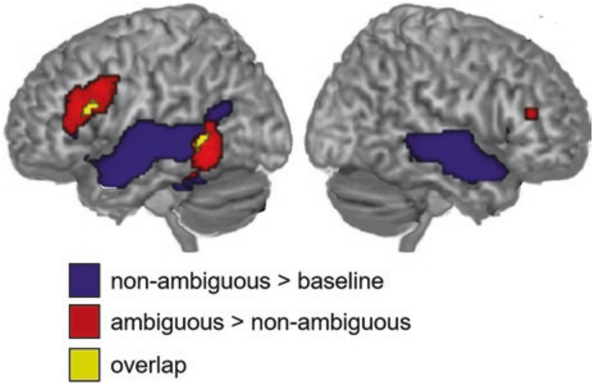

Fig. 6.3 Functional magnetic resonance imaging (fMRI) activation in response to spoken sentences with or without lexical ambiguity, shown superimposed on a brain structural image. The left hemisphere of the brain is shown on the left (front of the brain nearest left margin) and the right hemisphere on the right (front of the brain nearest right margin). Comparison between sentences without ambiguous words (e.g., “her secrets were written in her diary”) and a baseline, energy-matched, noise condition revealed a large area of greater activation for the former condition (in blue) in left and right superior and middle temporal gyri, extending in the left hemisphere into posterior inferior temporal cortex and the left fusiform gyrus. Greater activation for this intelligible speech, compared to noise baseline, was also observed in both hemispheres in lingual gyrus, and in the dorsal part of the inferior frontal gyrus (pars triangularis). Greater activation for sentences with ambiguous words (e.g., “the *shell* was *fired* towards the *tank*”) compared to matched sentences without (in red) was observed in left and right inferior frontal gyrus (IFG) (pars triangularis), and a region of the left posterior inferior temporal cortex. The yellow area indicates overlap between the two contrasts. (Adapted from Rodd et al. (2005))

activation clustered in seven distinct regions (Binder et al. 2009). Active regions included the inferior parietal lobule (the angular gyrus and some of the supramarginal gyrus); middle temporal gyrus; fusiform and adjacent parahippocampal regions; IFG, ventral and dorsal medial prefrontal cortex; and retrosplenial cortex.

Neuropsychological data are also consistent with this hierarchical framework. Damage from conditions such as stroke, in or near auditory cortex (particularly in the left hemisphere) in humans, can result in a condition called “word deafness,” a type of agnosia in which spoken words are no longer recognized (Phillips and Farmer 1990). It is doubtful, however, that “word deafness” is entirely specific to speech, as it may also apply to some nonverbal sounds.

Farther from auditory cortex, damage results in language deficits at a higher linguistic or conceptual level. Lesion-symptom mapping (Bates et al. 2003) is a technique that allows researchers to combine behavioral and brain imaging maps of lesions from individuals with brain damage to identify the brain regions that, if damaged, are most likely to result in deficits on specific behavioral tasks. For example, using this technique in 64 individuals with left-hemisphere cortical damage, Dronkers et al. (2004; Turken and Dronkers 2011) established that a number of regions outside of primary auditory cortex, in the middle and superior temporal gyri and in the inferior frontal cortex, were commonly damaged in individuals who had difficulty understanding spoken sentences (Fig. 6.4). Again, areas that process

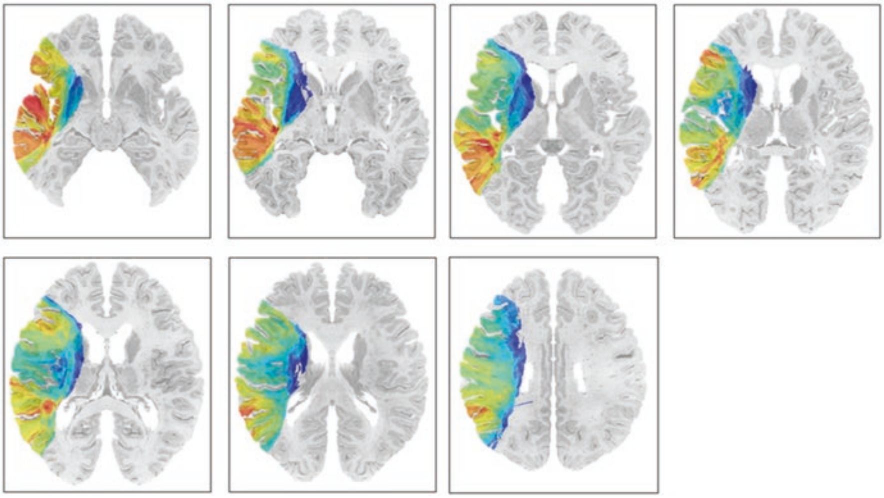

Fig. 6.4 Regions related to comprehension of spoken sentences (in red). The region of brain damage was mapped in each of 64 individuals with language disruption (aphasia) as a result of stroke in the left hemisphere of the brain. Different individuals showed difficulty with different aspects of language, depending on the location of the lesion. Areas in which damage related to impairment in the comprehension of spoken sentences are shown hot colors, with the strongest relationship shown in red. These data are superimposed on horizontal brain slices (in gray). In these images, left is on the left, and the front of the brain is at the top of each image. From left to right, and top to bottom, slices are progressively closer to the top of the brain. The red regions cover middle and superior temporal gyri and in the inferior frontal gyrus (IFG). (From Turken and Dronkers (2011); Fig. 1, panels 3–9)

meaning seem to be quite distant from auditory cortex. This is also demonstrated by a lesion-symptom mapping study conducted by Mesulam et al. (2012) on individuals with primary progressive aphasia, a neurodegenerative condition that presents as a loss of word meaning. They administered a comprehensive battery of language tests and examined the correlation between regional cortical atrophy and the magnitude of impairment on different tests. Impairment in auditory word comprehension correlated with atrophy in the anterior temporal region bilaterally, whereas impairment in sentence comprehension correlated with atrophy in orbitofrontal and lateral frontal regions, and in the inferior parietal lobule, all areas well away from auditory regions. To date, most work exploring the neurobiology of speech and language processing has examined responses to words or sentences presented in quiet conditions. How this network is altered when listening conditions are challenging will be discussed in Sect. 6.4.

6.3 The Cognitive Resources Recruited to Meet Challenges Resulting from Different Types of Adversity

The listening conditions of everyday life are highly variable. Sometimes speech is heard in quiet. More often, however, it is degraded or masked by other sounds. Such challenging situations increase processing demand (also referred to as processing load) when, for example, the stimulus is masked by interfering background noise or by speech from other talkers, or because the stimulus is degraded due to peripheral hearing loss. No specialized cognitive module fully accommodates the myriad of challenges one might encounter in everyday listening conditions – the mechanisms underlying the perception and understanding of speech are simply too distributed, and different challenges are met in different ways. Thus, while Sect. 6.3 discusses different types of adverse listening conditions separately for the sake of tractability, it should be remembered that in many real-world listening environments, more than one kind of listening challenge may be present at one time.

6.3.1 Masking

Masking can be defined as “the process by which the threshold of hearing for one sound is raised by the presence of another” (ANSI 2013). For example, the amplitude threshold for understanding a friend’s speech will be increased if they are talking over a roaring waterfall or over a professor delivering a lecture, relative to a quiet environment. Yet, as this example highlights, the “masking sound” is always defined relative to the target speech and thus can be acoustically highly variable, ranging from broadband noise (as is the case with the waterfall) to a single talker (as is the case with the professor). As such, researchers have drawn a conceptual distinction between types of masking sounds to clarify whether the masking is *energetic* or *informational* in nature (Brungart et al. 2001). These categories of masking, in addition to the mechanisms required to overcome them, are considered in Sects. 6.3.1.1, 6.3.1.2, and 6.3.1.3.

6.3.1.1 Energetic Masking

Energetic masking is thought to occur when the target sound and interfering sound overlap in time and frequency in the cochlea (e.g., Culling and Stone 2017), such as the detection of speech in broadband noise (like the waterfall example provided in the previous paragraph). Energetic masking poses a challenge for listeners because the masking noise interferes with the target speech at the level of the auditory nerve. Thus, energetic masking, as well as the mechanisms thought to provide a release from energetic masking, is typically discussed in terms of the auditory periphery, though in some cases the proposed explanations require some consideration of cognitive mechanism.

The effects of energetic masking also appear to be lessened when the masking noise is amplitude modulated, with optimal target speech intelligibility occurring around a 10 Hz modulation rate (Miller and Licklider 1950). This relative benefit of modulating the masker noise is presumably due to listeners being able to selectively process the target stimulus in the low-amplitude periods of the masker noise, which has been referred to as “dip listening” (Culling and Stone 2017). Importantly, with respect to a discussion of cognitive mechanism, dip listening appears to relate to masker familiarity, suggesting an influence of learning and memory on selective processing. Specifically, Collin and Lavandier (2013) demonstrated that masker modulations based on the same speech token are easier to cope with compared to masker modulations based on variable speech tokens. These findings suggest that the predictability of amplitude modulation in the masker stimulus is informative in modeling the relative benefit of dip listening, which points to a role of learning-driven familiarity on dip listening efficacy.

6.3.1.2 Informational Masking

Informational masking is the term for all other forms of masking that are not energetic. As the signal is physically not interfered with at the periphery, informational masking is thought to operate at a more central (rather than peripheral) level. Consequently, it is more frequently discussed in terms of underlying cognitive mechanisms. Research has indicated the conditions under which informational masking is thought to occur (e.g., see Kidd and Colbourn 2017). Broadly defined, informational masking can be thought of as an increased challenge in understanding due to the perceived similarities between a target and masker stimulus, even when the target and masker stimuli do not overlap in frequency or time. As such, cognitive processes such as selective attention, divided attention, and working memory are important factors in understanding both informational masking and how to mitigate it.

One of the clearest demonstrations of informational masking comes from studies in which listeners misattribute entire words or phrases spoken by a masker talker to the target talker (Brungart et al. 2001), as this kind of pattern cannot be explained in terms of poor audibility resulting from energetic masking. For example, in a popular paradigm known as the coordinate response measure (CRM; Bolia et al. 2000), listeners hear two or more talkers simultaneously say a sentence with the structure "Ready [call sign] go to [color] [number] now." Participants must listen for their designated call sign on each trial and then navigate to the appropriate coordinate (in a color-number grid). In order to succeed at the task, participants must not confuse the coordinates of the target talker with those spoken by the masker(s).

In these kinds of listening situations, research has established a relationship between accurate speech recognition and cognitive functioning, at least for older individuals with hearing loss. For example, Humes et al. (2006) investigated younger (non-hearing-impaired) and older (hearing-impaired) listeners' performance on the CRM, finding that situations in which listeners were required to divide attention resulted in consistently worse performance for the older, compared to younger, listeners. Moreover, individual differences in short-term memory and working memory (operationalized as an average score of forward and backward digit span) were related to accurate speech recognition among the older listeners. These results suggest that the attentional demands of informational masking may lead to population differences among older and younger listeners who differ in hearing impairment, although individual differences in short-term and working memory may provide a particular benefit for hearing-impaired older listeners, presumably due to a better control of attention and an ability to actively maintain a greater number of hypotheses about what was said by each talker, which may help to resolve ambiguity.

More broadly, working memory appears to be important for release from informational masking when the linguistic content of the target and masker are semantically confusable. In an experiment by Zekveld et al. (2013), listeners had to detect a target sentence that was played simultaneously with stationary noise, amplitude-modulated noise modeled on a speech envelope, or a single talker. The target sentence, moreover, could be preceded by a word that was semantically related to the sentence or an unrelated nonword. Results demonstrated that working memory positively related to sentence comprehension under specific conditions – namely, when the target sentence was preceded by a semantically related word and when the masker stimulus was a single talker. These findings highlight how higher working memory may help listeners to more effectively use a meaningful cue to attend to a target talker, at least in situations where the masker is easily confusable with the target.

Given the demands of informationally masked speech on working memory and attention, it is possible that interventions aimed at improving the functioning of these cognitive constructs may result in better speech comprehension. In support of this framework, Ingvalson et al. (2015) found improvements in both reading span (a memory span task thought to index working memory) and speech-in-noise perception after 10 days of training on backward digit span, although the speech-in-noise tasks used nonspeech environmental sounds (e.g., dog barks) as an informational masker, which may have different properties than speech used as an informational masker. In contrast, other work did not reveal a benefit of working memory training on speech-in-noise performance (Wayne et al. 2016). In this experiment, the masker stimulus was another talker. More research in this area is clearly needed.

Another means of improving listeners' abilities to understand informationally masked speech, which has received considerable empirical support, is to increase the perceptual familiarity with one of the talkers. Listeners can learn talker characteristics that lead to advantages in understanding informationally masked speech (Nygaard and Pisoni 1998). Importantly, this talker-familiarity advantage is not simply driven by heightened attention to, or salience of, the familiar talker; listeners also show enhanced performance when a novel talker is the target stimulus and the familiar talker is the masking stimulus. This suggests that familiarity may more broadly allow for the segregation of similar talkers into distinct auditory streams (Johnsrude et al. 2013).

To conclude, informational masking appears to pose a problem for listeners because the target and masker signals are often confusable in terms of linguistic content, which places particular demands on listeners' working memory and attention abilities to successfully parse these signals. Training programs that specifically target working memory have shown some transfer to perceiving informationally masked speech, but the evidence for this transfer is mixed. Long-term familiarity with a talker may improve speech intelligibility and reduce the demands of working memory and attention in part because listeners are more effectively able to orient their attention toward (or away from) the familiar talker, allowing greater segregation of auditory streams.

6.3.1.3 Spatial Release from Masking

One well-studied means of dealing with both energetic and informational masking is to use spatial cues to segregate the target speech from the masker stimulus, assuming such cues are present. Revisiting the scenario of conversing with a cashier in a crowded grocery store, spatial release from masking would help one differentiate the speech of one's cashier from, say, a cashier at another register, simply because these two sound sources are physically separated in space. Spatial release from masking is a particularly effective means of improving speech intelligibility across a wide range of masker stimuli. This is because the spatial separation between a talker and masker signal will result in both sounds reaching one's ears at slightly different times, with different loudness levels, and even different distributions of frequency components, due to the fact that sounds may be altered by the "acoustic shadow" of one's head, as well as by the shape of one's ears. These differences provide several cues that listeners may use to effectively segregate sound sources and improve comprehension.

For example, if a target and a masker sound are spatially separated, one ear may receive a more favorable SNR than the other. Listeners appear to be able to select the ear with a higher SNR – an ability also referred to as “better-ear listening” (Edmonds and Culling 2006). The precise way in which listeners are able to ultimately select the more favorable ear is not completely understood, though it appears to be a “sluggish” process, meaning listeners cannot rapidly shift to take advantage of the relatively more favorable SNR (Culling and Mansell 2013). Further, selective attention to a given ear may alter the physiological response of the outer hair cells in the unattended ear (Srinivasan et al. 2014), altering the effective SNR in that ear.

A second way listeners can separate sounds based on spatial location is through binaural unmasking. Binaural unmasking occurs because, if the target and masker are at different locations, the phase or level difference between the two ears will be different for the two different sounds.

A study by Kidd et al. (2010) examined the acoustic factors that influence spatial release from both informational and energetic masking. In their paradigm, which assessed speech intelligibility using the CRM (see Sect. 6.3.1.2), the target speech stimuli were filtered into several frequency bands. Importantly, the authors found the greatest spatial release from masking when the stimulus was presented at full bandwidth (not filtered), suggesting an integration of binaural cues (phase and level differences) across different frequency regions help to improve performance. The next best spatial release from masking, however, was found for low-frequency components, suggesting that phase differences may be more important than level differences. In a second study, in which energetic and informational masking were varied and listeners could only rely on timing differences between the ears, the authors (Kidd et al. 2010) found large spatial release from masking only when there was significant informational masking. Taken together, these results highlight the importance of considering the extent to which a masker is energetic or informational, as well as the relative contribution of different cues to spatial localization in characterizing the speech intelligibility benefits that may arise from spatial release of masking.

6.3.2 Unfamiliar Talker

Even in favorable listening environments, with little to no masking noise, speech perception can pose a challenge if the talker is unfamiliar. The extent to which understanding an unfamiliar talker poses a challenge, in many cases, depends on the relative difference in accent between the speaker and the listener (Adank et al. 2009). This is because listeners who encounter a nonnative accent or an unfamiliar native accent must rapidly adapt to this variation in speaking, which often permeates multiple levels of the hierarchy of speech. For example, perceiving nonnative accented speech can be challenging when speakers produce contrasts that are not present in their native language, such as the /r/–/l/ contrast in English for native Japanese speakers (Bradlow et al. 1997). At a more suprasegmental level, nonnative speakers sometimes cannot produce the native stress and intonation patterns that help listeners parse the speech signal into meaningful words and phrases (Guion et al. 2004). However, it should be noted that improvements in the production of native-like speech can be observed in adults after training (Lim and Holt 2011), highlighting the importance of learning and plasticity in the sensorimotor representations of nonnative speech categories.

Despite these challenges, listeners are often able to adapt to accented speech rather quickly, at least in specific listening environments. In a speeded word comprehension task, Clarke and Garrett (2004) found that listeners were initially slower to respond to nonnative accented speech, suggesting that there is an additional processing cost for comprehending unfamiliar speech. This relative slowdown, however, was rapidly attenuated (but not eliminated) over the course of just a few trials. This rapid accommodation, however, has been found to interact with the background noise of the listening environment. Under "quiet" listening conditions, the relative processing cost between accented and non-accented speech is small or sometimes not observed at all (Floccia et al. 2006) and can be mitigated through relatively little experience with the unfamiliar talker. Yet, in more adverse listening situations (such as the introduction of energetic or informational maskers), the relative difference between familiar and unfamiliar accented speech becomes significantly more pronounced.

This interaction between the noisiness of the listening environment and the understanding of unfamiliar speech has also been found with computer-synthesized speech, suggesting that it reflects a broader principle of unfamiliarity with the particular phonetic variation of a given talker rather than specific idiosyncrasies with a particular type of accent (Pisoni et al. 1985). In this experiment, the researchers compared several text-to-speech synthesizers to natural speech, finding that the relative difference in comprehension between synthetic and natural speech was magnified under more adverse (noisier) listening conditions. Moreover, the authors found that semantic and syntactic contexts were important components of intelligibility, which means that the relative challenges to comprehension posed by speaker unfamiliarity can be reduced by constraining the possibilities of a given speech token.

What cognitive mechanisms allow listeners to adapt to unfamiliar talkers in these listening situations? The observation that listeners show rapid improvements in understanding an unfamiliar talker suggests a kind of internal calibration, dependent on the degree to which stored phonological and lexical representations overlap with the incoming speech signal (Van Engen and Peelle 2014). This internal calibration may depend in part on working memory (Janse and Adank 2012), but the generalizability of this claim is unclear given that it was supported by a study using older listeners as participants, who may face unique challenges in speech perception and thus may recruit cognitive resources differently (see Sect. 6.3.3). Indeed, among younger listeners, the role of working memory has been less strongly supported in unfamiliar speech recognition and may be mediated by general vocabulary knowledge (Banks et al. 2015).

Successful adaptation to an unfamiliar talker may require inhibitory mechanisms. This is because an unfamiliar talker may pronounce a given word in a manner that more strongly aligns with a different representation for a listener. For example, using the /r/-/l/ contrast from above, a native Japanese speaker may pronounce “rake” closer to “lake,” and thus a listener must inhibit “lake” to facilitate understanding, especially in situations where the context of the accompanying speech does not clearly constrain the interpretation of the word (e.g., the sentence “The rake/lake is big”). In support of this hypothesis, Banks et al. (2015) demonstrated that inhibitory control – measured through a Stroop Task – predicted the speed and efficacy of adapting to an unfamiliar talker, though in their paradigm the unfamiliar speech was simultaneously presented with speech-shaped background noise. Although this choice in experimental design is certainly justified, especially given the augmented effects of unfamiliar talker adaptation when listening in a noisy environment, an important consideration in any discussion of mechanism is whether adaptation to an unfamiliar talker in noise reflects the same cognitive processes as those required in less adverse listening conditions. As such, it will be important to clarify in future research whether the cognitive mechanisms that allow an individual to adapt to an unfamiliar talker are identical (and just more heavily recruited) in noisy environments, or whether the presence of an unfamiliar talker in conjunction with noise results in an emergent set of required cognitive processes.

6.3.3 The Effect of Aging

The discussion of adverse listening conditions thus far highlights that both external factors (such as the presence of energetic or informational maskers) and internal factors (such as the degree of overlap in accented speech with one’s mental representations) can influence the ease with which speech can be understood. This illustrates the importance of considering the interaction of the individual’s cognitive resources – knowledge and abilities – with the challenges imposed by the listening environment when discussing speech perception in adverse conditions.

Yet, in this framework one cannot simply assume that the individual’s abilities remain constant across the lifespan. For example, older listeners often have difficulties in understanding speech, especially when it is heard in a noisy environment (see Rogers and Peele, Chap. 9). A detailed discussion of how aging influences speech perception is beyond the scope of this chapter; however, research in this area has highlighted that the relationship between aging and speech perception is likely grounded in changes to both perceptual and cognitive processes. More specifically, age-related declines in sensory processing may increase the perceptual challenge of any given listening environment, which in turn may place greater demands on cognitive processes, such as selective attention and working memory, for successful comprehension (see Wayne and Johnsrude 2015 for a review). However, given that aging is also associated with declines in cognitive functioning, older listeners may have increased difficulties engaging these cognitive processes in service of speech understanding. Training programs designed to improve cognitive processes such as working memory among older listeners have generally produced null or minimal transfers to speech perception in adverse conditions (e.g., Wayne et al. 2016), and consequently the best approach to reducing listening effort and increasing speech comprehension among elderly listeners is still actively debated.

6.4 Neuroimaging Evidence That Different Demands Recruit Different Systems

Just as the cognitive mechanisms that help listeners cope with adverse conditions depend on the particular elements of the listening environment, the neural mechanisms associated with speech recognition in adverse conditions depend on the specific factors that make a listening situation difficult. As such, it is inappropriate to think of any single brain area as responsible for accommodating "adverse conditions," broadly defined. Rather, neuroimaging research has identified consistent brain networks that are engaged as listeners cope with specific kinds of adverse conditions.

Before discussing these networks, it is important to highlight a methodological consideration in this research area. One of the most commonly used methods for investigating the neural underpinnings of speech perception in adverse conditions is fMRI, a noninvasive technique that provides relatively poor temporal resolution (given the lag of the hemodynamic response in response to neural activity) but good spatial resolution across the whole brain, making it particularly well-suited to studying networks serving complex behaviors such as speech perception. Yet, fMRI generates considerable acoustic noise that can energetically mask speech during image collection. To address this issue, researchers generally use a technique called "sparse scanning," in which the (noisy) process of image acquisition is confined to periods directly before and after, but not during, the presentation of speech (Hall et al. 1999).

Using fMRI sparse scanning, Davis and Johnsrude (2003) presented listeners with sentences that had three kinds of acoustic distortions (vocoded, interrupted, and energetically masked speech) applied to a varying degree, thus creating different levels of intelligibility. Whereas areas close to primary auditory cortex bilaterally were differentially activated for each of the acoustic distortion types – suggesting a kind of sound-form-based processing – the authors found several areas that were invariant to the acoustic distortions but sensitive to overall intelligibility, including the left IFG, hippocampus, and portions of the middle and superior temporal gyri. One explanation of these results, which supports a hierarchical view of speech processing, is that these acoustically invariant areas may modulate attention in service of understanding speech in adverse conditions. This hierarchy, however, does not necessarily imply "top-down" effects from frontal areas on auditory cortex; in fact, the timing of activation may be more parsimoniously understood as reflecting a feedforward process extending from auditory areas to a more distributed frontal and temporal network.

These findings highlight the importance of separating processing of the acoustic properties of distorted speech, from processing of intelligible speech. But speech may vary in intelligibility and comprehensibility in very different ways. As discussed previously, speech may be difficult to understand because it is masked by noise that competes with the speech signal at the level of the auditory nerve (energetic masking); because the accent of the talker is different from the listener (accented speech); or because there is a competing talker whose speech may be confusable with the target talker (informational masking). If, at the same time, speech is challenging to understand at a linguistic level because it, for example, incorporates words with multiple meanings, or complex syntactic structures, these further add to the demand on cognitive resources. Given the differences in perceptual and cognitive processing required to successfully accommodate all these challenges, it is reasonable to expect differential neural involvement (Scott and McGettigan 2013).

Energetic masking has been associated with the broad recruitment of frontal and parietal regions, including the IFG, frontal operculum, and angular gyrus (Adank et al. 2012). Moreover, individual differences in cognition modulate the degree to which contextual cues benefit speech-in-noise perception, which is associated with differential activation in IFG and angular gyrus (Zekveld et al. 2012). Taken together, these findings suggest that perceiving speech in noise involves an interaction between auditory and frontoparietal areas, with factors such as context and individual differences in frontally mediated executive functions influencing the way in which these areas interact.

Informational masking, on the other hand, most prominently appears to recruit superior temporal areas (Mattys et al. 2012). This pattern of activity largely overlaps with the areas that are involved in processing clear speech without a masker, which makes sense given the similarity of the masker to the target. However, more extended activation has also been observed in situations where the masking speech is highly similar to the target speech (Nakai et al. 2005), including dorsolateral prefrontal cortex, anterior cingulate, and premotor areas. This in turn suggests that nonauditory regions, thought to underlie executive functions such as cognitive control, may be recruited depending on the perceived challenge of the adverse listening situation, above and beyond its acoustic and linguistic factors. This will be discussed in more detail in Sect. 6.4.3.

A distributed network of brain regions appears to be involved in accommodating accented speech. The regions involved may look more varied than they actually are because of the methodological difficulties in equating acoustic factors and comprehension difficulty across different participant samples. Put another way, there are many ways to operationalize accented speech, and these might pose different kinds of challenges to listeners depending on the particular accent of the participant sample. With this caveat in mind, the neural areas implicated in accommodating accented speech partially overlap with areas implicated in both energetic and informational masking (see Adank et al. 2015). Similar to informational masking, listening to accented speech results in greater activation of bilateral superior temporal areas (Adank et al. 2012), presumably due to greater auditory and phonological processing demands. Accented speech also engages regions around the supplementary motor area (SMA), left IFG, and frontal operculum, which has at least two possible explanations, depending on the precise regions involved. One is that a network supporting cognitive control (the cinguloopercular network) has been activated due to the perceived difficulty of the task. This possibility will be discussed in Sect. 6.4.3. The other explanation is that, given the possible overlap between these areas and motor speech regions, listeners may recruit a speech motor network to simulate the production of the accented speech, direct attention to the most diagnostic features for successful recognition, and inhibit representations that may conflict with the auditory input (cf. Banks et al. 2015). This will be discussed further in Sect. 6.4.2.

6.4.1 Listening to Speech While Doing Something Else

When the sensory information at the ear is too ambiguous to support speech recognition by itself, knowledge-guided processes that help to interpret and repair the degraded signal are required. Many of these processes appear to be effortful and may not be recruited when attention is elsewhere. For example, imagine conversing with a friend at a hockey game. There are several potential energetic maskers (e.g., the synchronous roar of the crowd when a goal is scored) and informational maskers (e.g., the nearby conversations taking place), making speech comprehension more difficult and presumably effortful. Now, in this environment, imagine that, in the middle of the friend telling a story, one's attention is captured by the action of the hockey game. How well would the friend's speech be perceived? This is hard to study behaviorally, since it is difficult to measure perception of a stimulus to which a participant is not attending. Wild and colleagues (2012) used fMRI to compare processing of speech under full attention and under distraction. On every trial, young adult listeners with normal hearing attended to one of three simultaneously presented stimuli: an everyday, meaningful sentence (at one of four acoustic clarity levels), an auditory distracter, or a visual distracter. A post-scan recognition test showed that clear speech was processed even when not attended, but that attention greatly enhanced the processing of degraded speech. Furthermore, speech-sensitive cortex could be fractionated according to how speech-evoked responses were modulated by attention, and these divisions appeared to map onto the hierarchical organization of the auditory system, as discussed in Sect. 6.2.3. Only in middle temporal and frontal regions – regions corresponding to the highest stages of auditory processing – did activity appear to be enhanced by attention.

In a follow-up experiment, Ritz et al. (2016) pushed the paradigm, increasing the intelligibility of the degraded speech so that all words from sentences could be reported correctly when the sentences were attended (through the use of 12-band noise vocoding, referred to as NV12), and introducing a multiple object tracking (MOT) task with a parametrically varying number of moving dots to track (1, 3, 4, or 6). Both types of stimuli were presented on every trial, and the participant was cued at the beginning of each trial to attend to one or to the other. The results were striking and are shown in Fig. 6.5b. In anterior temporal cortex (yellow/orange

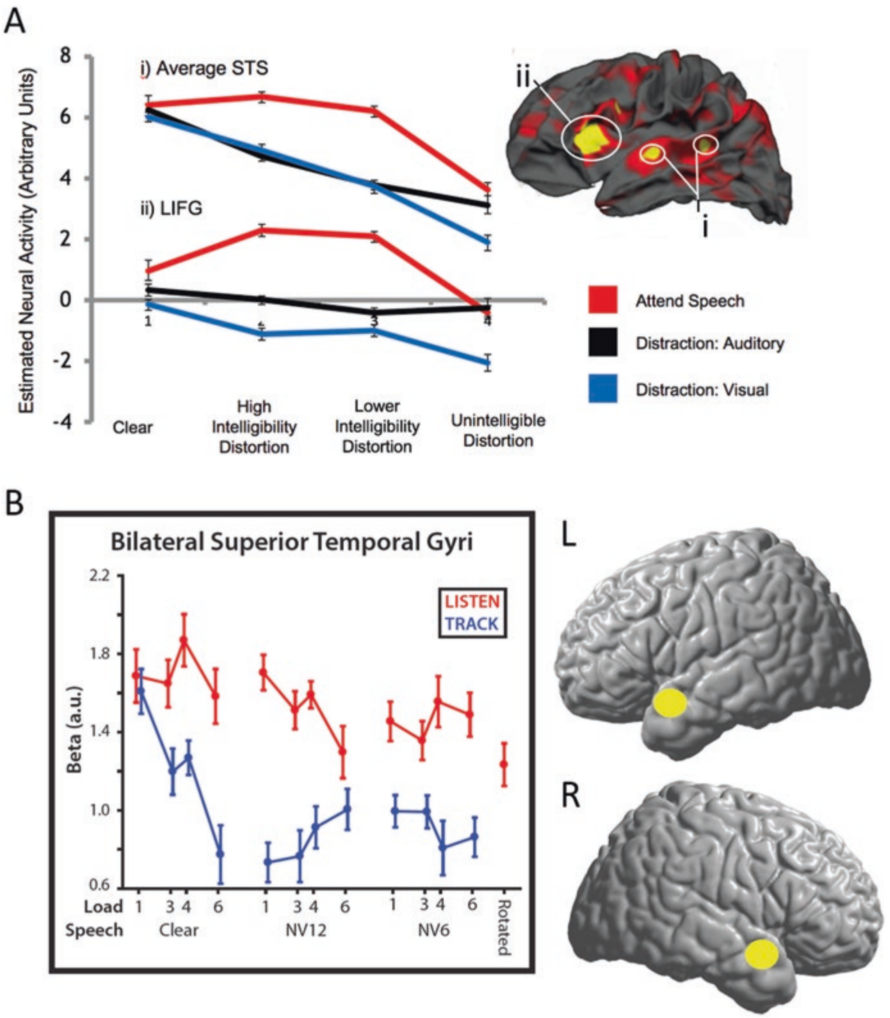

Fig. 6.5 (**a**) Activity in the left superior temporal sulcus (STS; i) and in left inferior frontal gyrus (IFG; ii) depends on attentional state and speech quality. In both regions, activity is enhanced when listeners attended to speech compared to when they attended to concurrently presented visual or auditory distractor, performing a target-detection task on these (Wild et al. 2012). Activity is particularly enhanced for degraded, but intelligible, distorted speech (the distortion was created through noise vocoding) (Shannon et al. 1995). When listening with no distractors, a pilot group could report 90% of the words from high-intelligibility distorted sentences, and 70% of the words from low-intelligibility distorted sentences. (Adapted from Wild et al. (2012)). (**b**) Activity in bilateral anterior temporal regions (shown schematically in yellow) depended on attentional state and speech quality in an unpublished study (Ritz et al., MSc thesis). As indicated by the red lines, activity was high when listeners were attending to speech, regardless of whether it was clear, very high-intelligibility 12-band noise-vocoded speech (NV12), or lower intelligibility 6-band NV speech (NV6). It was lower but still elevated when attending to "rotated speech" – this is completely unintelligible noise-vocoded speech. When listening with no distractors, a pilot group

(continued)

clusters in Fig. 6.5b), activity when attending to speech was uniformly quite high, whereas when the MOT task was attended, it was high only for clear speech, and only when one object was being tracked. The higher the MOT load, the lower the activity in this area. For degraded speech that was 100% intelligible (NV12 in Fig. 6.5b), a marked effect of attentional state was evident – even at the lowest MOT load (one object), activity when attending to speech was much higher than when attending to the MOT task. These results suggest that whereas these anterior temporal regions can process clear speech in the absence of attention, as long as the distractor task is not too demanding, processes involved in the comprehension of even lightly degraded speech critically require focused attention.

In a series of studies, Mattys and colleagues explored how additional concurrent processing load alters the processing of simultaneously presented spoken words. They found that processing speech under conditions of divided attention relies on different mechanisms compared to those involved in processing speech when attention is focused solely on speech. When listeners were required to listen to speech and perform a visual search task, they reweighted information in making perceptual decisions (Mattys et al. 2014). Moreover, they seemed to rely more on lexical semantic information for word segmentation, and on lexical knowledge for phoneme identification, than they would without a concurrent task. In contrast, they seemed to rely less on acoustic cues conveyed in fine phonetic detail. It may seem counterintuitive that, as load on central cognitive resources increases, listeners rely more, not less, on knowledge-guided factors (which presumably rely on the same central cognitive resources) for speech perception. This reweighting of cues may be due to poorer registration of the fine phonetic detail when distracted (Mattys and Palmer 2015). These studies are important because they indicate that attentional manipulations do not simply impair perception but instead qualitatively change perceptual decision criteria.

6.4.2 The Importance of Motor Representations

In a chapter focusing on speech perception in adverse listening conditions, it may seem initially inappropriate to devote a section to speech-motor representations. Yet, as briefly mentioned in the introduction of Sect. 6.4, certain kinds of adverse listening conditions (such as accented speech perception) have been associated with

could report 100% of the words from NV12 sentences, and 94% of the words from NV6 sentences. When sentences were heard while listeners focused on a distracting multiple object tracking (MOT) task (see text for details), activity was low even at the lowest level of MOT load (1 dot) when speech was even slightly degraded. When speech was clear and the MOT load was low (1 dot to track), there was no effect of attention in this area, and activity declined to the levels seen for degraded speech as tracking load increased. The y-axis is dimensionless beta weights (arbitrary units). On the x-axis, “load” (values 1–6) is the number of dots tracked during a concurrent MOT task. “Speech” is the speech type the activation of motor and premotor cortex, suggesting that the mechanisms underlying the planning and production of speech may improve speech perception at least in some listening situations.

The broader discussion of how motor representations specifically relate to speech perception has a long history. In its strongest form, the motor theory of speech perception asserts that speech is not understood through the perception of its auditory components but rather through more abstract and invariant articulatory gestures (Liberman and Mattingly 1985). Although the motor theory of speech perception has been the subject of considerable debate (e.g., see Lotto et al. 2009), an increasing body of research has supported at least some motor involvement in the perception of speech, particularly in adverse listening conditions. Researchers have used transcranial magnetic stimulation (TMS) to alter the excitability of motor cortex and have demonstrated enhancement of motor-evoked potentials (MEPs) from lip and tongue muscles when listening to speech (Fadiga et al. 2002). These studies were conducted using clear speech, but subsequent work demonstrates that motor activation may contribute to categorical speech perception under adverse listening conditions. In an fMRI study, Du et al. (2014) asked participants to identify phoneme tokens presented at different SNRs. Activity correlated negatively with perceptual accuracy in left ventral premotor cortex and a region anterior to it (anatomically defined Broca's area). Furthermore, pattern-information analysis revealed that whereas phonemes could not be reliably discriminated in patterns of activity in bilateral auditory cortex except when the noise level was very low, representations of phonemes remained robust in ventral premotor and Broca's areas at much higher levels of noise. This suggests a role for motor regions in categorical perception of degraded speech sounds.

The involvement of motor representations in speech perception appears to depend on attention. Using TMS to temporarily disrupt motor areas associated with lip movements, Möttönen et al. (2014) demonstrated that auditory representations of lip- and tongue-articulated speech sounds (/ba/, /da/, and /ga/) were differentially modulated based on attention. When the sounds were attended to, the TMS-related modulation in auditory cortex was relatively early and strongly left lateralized; when the sounds were not attended to, the modulation in auditory cortex was later and not lateralized. These results thus support the hypothesis that motor cortex can influence the response properties of auditory cortex in the context of speech perception, but the precise interaction between these areas may critically depend on attention.

6.4.3 The Cinguloopercular Network

For nearly 20 years, it has been clear that several distinct tasks recruit a common network involving dorsolateral prefrontal cortex and anterior insula, dorsal anterior cingulate cortex, and the adjacent pre-supplementary motor area (Duncan 2010). This cinguloopercular (CO) network appears to become active whenever cognitive demands are high, consistent with proposals that it is involved in cognitive control, specifically in performance monitoring (Dosenbach et al. 2006). This network appears to be recruited whenever a listener is attempting to understand speech that is challenging, either because the speech has been degraded or because a linguistic challenge, such as semantic ambiguity, has been imposed (Davis and Johnsrude 2003; Rodd et al. 2005). This elevated CO response, however, does not simply reflect challenge. Vaden et al. (2013) demonstrated that CO activity predicted word recognition on the next trial, which is similar to what has been noted in visuospatial tasks. The pattern of results suggests that the CO network is important for *adaptive* cognitive control. Furthermore, the results of Wild et al. (2012) indicate that this adaptive control may require focused attention on the difficult-to-understand speech signal.

6.5 Listening Effort

It has become increasingly evident to hearing-aid manufacturers and auditory researchers that "effortful listening" is an essential construct to consider. Two people might comprehend the same amount of speech in a given challenging listening situation, but one listener may feel that it was effortful and tiring, whereas another listener might have found it effortless. The first listener may alter their behavior to avoid such situations, or, if listening through a hearing prosthesis such as a hearing aid, may choose not to use it. Thus, the concept of "listening effort" may be a powerful predictor of behavior, independent of comprehension.

"Listening effort" is typically considered to be a unitary phenomenon and is studied as such. However, it has at least two different meanings. On the one hand, researchers write about listeners exerting effort. For example, Pichora-Fuller et al. (2016) define mental effort as the "deliberate allocation of mental resources to overcome obstacles in goal pursuit…" (p. 10S). In this sense, it is a process or brain activity. At the same time, listeners are aware of processing being fatiguing or effortful. In this sense, listening effort is a percept. Typically, listening effort is measuring using questionnaires – such subjective measures are focused on the explicit percept (e.g., Johnson et al. 2015). Physiological measures such as pupillometry and imaging (fMRI or EEG) have become more common tools (Peelle 2018). These may be sensitive either to mental exertion or the perception of difficulty or both; it is not presently clear.

As something that listeners perceive, listening effort may be most productively considered as an interaction between the perceptual, linguistic, or task challenges imposed by a listening situation and the cognitive resources that the listener brings to bear. Individual differences in cognitive resources (such as memory, perceptual learning, processing speed, fluid intelligence, and control processes) that permit one person to cope more efficiently or more successfully than another with the challenges imposed by a listening situation will have a strong influence on perceived effort (see Fig. 6.6). Although listening effort is usually measured in a

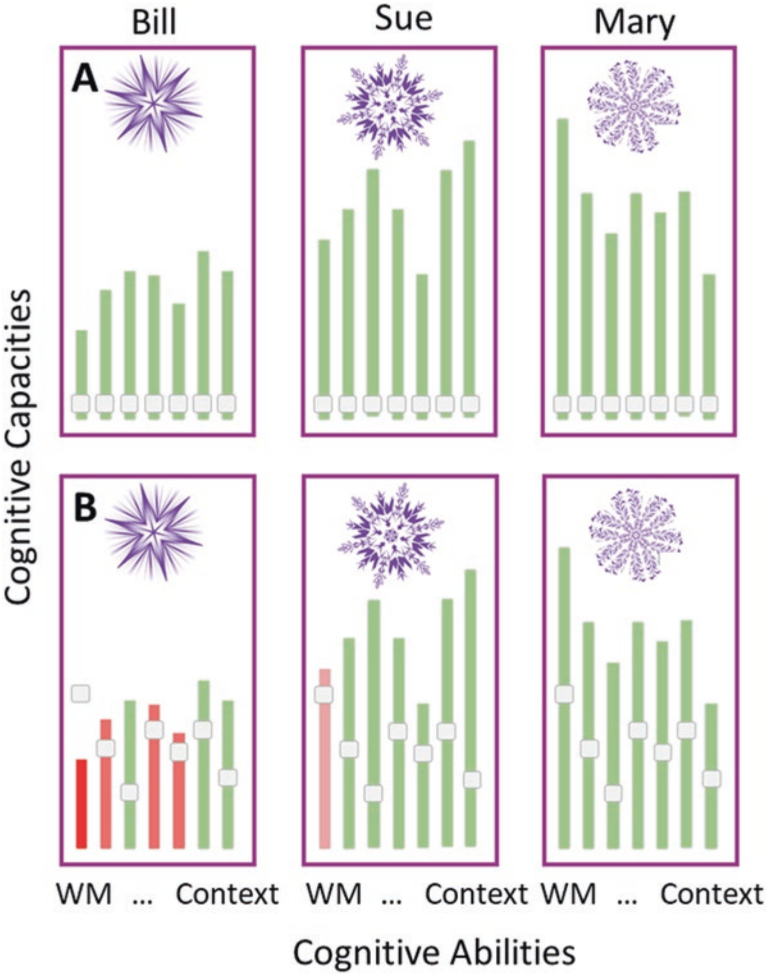

Fig. 6.6 (**a**) Three different individuals (represented by the unique purple snowflakes) and their distinct cognitive profile across seven putative abilities that are all relevant to speech perception in adverse conditions. Each bar is meant to represent an ability associated with speech perception, and the height of the bar indicates the strength of the ability. For example, the leftmost bar in each plot could be indexing working memory. (**b**) The seven white squares in each panel illustrate the cognitive demands imposed by a given listening situation. Note that the cognitive demands are the same across individuals. However, the degree to which each listener can respond to those demands depends on their individual cognitive profile. Demands fully occupy or outstrip several of the cognitive abilities for the listener on the left (highlighted in red). In contrast, the abilities of the listener on the right are more than adequate to cope with the demands – none of the squares are near the top of the ability bars (highlighted in green). The listener on the left will perceive effort (unless they give up), whereas the one on the left will find the listening situation effortless. This figure demonstrates how effort results from the interaction between the demands of a given listening situation (the position of the sliders) and an individual's cognitive abilities

unidimensional way (on a subjective questionnaire, or with pupillometry), it is probably not a unidimensional construct – different challenges are met in different ways. This framework enables researchers to cognitively and anatomically separate different processes, related to signal extraction, recovery, and repair that may contribute to the feeling of listening effort. At the same time, researchers can study factors that may alleviate listening effort, such as familiarity with someone's voice, or flexible and accurate use of meaningful context.

6.6 Chapter Summary

Speech is a complex and highly variable signal. Aspects of the speech signal itself (unfamiliar accents, semantic ambiguity, syntactic complexity), background signals (sound that either energetically or informationally masks a target speech signal), and listener-specific factors (selective attention and cognitive control abilities, familiarity with specific talkers or linguistic contexts) all contribute to how a given listening situation poses a challenge to recognition. Successful recognition of speech in adverse listening conditions therefore relies on interacting perceptual, cognitive, and linguistic factors. Some of these factors may be influenced considerably by learning, as seen with improved speech recognition for highly familiar talkers. Other factors, however, appear less susceptible to training, as seen with the mixed evidence of working memory training transferring to speech-in-noise perception. Although different adverse conditions place differential demands on cognitive resources, a consistent finding – supported behaviorally and neurally – is that adverse listening conditions place considerable demands on attention. Thus, compared to relatively clear listening condition, adverse listening conditions are served by the recruitment of additional brain networks – such as the CO network – even when both kinds of speech are equally intelligible. In this sense, the CO network may be viewed similarly to an "engine light" of a car, signaling an increase in mental effort but not specifically diagnosing the nature of the particular listening challenges in the moment. The emergence of "listening effort" as a construct, which represents the interaction between listening demands, and individual capacity across cognitive domains, may provide an important framework going forward for discussing speech perception in adverse listening conditions. Although the best operationalization of listening effort is still unclear and likely depends on the research question being addressed, it is clear that both listener-focused and signal-focused variables must be considered to fully understand speech perception in adverse listening conditions.

Compliance with Ethics Requirements Ingrid Johnsrude declares that she has no conflict of interest.

Stephen Van Hedger declares that he has no conflict of interest.

References

Adank P, Evans BG, Stuart-Smith J, Scott SK (2009) Comprehension of familiar and unfamiliar native accents under adverse listening conditions. J Exp Psychol Hum Percept Perform 35:520–529. https://doi.org/10.1037/a0013552

Adank P, Davis MH, Hagoort P (2012) Neural dissociation in processing noise and accent in spoken language comprehension. Neuropsychologia 50:77–84. https://doi.org/10.1016/j.neuropsychologia.2011.10.024

Adank P, Nuttall HE, Banks B, Kennedy-Higgins D (2015) Neural bases of accented speech perception. Front Hum Neurosci 9:1–7. https://doi.org/10.3389/fnhum.2015.00558

ANSI. (2013). *American National Standard Acoustical Terminology, ANSI S1.1-2013*. New York: American National Standards Institute.

Baddeley AD, Hitch G (1974) Working memory. Psychol Learn Motiv 8:47–89. https://doi.org/10.1016/S0079-7421(08)60452-1

Banks B, Gowen E, Munro KJ, Adank P (2015) Cognitive predictors of perceptual adaptation to accented speech. J Acoust Soc Am 137:2015–2024. https://doi.org/10.1121/1.4916265

Bates E, Wilson SM, Saygin AP et al (2003) Voxel-based lesion–symptom mapping. Nat Neurosci 6:448–450. https://doi.org/10.1038/nn1050

Binder JR, Desai RH, Graves WW, Conant LL (2009) Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. Cereb Cortex 19:2767–2796. https://doi.org/10.1093/cercor/bhp055

Bolia RS, Nelson WT, Ericson MA, Simpson BD (2000) A speech corpus for multitalker communications research. J Acoust Soc Am 107:1065–1066. https://doi.org/10.1121/1.428288

Bradlow AR, Pisoni DB, Akahane-Yamada R, Tohkura Y (1997) Training Japanese listeners to identify English / r / and / l /: IV. Some effects of perceptual learning on speech production. J Acoust Soc Am 101:2299–2310. https://doi.org/10.1121/1.418276

Brungart DS, Simpson BD, Ericson MA, Scott KR (2001) Informational and energetic masking effects in the perception of multiple simultaneous talkers. J Acoust Soc Am 110:2527–2538. https://doi.org/10.1121/1.1408946

Clarke CM, Garrett MF (2004) Rapid adaptation to foreign-accented English. J Acoust Soc Am 116:3647–3658. https://doi.org/10.1121/1.1815131

Collin B, Lavandier M (2013) Binaural speech intelligibility in rooms with variations in spatial location of sources and modulation depth of noise interferers. J Acoust Soc Am 134:1146–1159. https://doi.org/10.1121/1.4812248

Culling JF, Mansell ER (2013) Speech intelligibility among modulated and spatially distributed noise sources. J Acoust Soc Am 133:2254–2261. https://doi.org/10.1121/1.4794384

Culling JF, Stone MA (2017) Energetic masking and masking release. In: Middlebrooks J, Simon J, Popper A, Fay R (eds) The auditory system at the cocktail party. Springer handbook of auditory research, vol 60. Springer, Cham. https://doi.org/10.1007/978-3-319-51662-2_3

Cutler A, Norris D (1988) The role of strong syllables in segmentation for lexical access. J Exp Psychol Hum Percept Perform 14:113–121. https://doi.org/10.1037/0096-1523.14.1.113

Darwin CJ, Carlyon RP (1995) Auditory grouping. In: Moore BCJ (ed) The handbook of perception and cognition, vol 6, Hearing, 2nd edn. Academic Press, San Diego, pp 387–424

Davis MH, Johnsrude IS (2003) Hierarchical processing in spoken language comprehension. J Neurosci 23:3423–3431. https://doi.org/10.1523/JNEUROSCI.23-08-03423.2003

Davis MH, Johnsrude IS (2007) Hearing speech sounds: top-down influences on the interface between audition and speech perception. Hear Res 229:132–147. https://doi.org/10.1016/j.heares.2007.01.014

Denes PB, Pinson EN (1993) The speech chain: the physics and biology of spoken language. W.H. Freeman, New York

Dosenbach NUF, Visscher KM, Palmer ED et al (2006) A core system for the implementation of task sets. Neuron 50:799–812. https://doi.org/10.1016/j.neuron.2006.04.031

Dronkers NF, Wilkins DP, Van Valin RD et al (2004) Lesion analysis of the brain areas involved in language comprehension. Cognition 92:145–177. https://doi.org/10.1016/j.cognition.2003.11.002

Du Y, Buchsbaum BR, Grady CL, Alain C (2014) Noise differentially impacts phoneme representations in the auditory and speech motor systems. Proc Natl Acad Sci 111:7126–7131. https://doi.org/10.1073/pnas.1318738111

Duncan J (2010) The multiple-demand (MD) system of the primate brain: mental programs for intelligent behaviour. Trends Cogn Sci 14:172–179. https://doi.org/10.1016/j.tics.2010.01.004

Edmonds BA, Culling JF (2006) The spatial unmasking of speech: evidence for better-ear listening. J Acoust Soc Am 120:1539–1545. https://doi.org/10.1121/1.2228573

Fadiga L, Craighero L, Buccino G, Rizzolatti G (2002) Speech listening specifically modulates the excitability of tongue muscles: a TMS study. Eur J Neurosci 15:399–402. https://doi.org/10.1046/j.0953-816x.2001.01874.x

Floccia C, Goslin J, Girard F, Konopczynski G (2006) Does a regional accent perturb speech processing? J Exp Psychol Hum Percept Perform 32:1276–1293. https://doi.org/10.1037/0096-1523.32.5.1276

Giraud AL, Lorenzi C, Ashburner J et al (2000) Representation of the temporal envelope of sounds in the human brain. J Neurophysiol 84:1588–1598. https://doi.org/10.1152/jn.2000.84.3.1588

Guion SG, Harada T, Clark JJ (2004) Early and late Spanish–English bilinguals' acquisition of English word stress patterns. Biling (Camb Engl) 7:207–226. https://doi.org/10.1017/S1366728904001592

Hackett TA (2011) Information flow in the auditory cortical network. Hear Res 271:133–146. https://doi.org/10.1016/j.heares.2010.01.011

Hackett TA, de la Mothe LA, Camalier CR et al (2014) Feedforward and feedback projections of caudal belt and parabelt areas of auditory cortex: refining the hierarchical model. Front Neurosci. https://doi.org/10.3389/fnins.2014.00072

Hall DA, Haggard MP, Akeroyd MA et al (1999) "Sparse" temporal sampling in auditory fMRI. Hum Brain Mapp 7:213–223. https://doi.org/10.1002/(SICI)1097-0193(1999)7:3<213::AID-HBM5>3.0.CO;2-N

Hawkins S (2003) Roles and representations of systematic fine phonetic detail in speech understanding. J Phon 31:373–405. https://doi.org/10.1016/j.wocn.2003.09.006

Hickok G, Poeppel D (2015) Neural basis of speech perception. In: Aminoff MJ, Boller F, Swaab DF (eds) Handbook of clinical neurology, 129th edn. Elsevier, pp 149–160

Holmes E, Domingo Y, Johnsrude IS (2018) Familiar voices are more intelligible, even if they are not recognized as familiar. Psychol Sci 29:1575–1583. https://doi.org/10.1177/0956797618779083

Holt L (2005) Temporally nonadjacent nonlinguistic sounds affect speech categorization. Psychol Sci 16:305–312. https://doi.org/10.1111/j.0956-7976.2005.01532.x

Humes LE, Lee JH, Coughlin MP (2006) Auditory measures of selective and divided attention in young and older adults using single-talker competition. J Acoust Soc Am 120:2926–2937. https://doi.org/10.1121/1.2354070

Ingvalson EM, Dhar S, Wong PCM, Liu H (2015) Working memory training to improve speech perception in noise across languages. J Acoust Soc Am 137:3477–3486. https://doi.org/10.1121/1.4921601

Janse E, Adank P (2012) Predicting foreign-accent adaptation in older adults. Q J Exp Psychol 65:1563–1585. https://doi.org/10.1080/17470218.2012.658822

Johnson J, Xu J, Cox R, Pendergraft P (2015) A comparison of two methods for measuring listening effort as part of an audiologic test battery. Am J Audiol 24:419–431. https://doi.org/10.1044/2015_AJA-14-0058

Johnsrude IS, Mackey A, Hakyemez H et al (2013) Swinging at a cocktail party: voice familiarity aids speech perception in the presence of a competing voice. Psychol Sci 24:1995–2004. https://doi.org/10.1177/0956797613482467

Jones EG (2003) Chemically defined parallel pathways in the monkey auditory system. Ann N Y Acad Sci 999:218–233. https://doi.org/10.1196/annals.1284.033

Kidd G, Colbourn HS (2017) Informational masking in speech recognition. In: Middlebrooks J, Simon J, Popper A, Fay R (eds) The auditory system at the cocktail party, Springer handbook of auditory research, 60th edn. Springer International Publishing, Cham, pp 75–109

Kidd G, Mason CR, Best V, Marrone N (2010) Stimulus factors influencing spatial release from speech-on-speech masking. J Acoust Soc Am 128:1965–1978. https://doi.org/10.1121/1.3478781

Kraljic T, Brennan SE, Samuel AG (2008) Accommodating variation: dialects, idiolects, and speech processing. Cognition 107:54–81. https://doi.org/10.1016/j.cognition.2007.07.013

Liberman AM, Mattingly IG (1985) The motor theory of speech perception revised. Cognition 21:1–36. https://doi.org/10.1016/0010-0277(85)90021-6

Liberman AM, Cooper FS, Shankweiler DP, Studdert-Kennedy M (1967) Perception of the speech code. Psychol Rev 74:431–461. https://doi.org/10.1037/h0020279

Lim SJ, Holt LL (2011) Learning foreign sounds in an alien world: videogame training improves non-native speech categorization. Cogn Sci 35:1390–1405. https://doi.org/10.1111/j.1551-6709.2011.01192.x

Lotto AJ, Hickok GS, Holt LL (2009) Reflections on mirror neurons and speech perception. Trends Cogn Sci 13:110–114. https://doi.org/10.1016/j.tics.2008.11.008

Macleod A, Summerfield Q (1990) A procedure for measuring auditory and audiovisual speech-reception thresholds for sentences in noise: rationale, evaluation, and recommendations for use. Br J Audiol 24:29–43. https://doi.org/10.3109/03005369009077840

Mattys SL, Palmer SD (2015) Divided attention disrupts perceptual encoding during speech recognition. J Acoust Soc Am 137:1464–1472. https://doi.org/10.1121/1.4913507

Mattys SL, White L, Melhorn JF (2005) Integration of multiple speech segmentation cues: a hierarchical framework. J Exp Psychol Gen 134:477–500. https://doi.org/10.1037/0096-3445.134.4.477

Mattys SL, Davis MH, Bradlow AR, Scott SK (2012) Speech recognition in adverse conditions: a review. Lang Cogn Process 27:953–978. https://doi.org/10.1080/01690965.2012.705006

Mattys SL, Barden K, Samuel AG (2014) Extrinsic cognitive load impairs low-level speech perception. Psychon Bull Rev 21:748–754. https://doi.org/10.3758/s13423-013-0544-7

Mesulam MM, Wieneke C, Thompson C et al (2012) Quantitative classification of primary progressive aphasia at early and mild impairment stages. Brain 135:1537–1553. https://doi.org/10.1093/brain/aws080

Miller GA, Licklider JCR (1950) The intelligibility of interrupted speech. J Acoust Soc Am 22:167–173. https://doi.org/10.1017/S0031182000023970

Möttönen R, van de Ven GM, Watkins KE (2014) Attention fine-tunes auditory-motor processing of speech sounds. J Neurosci 34:4064–4069. https://doi.org/10.1523/JNEUROSCI.2214-13.2014

Nakai T, Kato C, Matsuo K (2005) An fMRI study to investigate auditory attention: a model of the cocktail party phenomenon. Magn Reson Med Sci 4:75–82. https://doi.org/10.2463/mrms.4.75

Norris D, Mcqueen JM, Cutler A, Butterfield S (1997) The possible-word constraint in the segmentation of continuous speech. Cogn Psychol 34:191–243. https://doi.org/10.1006/cogp.1997.0671

Nygaard LC, Pisoni DB (1998) Talker-specific learning in speech perception. Percept Psychophys 60:355–376. https://doi.org/10.3758/BF03206860

Okada K, Rong F, Venezia J et al (2010) Hierarchical organization of human auditory cortex: evidence from acoustic invariance in the response to intelligible speech. Cereb Cortex 20:2486–2495. https://doi.org/10.1093/cercor/bhp318

Peelle JE (2018) Listening effort: how the cognitive consequences of acoustic challenge are reflected in brain and behavior. Ear Hear 39:204–214. https://doi.org/10.1097/AUD.0000000000000494

Peelle JE, Johnsrude IS, Davis MH (2010) Hierarchical organization for speech in human auditory cortex and beyond. Front Hum Neurosci 4:1–3. https://doi.org/10.3389/fnhum.2010.00051

Phillips DP, Farmer ME (1990) Acquired word deafness, and the temporal grain of sound representation in the primary auditory cortex. Behav Brain Res 40:85–94. https://doi.org/10.1016/0166-4328(90)90001-U

Pichora-Fuller MK, Kramer SE, Eckert MA et al (2016) Hearing impairment and cognitive energy. Ear Hear 37:5S–27S. https://doi.org/10.1097/AUD.0000000000000312

Pisoni DB, Nusbaum HC, Greene BG (1985) Perception of synthetic speech generated by rule. Proc IEEE 73:1665–1676. https://doi.org/10.1109/PROC.1985.13346

Poremba A, Mishkin M (2007) Exploring the extent and function of higher-order auditory cortex in rhesus monkeys. Hear Res 229:14–23. https://doi.org/10.1016/j.heares.2007.01.003

Reisberg D, McLean J, Goldfield A (1987) Easy to hear but hard to understand: a lip-reading advantage with intact auditory stimuli. In: Dodd B, Campbell R (eds) Hearing by eye: the psychology of lip-reading. Lawrence Erlbaum Associates, Inc., Hillsdale, pp 97–113

Ritz H, Wild C, Johnsrude IJ (2016) The effects of concurrent cognitive load on the processing of clear and degraded speech. In: 22nd annual meeting of the Organization for Human Brain Mapping

Rodd JM, Gaskell G, Marslen-Wilson W (2002) Making sense of semantic ambiguity: semantic competition in lexical access. J Mem Lang 46:245–266. https://doi.org/10.1006/jmla.2001.2810

Rodd JM, Davis MH, Johnsrude IS (2005) The neural mechanisms of speech comprehension: fMRI studies of semantic ambiguity. Cereb Cortex 15:1261–1269. https://doi.org/10.1093/cercor/bhi009

Rodd JM, Johnsrude IS, Davis MH (2012) Dissociating frontotemporal contributions to semantic ambiguity resolution in spoken sentences. Cereb Cortex 22:1761–1773. https://doi.org/10.1093/cercor/bhr252

Scott SK, McGettigan C (2013) The neural processing of masked speech. Hear Res 303:58–66. https://doi.org/10.1016/j.heares.2013.05.001

Shannon RV, Zeng FG, Kamath V et al (1995) Speech recognition with primarily temporal cues. Science 270:303–304. https://doi.org/10.1126/science.270.5234.303

Srinivasan S, Keil A, Stratis K et al (2014) Interaural attention modulates outer hair cell function. Eur J Neurosci 40:3785–3792. https://doi.org/10.1111/ejn.12746

Sumby WH, Pollack I (1954) Visual contribution to speech intelligibility in noise. J Acoust Soc Am 26:212–215. https://doi.org/10.1121/1.1907309

Turken AU, Dronkers NF (2011) The neural architecture of the language comprehension network: converging evidence from lesion and connectivity analyses. Front Syst Neurosci 5:1–20. https://doi.org/10.3389/fnsys.2011.00001

Vaden KI, Kuchinsky SE, Cute SL et al (2013) The cingulo-opercular network provides word-recognition benefit. J Neurosci 33:18979–18986. https://doi.org/10.1523/JNEUROSCI.1417-13.2013

Van Engen KJ, Peelle JE (2014) Listening effort and accented speech. Front Hum Neurosci 8:1–4. https://doi.org/10.3389/fnhum.2014.00577

Wayne RV, Johnsrude IS (2015) A review of causal mechanisms underlying the link between age-related hearing loss and cognitive decline. Ageing Res Rev 23:154–166. https://doi.org/10.1016/j.arr.2015.06.002

Wayne RV, Hamilton C, Huyck JJ, Johnsrude IS (2016) Working memory training and speech in noise comprehension in older adults. Front Aging Neurosci 8:1–15. https://doi.org/10.3389/fnagi.2016.00049

Wild CJ, Yusuf A, Wilson DE et al (2012) Effortful listening: the processing of degraded speech depends critically on attention. J Neurosci 32:14010–14021. https://doi.org/10.1523/JNEUROSCI.1528-12.2012

Zekveld AA, Rudner M, Johnsrude IS et al (2012) Behavioral and fMRI evidence that cognitive ability modulates the effect of semantic context on speech intelligibility. Brain Lang 122:103–113. https://doi.org/10.1016/j.bandl.2012.05.006

Zekveld AA, Rudner M, Johnsrude IS, Rönnberg J (2013) The effects of working memory capacity and semantic cues on the intelligibility of speech in noise. J Acoust Soc Am 134:2225–2234. https://doi.org/10.1121/1.4817926

Chapter 7
Adaptive Plasticity in Perceiving Speech Sounds

Shruti Ullas, Milene Bonte, Elia Formisano, and Jean Vroomen

Abstract Listeners can rely on perceptual learning and recalibration in order to make reliable interpretations during speech perception. Lexical and audiovisual (or speech-read) information can disambiguate the incoming auditory signal when it is unclear, due to speaker-related characteristics, such as an unfamiliar accent, or due to environmental factors, such as noise. With experience, listeners can learn to adjust boundaries between phoneme categories as a means of adaptation to such inconsistencies. Recalibration experiments tend to use a targeted approach by embedding ambiguous phonemes into speech or speechlike items, and with continuous exposure, a learning effect can be induced in listeners, wherein disambiguating contextual information shifts the perceived identity of the same ambiguous sound. The following chapter will review current and past literature regarding lexical and audiovisual influences on phoneme boundary recalibration, as well as theories and neuroimaging data that potentially reveal what facilitates this perceptual plasticity.

Keywords Recalibration · Perceptual learning · Speech perception · Phonetic processing · Lexical processing · Audiovisual speech · Speech-reading

7.1 Introduction

Speech perception is seemingly easy and automatic to the listener, and for healthy young listeners, it requires little to no effort to accomplish in most circumstances. While it may appear straightforward, a great deal of variability exists in the quality

S. Ullas (✉) · M. Bonte · E. Formisano
Department of Cognitive Neuroscience, Faculty of Psychology and Neuroscience,
Maastricht University, Maastricht, The Netherlands

Maastricht Brain Imaging Center, Maastricht University, Maastricht, The Netherlands
e-mail: shruti.ullas@maastrichtuniversity.nl; m.bonte@maastrichtuniversity.nl;
e.formisano@maastrichtuniversity.nl

J. Vroomen
Department of Cognitive Neuropsychology, Tilburg University, Tilburg, The Netherlands
e-mail: j.vroomen@tilburguniversity.edu

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022

L. L. Holt et al. (eds.), *Speech Perception*, Springer Handbook of Auditory
Research 74, https://doi.org/10.1007/978-3-030-81542-4_7

of the speech signal, which requires the listener to adapt to the novel characteristics of the encountered speech. The acoustic signal can differ significantly across speakers, often due to unfamiliar accents, the presence of noise, or speech rate. The listener is able to easily resolve these inconsistencies and understand what is spoken. No two speakers will pronounce a phoneme in the exact same way, and even the same speaker may not produce a phoneme identically across multiple instances, yet listeners are effortlessly able to recognize what speakers are saying. Auditory quality can also vary within speakers, perhaps due to a cold or while speaking over the phone. Still, the listener is usually able to easily resolve these inconsistencies and understand what is spoken. In order to adapt to these irregularities, listeners can learn to reshape existing representations of speech sounds and categories to accommodate any possible variability.

Acoustics are not the only source of information capable of changing speech sound representations, as other contextual cues are also highly influential. Contextual features may be just as useful as auditory information, and possibly even more so. Winn (2018) introduces some non-acoustic cues that impact what listeners perceive to hear, including visual cues, such as the lip movements of a speaker, as well as the listener's own lexical knowledge. These non-acoustic sources can also enable processes known as recalibration and lexically guided perceptual learning. Contextual information can guide the retuning process of phoneme category boundaries, after continuous exposure to speech or videos of speechlike tokens, edited to contain ambiguous versions of a phoneme. Listeners can learn to incorporate these ambiguous sounds into the phoneme category itself, particularly when the sounds resemble already familiar phonemes.

Norris et al. (2003) termed this effect lexically guided perceptual learning, and observed that with the help of lexical knowledge, listeners could learn to adjust a perceptual boundary between two phonemes by hearing ambiguous phonemes embedded into words**.** Similarly, Bertelson et al. (2003) identified a comparable effect as recalibration, where listeners utilized visual or speech-reading information to adjust the perceptual boundary. The two discoveries were made close in time, and while Norris et al. (2003) used recordings of words as stimuli, Bertelson et al. (2003) relied on video recordings of syllables. Still, while the types of contextual information differed between the two studies, the experimental designs and stimuli constructions were remarkably similar. Since then, in the literature on lexical influences, the resulting aftereffect is often referred to as perceptual retuning or phoneme adaptation, while the studies on visual/speech-reading influences refer to the analogous effect as audiovisual recalibration.

In laboratory settings, recalibration and perceptual learning are typically measured in two phases, starting with an exposure phase and followed by a test phase (see Kraljic and Samuel 2009, for an overview). In the approach used to measure lexically guided perceptual learning, exposure stimuli are composed of audio recordings of words, whereas in audiovisual recalibration experiments, exposure stimuli comprise videos of a speaker's lip movements while pronouncing a syllable. Both types of stimuli contain edited audio, where one particular phoneme is replaced with an ambiguous sound halfway between two clear phonemes. For instance, speech stimuli containing /f/ sounds are replaced with a token halfway between /f/ and /s/. Listeners are presented with many examples of such edited stimuli in the exposure phase, with words such as "half" and "paragraph" edited to remove the clear /f/ and replaced with the ambiguous version. Because "half" and "paragraph" are real words in English, whereas "halss" and "paragrass" are not, listeners tend to perceive the ambiguous token as an /f/. During subsequent test phases, listeners hear the ambiguous sounds again, but without any lexical or visual context available, and respond with the phoneme they perceive to be hearing. Consequently, listeners become more likely to respond hearing the same phoneme that was replaced in the previously presented words or videos. In the case of the aforementioned example, the listener would now report hearing the ambiguous token as /f/ as well. This response pattern is understood to reflect recalibration or perceptual retuning, and is a result of the listeners learning to include the ambiguous sound as a part of that particular phoneme category.

Listeners in such experiments can also learn to perceive the same ambiguous phoneme, with no change in acoustic features, in opposing ways, depending on the bias of the surrounding context. A 50–50 /f/-/s/ blend can be learned as either /f/ or /s/ depending on the type of exposure the listener has undergone. Again, in the same example, if listeners were instead presented with speech stimuli that replaced all /s/ sounds with the same ambiguous token (the 50–50 blend of /f/ and /s/), listeners would be more likely to perceive the ambiguous sound as /s/ as well. With this approach, the contributions of visual and lexical information on speech perception can be disentangled from the auditory signal itself, as the exact same ambiguous tokens can be learned as different phonemes depending on the contextual cues. Perceptual retuning and recalibration studies (Bertelson et al. 2003; Norris et al. 2003; Kraljic and Samuel 2009) also reveal how flexible the units of speech are, and how they can be adapted depending on the surroundings of the listener. These experiments illuminate non-acoustic contributions to speech perception, and what listeners rely on in addition to the acoustic signal itself, which, again, tends to fluctuate greatly both within and across speakers.

With the advancement of neuroimaging technologies, the ways in which the brain incorporates these perceptual shifts have been explored with greater detail and have revealed the areas of the brain likely to be involved in these processes. Techniques such as functional MRI (fMRI; see Table 7.1 for abbreviations) and electrocorticography (ECoG) recordings have proven especially useful in elucidating the potential neural mechanisms (Hickok and Poeppel 2007; Mesgarani et al. 2014). These findings, combined with existing theories of speech perception, are useful for understanding how the brain adapts to unclear speech and how the necessary changes may be implemented at the neural level.

This chapter will present an overview of the current literature regarding lexical (Sect. 7.2.1) and audiovisual influences (Sect. 7.3.1) on phoneme boundary recalibration, as well as some related works on selective speech adaptation (Sect. 7.3.2). Changes over time (Sect. 7.2.2), generalization over speakers and sounds (Sects. 7.2.3 and 7.3.3), and other features (Sect. 7.2.4) will also be discussed, as well as a comparison between lexical and audiovisual perceptual learning (Sect. 7.4).

Table 7.1 Table of abbreviations

Abbreviation	Full name
ECoG	Electrocorticography
EEG	Electroencephalogram
fMRI	Functional MRI
IFS	Inferior frontal sulcus
IPL	Inferior parietal lobe
ITS	Inferior temporal sulcus
MEG	Magnetoencephalogram
MTG	Medial temporal gyrus
PT	Planum temporale
STG	Superior temporal gyrus
STS	Superior temporal sulcus
SWS	Sine-wave speech

Theories and neuroimaging studies that may explain the underlying mechanisms of recalibration will also be reviewed (Sect. 7.5), followed by a final conclusion and summary (Sect. 7.6).

7.2 Lexical Knowledge and Auditory Perception

7.2.1 Introduction to Lexically Guided Perceptual Learning

As mentioned earlier in the introduction (Sect. 7.1), top-down lexical knowledge can assist listeners in interpreting unclear speech. To investigate this, some researchers have used noise-vocoded or degraded speech stimuli that systematically distort frequency and amplitude components of the speech (Davis et al. 2005). Other researchers have studied how listeners adapt to accented speech (Clarke and Garrett 2004; Bradlow and Bent 2008), how listeners adapt to non-native speech in noise (Lecumberri et al. 2010), as well as how lexical knowledge supports understanding accented speech (Maye et al. 2008). A review by Holt and Lotto (2008) describes the various ways in which listeners can build links between acoustic information and linguistic representations. Prior to many of these studies, the discovery of what is now known as the Ganong effect (Ganong 1980) established a specific influence of lexical information on speech sound perception. Ganong (1980) showed that listeners were likely to report hearing words even when exposed to auditory stimuli that were edited to begin with ambiguous sounds. Listeners who heard the word "?eep," where the /?/ sound was acoustically halfway between /d/ and /t/, were likely to interpret the stimulus in the form of a word, such as "deep," rather than "teep." The same held true in the opposite direction, when the same ambiguous token replaced /t/ in recordings of words beginning with /t/, such as "?each." Again, listeners were likely to report hearing a word, such as "teach," rather than the nonword version, "deach." In essence, listeners were not hindered by the unclear auditory information and were still able to infer the intended words.

Extending further from the Ganong effect, the findings of Norris et al. (2003) revealed how lexical information could not only affect perception of speech stimuli but could also reshape speech sound representations. Native Dutch speakers performed a lexical decision task while listening to audio recordings of Dutch words, some of which typically ended in /f/, such as "witlo??" (*witlof*, meaning chicory) and "drui??" (*druif*, meaning grape), where all /f/ sounds were replaced with an ambiguous token halfway between /f/ and /s/. During the following test phase, where listeners responded to a continuum of sounds ranging from more /f/-like to more /s/-like, they were likely to report a significantly greater number of tokens as /f/ sounding. Another group of participants conducted the same lexical decision task while hearing words, but in contrast, these words typically contained /s/ (such as *radijs* and *relaas*, meaning radish and account) and were spliced with the same ambiguous token in the place of /s/, and the opposite pattern of results was found. These listeners responded to the same continuum of /f/ to /s/ sounds during the test phase, and were more likely to report hearing the sounds as /s/. A third control group heard pseudo-words containing the ambiguous phoneme to test whether the absence of any lexical information could impact subsequent categorization. This group showed no bias toward either phoneme during the test phase. An example of the pattern of results is shown in Fig. 7.1.

Together, these results built further upon the lexical effect first described by Ganong and illustrated how lexical knowledge impacted the participants' perception in two ways. First, during the exposure phase, the words containing the ambiguous sounds were still perceived as words and nearly indistinguishable from unedited words, and replicated the Ganong effect. Then, in the test phase, listeners categorized ambiguous sounds of a continuum and were prone to hearing the continuum sounds resembling the phoneme replaced in the prior exposure phase. That is, listeners were likely to perceive the ambiguous token as /f/ after exposure to f-final words containing the said token. Thus, phoneme category boundaries were found to be flexible, as listeners adjusted the boundary between two phonemes using their lexical knowledge. The authors proposed that the results mirrored what listeners may be doing in response to an unfamiliar accent, by shifting a category boundary to make room for the pronunciation of the newly encountered speaker (this will be discussed more in Sect. 7.2.3).

7.2.2 *Perceptual Learning Over Time*

Since Norris et al. (2003), later studies of perceptual learning explored the other attributes of this effect, such as the duration of time for which the retuning effects could last in the listener, as well as if these changes were permanent or if the categories returned to their previous state. Kraljic and Samuel (2005) used nearly the same

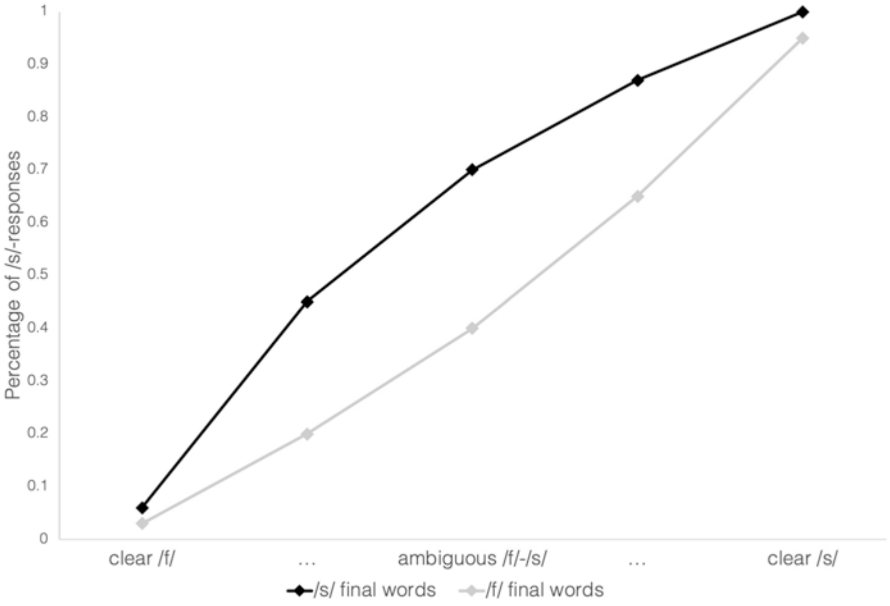

Fig. 7.1 Example graph of perceptual retuning results. After exposure to edited words, participants are presented with a continuum of sounds ranging from clear /f/ to clear /s/ in a test phase. Participants who hear words typically containing /f/ replaced with an ambiguous /f/-/s/ blend are likely to report hearing /f/ during the test phase (shown in gray), while participants who heard the same sound replacing /s/ in /s/-final words are likely to report hearing more /s (shown in black)

approach as Norris et al. (2003), testing native English speakers using words containing either /s/ or /ʃ/ (the "sh" sound in shoe), with items such as *eraser* and *publisher*. After a 25-minute delay, participants were tested on a continuum from /s/ to /ʃ/, and their responses reflected the shift induced by the preceding exposure phase (i.e., more /s/ responses after /s/-final words, or more /ʃ/ after /ʃ/-final words). Despite the delay, the listeners could still retain the newly learned phoneme boundary position. Eisner and McQueen (2006) also measured perceptual learning effects in subjects after a longer delay, where participants completed one test immediately after exposure, and also returned 12 hours after the exposure to complete the test phase again. The exposure phase was slightly altered from the original version by Norris et al. (2003) and consisted of words with ambiguous segments, all embedded into a short story. The potential confound of sleep was also accounted for, as one group waited 12 hours during the day to be retested, while another group waited 12 hours overnight, and returned for the second test phase after they had slept. Both groups still maintained retuning effects after the 12-hour delay, with or without sleeping. Perceptual learning is seemingly unaffected by long gaps between exposure and test, which suggests that lexically guided perceptual learning is largely stable over the order of hours.

7.2.3 Generalization of Perceptual Retuning

Although lexically driven perceptual learning appears to be quite robust, other investigators have identified the limitations of such learning. For example, perceptual learning tends to be restricted by the stimuli, particularly by the speakers of the tokens. The shift in perception resulting from experience with one phoneme pair by one speaker may not apply to the same pair produced by a new speaker. Eisner and McQueen (2005) had two groups of participants undergo exposure to Dutch words containing either an ambiguous /f/ or /s/ spoken by one speaker, but were tested on a continuum of /f/-/s/ sounds by a different speaker. Participants did not show the retuning effect when tested with the continuum by the novel speaker, so responses to the items on the continuum did not show a shift toward any particular phoneme. Thus, the authors concluded that the participants treated the sounds contained in the exposure stimuli as an idiosyncrasy, so it was tied specifically to the speaker of the ambiguous sounds and did not generalize to ambiguous sounds by a different speaker.

Kraljic and Samuel (2007) also addressed a possible discrepancy in generalization to new speakers based on phoneme types. Listeners who were exposed to words containing ambiguous /d/ or /t/ (plosives or stop consonants) sounds could generalize retuning to the same tokens of a new speaker during the test phase, translating to a shift in categorization responses toward the phoneme replaced in the prior exposure phase (i.e., more /d/ responses after exposure to /d/ words replaced with /d/-/t/ blend). However, those who were exposed to words spliced with ambiguous /s/ or /ʃ/ (fricatives) could not generalize any retuning to a new speaker, so no shift was found in categorization responses during the test phase. Evidently, perceptual learning may not always be constrained by the speaker, and depending on the type of phoneme pair used, it may also be token-specific.

Similarly, generalization to new speakers may also be dependent on the accent of the speaker. Kraljic et al. (2008a) compared effects of speaker characteristics on perceptual learning, with an idiosyncratic pronunciation versus an accent commonly known to the participants. The idiosyncrasy, or speaker-specific version, was designed by placing an ambiguous /s/-/ʃ/ sound before any consonants in the word stimuli, whereas the accented version only placed the ambiguous sound before an occurrence of /tr/ (such as /s/ in *string*), as is typical of some regional American accents. Phoneme boundary retuning was not successful in the latter group that was exposed to the tokens typical of the accented speech, but was detected in the non-accented group. Knowledge of reasonable and unrealistic deviations, which may be implicit or explicit, also seem to impact perceptual learning. In contrast, native English participants who heard exposure stimuli in English by a speaker with a Mandarin accent were more likely to generalize retuning to another acoustically similar Mandarin-accented speaker (Xie and Myers 2017), and to a lesser extent to speakers whose voices were acoustically more distant. The discrepancy in findings between Xie and Myers (2017) and Kraljic and Samuel (2008a) may once again reflect differences in learning effects due to the phoneme pair used.

Just as speaker specificity of perceptual learning is tied to the type of phoneme pairs, the same applies to generalization across phoneme pairs within a single speaker. Kraljic and Samuel (2006) saw that perceptual learning could generalize between pairs of plosives or stop consonants, particularly between /d/-/t/ and /b/-/p/. During the exposure phase, listeners heard words containing either an ambiguous /d/ or /t/, but during the test phase, they responded to both a /d/-/t/ continuum and a /b/-/p/ continuum. Participants were able to extend retuning to the /b/-/p/ continuum in the same direction of voicing, or the point in time at which the vocal folds vibrate, where /b/ and /d/ are voiced, whereas /d/ and /t/ are unvoiced. Participants who heard words with an ambiguous /b/ were more likely to report a greater amount of both /b/ along the /b/-/p/ continuum, as well as more /d/ during an additional test phase on a continuum of /d/-/t/. Mitterer et al. (2013) also explored phoneme specificity by creating exposure stimuli using Dutch words ending in an approximant /r/ (the /r/ in red) or dark /l/ (the /l/ in *pool*). Participants showed retuning effects during a test phase with a continuum of the versions of /r/ or /l/ they previously heard during exposure, but could not generalize to other allophones, or phonetic neighbors of /r/ and /l/, such as a trill /r/ (not in American English phonology but similar to the t-sound in *better*) or a light /l/ (the /l/ in *leaf*). Once again, the specificity of recalibration seems to be dependent on the acoustic features of the phoneme pair being learned.

Overall, it appears that retuning is often phoneme- and speaker-specific, but contingent on the specific phoneme pair used. Generalization to a new speaker is more likely to occur if the phoneme boundary is adjusted between two plosives and not between fricatives. Perceptual retuning effects upon plosives or stop consonants are also more likely to extend to other plosives, but, again, are unlikely to do so for fricatives or approximants. Acoustic similarity also plays an important role as to whether retuning effects can be applied to new sounds.

7.2.4 Other Attributes of Perceptual Learning

Most studies of the lexically guided perceptual learning studies described throughout Sect. 7.2 are twofold. They typically start with an exposure phase, with words containing one particular ambiguous phoneme, presented along with other filler words and pseudo-words. Listeners are also often asked to perform a lexical decision task during this exposure phase, in order to maintain their attention. This is followed by a categorization task, or the test phase, on a continuum between two clear phonemes with the aforementioned ambiguous phoneme in between. However, this design is not always used, and other similar designs can still lead to measurable retuning effects. McQueen et al. (2006b) concluded that perceptual learning is not dependent on a lexical decision task during the exposure phase. Instead, the lexical decision task was replaced with a simple counting task, and learning effects remained intact. However, a more recent study by Samuel (2016) suggested that targeted distractions during exposure that can prevent access to the lexicon are detrimental to perceptual retuning. In this study, listeners heard two voices only separated by 200 ms during exposure, of words containing an ambiguous /s/-/ʃ/ phoneme by a male speaker, and irrelevant words by a female speaker, and were asked to perform a lexical decision task on the male speaker, or to count the number of syllables spoken by the female speaker. Listeners who attended to the female speaker showed no recalibration during subsequent testing; however, when the voices were separated by 1200 ms, recalibration effects were reinstated. Similarly, listeners were also unable to undergo learning in the presence of background noise (Zhang and Samuel 2015), suggesting that recalibration cannot be performed automatically and requires attentional resources. But attention alone is also not enough to induce retuning, as can listeners still account for potentially transient characteristics of a speaker. In a creative design by Kraljic et al. (2008b), listeners viewed stimuli of a speaker with a pen in their mouth while pronouncing words dubbed with an ambiguous phoneme. These listeners did not show retuning during the subsequent test phase, implying that listeners also acknowledge temporary atypical pronunciations of a speaker before adjusting phoneme representations.

Attention aside, the prototypical test phase, most often a continuum of sounds between two phonemes, is also not a requisite to detect perceptual retuning effects. Effects were still preserved when test phase items were replaced with minimal word pairs ending in an ambiguous phoneme (McQueen et al. 2006a). Participants were then more likely to hear one of the two words of the pair, predicated by the prior exposure phase. For instance, after exposure to words with an ambiguous /f/ (such as *paragraph*, ending with an /f/-/s/ blend), participants were likely to hear "knife" rather than "nice" when presented with "kni-," ending in the same /f/-/s/ blend. The effect was observed in the opposite direction when listeners were presented with /s/ words ending in the ambiguous token during the exposure. In the same example, listeners were more likely to hear "nice."

Even fully intact lexical information is not a necessity for retuning to occur, and implicit knowledge of phonotactic information, or the rules within a language regarding allowable phoneme combinations, can be sufficient (Cutler et al. 2008). Here, exposure stimuli were phonotactically valid pseudo-words containing an ambiguous phoneme. Perceptual retuning can also be observed with other known phonemes that are acoustically related, such as /θ/ (represented as theta, or the "th" sound in thing) in place of /s/ or /f/, in place of the oft-mentioned ambiguous phoneme (Sjerps and McQueen 2010). Again, the acoustic or perceptual similarity can determine whether retuning is induced or not.

Thus, the exposure and test phases do not necessarily have to follow one particular procedure for phoneme boundary retuning, but all of the studies discussed within Sect. 7.2, as well as most of the classical studies of lexically driven perceptual retuning, have focused on native listeners. More recent works have also studied non-native listeners, and retuning can take place in these listeners as well. Native Dutch speakers with high proficiency in English also showed perceptual learning effects in response to English stimuli spoken by a British English speaker (Drozdova et al. 2015). Native German speakers of Dutch were also observed to undergo retuning effects in response to Dutch stimuli, at levels comparable to native Dutch speakers

(Reinisch et al. 2013). However, proficiency in the second language can also determine whether recalibration can occur, as a group of native Arabic speakers with lower English proficiency than another group of native Hebrew speakers showed no retuning effects with English phonemes, while the latter group did (Samuel and Frost 2015).

7.2.5 Summary of Lexically Driven Perceptual Learning

Section 7.2 summarized the seminal studies as well as some more recent findings about lexically guided perceptual learning. These effects are potentially long-lasting but may not generalize to new speakers. Non-native speakers are also capable of demonstrating learning effects, but this may be mitigated by the listener's proficiency in the second language. Generalization to new speakers and to other phonemes is mitigated by the type of phoneme category being adjusted. Retuning effects may be applied from stop consonants or plosives to other phonemes within this classification, but this is less likely for fricatives or approximants. While lexical knowledge is primarily driving the subsequent learning, acoustic features still place constraints on what can and cannot be extended to other speech sounds.

7.3 Audiovisual Information and Speech

7.3.1 Overview of Audiovisual Recalibration

Visual or speech-read information, much like lexical information, can also provide clarity when the available acoustics are unclear. Speech-reading can be relied upon if noise is present (Sumby and Pollack 1954), and also significantly alter what listeners perceive to hear. McGurk and MacDonald (1976) made the groundbreaking discovery that participants who viewed videos of a speaker pronouncing the syllable /gaga/, dubbed with audio of the syllable /baba/, perceived an entirely new percept, and reported hearing /dada/. Bertelson et al. (2003) extended this finding, and detected aftereffects on categorization responses following exposure to McGurk-like stimuli. Again, not only did speech-reading influence the perception of incongruent audiovisual tokens, but continuous exposure led to responses biased by the visual/speech-reading information. Much like the approach used by Norris et al. (2003) described in Sect. 7.2, participants first underwent an exposure phase, where they viewed audiovisual stimuli of a speaker's lip movements while pronouncing /aba/, dubbed with audio of an ambiguous phoneme halfway between /aba/ and /ada/. During a subsequent test phase, participants only heard the audio token of the ambiguous phoneme and its two neighbors from a continuum, and were more likely to report them as /aba/ sounding. Unlike Norris et al. (2003), a within-subjects design was used, and the same group of participants also viewed videos of the speaker pronouncing /ada/, but dubbed with the same ambiguous token. In this case, participants were more likely to report hearing the token as /ada/ during the test phase (Fig. 7.2).

In a follow-up experiment, listeners were exposed to congruent stimuli, or clear audio of /aba/ combined with lip movements of /aba/, and the same for an audio and video combination of /ada/. These unambiguous stimuli showed the reverse effect of the recalibration experiment and led to selective speech adaptation (Eimas and Corbit 1973). As a result of said selective speech adaptation, participants made fewer /aba/ responses to the ambiguous sounds if exposed to clear /aba/ tokens, and similarly gave fewer /ada/ responses after exposure to clear /ada/ tokens. This response is unlike recalibration, where participants who listen to ambiguous sounds during the exposure phase then become more likely to report hearing the phoneme being biased for by the lip movements of the speakers (i.e., ambiguous audio coupled with video of /aba/ leading to more /aba/ responses during the test phase). Selective speech adaptation will be discussed in more detail in Sect. 7.3.2.

7.3.2 Audiovisual Recalibration and Selective Speech Adaptation

Prior to studies of audiovisual recalibration, a perceptual learning effect known as selective speech adaptation was discovered (Eimas and Corbit 1973) and has also been helpful for understanding the building blocks of speech perception. Recalibration and selective speech adaptation share considerable overlap, especially in terms of their experimental design, but are also distinct in their interpretations. Both styles of experiments use a similar two-part procedure with an exposure and

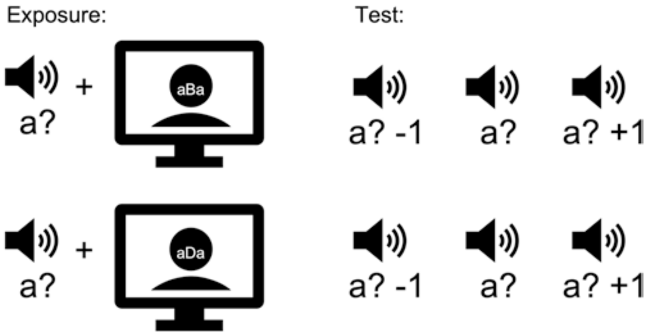

Fig. 7.2 A typical audiovisual recalibration procedure. Exposure phases pair ambiguous phoneme blends (such as an /aba/-/ada/ blend) with video of a speaker pronouncing one of the two phonemes (/aba/ or /ada/). Following exposure to these videos, listeners are then presented with the auditory items (the ambiguous /aba/-/ada/ blend, along with other similar sounds) and asked to respond with what they hear

test phase. Unlike recalibration, which typically uses ambiguous sounds, selective speech adaptation relies on exposure to clear sounds. While recalibration experiments lead to an increase in responses of the phoneme indicated by the videos during exposure, selective adaptation results in a reduction. For example, listeners repeatedly exposed to tokens of a clear /ba/ become less likely to perceiving /ba/ when given a categorization task on a /ba/-/da/ continuum. Selective speech adaptation is thought to reflect a fatigue effect, where listeners become desensitized to the auditory token during the exposure phase. The listener then becomes more sensitive to the acoustic differences in other similar sounds, thereby reports hearing the ambiguous tokens as the phoneme opposing the preceding exposure phase. The original study of selective speech adaptation (Eimas and Corbit 1973) relied on solely auditory stimuli, but later studies measured the same effects when exposure stimuli were coupled with videos of a speaker's lip movements, as Bertelson et al. (2003) reported. These unambiguous, or congruent, audiovisual stimuli also led to fewer responses of the phoneme presented in the test phase, as described in Sect. 7.3.1.

Selective speech adaptation and recalibration are often discussed together, as they both reflect a change in auditory perception, following an exposure phase to syllables or speech sounds. Just as the response patterns of the two phenomena go in opposite directions, the two differ in numerous other ways as well. Vroomen and colleagues have compared an audiovisual form of selective speech adaptation to recalibration and have found that the overall buildup and dissipation also tend to differ (Vroomen et al. 2006). The number of exposure trials has been found to share a log-linear relationship with selective speech adaptation, as the effect was observed to increase as exposure trials accumulate, whereas recalibration was found to have a curvilinear relationship in relation to the number of exposure trials, as it steadily increased until eight exposure trials, but reduced with additional exposure. Recalibration and selective speech adaptation are also differentially affected by the number of test trials, as audiovisual recalibration effects are short-lived and can be present only up until approximately 6 test trials, while selective speech adaptation effect can be continuously sustained for up to 60 test items (Vroomen et al. 2004).

Sine-wave speech (SWS) is constructed by starting from clear speech but stripped down until approximately three sinusoids that follow the central frequency and amplitude of the first three formants remain (Remez et al. 1981). These stimuli are often unintelligible unless listeners are explicitly told that the sounds have been extracted from actual speech. Vroomen and Baart (2009) also compared recalibration and selective speech adaptation in groups that viewed audiovisual SWS tokens as speechlike versus non-speechlike. In this experiment, all of the ambiguous and clear sounds typical of recalibration and selective speech adaptation studies were replaced with SWS versions, so a continuum including and between two clear phonemes was converted into SWS. For exposure phases, these SWS sounds were still paired with videos of a speaker's corresponding lip movements, but were presented without video for test phases. One "speech-mode" group viewed ambiguous SWS tokens paired with videos, which identified the tokens as /onso/ or /omso/, and showed recalibration effects. A "non-speech-mode" group viewed the same stimuli but categorized the ambiguous SWS tokens as “1” or “2,” and did not show a recalibration effect, so a “speech mode” did impact any possible recalibration. In contrast, for selective speech adaptation, participants viewed videos coupled with endpoint SWS tokens (rather than ambiguous), and adaptation effects were observed. In this instance, listeners who performed a categorization test on SWS versions of the ambiguous tokens heard them as the opposite phoneme to the one biased for by the preceding exposure (i.e., hearing more /omso/ after exposure to SWS versions of a clear /onso/ paired with video). Selective speech adaptation was still measurable in another non-speech-mode group, who underwent the same types of exposure, but categorized the subsequent test phase ambiguous sounds as 1 or 2. Essentially, selective speech adaptation was unaffected by either set of labels, so speech mode had no impact on perception and listeners still adapted accordingly. The awareness of speechlike qualities was crucial for successful recalibration, but selective speech adaptation was not hindered by this lack of this awareness. While recalibration and selective speech adaptation can reshape speech sound representations, based on these comparisons, it appears the two may be controlled by distinct but related substrates. The authors concluded that audiovisual recalibration may emerge from speech and language networks, while selective speech adaptation is purely a bottom-up process that does not require higher-level feedback. Potential neural mechanisms will be discussed in more detail in Sect. 7.5.

7.3.3 Specificity of Audiovisual Recalibration

Whether recalibration can be generalized has been addressed with regard to audiovisual information as well, just as it has with lexical context. While recalibration is robust enough to not depend on working memory (Baart and Vroomen 2010), audiovisual recalibration tends to be token-specific (Reinisch et al. 2014), as exposure to either visual /aba/ or /ada/ tokens dubbed with ambiguous audio had no effect on listeners’ categorization of continua of either /ibi/-/idi/ or /ama/-/ana/ sounds during test. Therefore, audiovisual recalibration appears to be constrained by the acoustics features, as learning could not extend to other phonemes, or even to the same phonemes paired with different vowels. The ear itself can limit recalibration (Keetels et al. 2016a, b), as the effect was optimal if exposure and test stimuli were presented into the same ear, but was diminished for test stimuli presented into the opposite ear, and locations in between resulted in a gradient of responses as the presentations moved further away from the original ear. The authors argue that this is further evidence that recalibration is strongly tied to the token and context, and the encoding process even accounts for the exact location of the presented sound (neural mechanisms will be addressed further in Sect. 7.5). Notably, listeners also have the capacity to recalibrate each ear in opposite directions using the same ambiguous sounds, e.g., one ear recalibrated toward /aba/, the other toward /ada/, with test sounds presented into the corresponding ears of the exposure phase (Keetels et al. 2015). Thus, phoneme representations may not be completely abstracted from the input received and can retain speaker- and context-specific details. Keetels et al. (2015) argue that this could be due to the perceptual system striking a balance between generalizing too often and too rarely. If recalibration is employed when speech is unclear, then it is may be only necessary to apply the newly learned boundary position to other instances that are similar both in acoustic and contextual features, so as to not unnecessarily overgeneralize.

While audiovisual recalibration may be restricted in some respects, it is not necessarily specific to the speaker, as listeners can recalibrate to another speaker's pronunciation of the same phoneme, although to a substantially lesser extent compared to the speaker during exposure (van der Zande et al. 2014). Recalibration is generally maximal in response to the sound used during exposure, which suggests that it generally tends to be constrained by the acoustic features of the exposure sound. Similarly, audiovisual recalibration is most often tested with consonant contrasts, but Franken et al. (2017) have found that recalibration is possible with a vowel contrast pair of /e/-/ø/. In addition, recalibration with a vowel pair and multiple speakers has also been observed, wherein the gender identity of the speakers combined with the visual cue indicated by the speech-reading information influenced listeners' categorization responses (Burgering et al. 2020).

The majority of the studies described have also been centered on adults, but audiovisual recalibration can also be adopted early in life and has been observed in children as young as 8 years old. Van Linden and Vroomen (2008) measured recalibration effects in two groups of children and determined that children at 8 years old could recalibrate with audiovisual stimuli, but children at 5 years old could not, so the ability may be developed within this window of 3 years. Dyslexia does not restrict the effect either (Baart et al. 2012), as adults with dyslexia were compared with fluently reading adults, and the dyslexic group showed no deficit in their ability to recalibrate. Even children with dyslexia are capable of undergoing recalibration driven by text (Romanovska et al. 2019), even though children with dyslexia often experience difficulties in speech-reading and letter-speech sound mappings (Snowling 1980; van Laarhoven et al. 2018).

7.3.4 Summary of Audiovisual Recalibration

Section 7.3 described audiovisual recalibration, originally described by Bertelson et al. (2003), and its various attributes. Later studies by Vroomen and colleagues have established the general buildup and dissipation, as well as similarities and differences with another perceptual learning effect, called selective speech adaptation. Audiovisual recalibration tends to both build up following a few exemplars during exposure and diminish with increasing numbers of test items as well. In contrast, selective speech adaptation requires much longer exposure phases, but subsequent effects can last for longer durations. Recalibration also tends to be token- and context-specific, even to the extent that listeners can recalibrate each ear in opposite directions. It also does not easily generalize to other speakers, phonemes, or other similar instances of the same phoneme, so it is considerably restricted by the acoustic features present during exposure. Nevertheless, it has shown to be utilized by a variety of listeners, including children and adults with dyslexia, and remains to be a helpful tool for listeners when the auditory signal is inadequate.

7.4 Comparison of Audiovisual Recalibration and Lexical Retuning

Sections 7.2 and 7.3 have discussed audiovisual recalibration and lexical retuning separately, but the two processes also share many common attributes. In realistic situations, listeners are likely to encounter lexical and visual information simultaneously, so it is possible that these two sources may interact while influencing speech perception. The designs of the two types of experiments share overlap in many respects, with exposure phases consisting of stimuli embedded with ambiguous phonemes, followed by forced-choice test phases where the ambiguous sounds are presented without lexical or speech-reading contextual cues. Even the response patterns between the two original studies by Bertelson et al. (2003) and Norris et al. (2003) paralleled each other, so it may appear that phoneme categories are affected comparably by both audiovisual and lexical information. Brancazio (2004) probed the influence of lexical and speech-reading information in audiovisual speech perception but found that speech-reading exerted a stronger influence on phoneme categorization. Audiovisual effects were similar irrespective of faster and slower response times, while lexical information showed a weaker effect overall and was associated with slower responses.

Based on this, van Linden and Vroomen (2007) proposed that audiovisual information may induce recalibration more effectively than lexical cues, and conducted a study comparing lexical and audiovisual recalibration to test this hypothesis. Two forms of recalibration were compared in native Dutch speakers using a /p/-/t/ phoneme contrast. One group was exposed to lexical stimuli, which consisted of audio Dutch words typically ending in either /op/ or /ot/ (such *bioscoop*, or movie theater, and *idioot*, or idiot*)*, with all endings replaced by an ambiguous token halfway between /op/ and /ot/. Another group was exposed to audiovisual stimuli, comprised of videos of pseudo-words, where lip movements indicated a /op/ or /ot/ ending, and were also dubbed with audio of the ambiguous phoneme at the end of the token. Participants were also exposed to both /op/- and /ot/-biased stimuli, to explore whether they could recalibrate in both directions of the phoneme pair, such that half of the exposure blocks would induce a bias toward /p/, and the remaining half were biased toward /t/. Test phase judgments indicated that recalibration was indeed successful in both groups and in response to both phonemes as well. As the authors originally proposed, audiovisual information was largely more effective in producing recalibration than lexical information. The discrepancy may have resulted from the inherent differences in the stimuli and the processing levels affected, as lexical information might only induce a phoneme preference with the help of top-down influences, whereas the incoming audiovisual information already contained a visual bias toward one phoneme. Theories of top-down and bottom-up processing will be discussed in more depth in Sect. 7.5.

In contrast to previous studies on lexical retuning, both audiovisual and lexical recalibration dissipated at the same rate. Although audiovisual recalibration has been known to dissipate relatively quickly (Vroomen et al. 2007b), other studies have found that lexically guided perceptual learning can be long-lasting (Eisner and McQueen 2006). Participants in the van Linden and Vroomen (2007) study were flexibly adjusting the phoneme boundary back and forth between the two phonemes, throughout the duration of the experiment, so the faster dissipation of lexical recalibration may have resulted from constant switching between the two phonemes. However, this was refuted in a follow-up experiment with a between-subjects design, where each group of participants were only exposed to one phoneme-modality combination, and no improvements to recalibration were found. Still, the chosen phoneme pair is also worth noting, as plosives or stop consonants such as /p/ and /t/ may be more amenable to adjustment than fricatives (as mentioned in Sect. 7.2), such as /f/ and /s/ (Kraljic and Samuel 2007). Overall, lexical and audiovisual recalibrations seem to be markedly similar, although the pathways supporting them may not be identical, and may only overlap.

The two types of retuning also tend to differ in their stability, as lexical retuning has been shown to be stable over time, but audiovisual recalibration can be more susceptible to decay with the passage of time. After a standard exposure phase, participants were tested after a 24-hour gap and effects had dissipated (Vroomen et al. 2007a), even if participants were tested both immediately after the exposure phase and again 24 hours later (Vroomen and Baart 2009). Audiovisual recalibration effects have also been shown to diminish within the test phase, as responses that corresponded with the preceding visual exposure (such as /b/ responses after viewing /aba/ videos) were maximal at the start of the test phase, but consistently decreased as the test phase progressed (Vroomen and Baart 2009). In contrast, lexical retuning effects can be preserved throughout longer testing sessions, often containing approximately 30 test items (Kraljic and Samuel 2009), or up to 12 hours later (Eisner and McQueen 2006). As mentioned earlier in Sect. 7.2, lexical retuning is capable of generalizing to new speakers and certain phonemes, while audiovisual recalibration is most often token-specific and may generalize if the critical phonemes are plosives/stop consonants.

More recently, studies comparing audiovisual recalibration and lexical retuning within both a single session and the same participants have found that the resulting effects were similar between the two, with similar patterns of dissipation as well (Ullas et al. 2020a). The simultaneous presentation of both audiovisual and lexical information within exposure (i.e., listeners presented with videos of words edited to contain an ambiguous final phoneme) also showed effects comparable to audiovisual recalibration alone, suggesting that the combination leads to no benefit in subsequent phoneme boundary retuning as a result of differences in the pathways involved in the two forms of perceptual learning (Ullas et al. 2020b). Overall, lexical retuning and audiovisual recalibration share many similarities in terms of how the subsequent effects are exhibited, how the experiments measuring them are designed, as well as the resulting response patterns to presentations of ambiguous sounds. Both approaches are useful for adapting to speech in noise, even if their origins and functions may differ.

7.5 Theoretical and Neural Explanations of Recalibration

7.5.1 Theories of Speech Perception

The mechanisms that enable the auditory system to adjust phoneme boundaries are often debated. Numerous theories of speech perception have been invoked in explanations of recalibration and perceptual retuning as well. Cutler, McQueen, Norris, and colleagues (Norris et al. 2000) originally proposed a feed-forward model of speech perception called Merge and argued that listeners can retune phoneme categories through a bottom-up abstraction process, which does not rely upon online feedback from the lexicon, not unlike the COHORT model which also states that word recognition primarily relies on bottom-up processes (Gaskell and Marslen-Wilson 1997). COHORT presents a modular, unidirectional explanation, where word recognition is initiated first by acoustic information, triggering a possible "cohort" of matches, and later, other features such as context and semantics allow the listener to narrow down the possibilities. Similarly, according to the Merge model, top-down feedback during speech recognition and phoneme categorization is not essential, so recognition and categorization operate at a pre-lexical level. Feedback during categorization could be time-consuming or lead to misinterpretations of the input, so interactions between lexical and pre-lexical processing would not be beneficial. Phonemic decisions can be made based on both lexical and pre-lexical information but do not necessitate interactions between the processes. Cutler et al. (2010) also emphasized that perceptual retuning cannot be explained purely by episodic information and that abstraction from such events must be involved as well. A more recent model by Norris et al. (2016) has been updated to include predictions of perception based on Bayesian inference, but still does not rely upon online feedback during phoneme processing. Acoustic information and lexical knowledge are combined to calculate probable phonemes, but again, the two processes are not proposed to interact.

Others have described top-down (Davis et al. 2005; Davis and Johnsrude 2007) and bidirectional influences on speech perception (McClelland and Elman 1986; McClelland et al. 2006). A classical, interactive model of speech perception, TRACE (McClelland and Elman 1986), derives its name from a structure called "The Trace," a perceptual processing tool. McClelland and Elman proposed that top-down feedback modulates connections between three layers, from words, to phonemes, down to features. Phoneme identification can be influenced by lexical and speech-reading contexts, and can also be improved through experience. According to TRACE, this influence is due to feedback from higher levels of processing. Similarly, McClelland et al. (2006) contend that both top-down and bottom-up information streams are essential for speech perception. Phoneme representations can be influenced by both lexical and acoustic features, and vice versa.

While most classical theories of speech perception have not accounted for the role of visual information, more recently, Kleinschmidt and Jaeger (2011) have put forth a belief-updating model based on Bayesian inference, by using data from previous studies of recalibration and selective speech adaptation to calculate probabilities of outcomes. This model, called the Ideal Adaptor Framework, is tailored to explain audiovisual recalibration and selective speech adaptation. As described in Sect. 7.3.2, audiovisual recalibration and selective speech adaptation are two forms of perceptual learning, but their response profiles are in direct contrast. According to the Ideal Adaptor Framework, both recalibration and selective speech adaptation are described as forms of statistical learning, as a result of exposure to various distributions of phonemes. Listeners can create speaker-specific models of phoneme categories which allow for initial speaker-level adaptation, but can eventually generalize to more speakers with additional experience and if they are also acoustically close. The authors also posit recalibration and selective speech adaptation as two response patterns along a continuum ranging from ambiguous to prototypical sounds. As mentioned earlier in Sect. 7.2.2, recalibration effects tend to peak after approximately eight exposure tokens and slowly diminish with additional exposures, while selective speech adaptation tends to continuously build in a linear manner with increasing exposure. According to the model, recalibration reflects a response to ambiguous sounds, but with increasing amounts of exposure tokens and as speech sounds become more prototypical, selective adaptation effects can be observed.

7.5.2 Neural Basis of Recalibration and Perceptual Learning

While theoretical frameworks and models have been useful in understanding recalibration and retuning, neuroimaging studies have shed additional light on areas of the brain where these changes occur and how they might explain the levels of processing involved. More general models of speech perception drawn from neuroimaging data and primate studies (Scott and Johnsrude 2003; Rauschecker and Scott 2009) have described the hierarchical and topographic nature of processing in the auditory cortex and surrounding areas.

Hickok and Poeppel (2007) proposed the dual-stream processing model of speech, with certain features equivalent to those found in visual-processing models. According to the model, areas of the brain along a ventral pathway, including medial temporal gyrus (MTG) and inferior temporal sulcus (ITS), are geared toward connecting phonological and lexical representations, while regions along a dorsal pathway, including parietal-temporal, (pre)motor, and inferior frontal regions, are geared toward connecting phonological with sensorimotor and articulatory representations. Adank and Devlin (2010) also explored how listeners adjust to recordings of unclear sentences and found activation patterns consistent with the Hickok and Poeppel (2007) model. Jäncke et al. (2002) also identified structures of the brain in the planum temporale (PT) and middle superior temporal gyrus (STG) that are specific to phoneme perception. STG and the primary auditory cortex can also encode fine-tuned phonetic information (Mesgarani et al. 2008, 2014), with evidence for speaker-invariant phoneme representations distributed across both of these regions (Formisano et al. 2008; Bonte et al. 2014). Other regions implicated in categorical perception of speech sounds include the inferior frontal gyrus (Rogers and Davis 2017) and the supramarginal gyrus (Raizada and Poldrack 2007; see Davis and Johnsrude 2007 for a review).

While these studies paved the way toward delineating a network of regions possibly implicated in recalibration, they may still be insufficient, as this process relies on the integration of both acoustic and contextual information, which are often lexical or visual. In light of this, Obleser and Eisner (2009) proposed a model of pre-lexical abstraction based on prior neuroimaging studies of speech perception, reminiscent of the Merge model (with similarities to TRACE as well, but this model focuses on word recognition and not on abstraction). Pre-lexical abstraction may appear to resemble recalibration, but it also implies that the phoneme representation can be fully disentangled from the acoustic input and thereby abstracted. Pre-lexical abstraction could be implemented probabilistically, primarily along the STG, resulting in phoneme likelihoods rather than definitive phoneme identification. Likelihoods could be calculated by weighing various acoustic features, first processed by primary auditory cortex, and could be updated with talker and context-specific information. Similarly, Holdgraf et al. (2016) have found evidence for acoustic updating, using spectro-temporal receptive field mapping on ECoG recordings of the auditory cortex. Responses of cortical populations were observed to have increased sensitivity to speechlike spectro-temporal features of degraded speech, after exposure to intact speech. This sensitivity could reflect how listeners encode rudimentary acoustic features that also allow the listener to interpret less intelligible speech, or how listeners “fill in the gaps.”

The merits of these models of speech perception can be reexamined in light of fMRI studies of recalibration and retuning. Kilian-Hütten et al. (2011b) had participants undergo audiovisual recalibration using the classic /aba/-/ada/ stimuli while fMRI data was collected. It was discovered that a higher-order network of areas in and around the auditory cortex, including bilateral inferior parietal lobe (IPL), inferior frontal sulcus (IFS), superior temporal sulcus and superior temporal gyrus (STS/STG), and posterior MTG, were all active in recalibration. These areas showed overlapping activation during both the exposure phase and the subsequent test phase. These regions are also known to be involved in audiovisual integration and constructive processes, which would account for their increased activation during recalibration. Kilian-Hütten et al. (2011a) were also able to investigate audiovisual recalibration using MVPA, or multivariate pattern analysis, a technique using fMRI data to train an algorithm to recognize differences in patterns of brain activity. They were successfully able to decode whether a participant perceived /aba/ or /ada/ while presented with the ambiguous sounds during the test phase of the same audiovisual recalibration experiment, solely using the activation patterns. Active clusters were found in and around left PT and left Heschl's gyrus and sulcus, which are typically viewed as low-level auditory areas, but they may have been influenced by information other than rudimentary acoustics features as they effectively predicted the percepts that were driven by the visual cue and not the auditory information alone.

More recently, Lüttke et al. (2016) investigated a form of adaptation induced by McGurk-style adaptors with fMRI. Exposure to McGurk adaptors, or clear auditory /aba/ paired with video of /aga/, resulted in the percept of /ada/. These stimuli led to an effect much like selective speech adaptation, where follow-up presentations of clear auditory /aba/ were incorrectly perceived as /ada/ as a result. This mistaken /ada/ percept showed closely related neural patterns to those elicited by correctly perceived auditory /ada/, and more so than to patterns associated with correct perception of clear /aba/ tokens. Again, neural activations echoed a shift in auditory perception due to adaptation through contextual cues.

fMRI has also been used to explore lexically driven perceptual learning and other related phenomena. Activation in posterior left STG and STS has been recorded in listeners receiving instructions to switch from an acoustic mode to speech mode while listening to SWS stimuli (Dehaene-Lambertz et al. 2005). While stimuli remained the same, instructions alone could induce a shift in both perception and the resulting activation patterns. Similarly, activity in left pSTS has also been associated with identification of nonphonemic, short-term sound categories, while left mSTS may store long-term representation of phoneme patterns already known to the listener (Liebenthal et al. 2010). Myers and Blumstein (2008) investigated the Ganong effect (described in Sect. 7.1), or the impact of lexical knowledge on perception of ambiguous speech tokens. Participants heard auditory items with ranging voice onset time (VOT) from *gift* to *kift* (i.e., word to nonword) and another continuum ranging from *giss* to *kiss* (from nonword to word). Activity in STG was modulated by the lexical effect, such that boundary tokens that were perceived as words showed higher activations compared to acoustically similar tokens from the other continuum that were not perceived as words. As STG was engaged in both phonological and lexical processing, the authors suggested that this was evidence in support of top-down models similar to TRACE that accommodate higher-level information during processing (Liebenthal et al. 2010).

Similarly, Myers and Mesite (2014) tested participants in a classic lexically guided perceptual retuning experiment with the addition of fMRI, alternating between exposure phases containing edited words ending in an ambiguous phoneme, followed by a forced-choice test phase on a continuum of the same ambiguous sounds. Participants were separated into two groups with the stimuli biased toward /s/ for one group, and toward /ʃ/ (the "sh" in shop) for the other. Behavioral results indicated a boundary shift, so over the course of the successive test phases, participants' perception of the ambiguous /s/-/ʃ/ phoneme had changed. Increased activity in left IFG and STG was measured with boundary shifted items. These items reflected the perceptual shift, and were categorized as the biasing phoneme in test blocks following the exposure, but not during the earlier blocks at the start of the experiment. Activity both within the auditory cortex and in higher-level cognitive areas suggests that top-down information may have influenced the learning process and may also have been responsible for creating connections between phonetic information and the speaker. Together, the results of these two studies of lexical context imply that perceptual learning involves areas responsible for both lower and higher levels of information processing in resolving the perception of these sounds. However, it remains unclear as to whether the flow of information is simply feed-forward or not, as the exact timing as to when each region is engaged is not yet understood. The authors suggest that initial processing of the unclear sounds relies on higher-level executive regions, but once the listener undergoes sufficient training and has shifted the perceptual boundary, then regions responsible for lower levels of processing, such as STG, can be activated in response to the ambiguous sound.

Combined magnetoencephalogram (MEG) and electroencephalogram (EEG) data have also confirmed that activity in STG reduced over time, as participants learned to improve in identification of degraded speech sounds combined with matching text (Sohoglu and Davis 2016). Furthermore, the results were framed within a model of predictive coding, not unlike Bayesian inference, such that the listener learns to reduce prediction errors as a consequence of learning. STG is proposed to process acoustic features and receives predictions of phonological categories from higher-level frontal areas, and predictions are continuously updated with experience.

While many of the studies discussed thus far have identified STG to be involved in perceptual learning or recalibration, a recent study has also found evidence from the cerebellum (Guediche et al. 2015). Listeners learned to identify words distorted by noise vocoding, and consequently, cerebellar regions showed changes, as well as functional connections to cortical language and auditory regions. Stemming in part from this finding, another model of speech adaptation has been proposed, also relying on a predictive coding mechanism, but supervised by the cerebellum (see Guediche et al. 2014, for a complete review). In contrast, some areas of the brain may be uniquely engaged by either recalibration or retuning. When compared directly using fMRI within the same participants, audiovisual recalibration and lexical retuning showed largely similar areas of activation, over temporal, parietal, and motor cortex areas, although audiovisual recalibration specifically seems to retrigger activation within areas of the visual cortex, despite the lack of visual stimuli during the recalibration test trials (Ullas et al. 2020).

7.5.3 Summary of Theories of Speech Perception

Section 7.5 detailed various theories of speech perception as well as supporting neuroimaging data that propose the channels through which recalibration and perceptual retuning may operate. Proponents of these speech perception theories have debated the nature of how phoneme categories can be reshaped, as some argue that this is a unidirectional, bottom-up abstraction process (Merge, COHORT), while others postulate that both top-down and bottom-up processes contribute (TRACE). Theories incorporating distributional and statistical learning, such as the Ideal Adaptor Framework (Kleinschmidt and Jaeger 2011), have also been useful for understanding how listeners adapt to variability. Neuroimaging data suggest that both top-down and bottom-up influences are involved, based on the areas of the brain that tend to be active during perception of ambiguous tokens, such as STS/STG and IFS/IFG. Sophisticated analysis techniques such as MVPA have also been useful for pinpointing specific patterns of neural activity associated with the shifts in perception, but the directionality of influences upon these percepts remains unclear and may require more advanced neuroscientific methods.

7.6 Conclusion and Future Directions

The literature described throughout this chapter has focused on lexical and audiovisual information as contextual influences on speech perception, as well as their dimensions and limitations. Section 7.2 highlighted the seminal findings regarding lexical retuning, starting from Norris et al. (2003) and the studies since then that have illuminated the strengths and drawbacks. Section 7.3 discussed audiovisual recalibration, first described by Bertelson et al. (2003) and expanded upon by others.

These two contextual sources can differ in terms of their impact on perception, as lexical information can potentially lead to more stable and longer-lasting shifts in perception, while audiovisual information results in adjustments in shorter durations that are not easily generalizable and are often either (or both) context- and token-dependent. The phoneme categories themselves can also impose restrictions, as plosives (also known as stop consonants) may allow for generalization to other speakers more so than other types of phonemes, such as fricatives or liquids. Evidently, contextual cues alone do not drive these phoneme boundary shifts, and acoustic information still modulates learning effects to a great extent. Theories of speech perception have also been helpful for understanding the basis of phoneme boundary adjustments, but disagreements exist with regard to the stages of processing that are thought to be involved.

Although questions remain in the field as to the precise details of retuning, researchers continue to pursue the answers with behavioral and neuroimaging studies. Related works may also shed light upon how exactly these perceptual shifts may occur. Recent studies have investigated another related form of text-based recalibration. Reading text of syllables while listening to ambiguous phonemes can also contribute to changes in phoneme categorization (Keetels et al. 2016a, b), and this has also been tested using fMRI (Bonte et al. 2017). Just as in audiovisual and lexical experiments, participants viewed either /aba/ or /ada/ written in text, while hearing an ambiguous blend of the two, and participants were able to effectively recalibrate depending on the text they viewed (Keetels et al. 2016a, b). In addition, fMRI results showed that text-based recalibration was linked to activity in posterior superior temporal cortex, and percepts of /aba/ and /ada/ during test could also be decoded with MVPA, primarily based on patterns of activity in left posterior STG and PT and right STS (Bonte et al. 2017). Functional connectivity was observed between IPL and left STG during exposure and may be indicative of higher-order influences leading to eventual retuning. While lexical and audiovisual recalibration studies have been useful for understanding how listeners adapt to ambiguity in speech, this new paradigm illuminates how mappings are acquired between auditory and written representations, and may also have the potential to detect disruptions of reading networks during development, particularly in individuals with dyslexia.

Together, these approaches using lexical and audiovisual information, and more recently with text, have proven useful in understanding the plasticity of speech sounds. These non-acoustic sources of information can not only sway how speech tokens are perceived but, moreover, can restructure the units of speech. Evidently, these units are malleable and are continuously updated with experience; they are susceptible to change even within short windows of time and with relatively little input required to do so. This adaptive tool is beneficial for adjusting to speakers, noise, or other obstacles that could impede successful speech comprehension, although the acoustic features of the input may restrict the extent to which recalibration can be generalized. Still, stimulus specificity may be advantageous, as a complete overhaul of speech sounds in response to deviations from the norm would be impractical. Speech perception theories and neuroimaging studies have highlighted the possible processing streams involved, and both lexical and speech-reading influences appear to share significant similarities in terms of the brain areas being recruited. The relative contributions of top-down and bottom-up information in processing the acoustic input are still hotly debated, but the continued application of advanced neuroimaging techniques, as well as statistical modeling, may aid in building a more cohesive picture of perceptual retuning.

Compliance with Ethics Requirements Shruti Ullas declares no conflict of interest.
Milene Bonte declares no conflict of interest.
Elia Formisano declares no conflict of interest.
Jean Vroomen declares no conflict of interest.

References

Adank P, Devlin JT (2010) On-line plasticity in spoken sentence comprehension: adapting to time-compressed speech. NeuroImage 49(1):1124–1132. https://doi.org/10.1016/j.neuroimage.2009.07.032

Baart M, Vroomen J (2010) Phonetic recalibration does not depend on working memory. Exp Brain Res 203:575–582. https://doi.org/10.1007/s00221-010-2264-9

Baart M, de Boer-Schellekens L, Vroomen J (2012) Lipread-induced phonetic recalibration in dyslexia. Acta Psychol 140(1):91–95. https://doi.org/10.1016/j.actpsy.2012.03.003

Bertelson P, Vroomen J, De Gelder B (2003) Visual recalibration of auditory speech identification: a McGurk aftereffect. Psychol Sci 14(6):592–597. https://doi.org/10.1046/j.0956-7976.2003.psci_1470.x

Bonte M, Hausfeld L, Scharke W, Valente G, Formisano E (2014) Task-dependent decoding of speaker and vowel identity from auditory cortical response patterns. J Neurosci 34(13):4548–4557. https://doi.org/10.1523/JNEUROSCI.4339-13.2014

Bonte M, Correia JM, Keetels M, Vroomen J, Formisano E (2017) Reading-induced shifts of perceptual speech representations in auditory cortex. Sci Rep 7:1–11. https://doi.org/10.1038/s41598-017-05356-3

Bradlow AR, Bent T (2008) Perceptual adaptation to non-native speech. Cognition 106(2):707–729. https://doi.org/10.1016/j.cognition.2007.04.005

Brancazio L (2004) Lexical influences in audiovisual speech perception. J Exp Psychol Hum Percept Perform 30(3):445–463. https://doi.org/10.1037/0096-1523.30.3.445

Burgering M, van Laarhoven T, Baart M, Vroomen J (2020) Fluidity in the perception of auditory speech: cross-modal recalibration of voice gender and vowel identity by a talking face. Q J Exp Psychol (Hove) 73(6):957–967. https://doi.org/10.1177/1747021819900884

Clarke CM, Garrett MF (2004) Rapid adaptation to foreign-accented English. J Acoust Soc Am 116(6):3647–3658. https://doi.org/10.1121/1.1815131

Cutler A, McQueen JM, Butterfield S, Norris D (2008) Prelexically-driven perceptual retuning of phoneme boundaries. In: Fletcher J, Loakes D, Goecke R, Burnham D, Wagner M (eds) Proceedings of Interspeech, Brisbane, 2008

Cutler A, Eisner F, McQueen JM, Norris D (2010) How abstract phonemic categories are necessary for coping with speaker-related variation. In: Fougeron C, Kühnert B, D'Imperio M, Vallée N (eds) Laboratory phonology, vol 10. de Gruyter, Berlin, pp 91–111

Davis MH, Johnsrude IS (2007) Hearing speech sounds: top-down influences on the interface between audition and speech perception. Hear Res 229(1–2):132–147. https://doi.org/10.1016/j.heares.2007.01.014

Davis MH, Johnsrude IS, Hervais-Adelman AG, Taylor K, McGettigan C (2005) Lexical information drives perceptual learning of distorted speech: evidence from the comprehension of noise-vocoded sentences. J Exp Psychol Gen 134(2):222–241. https://doi.org/10.1037/0096-3445.134.2.222

Dehaene-Lambertz G, Pallier C, Serniclaes W, Sprenger-Charolles L, Jobert A, Dehaene S (2005) Neural correlates of switching from auditory to speech perception. NeuroImage 24(1):21–33. https://doi.org/10.1016/j.neuroimage.2004.09.039

Drozdova P, van Hout R, Scharenborg O (2015) Lexically-guided perceptual learning in non-native listening. Biling (Camb Engl) 19(5):914–920. doi: https://doi.org/10.1017/S136672891600002X

Eimas PD, Corbit JD (1973) Selective adaptation of linguistic feature detectors. Cogn Psychol 4:99–109. https://doi.org/10.1016/0010-0285(73)90006-6

Eisner F, McQueen JM (2005) The specificity of perceptual learning in speech processing. Atten Percept Psychophys 67:224–238. https://doi.org/10.3758/BF03206487

Eisner F, McQueen JM (2006) Perceptual learning in speech: stability over time. J Acoust Soc Am 119:1950–1953. https://doi.org/10.1121/1.2178721

Formisano E, De Martino F, Bonte M, Goebel R (2008) "Who" is saying "what"? Brain based decoding of human voice and speech. Science 322(5903):970–973. https://doi.org/10.1126/science.1164318

Franken MK, Eisner F, Schoffelen JM, Acheson DJ, Hagoort P, McQueen JM (2017) Audiovisual recalibration of vowel categories. In: Proceedings of Interspeech, Stockholm, pp 655–658. https://doi.org/10.21437/Interspeech.2017-122

Ganong WF (1980) Phonetic categorization in auditory word perception. J Exp Psychol Hum Percept Perform 6(1):110–125. https://doi.org/10.1037/0096-1523.6.1.110

Gaskell MG, Marslen-Wilson WD (1997) Integrating form and meaning: a distributed model of speech perception. Lang Cogn Process 12(5–6):613–656. https://doi.org/10.1080/016909697386646

Guediche S, Blumstein SE, Fiez JA, Holt LL (2014) Speech perception under adverse conditions: insights from behavioral, computational, and neuroscience research. Front Syst Neurosci 7:1–16. https://doi.org/10.3389/fnsys.2013.00126

Guediche S, Holt LL, Laurent P, Lim S, Fiez JA (2015) Evidence for cerebellar contributions to adaptive plasticity in speech perception. Cereb Cortex 25:1867–1877. https://doi.org/10.1093/cercor/bht428

Hickok G, Poeppel D (2007) The cortical organization of speech processing. Nat Rev Neurosci 8:393–402. https://doi.org/10.1038/nrn2113

Holdgraf CR, de Heer W, Pasley B, Rieger J, Crone N, Lin JJ, Knight RT, Theunissen FE (2016) Rapid tuning shifts in human auditory cortex enhance speech intelligibility. Nat Commun 7:13654. https://doi.org/10.1038/ncomms13654

Holt LL, Lotto AJ (2008) Speech perception within an auditory cognitive science framework. Curr Dir Psychol Sci 17(1):42–46. https://doi.org/10.1111/j.1467-8721.2008.00545.x

Jäncke L, Wüstenberg T, Scheich H, Heinze HJ (2002) Phonetic perception and the auditory cortex. NeuroImage 15(4):733–746. https://doi.org/10.1006/nimg.2001.1027

Keetels MN, Pecoraro M, Vroomen J (2015) Recalibration of auditory phonemes by lipread speech is ear-specific. Cognition 141:121–126. https://doi.org/10.1016/j.cognition.2015.04.019

Keetels MN, Schakel L, Bonte M, Vroomen J (2016a) Phonetic recalibration of speech by text. Atten Percept Psychophys 78:938–945. https://doi.org/10.3758/s13414-015-1034-y

Keetels MN, Stekelenburg JJ, Vroomen J (2016b) A spatial gradient in phonetic recalibration by lipread speech. J Phon 56:124–130. https://doi.org/10.1016/j.wocn.2016.02.005

Kilian-Hütten N, Valente G, Vroomen J, Formisano E (2011a) Auditory cortex encodes the perceptual interpretation of ambiguous sound. J Neurosci 31(5):1715–1720. https://doi.org/10.1523/JNEUROSCI.4572-10.2011

Kilian-Hütten N, Vroomen J, Formisano E (2011b) Brain activation during audiovisual exposure anticipates future perception of ambiguous speech. NeuroImage 57(4):1601–1607. https://doi.org/10.1016/j.neuroimage.2011.05.043

Kleinschmidt DF, Jaeger TF (2011) Robust speech perception: recognize the familiar, generalize to the similar, and adapt to the novel. Psychol Rev 122(2):148–203. https://doi.org/10.1037/a0038695

Kraljic T, Samuel AG (2005) Perceptual learning for speech: is there a return to normal? Cogn Psychol 51:141–178. https://doi.org/10.1016/j.cogpsych.2005.05.001

Kraljic T, Samuel AG (2006) Generalization in perceptual learning for speech. Psychon Bull Rev 13:262–268. https://doi.org/10.3758/BF03193841

Kraljic T, Samuel AG (2007) Perceptual adjustments to multiple speakers. J Mem Lang 56:1–15. https://doi.org/10.1016/j.jml.2006.07.010

Kraljic T, Samuel AG (2009) Perceptual learning for speech. Atten Percept Psychophys 71(3):1207–1218. https://doi.org/10.3758/APP.71.6.1207

Kraljic T, Brennan SE, Samuel AG (2008a) Accommodating variation: dialects, idiolects, and speech processing. Cognition 107:51–81. https://doi.org/10.1016/j.cognition.2007.07.013

Kraljic T, Samuel AG, Brennan SE (2008b) First impressions and last resorts: how listeners adjust to speaker variability. Psychol Sci 19:332–338. https://doi.org/10.1111/j.1467-9280.2008.02090.x

Lecumberri MLG, Cooke M, Cutler A (2010) Non-native speech perception in adverse conditions: a review. Speech Commun 52(11–12):864–886. https://doi.org/10.1016/j.specom.2010.08.014.

Liebenthal E, Desai R, Ellingson MM, Ramachandran B, Desai A, Binder JR (2010) Specialization along the left superior temporal sulcus for auditory categorization. Cereb Cortex 20(12):2958–2970. https://doi.org/10.1093/cercor/bhq045

Lüttke C, Ekman M, van Gerven M, de Lange FP (2016) McGurk illusion recalibrates subsequent auditory perception. Sci Rep 6:32891. https://doi.org/10.1038/srep32891

Maye J, Aslin RN, Tanenhaus MK (2008) The Weckud Wetch of the Wast: Lexical adaptation to a novel accent. Cogn Sci 32(3):543–562. https://doi.org/10.1080/03640210802035357

McClelland JL, Elman JL (1986) The TRACE model of speech perception. Cogn Psychol 18:1–86. https://doi.org/10.1016/0010-0285(86)90015-0

McClelland JL, Mirman D, Holt LL (2006) Are there interactive processes in speech perception? Trends Cogn Sci 10(8):363–369. https://doi.org/10.1016/j.tics.2006.06.007

McGurk H, MacDonald J (1976) Hearing lips and seeing voices. Nature 264:746–748. https://doi.org/10.1038/264746a0

McQueen JM, Cutler A, Norris D (2006a) Phonological abstraction in the mental lexicon. Cogn Sci 30:1113–1126. https://doi.org/10.1207/s15516709cog0000_79

McQueen JM, Norris D, Cutler A (2006b) The dynamic nature of speech perception. Lang Speech 49(1):101–112. https://doi.org/10.1177/00238309060490010601

Mesgarani N, David SV, Fritz JB, Shamma SA (2008) Phoneme representation and classification in primary auditory cortex. J Acoust Soc Am 123(2):899–909. https://doi.org/10.1121/1.2816572

Mesgarani N, Cheung C, Johnson K, Chang EF (2014) Phonetic feature encoding in human superior temporal gyrus. Science 343(6174):1006–1010. https://doi.org/10.1126/science.1245994

Mitterer H, Scharenborg O, McQueen JM (2013) Phonological abstraction without phonemes in speech perception. Cognition 129:356–261. https://doi.org/10.1016/j.cognition.2013.07.011

Myers EB, Blumstein SE (2008) The neural basis of the lexical effect: an fMRI investigation. Cereb Cortex 18:278–288. https://doi.org/10.1093/cercor/bhm053

Myers EB, Mesite LM (2014) Neural systems underlying perceptual adjustment to non-standard speech tokens. J Mem Lang 76:80–93. https://doi.org/10.1093/cercor/bhm053

Norris D, McQueen JM, Cutler A (2000) Merging information in speech recognition: feedback is never necessary. Behav Brain Sci 23:299–325. https://doi.org/10.1017/S0140525X00003241

Norris D, McQueen JM, Cutler A (2003) Perceptual learning in speech. Cogn Psychol 47:204–238. https://doi.org/10.1016/S0010-0285(03)00006-9

Norris D, McQueen JM, Cutler A (2016) Prediction, Bayesian inference and feedback in speech recognition. Lang Cogn Neurosci 31(1):4–18. https://doi.org/10.1080/23273798.2015.1081703

Obleser J, Eisner F (2009) Pre-lexical abstraction of speech in the auditory cortex. Trends Cogn Sci 13(1):14–19. https://doi.org/10.1016/j.tics.2008.09.005

Raizada RD, Poldrack RA (2007) Selective amplification of stimulus differences during categorical processing of speech. Neuron 56(4):726–740. https://doi.org/10.1016/j.neuron.2007.11.001

Rauschecker JP, Scott SK (2009) Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. N Neurosci 12(6):718–724. https://doi.org/10.1038/nn.2331

Reinisch E, Weber A, Mitterer H (2013) Listeners retune phoneme categories across languages. J Exp Psychol Hum Percept Perform 39:75–86. https://doi.org/10.1037/a0027979

Reinisch E, Wozny D, Mitterer H, Holt LL (2014) Phonetic category recalibration: what are the categories? J Phon 45:91–105. https://doi.org/10.1016/j.wocn.2014.04.002

Remez RE, Rubin PE, Pisoni DB, Carell TD (1981) Speech perception without traditional speech cues. Science 212:947–950

Rogers JC, Davis MH (2017) Inferior frontal cortex contributions to the recognition of spoken words and their constituent speech sounds. J Cogn Neurosci 29(5):919–936. https://doi.org/10.1162/jocn_a_01096

Romanovska L, Janssen R, Bonte M (2019) Reading-induced shifts in speech perception in dyslexic and typically reading children. Front Psychol 10:221. https://doi.org/10.3389/fpsyg.2019.00221

Samuel AG, Frost R (2015) Lexical support for phonetic perception during non-native spoken word recognition. Psychon Bull Rev 22(6):1746–1752. https://doi.org/10.3758/s13423-015-0847-y

Samuel AG (2016) Lexical representations are malleable for about one second: Evidence for the non-automaticity of perceptual recalibration. Cognitive Psychol 88:88–114. https://doi.org/10.1016/j.cogpsych.2016.06.007

Scott SK, Johnsrude IS (2003) The neuroanatomical and functional organization of speech perception. Trends Neurosci 26(2):100–7. https://doi.org/10.1016/S0166–2236(02)00037-1

Sjerps MJ, McQueen JM (2010) The bounds on flexibility in speech perception. J Exp Psychol Hum Percept Perform 36:195–211. https://doi.org/10.1037/a0016803

Snowling MJ (1980) The development of grapheme-phoneme correspondence in normal and dyslexic readers. J Exp Child Psychol 29:294–305. https://doi.org/10.1016/0022-0965(80)90021-1

Sohoglu E, Davis MH (2016) Perceptual learning of degraded speech by minimizing prediction error. Proc Natl Acad Sci USA 113(12):1747–1756. https://doi.org/10.1073/pnas.1523266113

Sumby WH, Pollack I (1954) Visual contribution to speech intelligibility in noise. J Acoust Soc Am 26:212–215. https://doi.org/10.1121/1.1907309

Ullas S, Hausfeld L, Cutler A, Eisner F, Formisano E (2020) Neural correlates of phonetic adaptation as induced by lexical and audiovisual context. J Cogn Neurosci:1–14. https://doi.org/10.1162/jocn_a_01608

Ullas S, Formisano E, Eisner F, Cutler A (2020a) Interleaved lexical and audiovisual information can retune phoneme boundaries. Atten Percept Psychophys 82:2018–2026. https://doi.org/10.3758/s13414-019-01961-8

Ullas S, Formisano E, Eisner F, Cutler A (2020b) Audiovisual and lexical cues do not additively enhance perceptual adaptation. Psychon Bull Rev 27:707–715. https://doi.org/10.3758/s13423-020-01728-5

Van der Zande P, Jesse A, Cutler A (2014) Hearing words helps seeing words: a cross-modal word repetition effect. Speech Commun 59:31–43. https://doi.org/10.1016/j.specom.2014.01.001

Van Laarhoven T, Keetels M, Schakel L, Vroomen J (2018) Audio-visual speech in noise perception in dyslexia. Dev Sci 21(1):e12504. https://doi.org/10.1111/desc.12504

Van Linden S, Vroomen J (2007) Recalibration of phonetic categories by lipread speech versus lexical information. J Exp Psychol Hum Percept Perform 33(6):1483–1494. https://doi.org/10.1037/0096-1523.33.6.1483

Van Linden S, Vroomen J (2008) Audiovisual speech recalibration in children. J Child Lang 35(4):809–822. https://doi.org/10.1017/S0305000908008817

Vroomen J, Baart M (2009) Recalibration of phonetic categories by lipread speech: measuring aftereffects after a twenty-four hours delay. Lang Speech 52:341–350. https://doi.org/10.1177/0023830909103178

Vroomen J, van Linden S, Keetels M, de Gelder B, Bertelson P (2004) Selective adaptation and recalibration of auditory speech by lipread information: dissipation. Speech Commun 44:55–61. https://doi.org/10.1016/j.specom.2004.03.009

Vroomen J, van Linden S, Baart M (2007a) Lipread aftereffects in auditory speech perception: measuring aftereffects after a twenty-four hours delay. In: Vroomen J, Swerts M, Krahmer E (eds) Auditory-visual speech processing, Hilvarenbeek, p P05

Vroomen J, van Linden S, de Gelder B, Bertelson P (2007b) Visual recalibration and selective adaptation in auditory-visual speech perception: contrasting build-up courses. Neuropsychologia 45(3):572–577. https://doi.org/10.1016/j.neuropsychologia.2006.01.031

Winn M (2018) Speech: it's not as acoustic as you think. Acoust Today 14(2):43–49

Xie X, Myers EB (2017) Learning a talker or learning an accent: acoustic similarity constrains generalization of foreign accent adaptation to new talkers. J Mem Lang 97:30–46. https://doi.org/10.1016/j.jml.2017.07.005

Zhang X, Samuel AG (2015) Perceptual learning of speech under optimal and adverse condition. J Exp Psychol Hum Percept Perform 40(1):200–217. https://doi.org/10.1037/a0033182

Chapter 8
Development of Speech Perception

Judit Gervain

Abstract Infants start their journey into language as universal listeners but by the end of the first year of life they become native language experts, as their perceptual systems and brains attune to the sound patterns of their native language(s). This chapter describes this attunement process and its neural correlates. Speech is the auditory medium that allows us to externalize language. Speech perception and language acquisition are thus tightly connected, especially during development. While focusing primarily on the development of speech perception, this chapter, therefore, necessarily touches upon the growth of language more generally. It discusses the major milestones of this developmental trajectory in chronological order, starting out with prenatal experience and newborns' speech perception abilities, and following the attunement process in phoneme and tone perception during the first year of life, early word learning and the prosodic bootstrapping of grammar during the toddler years.

Keywords Newborn · Infant · Prenatal experience · Postnatal experience · Perceptual attunement · Perceptual reorganization · Native language · Critical period · Neural plasticity · Language input

8.1 Introduction

Speech perception undergoes dramatic changes during the first years of human development. Infants are born with speech perception abilities that allow them to acquire any of the world's languages. After months of experience, these initially broadly based, universal abilities get tuned to the native language(s). This attunement process implies a reorganization and/or narrowing of perceptual categories, with the maintenance or refinement of native sound categories, and a loss or decrease

J. Gervain (✉)
DPSS, Università degli Studi di Padova, Padova, Italy

INCC, CNRS and Université de Paris, Paris, France
e-mail: judit.gervain@unipd.it

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022

L. L. Holt et al. (eds.), *Speech Perception*, Springer Handbook of Auditory Research 74, https://doi.org/10.1007/978-3-030-81542-4_8

in sensitivity to non-native ones. At the neural level, it is accompanied by an increasingly focal brain specialization for native language processing.

This chapter describes this language attunement process and its neural correlates. Importantly, speech is the auditory medium that allows us to externalize language. Speech perception cannot thus be described without reference to language, the representation and rule system humans possess. This is particularly true in development. Hearing infants only have access to speech to learn language; they receive no formal or explicit instructions about the words or rules of their native languages. Yet, they successfully acquire the lexicon and the grammar of their mother tongue in the span of just a few years with amazing ease and efficiency. This fact has led to the assumption that the sound patterns of language are intimately intertwined with the other levels of language such as grammar and lexicon. Correlations between the sound patterns and abstract linguistic regularities allow infants to use speech to learn about or "bootstrap" the grammar and the lexicon (Morgan and Demuth 1996). These abstract acquisitions in turn help infants further fine-tune their perception of the speech signal.

Speech perception and language acquisition are thus tightly connected and interact synergistically. Empirical evidence for these connections is steadily increasing (Werker 2018; Swingley 2021). Given this interactive view of speech and language, this chapter, while focusing primarily on the development of speech perception, will necessarily touch upon the growth of language more generally.

The ultimate mechanisms of language development have long been debated, with some theories arguing for genetically endowed factors (Lenneberg 1967; Chomsky et al. 2002), and others for experiential and learning-based ones (Elman et al. 1997; Tomasello 2000). With the advent of brain imaging and especially genetic and epigenetic studies (Werker and Hensch 2015), it is becoming increasingly clear that biologically endowed and experiential factors are likely to act synergistically and rely on each other to bring about language development. The strict binary dichotomy of the traditional nature-nurture debate is thus replaced by a more integrative view of the factors that contribute to the developmental changes in speech perception and language acquisition.

Related to this new perspective, the notion of critical periods in speech perception and language acquisition has been revisited. The original proposal (e.g., Lenneberg 1967) was based on observations about language development failing to reach native-like competence when acquisition starts late, typically after puberty. One example comes from cases of feral children. These children are raised in social deprivation and thus not spoken to. They only recover language if they are discovered and introduced to language before puberty (Curtiss et al. 1974). Immigrants to the USA constitute another example. They have been observed to achieve native competence in English if they arrived before age 8–10 years (Johnson and Newport 1989). But since language learning remains possible throughout the life span, with large individual variations in outcome, the notion of critical periods was sometimes debated. With a better understanding of the experiential, molecular, and neural mechanisms controlling critical period phenomena, both in humans (Weikum et al. 2012; Gervain et al. 2013) and in animals (Weaver et al. 2004; Hensch 2005), where invasive studies can be carried out to close or re-open critical periods, the notion of critical periods has taken on a new, biologically better-defined meaning. How brain plasticity changes during speech perception and language development as a result of the closure of the critical period has thus recently received considerable attention (Werker and Hensch 2015).

This chapter follows the development of speech perception chronologically. It starts by reviewing newborn infants' universal abilities and then following how these abilities narrow down and attune to the native language. Such attunement processes operate at different levels of phonological organization, from global ones such as rhythm to smaller units such as phonemes, syllables, tones, or words.

8.2 Newborns' Speech Perception Abilities

In the light of the theoretical debates on the role of innate and learned factors in language acquisition, newborn infants' abilities have received considerable attention. These abilities were viewed as the best approximation we can methodologically get of the "initial state," that is, the state of the perceptual and language learning system before experience begins. Since then, evidence has gathered that fetuses learn from the speech input they receive in utero, as hearing becomes operational between the 24th and 28th week of gestation (Eggermont and Moore 2012). Newborns thus show universal, broadly based speech perception abilities not specific to any language yet, as well as abilities that are already shaped by prenatal experience.

8.2.1 Newborns' Universal Speech Perception Abilities

The auditory system is immature at birth and continues developing into late childhood/early adolescence (Moore 2002; Eggermont and Moore 2012). Yet, newborns show a variety of speech perception abilities, many of which are universal and broadly based, allowing newborns to discriminate most sound patterns found in the world's languages and thus enabling them to start learning any language.

Newborns' first task is to identify speech among the sounds present in their environment. Newborns and 2-month-old[1] infants can indeed recognize speech, and show a strong preference for it over equally complex sine wave analogs (Vouloumanos and Werker 2004). However, the category "speech" may be relatively broad at birth, roughly corresponding to primate vocalizations, as newborns show equal preference for human speech and rhesus monkey vocalizations (Vouloumanos et al. 2010). Yet,

[1] Throughout this chapter, the ages specified indicate the ages of infants tested in the cited studies. While an individual infant of a given age may not show a specific ability, on average, infants as a group do so at the age indicated.

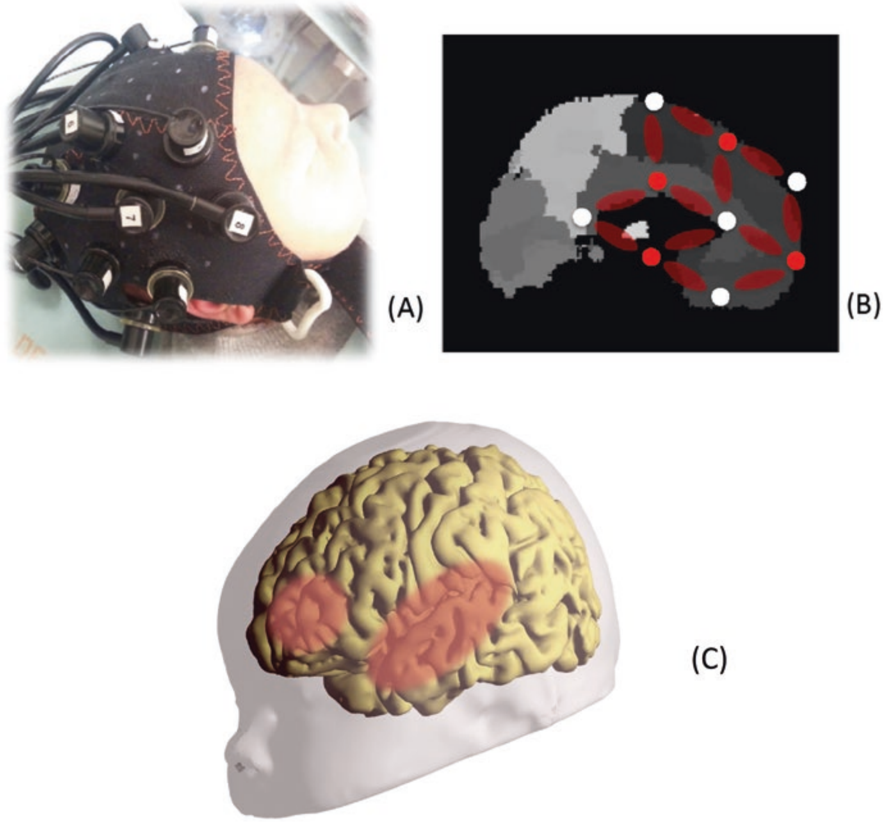

Fig. 8.1 Language in the newborn brain. (**a**) A near-infrared spectroscopy (NIRS) brain imaging cap on a newborn infant head, and (**b**) the corresponding sensor space overlaid on a newborn structural scan. NIRS is a commonly used imaging technique to localize the language network in infants' and young children's brains. (Images adapted from Abboub et al. (2016)). (**c**) A schematic illustration of the brain areas that have been found to be involved in a variety of speech and language processing tasks in young infants using brain imaging

by 3 months, infants show a unique preference for speech over both sine wave analogs and monkey calls (Vouloumanos et al. 2010).

Analogously with this behavioral preference for speech, the brains of young infants are specialized for speech processing. Three-month-old infants, full-term neonates, and even premature newborns activate similar brain networks as adults (the superior and middle temporal gyri, the inferior parietal cortex, and the inferior frontal gyrus, including Broca's area; Fig. 8.1) in response to language, but not to non-linguistic controls such as backward speech (Dehaene-Lambertz et al. 2002; Peña et al. 2003; Mahmoudzadeh et al. 2013). As discussed in Sect. 8.2.2, prenatal experience may already shape this specialization.

In addition to identifying speech in their environment, newborns are able to discriminate languages from one another, even if they never heard them before, on the basis of the languages' different rhythms (Mehler et al. 1988; Nazzi et al. 1998). Language rhythm was first quantified along three acoustic dimensions (Ramus et al. 1999): %V, which is the relative proportion of vowels in the speech signal as well as ΔC and ΔV, which are the variability in the length of consonant and vowel clusters, respectively. In the space defined by these variables (Fig. 8.2), languages cluster together into what was traditionally called the rhythm classes of languages. While language rhythmic is best understood as a continuum (Nespor 1990), the classes are still often used. They are named after the time unit that was once believed to be the basic isochronous element in the languages belonging to a given class (Abercrombie 1967). Japanese is thus a mora-timed language the mora is a unit larger than the phoneme, but smaller than the syllable). (Mora-timed language have the highest %V and the lowest ΔC values. Syllable-timed languages, such as French or Italian, still have relatively high %V, but medium ΔC values. Stress-timed languages such as English and Polish, in which the unit was believed to be the interval between stressed syllables, have lower %V and higher ΔC. Subsequently, other metrics have also been proposed to quantify rhythm (Grabe and Low 2002; Dellwo 2006). They are better at accounting for speech rate differences across speakers.

Rhythmic discrimination does not require familiarity with the languages. Newborns prenatally exposed to French are able to discriminate between English and Japanese, for instance, as can tamarin monkeys (Ramus et al. 2000). This

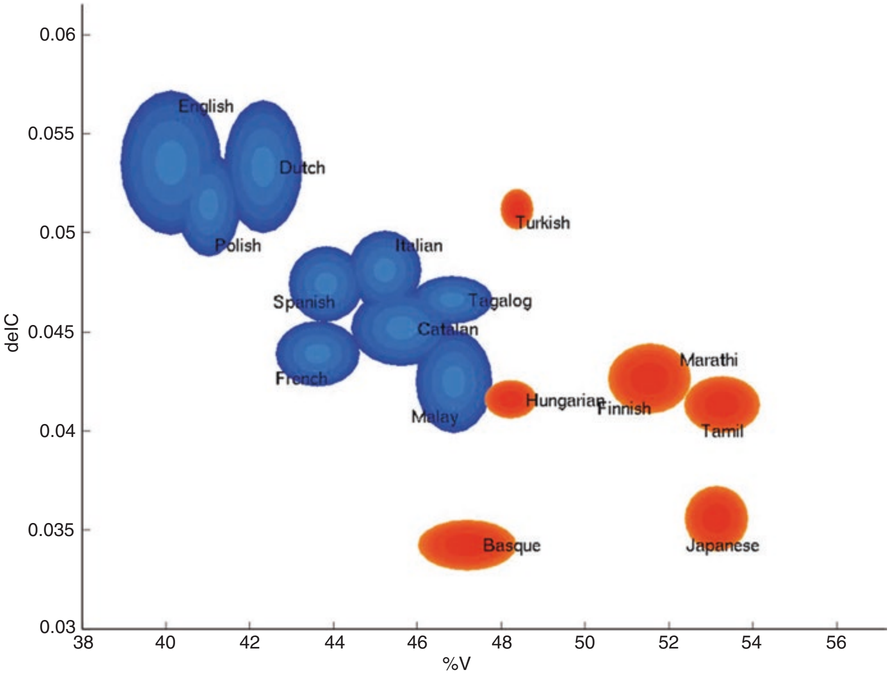

Fig. 8.2 Different languages in the space defined by %V and ΔC, two measures of speech rhythm. (Adapted from Mehler et al. (2004))

finding suggests that rhythmic discrimination might be a general property of the primate or mammalian auditory system, independent of experience with language or the ability to acquire it.

One important implication of newborns' ability to discriminate languages on the basis of rhythm is that infants born into a multilingual environment can immediately detect that they are being exposed to different languages, at least if the languages are rhythmically different. Bilingual newborns have indeed been shown to be able to discriminate their two languages from a third, rhythmically different language (Byers-Heinlein et al. 2010).

In addition to their abilities to identify speech in different languages in their environment, newborns are also able to process smaller units within the speech signal. Behavioral results show, for instance, that infants readily detect the acoustic cues correlated with the beginnings and ends of words (Christophe et al. 1994). They have also been found to be sensitive to syllables within words (Sansavini et al. 1997), readily discriminating words in which the stress is on the first syllable (e.g., ***doc**tor*) vs. those in which it is on the final one (e.g., *gui**tar***). Interestingly, however, infants cannot tell apart words with different numbers of phonemes if the number of syllables is the same.

During the first months of life, infants can also discriminate many of the phonemes appearing in the world's languages, as has been shown both behaviorally (Eimas et al. 1971) and electrophysiologically (Dehaene-Lambertz and Baillet 1998). This universal discrimination repertoire is one of the hallmarks of young infants' broad-based abilities, allowing them to learn any language to which they are exposed. Interestingly, chinchillas and songbirds can also discriminate phonemes at similar acoustic boundaries (Kuhl 1981, 1986), suggesting that phoneme perception builds on evolutionarily available perceptual abilities. It is, therefore, available to young infants prior to experience. How this ability is then shaped by language experience will be discussed in Sect. 8.3.

What features of the acoustic signal of speech newborns rely on to discriminate phonemes is only now starting to be investigated. When processing speech presented in silence (Chap. 4, Tune and Obleser; Chap. 7, Ullas, Bonte, Formisano, and Vroomen), adults can discriminate phonemes even on the basis of a strongly impoverished speech signal retaining only the slowest modulations (<16 Hz) of the amplitude envelope (Drullman 1995; Shannon et al. 1995), mimicking the signal available to cochlear implant users. A brain imaging study (Cabrera and Gervain 2020) investigated how newborns process consonant contrasts in three acoustic conditions. One condition consisted of the intact speech signal. In the second condition, the full envelope was preserved, but the temporal fine structure was suppressed. In the third condition, only the slowest modulations of the envelope were preserved. This study showed that newborns were able to discriminate consonants in all three conditions, suggesting that, like for adults, the slowest modulation cues of the speech envelope are sufficient for young infants to process the finest details of the speech signal. Interestingly, however, the three conditions activated different brain areas, suggesting early neural specialization for different aspects of the speech signal. Specifically, the condition containing the full envelope evoked a more left-lateralized activation than the slow envelope condition, suggesting adult-like brain specialization for the different aspects of the speech signal early in life.

Newborns are sensitive not only to sound patterns, but also to structural regularities in the speech input. Thus, they can detect repetition-based patterns such as ABB (e.g., “mu ba ba,” “pe na na,” etc.) or AAB (e.g., “ba ba mu,” “na na pe,” etc.), and discriminate them from otherwise similar random sequences such as ABC (e.g., “mu ba ge,” “pe na ku,” etc.; Gervain et al. 2008), or from one another (e.g., ABB vs. AAB; Gervain et al. 2012). Furthermore, this ability involves the bilateral temporal and left frontal areas, including Broca’s area, implying that the infant language network is already similar to the adult one.

In sum, newborns already possess a repertoire of basic auditory, speech perception, and learning mechanisms, many of them shared with chinchillas or songbirds, that allow them to crack the linguistic code in any language they encounter in their environment, independently of prenatal speech experience.

8.2.2 Newborns’ Speech Perception Abilities Shaped by Prenatal Experience

A growing number of studies suggests that newborns also have abilities shaped by experience with speech heard in the womb in addition to their universal perceptual sensitivities. Auditory experience with speech starts in the womb. But the intrauterine speech signal is different from the signal heard outside of the womb. Maternal tissues and the amniotic fluid act as low-pass filters at about ~400–800 Hz, although the exact values can only be estimated from computational simulations and recordings in pregnant sheep models (Gerhardt et al. 1990; Lecanuet and Granier-Deferre 1993; DeCasper et al. 1994). As a result of this low-pass filtering, the melody and rhythm of speech, which jointly define the prosody of a language, are preserved. At the same time, the fine details necessary to identify individual sounds, especially consonants, are suppressed. As a result, words are mostly unintelligible. Infants’ first experience with speech thus consists mainly of prosodic information (Gervain 2015, 2018).

This prenatal experience already shapes fetuses’ speech perception abilities. Newborns recognize and prefer their mother’s voice over other female voices (DeCasper and Fifer 1980). They also show a preference for their native language over other languages (Mehler et al. 1988; Moon et al. 1993) and a story heard frequently in the womb over other stories (DeCasper et al. 1994; Kisilevsky et al. 2009).

Relevant to language acquisition, fetuses learn even more specific details about their native language. Since vowels have the highest energy in the speech signal and are the main carriers of prosody, some vowels seem to be already learned in part prenatally. Indeed, newborns show a preference for a vowel they did not hear prenatally over a native one (Moon et al. 2013). Fetuses can also learn word-level prosodic information (Partanen et al. 2013), readily detecting a change in lexical pitch trained prenatally, which untrained newborns do not recognize.

Infants also show evidence of learning prenatally about the prosody of larger linguistic units, such as utterances. Languages vary as a function of what acoustic cues mark prosodic prominence in their phonological phrases. In some languages, such as French or English, prominence is carried by a durational contrast, meaning that the prominent element is lengthened as compared to the non-prominent one (e.g., in the phrase *to Rome*, the vowel of the prominent content word *Rome* is longer than the vowel of the non-prominent preposition). In these languages, the prominent element typically occupies a phrase-final position, so phrases have an iambic prosodic pattern.

In other languages, like Japanese or Turkish, prominence is indicated by a pitch/intensity contrast. In these languages, prominence is phrase-initial (i.e., trochaic), so the higher or louder element is at the phrase onset (e.g., in the Japanese phrase, *Tokyo kara*, which literally translates as “Tokyo to” and means “to Tokyo,” the first vowel of *Tokyo* is higher than the vowel of the word *kara*). This alternation of prominent and non-prominent elements creates a rhythmic prosodic pattern readily perceivable even by listeners who are unfamiliar with a given language (Langus et al. 2016).

Newborn infants also seem to pick up on this pattern from their prenatal exposure (Abboub et al. 2016). Newborns were presented with pairs of pure tones contrasting in duration, pitch, or intensity. In one condition, the pairs were consistent with the patterns found in natural languages. Specifically, they were iambic pairs (e.g., short-long) for the durational contrast, like in the English example *to Rome*, and trochaic pairs for the pitch (e.g., high-low) and intensity contrasts (e.g., loud-soft), like in the Japanese example *Tokyo kara*. In the other condition, the pairs were inconsistent with these patterns, so trochees (e.g., long-short) for duration, iambs for pitch/intensity (e.g., low-high/soft-loud). The newborn brain showed a greater response to the inconsistent patterns, but only for the acoustic cue that marks prosodic prominence in the language the infants were exposed to prenatally.

Newborns’ knowledge of the native prosody might even go beyond perception. It has been suggested that newborns’ communicative cries reproduce the prosodic patterns of their native language (Mampe et al. 2009). Indeed, German newborns’ cry patterns were found to have initial prominence, just like typical declarative utterances in German do. By contrast, French babies’ cries were prominence-final mimicking the prosodic contour characteristic of French utterances. Recently, these findings received some criticism on the basis of the statistical analyses used (Gustafson et al. 2017). But in a subsequent study, automated classification algorithms could separate cries from French-, Arabic-, and Italian-exposed newborns according to native language (Manfredi et al. 2019). If confirmed to be true, these findings would indicate that prenatal experience is sufficiently strong to shape even production.

Prenatal experience also shapes the brain specialization for language processing. Newborns’ brain responses to speech in the native language are different from responses to non-native languages. Some studies find stronger left-lateralization for the native language played forward than backward. The response involves the same regions as in adults, mainly the middle and superior temporal areas and the inferior frontal regions, including Broca's area (Peña et al. 2003). When directly comparing neonates' responses to their native language vs. a non-native tongue, some studies reproduced the left hemisphere advantage for forward vs. backward speech in the native language, but no hemispheric difference in a non-native language (Sato et al. 2012; May et al. 2017). However, other studies found no hemispheric differences for either language, but an overall advantage for the native language over the non-native one (May et al. 2011). The lateralization issue notwithstanding, all studies found a difference between the responses to the native language and unfamiliar languages, strongly suggesting that the brain network for speech processing is sculpted by prenatal experience. Furthermore, this network is already specialized for processing speech, as a whistled language does not activate it despite being a human communication system (May et al. 2017).

In sum, despite their immature auditory system, newborns show sophisticated speech perception abilities. Some of these abilities are universal, allowing infants to start acquiring any language. Others, by contrast, are already tuned to the prenatal experience with speech, especially prosody, infants received in the womb.

8.3 Perceptual Attunement to the Native Language

After birth, experience with the full-band speech signal begins and infants start to learn about the sound patterns specific to their native language(s). The experience induces a perceptual reorganization or attunement to the native language, whereby the ability to discriminate linguistic contrasts found in the language(s) heard is maintained or even improved, while the ability to distinguish most contrasts that do not appear in the input decreases (Werker and Tees 1984; Kuhl 2004). This reorganization may show different developmental trajectories in different areas of language. In some, a simple decrease in non-native discrimination (with a concomitant improvement in native discrimination) is observed (Werker and Tees 1984). Other areas are characterized by a U-shaped trajectory, where after the initial ability to discriminate certain contrasts and a subsequent decline, the ability re-emerges (Weikum et al. 2007, 2013). This newly emerging ability is sometimes underpinned by mechanisms that are different in nature than those underlying the initial ability. The initial ability is acoustic, closely linked to acoustic discriminability, whereas the emerging one is shaped by native language experience.

Attunement to the native language comes about as an intricate interplay between experience and perceptual/cognitive mechanisms. Attunement is accompanied by reorganization at the neural level, with increasingly focal, lateralized brain specialization for native language processing. This, in turn, is tied to developmental changes in brain plasticity, brought about by the changing balance of inhibitory and excitatory connections, ultimately linked to synaptogenesis, myelination, and pruning (Casey et al. 2000; Tierney and Nelson 2009; Haartsen et al.

2016)—neurophysiological mechanisms that are particularly active between the prenatal period and adolescence (although they remain operational throughout the lifespan).

It is not surprising, therefore, that attunement to experience is not unique to speech and language. Similar phenomena have been observed in other perceptual domains, for instance, in face perception (Pascalis et al. 2002; Maurer and Werker 2014).

The general principle of attunement to the native language(s) notwithstanding, different areas of speech perception undergo different narrowing trajectories, and some non-native contrasts, such as click consonants or some tonal contrasts, remain discriminable throughout life. The following sections discuss each of these developmental trajectories in turn.

8.3.1 Linguistic Rhythm

The rhythmic discrimination ability observed in newborns provides a good explanation of how bilinguals exposed to rhythmically different languages may distinguish their languages from birth. However, some bilinguals are exposed to rhythmically similar languages, and newborns cannot discriminate these from one another at birth. From what age and on what basis do bilinguals of rhythmically similar languages start distinguishing between their mother tongues? Bilingual infants growing up with Spanish and Catalan, two rhythmically similar languages, were found to succeed on this discrimination task at 4 months of age (Bosch and Sebastian-Galles 1997). Monolingual Spanish and monolingual Catalan infants also performed similarly. Basque-Spanish bilinguals were also shown to distinguish the two languages between 3.5 and 4 months (Molnar et al. 2013). While monolingual Basque infants behaved similarly, interestingly, the monolingual Spanish infants in this study only discriminated the two languages when habituated to Basque, but not when habituated to Spanish. This asymmetry may be related to the geopolitical dominance of Spanish in the Spanish Basque Country, the location of the study.

Taken together, the above results suggest that familiarity and experience with at least one of the languages allow discrimination even within the rhythmic group after 3–4 months of experience. Specifically, this discrimination ability may rely on familiarity with the phoneme repertoire, syllable structure, and/or phonotactic regularities of at least one of the languages.

8.3.2 Audio-Visual Speech Perception

Speech is not only heard, but also seen. A considerable amount of visual information is available in the speaker's face/head when producing speech. This information includes the position and movement of the lips, and the tongue, as well as of the eyes, eyebrows, and head, about the global features of different languages, their prosody, as well as individual phonemes (Guellaï et al. 2014; Wagner et al. 2014; de la Cruz-Pavía et al. 2020a). Adults have been shown to readily integrate such visual information with the auditory signal while processing speech (McGurk and MacDonald 1976). They can also use it to discriminate languages presented only visually (Soto-Faraco et al. 2007; Weikum et al. 2013). Furthermore, visual information also supports and augments speech perception in a non-native language or when the signal is degraded (Birulés et al. 2020). It also helps listeners segment out words from continuous speech (Mitchel and Weiss 2014) or parse the speech input to larger prosodic units (de la Cruz-Pavía et al. 2020b).

How infants use the visual correlates of speech has received increasing attention. Both monolingual and bilingual infants can readily discriminate two languages on the basis of visual speech alone at 4 and 6 months if at least one is their native language. By 8 months, only bilingual infants continue to do so (Weikum et al. 2007). This suggests that maintaining visual sensitivity helps infants in their daily task of discriminating between their two languages, a challenge that monolinguals do not face. Interestingly, this maintained perceptual sensitivity is general, as familiarity with the languages is not necessary to show successful discrimination. Indeed, both English-French and Spanish-Catalan bilinguals discriminate visual French and visual English at 8 months (Sebastián-Gallés et al. 2012). During the first 6 months of life, while their audio-visual sensitivity to speech is still broadly based, infants can also match talking faces to speech in languages that are unfamiliar to them. This ability weakens by 12 months of age when speech is adult-directed, showing perceptual narrowing (Kubicek et al. 2014a). Interestingly, 12-month-olds still succeed if the auditory stimuli used are infant-directed (Kubicek et al. 2014b).

The prosody of speech also has its visual correlates: speakers of Japanese and English produce eyebrow movements and head nods to mark phrase boundaries (de la Cruz-Pavía et al. 2020a), which adult listeners can use, in conjunction with auditory information, to parse speech into phrasal units (de la Cruz-Pavía et al. 2019). Eight-month-old, but not yet 4-month-old infants, also start to show sensitivity to these visual cues, and can integrate them with auditory prosodic information and word frequency. However, this integration process is not yet adult-like, in particular in the temporal asynchrony that infants expect and tolerate between the different cues (de la Cruz-Pavía et al. 2019).

Like in adults, infants' perception of speech in noise improves when they are provided with additional visual information (Hollich et al. 2005). A large body of work indicates that this facilitatory effect is based on infants' ability to match the auditory and visual signals at the syllable/phoneme level. Infants, for instance, can choose which of two silently talking faces articulates a syllable heard auditorily (Kuhl and Meltzoff 1982; MacKain et al. 1983; Patterson and Werker 1999, 2002). Audio-visual matching also undergoes perceptual narrowing, similarly to auditory phoneme perception (see Sect. 8.3.3). By 11–12 months of age, infants no longer match the auditory and visual signals of phonemes if those are not found in their native language (Pons et al. 2009; Danielson et al. 2017).

Interestingly, what cues infants rely on in a talking face also changes during development, reflecting the underlying perceptual reorganization. While infants and adults mainly look at the eyes of a (talking) face (Hunnius and Geuze 2004; Viola Macchi et al. 2004), around 8–12 months of age, infants shift their attention to the mouth region, and this shift is more pronounced if infants hear non-native speech (Lewkowicz and Hansen-Tift 2012; Kubicek et al. 2013) or if they are bilingual (Pons et al. 2015). This shift corresponds to the developmental timeline of perceptual narrowing to the native language, and might thus reflect infants' strategy to seek out visual information that supports the attunement process. Indeed, by 12 months of age, infants only look to the mouth region when hearing non-native speech, but not when hearing their native language (Lewkowicz and Hansen-Tift 2012). This audio-visual reorganization may be a crucial milestone in speech perception, as children with language impairment show reduced attention to the mouth (Pons et al. 2018).

8.3.3 Phoneme Perception

Very young infants, up to about 4–6-months of age, can discriminate almost all phonemes appearing in the world's languages, even those that do not appear in their native language and that adult speakers of a different language are unable to discriminate, as has been shown both behaviorally (Eimas et al. 1971; Werker and Curtin 2005) and electrophysiologically (Dehaene-Lambertz and Baillet 1998; Kujala et al. 2004). Infants' phoneme perception, like that of adults, is categorical, especially for consonants, possibly less so for vowels (Swingley 2021). Perception is categorical when a given acoustic difference between two sounds is discriminated and treated as contrastive if it spans a phoneme boundary, but not discriminated if it falls within the boundaries of a phoneme category (even though infants are able to perceive the acoustic difference; McMurray and Aslin 2005). This universal discrimination repertoire is one of the hallmarks of young infants' broad-based abilities, allowing infants to learn any language they are exposed to.

After several months of experience with the native language, non-native sound discrimination declines (Werker and Tees 1984), while the discrimination of contrasts found in the native language is maintained or even improves (Kuhl et al. 2006; Narayan et al. 2010). This perceptual attunement toward the native sound repertoire takes place around 4–6 months for vowels (Kuhl et al. 1992) and 10–12 months for consonants (Werker and Tees 1984). The system nevertheless remains plastic for several years after attunement. It is thus possible to learn the phoneme inventory of another language until age 6–8 years (or the onset of puberty the latest), as studies with immigrants (Johnson and Newport 1989) and international adoptees suggest (Ventureyra et al. 2004; Pierce et al. 2014). Infants growing up multilingually go through the same perceptual narrowing, although for some sounds, they also show different developmental patterns (Byers-Heinlein and Fennell 2014). For instance, when a phoneme pair is distinguished in one of their languages, but not in the other, bilingual infants may go through a phase when they do not discriminate between the two sounds.

Interestingly, the ability to discriminate non-native contrasts does not always get lost. Some features of click sounds, found for example in the African language Zulu, remain discriminable to non-native adults (Best et al. 1988). This has been explained by the unusual, almost non-linguistic nature of these sounds.

Phoneme discrimination may be facilitated by systematic associations between sounds and objects, implying that the relationship between phoneme perception and word learning is mutual (Werker and Yeung 2005). Thus, 9-month-old infants can successfully discriminate a non-native sound contrast if each phoneme occurs in a nonword that is associated with an object (Yeung and Werker 2009), whereas at this age, they would already fail without the association with objects, due to perceptual attunement.

This relationship between word learning and perceptual attunement notwithstanding, the lexicon is still relatively small between 4 and 12 months of life, when perceptual attunement takes place. To explain how phonetic perception changes without a sizeable lexicon, different mechanisms based on similarity-matching and distributional learning have been proposed (Kuhl 2004; Werker and Curtin 2005). These models assume that the distributional characteristics of different phonemes in the speech signal reflect those perceptible differences that are contrastive in the language and de-emphasize differences that are not.

Existing results also point toward another factor in the development of early phoneme perception, the contribution of the motor system. In adult speech perception, a long tradition has argued for the key role of the motor system in phoneme perception (e.g., Liberman and Mattingly 1985). According to the motor theory of speech perception, the motor schemes necessary to produce a speech sound play an important role in its identification, when the sound is perceived. Whereas the original conception of a necessary role for the motor system in speech perception is not supported empirically, there is a strong case to be made for the interplay of speech perception and production in adults (Hickok et al. 2003).

This theory received relatively little attention in developmental research, since infants' motor and production skills so clearly lag behind their perceptual skills, although correlational evidence between infants' babbling/production and phoneme perception abilities has been reported (Guellaï et al. 2014; Majorano et al. 2014; Vilain et al. 2019). However, a study by Bruderer et al. (2015) has provided direct experimental evidence that the position of infants' tongue and lips may impact how they perceive speech sounds. When 6-month-old English-learning infants were tested on a non-native speech contrast produced with movement of the tongue tip, they showed successful discrimination, replicating earlier results about infants' quasi-universal ability to discriminate consonant contrasts before about 10 months (Werker and Tees 1984). However, when the same infants had to accomplish the same task with a teething toy in their mouth that specifically inhibited tongue tip movements, infants failed. This effect was specific as teething toys with other shapes not impacting the position of the tongue tip, but that of the lips (lip spreading), did not prevent infants from making the discrimination. These results are remarkable in that the infants tested were preverbal, barely starting to babble, and had no experience with the phoneme contrast tested, yet showed the influence of the position of the articulators on their discrimination abilities, providing experimental evidence for the auditory-motor link at the earliest age in development (Bruderer et al. 2015).

8.3.4 Tone Perception

Perceptual attunement to the native language has been observed not only for phonemes, but also for lexical tones, the linguistic function of which is similar to that of phonemes in that they are minimal units of distinguishing meaning in tonal languages like Thai or Mandarin Chinese. The perception of lexical tone follows a similar attunement pattern to phonemes, with infants exposed to tone languages maintaining discrimination, and unexposed infants losing it over the second half of the first year of life, although some studies paint a more complex picture. The acoustic distance between the tested tone pairs seem to play a role, and some studies have also shown U-shaped developmental patterns whereby the ability to discriminate non-native tones returns after a drop even in non-exposed infants (Mattock and Burnham 2006; Liu and Kager 2012).

8.3.5 Increasing Brain Specialization as a Correlate of Perceptual Attunement

Perceptual attunement observed behaviorally is paralleled by increasing brain specialization at the neural level. Brain activation in response to language features found in the native language becomes more focal and more lateralized, with phoneme discrimination lateralizing to the left hemisphere and prosody-related discrimination lateralizing to the right (Minagawa-Kawai et al. 2011). As an example, 3-month-old Parisian infants respond bilaterally to Parisian French, their native dialect, and Quebecois French, a non-native regional dialect. Their brain responses to the two dialects are similar. By 5 months of age, however, Parisian infants show a differential response to the native dialect, which is left lateralized and more focal than 3-month-olds' responses (Cristia et al. 2014).

The processing of smaller linguistic units also gets lateralized. Lexical pitch accent contrasts, such as high-low vs. low-high, are readily discriminated both by 4-month-old and 10-month-old Japanese infants behaviorally, but brain imaging reveals important underlying differences in processing (Sato et al. 2010). The younger infants process the contrast bilaterally, with the activation patterns closely resembling their brain responses to pure tones, suggesting that processing is mostly based on the acoustic properties of the stimuli. The older infants, by contrast, show a left-lateralized discrimination response to the pitch accent contrast, the intensity of which is greater than that of the response to pure tones, indicating more specialized and more linguistically based processing. Similarly, Japanese infants have been found to discriminate the vowel duration contrast such as the short and long /a/ in Japanese (Minagawa-Kawai et al. 2007) at 6–7 months, not at 10–11 months, and then again from 13 to 14 months onward until adulthood. The initial discrimination response at 6–7 months is bilateral, whereas it becomes left lateralized from 13 to 14 months on, after reorganization.

These results clearly illustrate the development of the brain specialization for the native language. Processing and discrimination are initially acoustically based, and hence more bilateral. During reorganization, response patterns may vary or even weaken, and then re-emerge as more linguistic in nature, indexed by their more focal and lateralized location.

8.4 Learning Word Forms

As infants attune to their native sound repertoire, they also start acquiring their first words (Jusczyk and Aslin 1995; Tincoff and Jusczyk 1999; Bergelson and Swingley 2012). Speech is a continuous signal in which words are not systematically separated by pauses or other acoustic cues in a fully reliable manner. Thus, one challenge infants face when learning words is to segment out the possible word form candidates from the speech stream so that they can associate them with appropriate meanings. Here, we will only be addressing the word segmentation problem. How infants associate the extracted word forms with meaning goes beyond speech perception; the reader is, therefore, referred to existing overviews on this issue (Markman 1994; Golinkoff et al. 2000).

What cues do infants rely on to identify possible word forms? Several types of cues have been identified and statistical cues have received considerable attention. It has long been recognized that the statistical regularities of phoneme co-occurrences are also reliable indicators of word boundaries (Harris 1955; Brent and Cartwright 1996). Thus, the syllable /ti/ follows the syllable /pri/ with a greater probability than /bei/ follows /ti/, for example, as in the sequence *pretty baby*, because /pri/ and /ti/ frequently co-occur in the same word, while the adjective *pretty* might be followed by a large number of other words; thus, /ti/ and /bei/ do not necessarily co-occur. Saffran et al. (1996) and much subsequent work have shown that infants are able to pick up such regularities and use them to segment speech. Thus, infants expect a boundary between words when the probabilities between syllables are low.

Other segmentation cues have also been proposed in the literature. First, infants might rely on typical stress patterns, such as the strong-weak (trochaic) pattern commonly found in English content words (e.g., *‘**doc**tor*). This is plausible, because infants have been shown to develop sensitivity to the stress patterns typical of their native language between 6 and 9 months (Jusczyk and Aslin 1995; Morgan and Saffran 1995; Morgan 1996). Such a stress-based segmentation mechanism, called the Metrical Segmentation Strategy (Cutler and Carter 1987; Cutler 1994), has been shown to underlie 7.5-month-old English-learning infants' recognition of familiar words. In a series of studies, Jusczyk et al. (1999b) have shown that when familiarized with trochaic English words (e.g., ***'doc****tor*, ***'can****dle*), 7.5-month-olds prefer passages containing these words over passages that do not contain them. This preference is specific to the trochaic word form, because passages containing only the first strong syllables of the words (e.g., *dock*, *can*) did not give rise to a similar preference.

Moreover, by this age, English infants use language-specific stress cues to segment words from the ongoing speech stream. When presented with a continuous stream of syllables consisting of a consonant and a vowel, where every third syllable was stressed, 7- and 9-month-olds treated as familiar only those trisyllabic sequences that had initial stress (soft-weak-weak). Infants showed no recognition of trisyllabic sequences that were not trochaic (weak-soft-weak or weak-soft-soft; Curtin et al. 2001). The Metrical Segmentation Strategy also predicts that weak-strong, that is, iambic, words (e.g., *gui****'tar***) might initially be missegmented, which turns out to be the case (Jusczyk et al. 1999b).

A second possible language-specific cue to segmentation is phonotactics, that is, the regularities of how phonemes can be combined in a language. Knowing that the sequence /br/ is frequent in the initial positions of English words, while /nt/ typically appears at the end can help the learner posit word boundaries. Indeed, Saffran and Thiessen (2003) tested the acquisition of phonotactic constraints using segmentation as the experimental task. Using a different approach, Mattys et al. (1999) explored how 9-month-old English-learning infants' knowledge of English phonotactics helps them posit word boundaries. They familiarized infants with non-sense words consisting of a sequence of consonants (C) and vowels (V) in the following order CVCCVC. The CC cluster in the middle was either frequent word-internally in English, but infrequent across word boundaries (e.g., /nk/) or vice versa (e.g., /nt/). Infants segmented the non-sense words into two monosyllables when the CC cluster was infrequent word-internally and frequent across word boundaries. No segmentation was observed for the other type of CC clusters, indicating that 9-month-old infants can use their phonotactic knowledge to assist them in word segmentation (Mattys and Jusczyk 2001). Phonotactic biases, that is frequent, typical phonotactic patterns that appear in a language, can also aid segmentation. Thus, infants have been found to be perceptually sensitive to the Labial-Coronal bias by 10 months of age in languages, like French, in which this bias is present in the lexicon. The Labial-Coronal bias means that in the vocabulary of many languages, words with two consonants in them are such that the initial consonant is labial and the subsequent one is coronal, rather than the other way round. Studies suggest that infants show a preference for words that are Labial-Coronal over words with the opposite pattern (Nazzi et al. 2009). Similarly, infants learning languages with vowel harmony, but not those exposed to a non-harmonic language, are sensitive to this property of their native language by about 7–13 months of life (Altan et al. 2016; Gonzalez-Gomez et al. 2019). Vowel harmony is the tendency found in some languages for vowels within a word to be similar to one another in some feature. For instance, in Hungarian, vowels harmonize in frontness/backness (e.g., the word *ajtó* "door" only has back vowels, while the word *edény* "dish" only has front vowels). Sensitivity to such biases can help infants identify possible word forms in the input, and thus contribute to segmentation.

A third segmentation cue comes from the distributions of allophones, different realizations of the same phoneme in different positions within words. In English, aspirated stop consonants, consonants produced with a small puff of air, appear in the initial positions of stressed syllables (Church 1987), their unaspirated allophones appear elsewhere. Consequently, aspirated stops are good cues to word onsets. Because infants as young as 2 months are able to discriminate the different allophones of a phoneme (Hohne and Jusczyk 1994), it is not implausible to assume that they might use them as cues for segmentation. Indeed, Jusczyk et al. (1999a) have shown that at 9 months, infants are able to posit word boundaries (e.g., *night rates* vs. *nitrates*) based on allophonic and distributional cues, and at 10.5 months, allophonic cues alone are sufficient for successful segmentation.

In the speech input that infants receive, the above cues never occur in isolation. Therefore, it is important to understand how these cues interact during the actual process of language acquisition. Work by Mattys, Jusczyk, and colleagues (Mattys et al. 1999; Mattys and Jusczyk 2001) has shown that when stress and phonotactic cues are pitted against each other, that is, provide conflicting information about word boundaries, 9-month-old infants prefer to rely on stress cues. When stress and statistical information are contrasted, 6-month-olds follow the statistical information (Saffran and Thiessen 2003), while 8-month-olds rely more on stress (Johnson and Jusczyk 2001). This developmental trajectory might indicate a shift from universal to more language-specific strategies, reflecting infants' growing knowledge of the specifics of their native phonology.

By the end of the first year of life, infants thus develop powerful strategies to segment the continuous speech stream into words and start building a small vocabulary of candidate word forms. This development happens in parallel with the attunement to the native phoneme repertoire, and the two processes mutually influence each other. As a consequence, the native phoneme categories only become stable enough to support word learning in highly demanding contexts by about 18 months of age, but not yet at 14 months (although they are sufficiently reliable to allow word learning when context and task demands are low). Indeed, while 14-month-old infants can reliably learn to associate one non-sense word with a novel object and another non-sense word with another novel object when the non-sense words are phonologically distinct, such as "lif" and "neem," they have difficulty with minimal pairs. Minimal pairs are word that differ in a single phoneme, such as "bih" and "dih," and succeed in the latter task only by 18 months (Stager and Werker 1997). By about this age, they seem to encode even subsegmental detail in word forms (White and Morgan 2008).

Infants thus first show evidence of recognizing some word forms and reliably associate them with possible meanings between 6 and 9 months. Between this age and about 18 months, as their native phoneme repertoire stabilizes and they develop language-specific strategies for segmenting words, they start to build a sizeable lexicon as they become expert word learners during the second year of life.

8.5 Prosodic Perception

Infants' first linguistic experience largely consists of the rhythm and melody of the language(s) spoken by their mothers before birth (Gervain 2018). Throughout early language acquisition, prosody continues to play an important role in scaffolding language learning—this is known as prosodic bootstrapping.

Many lexical and grammatical properties of language are accompanied by characteristic prosodic patterns. The theory of prosodic bootstrapping (Morgan and Demuth 1996) holds that young learners can exploit the prosodic cues that are directly available in their input to learn about the perceptually unavailable, abstract lexical, and grammatical properties with which those cues are correlated. In English, for instance, bisyllabic nouns (N) and verbs (V) with the same segmental make-up are distinguished by lexical stress: nouns tend to have initial stress, verbs final stress, such as the noun *record* /ˈrekə(r)d/ vs. the verb *record* /riˈko(r)d / (Cutler and Carter 1987). Knowing this regularity, a learner is able to categorize novel words as nouns or verbs even if she does not know their meanings.

Experimental findings over the past two decades suggest that infants are indeed able to exploit such correlations to break into the lexicon and grammar of their native language(s), thus alleviating the learning problem they face when confronted with the acquisition of abstract linguistic properties (Gervain et al. 2021).

As reviewed in Sect. 8.2, many of newborns' speech perception abilities rely on prosody. These sensitivities constitute the basis of the subsequent bootstrapping role of prosody. One area in which this has been extensively documented is word learning (Sect. 8.4). Once infants learn the lexical stress pattern typical of their native language on the basis of the first few words they encounter, they can then use this knowledge to constrain and support further learning.

Another important mechanism of prosodic bootstrapping is prosodic grouping, also known as the Iambic-Trochaic Law (ITL) (Hayes 1995), which states that sound sequences contrasting in duration are naturally perceived iambically (e.g., as forming pairs in which the first sound is short, the second one is long), whereas sound sequences that contrast in pitch or intensity are perceived trochaically (e.g., as forming pairs in which the first sound is high/loud, the second one is low/soft). The position as well as the acoustic realization of phrase-level prosodic prominence co-varies with word order (Nespor et al. 2008; Gervain and Werker 2013). In languages in which phrases start with grammatical words called functors, (e.g., *in Rome*), such as English or Italian, prosodic prominence in phonological phrases, which falls on the content word, is phrase-final (i.e., iambic) and is realized as a durational contrast—that is, as the lengthening of the stressed vowel of the content word (e.g., *in* ***Ro****me*). By contrast, languages, such as Japanese, Turkish, or Basque, where grammatical words appear at the end of phrases, the prominence is initial (i.e., trochaic) and is realized as increased pitch or intensity (e.g., Japanese: ***To****kyo ni* "to Tokyo"). While other cues may accompany prominence in any language, pitch or intensity serves as the contrastive cue in languages with final grammatical functors, whereas duration plays this role in functor-initial languages. Infants as young as 8–9 months of age can align phrasal prosody with the underlying syntactic pattern within phrases, as they expect functors to be non-prominent and content words to be prominent (Bernard and Gervain 2012). Even more importantly, 7-month-old bilinguals exposed to a functor-initial and a functor-final language use the different prosodic realizations to select the relevant word order (Gervain and Werker 2013). Upon hearing a durational contrast (short-long), they select sequences with a functor-initial order, while, when presented with a pitch contrast (high-low), they prefer functor-final sequences. This is strong evidence that infants start using prosody to bootstrap syntax even before they have a sizeable lexicon, suggesting that they set abstract syntactic parameters rather than memorize or rote-learn lexical patterns or item-based expressions. In this regard, the role of the ITL is particularly relevant. As mentioned before, newborns already show familiarity with the predominant iambic or trochaic prosodic patterns of their native language from prenatal experience (Abboub et al. 2016). This knowledge may guide young infants from very early on in how they segment and parse the language input, and allow them to determine basic properties of their native grammar such as its word order. For instance, an infant expecting a functor-content word order on the basis of prosody will be able to directly assign the correct lexical category to the novel words she encounters in an input sentence. This is further aided by young infants' ability to distinguish functors and content words on the basis of their phonological differences (Shi et al. 1999). Thus, on the basis of auditory cues alone, infants may be able to already build a rudimentary representation of the basic word order of functors and content words, which then further correlates with other word order phenomena, such as the relative order of verbs and their objects, or main clauses and subordinate clauses, etc., (Dryer 1992), providing infants with a powerful strategy to break into the grammar of their native language.

Later, children can also use prosody to support the processing of syntactic structures (Christophe et al. 2008, 2016; Hawthorne and Gerken 2014). Infants perceive intonational phrase boundaries from 5 months of age (Hirsh-Pasek et al. 1987; Männel and Friederici 2009). To test the effect of phrasal prosody on syntactic analysis, sentences with syntactically ambiguous phrases were presented to toddlers such as *the baby flies*, which can be interpreted as a noun phrase as in *The baby flies hide in the shadows*, or as a noun and a verb as in *The baby flies her kite*. In these sentences, prosody disambiguates the two possibilities, as in one sentence there is a prosodic boundary before the ambiguous *word fly*, in the other case, the boundary follows *fly*. When listening to the critical phrase in such sentences (with the end of the sentence being masked by noise), toddlers as young as 20 months of age are able to exploit the prosodic information, and looked at the picture depicting the intended meaning (Carvalho et al. 2016).

Children thus use prosody from the very beginning of language development starting with their prenatal experience with speech to identify and break into the native language, relying on prosodic cues to extract words from the input, learn the basic word order of the native language, and subsequently to constrain syntactic parsing.

8.6 Chapter Summary

Infants start their journey into language as universal listeners, but by the end of the first year of life they become native language experts, as their perceptual systems and brains reorganize to better perceive those linguistic contrasts that they encounter in the native language, losing sensitivity to non-native sound patterns. Attunement to the native language starts prenatally, as infants first experience speech in the womb. Accordingly, newborns possess speech perception abilities, some of which already show the impact of prenatal experience, while many others are universal and broadly based, allowing infants to learn any of the world's languages. After several months of experience with their native languages, infants start to lose these universal abilities, becoming unable to discriminate most contrasts (phonemes, tones, etc.) that are not used in the native language(s), while improving and fine-tuning their native sound categories. This perceptual attunement is accompanied by an increasing hemispheric specialization for language at the neural level. In parallel with the perceptual reorganization, infants also start learning their first words and the basics of their native grammar. The acquisition of the different levels of language thus proceeds in parallel and interacts with one another in synergistic ways.

Acknowledgment This work was funded by the ERC Consolidator Grant "BabyRhythm" (no. 773202) to J.G.

Compliance with Ethics Requirements Judit Gervain declares that she has no conflict of interest.

References

Abboub N, Nazzi T, Gervain J (2016) Prosodic grouping at birth. Brain Lang 162:46–59

Abercrombie D (1967) Elements of general phonetics. Edinburgh University Press, Edinburgh

Altan A, Kaya U, Hohenberger A (2016) Sensitivity of Turkish infants to vowel harmony in stem-suffix sequences: preference shift from familiarity to novelty. BUCLD 40 Online Proceedings Supplements

Bergelson E, Swingley D (2012) At 6–9 months, human infants know the meanings of many common nouns. Proc Natl Acad Sci USA 109:3253–3258

Bernard, C., & Gervain, J. (2012). Prosodic cues to word order: What level of representation? *Frontiers in Language Sciences, 3*:, 451. https://doi.org/10.3389/fpsyg.2012.00451

Best CT, McRoberts GW, Sithole NM (1988) Examination of perceptual reorganization for nonnative speech contrasts: Zulu click discrimination by English-speaking adults and infants. J Exp Psychol Hum Percept Perform 14:345–360

Birulés J, Bosch L, Pons F, Lewkowicz DJ (2020) Highly proficient L2 speakers still need to attend to a talker's mouth when processing L2 speech. Lang Cogn Neurosci 35(10):1314–1325

Bosch L, Sebastian-Galles N (1997) Native-language recognition abilities in 4-month-old infants from monolingual and bilingual environments. Cognition 65:33–69

Brent MR, Cartwright TA (1996) Distributional regularity and phonotactic constraints are useful for segmentation. Cognition 61:93–125

Bruderer AG, Danielson DK, Kandhadai P, Werker JF (2015) Sensorimotor influences on speech perception in infancy. Proc Natl Acad Sci USA 112:13531–13536

Byers-Heinlein K, Fennell CT (2014) Perceptual narrowing in the context of increased variation: insights from bilingual infants. Dev Psychobiol 56:274–291

Byers-Heinlein K, Burns TC, Werker JF (2010) The roots of bilingualism in newborns. Psychol Sci 21:343–348

Cabrera L, Gervain J (2020) Speech perception at birth: the brain encodes fast and slow temporal information. Sci Adv 6:eaba7830

Carvalho A, Dautriche I, Christophe A (2016) Preschoolers use phrasal prosody online to constrain syntactic analysis. Dev Sci 19:235–250

Casey BJ, Giedd JN, Thomas KM (2000) Structural and functional brain development and its relation to cognitive development. Biol Psychol 54:241–257

Chomsky N, Belletti A, Rizzi L (2002) On nature and language. Cambridge University Press, Cambridge

Christophe A, Dupoux E, Bertoncini J, Mehler J (1994) Do infants perceive word boundaries? An empirical study of the bootstrapping of lexical acquisition. J Acoust Soc Am 95:1570–1580

Christophe A, Millotte S, Bernal S, Lidz J (2008) Bootstrapping lexical and syntactic acquisition. Lang Speech 51:61–75

Christophe A, de Carvalho A, Dautriche I et al (2016) Synergies in early language acquisition. Int J Psychol 51:811

Church KW (1987) Phonological parsing and lexical retrieval. Cognition 25:53–69

Cristia A, Minagawa-Kawai Y, Egorova N et al (2014) Neural correlates of infant accent discrimination: an fNIRS study. Dev Sci 17:628–635

Curtin S, Mintz TH, Byrd D (2001) Coarticulatory cues enhance infants' recognition of syllable sequences in speech. In: Proceedings of the 25th annual Boston University conference on language development. Cascadilla Press, Somerville, pp 190–201

Curtiss S, Fromkin VA, Krashen S et al (1974) The linguistic development of Genie. Language 50:528–554

Cutler A (1994) Segmentation problems, rhythmic solutions. Lingua 92:81–104

Cutler A, Carter DM (1987) The predominance of strong initial syllables in the English vocabulary. Comput Speech Lang 2:133–142

Danielson DK, Bruderer AG, Kandhadai P et al (2017) The organization and reorganization of audiovisual speech perception in the first year of life. Cogn Dev 42:37–48

de la Cruz-Pavía I, Gervain J, Vatikiotis-Bateson E, Werker JF (2019) Finding phrases: on the role of co-verbal facial information in learning word order in infancy. PLoS One 14:e0224786

de la Cruz-Pavía I, Gervain J, Vatikiotis-Bateson E, Werker JF (2020a) Coverbal speech gestures signal phrase boundaries: a production study of Japanese and English infant-and adult-directed speech. Lang Acquis 27:160–186

de la Cruz-Pavía I, Werker JF, Vatikiotis-Bateson E, Gervain J (2020b) Finding phrases: the interplay of word frequency, phrasal prosody and co-speech visual information in chunking speech by monolingual and bilingual adults. Lang Speech 63:264–291

DeCasper AJ, Fifer WP (1980) Of human bonding: newborns prefer their mothers' voices. Science 208:1174–1176

DeCasper AJ, Lecanuet J-P, Busnel MC, Granier-Deferre C (1994) Fetal reactions to recurrent maternal speech. Infant Behav Dev 17:159–164

Dehaene-Lambertz G, Baillet S (1998) A phonological representation in the infant brain. Neuroreport 9:1885–1888

Dehaene-Lambertz G, Dehaene S, Hertz-Pannier L (2002) Functional neuroimaging of speech perception in infants. Science 298:2013–2015

Dellwo V (2006) Rhythm and speech rate: a variation coefficient for ΔC. In: Language and language-processing. Peter Lang, Frankfurt/Main, pp 231–241

Drullman R (1995) Temporal envelope and fine structure cues for speech intelligibility. J Acoust Soc Am 97:585–592

Dryer MS (1992) The Greenbergian word order correlations. Language 68:81–138

Eggermont JJ, Moore JK (2012) Morphological and functional development of the auditory nervous system. In: Human auditory development. Springer, New York, pp 61–105

Eimas PD, Siqueland ER, Jusczyk PW, Vigorito J (1971) Speech perception in infants. Science 171:303–306

Elman J, Bates E, Johnson M et al (1997) Rethinking innateness: a connectionist perspective on development. The MIT Press, Boston

Gerhardt KJ, Abrams RM, Oliver CC (1990) Sound environment of the fetal sheep. Am J Obstet Gynecol 162:282–287

Gervain J (2015) Plasticity in early language acquisition: the effects of prenatal and early childhood experience. Curr Opin Neurobiol 35:13–20

Gervain J (2018) The role of prenatal experience in language development. Curr Opin Behav Sci 21:62–67

Gervain J, Werker JF (2013) Prosody cues word order in 7-month-old bilingual infants. Nat Commun 4:1490

Gervain J, Macagno F, Cogoi S et al (2008) The neonate brain detects speech structure. Proc Natl Acad Sci USA 105:14222–14227

Gervain J, Berent I, Werker JF (2012) Binding at birth: the newborn brain detects identity relations and sequential position in speech. J Cogn Neurosci 24:564–574

Gervain J, Vines BW, Chen LM et al (2013) Valproate reopens critical-period learning of absolute pitch. Front Syst Neurosci 7:102

Gervain J, Christophe A, Mazuka R (2021) Prosodic bootstrapping. In: The handbook of prosody. Oxford University Press, Oxford

Golinkoff RM, Hirsh-Pasek K, Bloom L et al (2000) Becoming a word learner: a debate on lexical acquisition. Oxford University Press, New York

Gonzalez-Gomez N, Schmandt S, Fazekas J et al (2019) Infants' sensitivity to nonadjacent vowel dependencies: the case of vowel harmony in Hungarian. J Exp Child Psychol 178:170–183

Grabe E, Low EL (2002) Durational variability in speech and the rhythm class hypothesis. In: Papers in laboratory phonology. Mouton de Gruyter, Berlin, pp 515–546

Guellaï B, Langus A, Nespor M (2014) Prosody in the hands of the speaker. Front Psychol 5:700

Gustafson GE, Sanborn SM, Lin H-C, Green JA (2017) Newborns' cries are unique to individuals (but not to language environment). Infancy 22:736–747

Haartsen R, Jones EJ, Johnson MH (2016) Human brain development over the early years. Curr Opin Behav Sci 10:149–154

Harris Z (1955) From phoneme to morpheme. Language 31:190–222

Hawthorne K, Gerken L (2014) From pauses to clauses: prosody facilitates learning of syntactic constituency. Cognition 133:420–428

Hayes B (1995) Metrical stress theory: principles and case studies. University of Chicago Press, Chicago

Hensch TK (2005) Critical period plasticity in local cortical circuits. Nat Rev Neurosci 6:877–888

Hickok G, Buchsbaum B, Humphries C, Muftuler T (2003) Auditory–motor interaction revealed by fMRI: speech, music, and working memory in area Spt. J Cogn Neurosci 15:673–682

Hirsh-Pasek K, KemlerNelson DG, Jusczyk PW et al (1987) Clauses are perceptual units for young infants. Cognition 26:269–286

Hohne EA, Jusczyk PW (1994) Two-month-old infants' sensitivity to allophonic differences. Percept Psychophys 56:613–623

Hollich G, Newman RS, Jusczyk PW (2005) Infants' use of synchronized visual information to separate streams of speech. Child Dev 76:598–613

Hunnius S, Geuze RH (2004) Developmental changes in visual scanning of dynamic faces and abstract stimuli in infants: a longitudinal study. Infancy 6:231–255

Johnson EK, Jusczyk PW (2001) Word segmentation by 8-month-olds: when speech cues count more than statistics. J Mem Lang 44:548–567

Johnson JS, Newport EL (1989) Critical period effects in second language learning: the influence of maturational state on the acquisition of English as a second language. Cogn Psychol 21:60–99

Jusczyk PW, Aslin RN (1995) Infants' detection of the sound patterns of words in fluent speech. Cogn Psychol 29:1–23

Jusczyk PW, Hohne EA, Bauman A (1999a) Infants' sensitivity to allophonic cues to word segmentation. Percept Psychophys 61:1465–1476

Jusczyk PW, Houston DM, Newsome MR (1999b) The beginnings of word segmentation in English-learning infants. Cogn Psychol 39:159–207

Kisilevsky BS, Hains SMJ, Brown CA et al (2009) Fetal sensitivity to properties of maternal speech and language. Infant Behav Dev 32:59–71

Kubicek C, De Boisferon AH, Dupierrix E et al (2013) Face-scanning behavior to silently-talking faces in 12-month-old infants: the impact of pre-exposed auditory speech. Int J Behav Dev 37:106–110

Kubicek C, De Boisferon AH, Dupierrix E et al (2014a) Cross-modal matching of audio-visual German and French fluent speech in infancy. PLoS One 9:e89275

Kubicek C, Gervain J, Hillairet de Boisferon A et al (2014b) The influence of infant-directed speech on 12-month-olds' intersensory perception of fluent speech. Infant Behav Dev 37:644–651

Kuhl PK (1981) Discrimination of speech by nonhuman animals: basic auditory sensitivities conducive to the perception of speech-sound categories. J Acoust Soc Am 70:340–349

Kuhl PK (1986) Theoretical contributions of tests on animals to the special-mechanisms debate in speech. Exp Biol 45:233–265

Kuhl PK (2004) Early language acquisition: cracking the speech code. Nat Rev Neurosci 5:831–843

Kuhl PK, Meltzoff AN (1982) The bimodal perception of speech in infancy. Science 218:1138–1141

Kuhl PK, Williams KA, Lacerda F et al (1992) Linguistic experience alters phonetic perception in infants by 6 months of age. Science 255:606–608

Kuhl PK, Stevens E, Hayashi A et al (2006) Infants show a facilitation effect for native language phonetic perception between 6 and 12 months. Dev Sci 9:F13–f21

Kujala A, Huotilainen M, Hotakainen M et al (2004) Speech-sound discrimination in neonates as measured with MEG. Neuroreport 15:2089–2092

Langus A, Mehler J, Nespor M (2016) Rhythm in language acquisition. Neurosci Biobehav Rev 81:158–166

Lecanuet JP, Granier-Deferre C (1993) Speech stimuli in the fetal environment. In: Developmental neurocognition: speech and face processing in the first year of life. Kluwer Academic Press, Dordrecht, pp 237–248

Lenneberg EH (1967) The biological foundations of language. Wiley Press, New York

Lewkowicz DJ, Hansen-Tift AM (2012) Infants deploy selective attention to the mouth of a talking face when learning speech. Proc Natl Acad Sci USA 109:1431–1436

Liberman AM, Mattingly IG (1985) The motor theory of speech perception revised. Cognition 21:1–36

Liu L, Kager R (2012) Non-native tone perception from infant to adult: how consistent and flexible is it? In: Proceedings of speech prosody

MacKain KS, Studdert-Kennedy M, Spieker S, Stern DN (1983) Infant intermodal speech perception is a left-hemisphere function. Science 219:1347–1349

Mahmoudzadeh M, Dehaene-Lambertz G, Fournier M et al (2013) Syllabic discrimination in premature human infants prior to complete formation of cortical layers. Proc Natl Acad Sci USA 110:4846–4851

Majorano M, Vihman MM, DePaolis RA (2014) The relationship between infants' production experience and their processing of speech. Lang Learn Dev 10:179–204

Mampe B, Friederici AD, Christophe A, Wermke K (2009) Newborns' cry melody is shaped by their native language. Curr Biol 19:1994–1997

Manfredi C, Viellevoye R, Orlandi S et al (2019) Automated analysis of newborn cry: relationships between melodic shapes and native language. Biomed Signal Process Control 53:101561

Männel C, Friederici AD (2009) Pauses and intonational phrasing: ERP studies in 5-month-old German infants and adults. J Cogn Neurosci 21:1988–2006

Markman EM (1994) Constraints on word meaning in early language acquisition. Lingua 92:199–227

Mattock K, Burnham D (2006) Chinese and English infants' tone perception: evidence for perceptual reorganization. Infancy 10:241–265

Mattys SL, Jusczyk PW (2001) Phonotactic cues for segmentation of fluent speech by infants. Cognition 78:91–121

Mattys SL, Jusczyk PW, Luce PA, Morgan JL (1999) Phonotactic and prosodic effects on word segmentation in infants. Cogn Psychol 38:465–494

Maurer D, Werker JF (2014) Perceptual narrowing during infancy: a comparison of language and faces. Dev Psychobiol 56:154–178

May L, Byers-Heinlein K, Gervain J, Werker JF (2011) Language and the newborn brain: does prenatal language experience shape the neonate neural response to speech? Front Lang Sci 2:222

May L, Gervain J, Carreiras M, Werker JF (2017) The specificity of the neural response to speech at birth. Dev Sci 21:e12564

McGurk H, MacDonald J (1976) Hearing lips and seeing voices. Nature 264:746–748

McMurray B, Aslin RN (2005) Infants are sensitive to within-category variation in speech perception. Cognition 95:B15–B26

Mehler J, Jusczyk PW, Lambertz G et al (1988) A precursor of language acquisition in young infants. Cognition 29:143–178

Mehler J, Sebastian-Galles N, Nespor M (2004) Biological foundations of language: language acquisition, cues for parameter setting and the bilingual infant. In: The new cognitive neuroscience. MIT Press, Cambridge, MA, pp 825–836

Minagawa-Kawai Y, Mori K, Naoi N, Kojima S (2007) Neural attunement processes in infants during the acquisition of a language-specific phonemic contrast. J Neurosci 27:315–321

Minagawa-Kawai Y, Cristia A, Dupoux E (2011) Cerebral lateralization and early speech acquisition: a developmental scenario. Dev Cogn Neurosci 1:217–232

Mitchel AD, Weiss DJ (2014) Visual speech segmentation: using facial cues to locate word boundaries in continuous speech. Lang Cogn Process 29:771–780

Molnar M, Gervain J, Carreiras M (2013) Within-rhythm class native language discrimination abilities of Basque-Spanish monolingual and bilingual infants at 3.5 months of age. Infancy 19:326–337

Moon C, Cooper RP, Fifer WP (1993) Two-day-olds prefer their native language. Infant Behav Dev 16:495–500

Moon C, Lagercrantz H, Kuhl PK (2013) Language experienced in utero affects vowel perception after birth: a two-country study. Acta Paediatr 102:156–160

Moore DR (2002) Auditory development and the role of experience. Br Med Bull 63:171–181

Morgan JL (1996) A rythmic bias in preverbal speech segmentation. J Mem Lang 35:666–688

Morgan JL, Demuth K (1996) Signal to syntax: bootstrapping from speech to grammar in early acquisition. Lawrence Erlbaum Associates, Inc, Hillsdale

Morgan JL, Saffran JR (1995) Emerging integration of sequential and suprasegmental information in preverbal speech segmentation. Child Dev 66:911–936

Narayan CR, Werker JF, Beddor PS (2010) The interaction between acoustic salience and language experience in developmental speech perception: evidence from nasal place discrimination. Dev Sci 13:407–420

Nazzi T, Bertoncini J, Mehler J (1998) Language discrimination by newborns: toward an understanding of the role of rhythm. J Exp Psychol Hum Percept Perform 24:756–766

Nazzi T, Bertoncini J, Bijeljac-Babic R (2009) A perceptual equivalent of the labial-coronal effect in the first year of life. J Acoust Soc Am 126:1440

Nespor M (1990) On the rhythm parameter in phonology. In: Logical issues in language acquisition. Foris, Dordrecht, pp 157–175

Nespor M, Shukla M, van de Vijver R et al (2008) Different phrasal prominence realization in VO and OV languages. Lingue E Linguaggio 7:1–28

Partanen E, Kujala T, Näätänen R et al (2013) Learning-induced neural plasticity of speech processing before birth. Proc Natl Acad Sci USA 110:15145–15150

Pascalis O, deHaan M, Nelson CA (2002) Is face processing species-specific during the first year of life? Science 296:1321–1323

Patterson ML, Werker JF (1999) Matching phonetic information in lips and voice is robust in 4.5-month-old infants. Infant Behav Dev 22:237–247

Patterson ML, Werker JF (2002) Infants' ability to match dynamic phonetic and gender information in the face and voice. J Exp Child Psychol 81:93–115

Peña M, Maki A, Kovacic D et al (2003) Sounds and silence: an optical topography study of language recognition at birth. Proc Natl Acad Sci USA 100:11702–11705

Pierce LJ, Klein D, Chen J-K et al (2014) Mapping the unconscious maintenance of a lost first language. Proc Natl Acad Sci USA 111:17314–17319

Pons F, Lewkowicz DJ, Soto-Faraco S, Sebastián-Gallés N (2009) Narrowing of intersensory speech perception in infancy. Proc Natl Acad Sci USA 106:10598

Pons F, Bosch L, Lewkowicz DJ (2015) Bilingualism modulates infants' selective attention to the mouth of a talking face. Psychol Sci 26:490–498

Pons F, Sanz-Torrent M, Ferinu L et al (2018) Children with SLI can exhibit reduced attention to a talker's mouth. Lang Learn 68:180–192

Ramus F, Nespor M, Mehler J (1999) Correlates of linguistic rhythm in the speech signal. Cognition 73:265–292

Ramus F, Hauser MD, Miller C et al (2000) Language discrimination by human newborns and by cotton-top tamarin monkeys. Science 288:349–351

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science, 274*(5294), 1926–1928.

Saffran JR, Thiessen ED (2003) Pattern induction by infant language learners. Dev Psychol 39:484–494

Sansavini A, Bertoncini J, Giovanelli G (1997) Newborns discriminate the rhythm of multisyllabic stressed words. Dev Psychol 33:3–11

Sato Y, Sogabe Y, Mazuka R (2010) Development of hemispheric specialization for lexical pitch-accent in Japanese infants. J Cogn Neurosci 22:2503–2513

Sato H, Hirabayashi Y, Tsubokura H et al (2012) Cerebral hemodynamics in newborn infants exposed to speech sounds: a whole-head optical topography study. Hum Brain Mapp 33:2092–2103

Sebastián-Gallés N, Albareda-Castellot B, Weikum WM, Werker JF (2012) A bilingual advantage in visual language discrimination in infancy. Psychol Sci 23:994–999

Shannon RV, Zeng F-G, Kamath V et al (1995) Speech recognition with primarily temporal cues. Science 270:303–304

Shi R, Werker JF, Morgan JL (1999) Newborn infants' sensitivity to perceptual cues to lexical and grammatical words. Cognition 72:B11–B21

Soto-Faraco S, Navarra J, Weikum WM et al (2007) Discriminating languages by speech-reading. Percept Psychophys 69:218–231

Stager CL, Werker JF (1997) Infants listen for more phonetic detail in speech perception than in word-learning tasks. Nature 388:381–382

Swingley D (2021) Infants' learning of speech sounds and word forms. In: Oxford handbook of the mental lexicon. Oxford University Press, Oxford

Tierney AL, Nelson CA (2009) Brain development and the role of experience in the early years. Zero Three 30:9

Tincoff R, Jusczyk PW (1999) Some beginnings of word comprehension in 6-month-olds. Psychol Sci 10:172–175

Tomasello M (2000) Do young children have adult syntactic competence? Cognition 74:209–253

Ventureyra VA, Pallier C, Yoo H-Y (2004) The loss of first language phonetic perception in adopted Koreans. J Neurolinguistics 17:79–91

Vilain A, Dole M, Lœvenbruck H et al (2019) The role of production abilities in the perception of consonant category in infants. Dev Sci 22:e12830

Viola Macchi C, Turati C, Simion F (2004) Can a nonspecific bias toward top-heavy patterns explain newborns' face preference? Psychol Sci 15:379–383

Vouloumanos A, Werker JF (2004) Tuned to the signal: the privileged status of speech for young infants. Dev Sci 7:270

Vouloumanos A, Hauser MD, Werker JF, Martin A (2010) The tuning of human neonates' preference for speech. Child Dev 81:517–527

Wagner P, Malisz Z, Kopp S (2014) Gesture and speech in interaction: an overview. Elsevier

Weaver IC, Cervoni N, Champagne FA et al (2004) Epigenetic programming by maternal behavior. Nat Neurosci 7:847–854

Weikum WM, Vouloumanos A, Navarra J et al (2007) Visual language discrimination in infancy. Science 316:1159

Weikum WM, Oberlander TF, Hensch TK, Werker JF (2012) Prenatal exposure to antidepressants and depressed maternal mood alter trajectory of infant speech perception. Proc Natl Acad Sci USA 109:17221–17227

Weikum WM, Vouloumanos A, Navarra J et al (2013) Age-related sensitive periods influence visual language discrimination in adults. Front Syst Neurosci 7:86

Werker JF (2018) Perceptual beginnings to language acquisition. Appl Psycholinguist 39:703–728

Werker JF, Curtin S (2005) PRIMIR: a developmental model of speech processing. Lang Learn Dev 1:197–234

Werker JF, Hensch TK (2015) Critical periods in speech perception: new directions. Annu Rev Psychol 66:173–196

Werker JF, Tees RC (1984) Cross-language speech perception: evidence for perceptual reorganization during the first year of life. Infant Behav Dev 7:49–63

Werker JF, Yeung HH (2005) Infant speech perception bootstraps word learning. Trends Cogn Sci 9:519–527

White KS, Morgan JL (2008) Sub-segmental detail in early lexical representations. J Mem Lang 59:114–132

Yeung HH, Werker JF (2009) Learning words' sounds before learning how words sound: 9-month-olds use distinct objects as cues to categorize speech information. Cognition 113:234–243

Chapter 9
Interactions Between Audition and Cognition in Hearing Loss and Aging

Chad S. Rogers and Jonathan E. Peelle

Abstract Successful speech understanding relies not only on the auditory pathway, but on cognitive processes that act on incoming sensory information. One area in which the importance of cognitive factors is particularly striking during speech comprehension is when the acoustic signal is made more challenging, which might happen due to background noise, talker characteristics, or hearing loss. This chapter focuses on the interaction between hearing and cognition in hearing loss in older adults. The chapter begins with a review of common age-related changes in hearing and cognition, followed by summary evidence from pupillometric, behavioral, and neuroimaging paradigms that elucidate the interplay between hearing and cognition. Across a variety of experimental paradigms, there is compelling evidence that when listeners process acoustically challenging speech, additional cognitive effort is required compared to acoustically clear speech. This increase in cognitive effort is associated with specific brain networks, with the clearest evidence implicating cingulo-opercular and executive attention networks. Individual differences in hearing and cognitive ability thus determine the cognitive demand faced by a particular listener, and the cognitive and neural resources needed to aid in speech perception.

Keywords Listening effort · Background noise · Speech perception · Cognitive aging · Sentence comprehension · Neuroimaging · Cingulo-opercular network · Executive attention · Pupillometry · fMRI

C. S. Rogers (✉)
Department of Psychology, Union College, Schenectady, NY, USA
e-mail: rogersc@union.edu

J. E. Peelle
Department of Otolaryngology, Washington University in St. Louis, St. Louis, MO, USA
e-mail: jpeelle@wustl.edu

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022

L. L. Holt et al. (eds.), *Speech Perception*, Springer Handbook of Auditory
Research 74, https://doi.org/10.1007/978-3-030-81542-4_9

9.1 Introduction

It goes without saying that the auditory system is of key importance for speech perception. However, a number of recent frameworks for speech understanding have emphasized the important additional contributions of cognitive factors (Wingfield et al. 2005; Peelle 2018). Although there are many reasons to consider cognitive processing in speech perception, one especially important catalyst has been the longstanding realization that hearing sensitivity, alone, is unable to fully account for challenges faced by listeners, particularly in noise (Plomp and Mimpen 1979; Humes et al. 2013). One explanation for this finding is that current tests of auditory function may be lacking, and that more informative tests are needed. However, another possibility is that individual differences in cognitive ability contribute to a listener's success understanding speech. Thus, clarifying the cognitive challenges associated with understanding acoustically challenging speech is not only of theoretical importance, but may significantly improve our understanding of communication in everyday situations.

Hearing loss affects listeners of all ages, and in the United States, it is estimated to affect 23% of those aged 12 or older (Goman and Lin 2016). The focus of this chapter is primarily on age-related hearing loss, given the particularly high incidence of hearing loss in adults over the age of 65 (Cruickshanks et al. 1998; Mitchell et al. 2011). Healthy older adults are also a prime group of listeners in whom to study the interactions of sensory and cognitive factors, given that changes in both of these areas are frequently seen as we age.

Figure 9.1 shows a schematic of speech comprehension that includes processing related to auditory, language, and memory systems (because listeners frequently would like to remember what they have heard), as well as domain-general cognitive processes that act on one or more of these stages. The following sections cover a number of these areas where age-related changes are reported, as well as others that seem to be relatively preserved in older age. An important point to keep in mind is the significant variability in individual ability in all of these domains.

Section 9.2 of this chapter highlights the most salient age-related changes in hearing and cognition. Following this, the ways in which these changes manifest during speech comprehension and inform broader understanding of auditory-cognitive interaction are examined.

9.2 Age-Related Hearing Loss

Age-related hearing loss is extremely common: Although estimates vary, some 40–50% of adults over the age of 65 years have a measurable hearing impairment, with this number rising to greater than 80% of those over the age of 70 years (Cruickshanks et al. 1998). Age-related changes in hearing occur at every level of the auditory system (Peelle and Wingfield 2016), and the specific etiology is likely to have consequences for information processing. Age-related hearing loss can be broadly categorized into

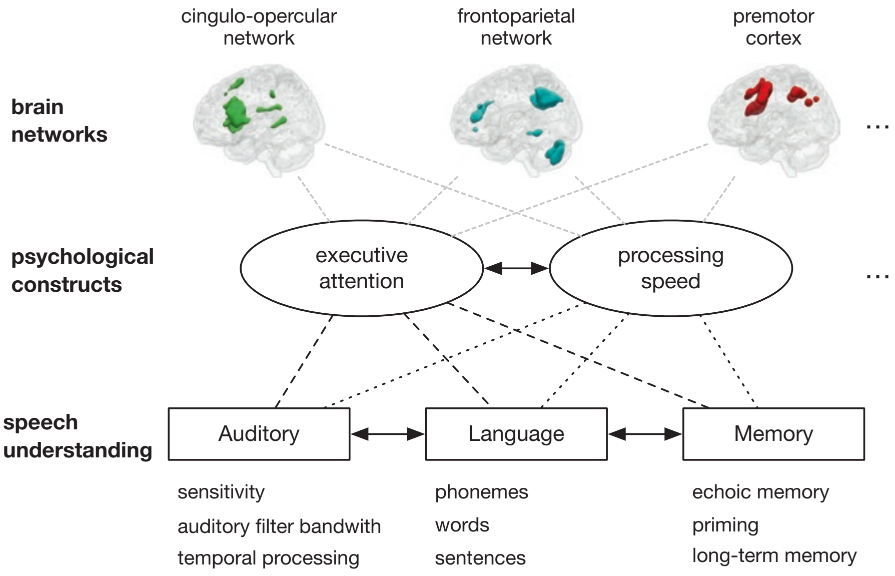

Fig. 9.1 Schematic of domains involved in speech processing. At bottom are stages of processing from low-level auditory perception through speech understanding, and on to tasks that might be done with the heard speech (such as remembering it). Bidirectional arrows signal interactivity between these levels (e.g., linguistic factors can bias auditory perception). At top are some example cognitive processes and corresponding brain networks that are used in understanding speech. The interaction between auditory and cognitive factors is thus a complex and highly interactive process spanning multiple levels of representation. (Figure available via a CC-BY4.0 license from https://osf.io/mv95h/)

peripheral hearing loss (having to do with the cochlea) or central hearing loss (having to do with the auditory nerve, subcortical structures, or cortex).

9.2.1 *Peripheral Age-Related Changes in Hearing*

A major cause of age-related hearing loss is a reduction in the number of outer hair cells of the cochlea. For reasons that are still not entirely clear, hair cell loss occurs primarily in the basal end of the cochlea responsible for encoding high frequency information (Merchant and Nadol 2010). Thus, the most striking characteristic of age-related hearing loss is a decrease in sensitivity to higher frequencies. Figure 9.2 shows median pure-tone sensitivity for adults of different ages, illustrating this characteristic pattern, as well as the characteristically poorer hearing of men relative to women in older age. Fortunately, hearing in the range most important for speech information (4 kHz and below) is generally relatively well preserved in cases of mild age-related hearing loss. However, some speech information (such as fricatives) can still be lost (Bilger and Wang 1976; Scharenborg et al. 2015), and as hearing sensitivity worsens speech intelligibility may decline.

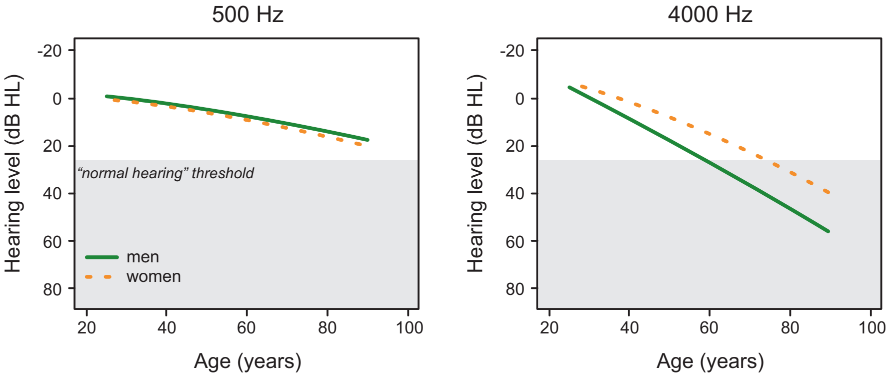

Fig. 9.2 Median pure-tone hearing levels for adult men and women at 500 Hz and 4000 Hz (cartoon based on Morrell et al. 1996). The shaded region indicates a typical cutoff for clinically normal hearing of 25 dB HL. (Figure available via a CC-BY4.0 license from https://osf.io/mv95h/)

Changes to peripheral hearing can also be caused by synaptic dysfunction and degeneration of cochlear nerve axons. In animal models, Kujawa and Liberman (2009) found that a single noise exposure can weaken cochlear afferent nerve terminals. This weakening was observed even when there was no apparent damage to hair cells, or evidence of a long-term threshold shift (i.e., differing sensitivity to pure tones—a change in the audiogram illustrated in Fig. 9.2). Because these changes are not always evident in pure-tone thresholds, they are sometimes referred to as "hidden" hearing loss. Although still a relatively new area, there is evidence suggesting that hidden hearing loss contributes to deficits in amplitude modulation coding (Paul et al. 2017) and temporal processing (Bharadwaj et al. 2015).

9.2.2 Central Age-Related Changes in Hearing

Data from both animal and human studies suggest age-related changes in spiral ganglion neurons (Bao and Ohlemiller 2010), cochlear nuclei, the superior olivary complex, and inferior colliculus (Caspary et al. 2008; Engle et al. 2014). In humans, much work has focused on age-related changes to the auditory brainstem response (ABR). The ABR is a time-locked electrophysiological response elicited by brief acoustic stimuli (e.g., clicks, or a phoneme), typically recorded from electroencephalographic (EEG) electrodes placed on the scalp. The amplitude and timing of the peaks of the ABR can thus be used to infer the fidelity of subcortical auditory processing. With advancing age, the peaks of the ABR show reduced amplitude, and some peaks show additional delays in their timing relative to stimulus onset (Skoe et al. 2015). Aging is associated with decreased precision of the ABR including longer delays and greater variability across trials (Anderson et al. 2012). Changes in speech-evoked ABRs are also evident in listeners with hearing loss, suggesting that changes subcortical representations may contribute to difficulties with speech-in-noise perception (Anderson et al. 2013).

Age-related changes are also evident in primary auditory cortex, reflected in both anatomy and electrophysiology. Parvalbumin is a calcium-binding protein expressed in auditory interneurons that play in important role in novelty detection and stimulus sensitivity. The number of parvalbumin-containing neurons is reduced in aging and is accompanied by reduced myelin (de Villers-Sidani et al. 2010; del Campo et al. 2012). Animal studies show that aging is also associated with a reduction in GAD, a GABA synthetic enzyme, in layers II–IV, probably reflecting a reduction in levels of GABA (an inhibitory neurotransmitter) (Ling et al. 2005; Burianova et al. 2009). Reports using magnetic resonance spectroscopy suggest some evidence consistent with the animal literature (Profant et al. 2013; Gao et al. 2015).

Noninvasive electrophysiological studies in humans suggest numerous age-related changes in the function of auditory cortex, including the magnitude of auditory evoked responses (Alain et al. 2014) and altered dynamics of stimulus adaptation (Herrmann et al. 2016). Finally, on a gross anatomical level, the volume of gray matter in auditory cortex is reduced in older adults with poorer hearing compared to those with better hearing (Peelle et al. 2011; Eckert et al. 2012). These macroanatomical structural changes may reflect changes in lower-level physiology associated with altered auditory processing.

9.3 Cognition in Older Adulthood

In addition to age-related changes to peripheral and central auditory structures that impact hearing, age-related changes to cognitive abilities are well documented (Salthouse 1991; Gordon-Salant et al. 2020). However, it is crucial to emphasize the significant variability in age-related changes to cognitive systems supporting speech understanding (Park et al. 1996). The differing consequences of aging for dissociable cognitive systems can elucidate the consequences of aging on speech understanding, and also help us better understand the underlying bases for speech processing in other populations of listeners (such as healthy young adults). The following sections highlight several cognitive systems closely related to speech understanding, and their potential for age-related decline: processing speed, inhibition and cognitive control, working memory capacity, episodic memory, and metacognition.

9.3.1 Age-Related Cognitive Decline

Age-related changes are observed in many domains of cognition, only a few of which are reviewed here. Because age-related changes are correlated across a large number of tasks, age-related change is frequently thought of in a factor-analytic framework, in which common variance across many tasks can be reduced to a smaller number of common factors. In the context of age-related cognitive decline, performance on many specific tasks might be better explained by declines in domain-general processing. Common domain-general constructs have included executive attention and processing speed (Salthouse 1996b; McCabe et al. 2010). As discussed throughout this section, changes in broad domain-general areas can also impact specific domains such as episodic memory or metacognition.

9.3.1.1 Processing Speed

Age-related changes to processing speed, or the rate at which tasks are performed, are ubiquitous in adult aging. Meta-analyses have revealed that older adults have slower reaction times than young adults in virtually every timed task (Cerella 1985; Verhaeghen and De Meersman 1998). These findings have led to an influential theory of general slowing in the cognitive aging literature (Salthouse 1996b), whereby age-related decline on cognitive tasks results from cascading failures of cognitive operations to complete on time. Evidence in support of this theory includes large-scale psychometric studies and meta-analyses that assess covariation among tasks assessing processing speed, memory, and executive attention, and find that the two-way relationships between aging and memory or aging and executive attention are minimized or eliminated after statistically taking into account the relationship between aging and processing speed (Salthouse 1996a; Verhaeghen and Salthouse 1997). Measures of processing speed have also been shown to correlate with neurobiological markers of aging (Eckert 2011). While this framework provides a powerful and parsimonious account for changes that occur as people grow older, one criticism is that it does not provide a straightforward account for the entire pattern of behaviors observed in the literature, particularly in areas in which age-related declines are not observed (e.g., Balota et al. 2000; McCabe et al. 2010).

9.3.1.2 Inhibition and Cognitive Control

In daily experience, people frequently need to ignore, or inhibit, information that is not relevant for a current task. A laboratory task which is a favorite among cognitive psychologists is the Stroop task (Stroop 1935). Participants are shown various words, including color words (e.g., “red”), that are written in different colors, and instructed to indicate the color the word is written in. For literate participants, the actual word (“red”) is processed automatically, but does not help with performance on the task (thus, “irrelevant”). When the word and its color agree (“red” written in red), participants respond more quickly; however, when the word and color disagree (“red” written in blue), participants are slower to respond. This general pattern is interpreted to reflect participants’ inability to ignore or inhibit the word information, which is irrelevant to the current task of naming the color a word is written in.

Older adults have a well-documented deficit in the ability to inhibit irrelevant stimuli (Hasher and Zacks 1988), which has been demonstrated in memory tasks (Gazzaley et al. 2005), visual tasks such as the Stroop task (Bugg et al. 2007), and auditory tasks such as dichotic listening (Rogers et al. 2018). In this last example, participants listened to simultaneously presented auditory streams of words, one stream to each ear, preceded by a visual arrow that indicated the ear to which participants should attend. Afterward, participants were shown a list of words—which could be from the attended ear, the unattended ear, or unrelated—and asked to indicate words that had been in the attended ear. Performance between young and older adults was equivalent in all conditions except for when recognition probes were words from the unattended stream. In those cases, older adults were more likely than the young adults to (erroneously) endorse those items, indicating a deficit in suppression of the unattended ear (Tun et al. 2002). This kind of deficit in selective attention suggests that older adults may actually encode more, not less, than young adults, but include both relevant and irrelevant information (Weeks and Hasher 2017).

9.3.1.3 Working Memory Capacity

Older adults also have deficits in working memory. In the classic model of Baddeley (1986), working memory contains both a verbal memory buffer (the phonological loop) and a visual memory buffer (the visuospatial sketchpad). These buffers allow for auditory and visual information, respectively, to be maintained for in an active state for a finite period of time. Baddeley's (1986) model also includes a central executive component which allows for manipulation and processing of information contained within these buffer and long-term memory systems in order to achieve goals of a specific task (Rudner and Ronnberg 2008). For example, retaining the digits of a phone number relies on short-term memory, whereas separating the digits into odd and even numbers then reciting them in ascending order would likely tap central executive processes (Belleville et al. 1998). The central executive component is considered to be the locus of age-related declines in working memory (Rose et al. 2009), specifically with regard to its role of suppression of task-irrelevant stimuli (Gazzaley et al. 2005).

9.3.1.4 Episodic Memory

Episodic memory is the capacity for memory for specific past events including details about where and when the event occurred (Tulving and Szpunar 2009), and allows people to remember specific autobiographical information about their own lives. For example, whereas semantic memory allows one to know what kind of dinner is served at an Italian restaurant, episodic memory allows for one to remember the last time they ate Italian food, who they ate with, and how good the food was.

Episodic memory is generally considered to decline as a function of age (Craik 2008), probably due to age-related changes in a number of related cognitive processes. Memory is often conceptualized as relying on at least three stages: encoding (when events are initially stored into memory), retention (when memories are held for a duration of time), and retrieval (when memories are accessed) (Melton 1963). Older adults have been shown to have declines in the ability to encode events into long-term memory (Craik and Rose 2012), with older adults less likely than young adults to engage in self-initiated elaborative encoding strategies that are beneficial for long-term retention (Craik and Rabinowitz 1985). In addition, older adults may have more difficulty than young adults encoding specific temporal associations, as is needed to recall a list of unrelated words (Golomb et al. 2008). When controlling for initial encoding, older and young adults are generally assumed to have relatively similar rates of memory retention (Park et al. 1988). However, older adults show deficits relative to young adults in the ability to retrieve memories when prompted (Craik and McDowd 1987; Wingfield and Kahana 2002). The largest evidence of age-related changes in retrieval is observed in free recall; when older adults are given helpful cues to facilitate retrieval, differences between young and older adults are minimized (Smith 1977).

9.3.1.5 Metacognition

One particularly interesting aspect of age-related cognitive change is the extent to which people have insight over their own cognitive states. For example, are listeners aware of the extent their hearing or cognitive abilities? Nelson and Narens' (1990) seminal framework describes metacognition as two processes that operate between the object level (e.g., *actual* cognitive processing) and the meta level (e.g., *awareness* of this processing). The flow of information from the object level to the meta level is called monitoring, and the flow of information from the meta level to the object level is called control. For example, monitoring occurs when one notices that they are having a hard time hearing the television after the air conditioner kicks on, and control occurs when one increases the television volume in response to that awareness.

Hertzog and Dunlosky (2011) report that older adults' monitoring is preserved relative to young adults in episodic memory tasks, and that their predictions and post-dictions of future and past performance are well calibrated. However, this pattern of age invariance does not appear to hold when tasks require executive attention at the object level (Souchay and Isingrini 2004). For example, in a study by Kelley and Sahakyan (2003), young and older adults studied pairs of words (e.g., CLOCK-DOLLAR), and were tested using a cued recall test (e.g., CLOCK-DO_ _ _ R). The pair CLOCK-DOLLAR is an example of a baseline item in which pairs of words were not semantically associated. Kelley and Sahakyan (2003) also used deceptive items in which the words pairs were not semantically associated at study (e.g., NURSE-DOLLAR) but the cue fragment at test could erroneously lead to a semantic associate of the first word (e.g., NURSE-DO_ _ _ R). Note that successful performance on these deceptive items requires inhibiting the semantic associate DOCTOR to respond correctly with the studied item DOLLAR. During each trial at test, participants made a cued recall attempt, then rated their confidence in their memory (e.g., monitoring), and then decided if they wanted their response to be scored for a later monetary reward (e.g., control). The results of that study showed that while cued recall for baseline items was poorer for older adults than young adults overall, metacognitive judgments by older adults in terms of their confidence and willingness to have their responses scored were just as well calibrated to their actual performance as young adults' metacognitive judgments. However, on deceptive items, where participants had to inhibit the semantic associate of the first word at recall, older adults' metacognitive judgments were poorly calibrated relative to young adults—older adults were more confident in their errors and more likely to volunteer to have their errors scored than were young adults. This study provides strong evidence that older adults have intact metacognition relative to young adults only to the extent that input to the monitoring process does not require executive attention.

9.3.2 Resilience to Age-Related Decline in Some Memory Systems Important for Speech Perception

Despite widespread findings of age-related cognitive change in many areas, there are others in which older adults perform very similarly to young adults. Areas of preserved performance in older adulthood are important because they may provide means for older adults to compensate for declines to hearing and cognition in service of speech understanding.

9.3.2.1 Echoic Memory

Echoic memory is a short-term auditory store that holds sensory-based auditory information for a very short period of time, probably on the order of hundreds of milliseconds (Cowan 1984). For example, during a telephone conversation where the listener asks the speaker to repeat themselves, and yet remembers what the speaker initially said before the speaker even attempts to repeat, it is likely that the listener was able to retrieve the spoken information from echoic memory. As with its visual analog, iconic memory (Sperling 1960), echoic memory has been shown to be invariant with age (Parkinson and Perey 1980) and is not typically considered to be a likely locus of age-related decrements to speech perception. The same age invariance has been found with an electrophysiological correlate of echoic memory derived from the mismatch negativity (Näätänen et al. 2007) wave of event-related potentials (Alain and Woods 1999).

9.3.2.2 Short-Term Memory

Short-term memory, also sometimes known as primary memory, is the capacity to maintain small quantities of information in the focus of immediate awareness for a short period of time (Waugh and Norman 1965), for example, holding a telephone number in mind long enough to enter it into a phone. This type of processing is reflected in tasks such as forward digit span, where participants listen to and repeat aloud a string of digits in the same order they were presented. While individual studies of forward digit span have revealed modest or no effects of age (Craik 1977), a meta-analysis by Verhaeghen et al. (1993) revealed that older adults do less well on forward digit span tasks than young adults. However, some have argued that the observed age difference in forward digit span is more likely to reflect age differences in long-term memory and speaking articulation rate, which is slower in older adults (Multhaup et al. 1996; Zacks et al. 2000).

9.3.2.3 Repetition Priming

Priming is a nonconscious form of memory typically associated with the perceptual identification of stimuli (Tulving and Schacter 1990). Studies of repetition priming commonly involve initial exposure of a target stimulus, and after a delay period that could vary from seconds to years, a test exposure of the same stimulus, albeit under some form of degradation, obliteration, or compression. A common example from the auditory domain is auditory noise masking, where a word could be spoken clearly, and then in a later test phase of the experiment, presenting that same word with a significant degree of noise. Typically, older adults show similar repetition priming to young adults, although older adults with Alzheimer's disease show impaired repetition priming (Fleischman and Gabrieli 1998).

9.3.2.4 Semantic Priming

In semantic priming, the accessibility of a target item can be changed by prior exposure to a different but conceptually related stimulus. A common example of semantic priming is that of paired associates, where a word comes to mind more easily in the presence of a semantically related cue word (e.g., in *dog*-*cat* or *ocean*-*water* the second word is more readily accessed than *dog*-*water* or *ocean*-*cat* because of the conceptual relationship of the pair). Studies of semantic priming have been used to understand how concept knowledge is organized. The most common form of semantic priming paradigm is when the relationship between the prime (e.g., *dog*) and the target (e.g., *cat*) is manipulated. Given *dog* as a prime, participants are quicker to identify the target word *cat* than when given *corn* as a prime (Neely 1977). To the extent that timing of these responses reflects the underlying semantic network, we can understand the spreading of activation from one lexical entry to another, and the relative integrity of the semantic system. Typically, older adults report equal or stronger semantic priming effects to that of young adults (Burke et al. 1987; Laver and Burke 1993).

9.4 Behavioral and Pupillometric Evidence for Interactions Between Hearing and Cognition

This section reviews studies that have highlighted the interactivity of different sensory and cognitive systems that operate in the service of speech perception and language understanding. While age-related declines to hearing loss and cognition are often studied independently, there is an increasing trend for researchers to study the interactions between hearing loss and cognitive decline as a way to understand the underlying basis for language comprehension. For example, to investigate the question, “How does attention enhance auditory perception?” it may be helpful to study older adults who have declines in attentional control. Conversely, researchers interested in the impacts of hearing on attention may study populations with hearing loss as a way of understanding the input to attentional control systems.

9.4.1 Pupillometric Measures

An exciting development in the understanding of the cognitive demands placed on listeners in difficult auditory environments arises from studies using pupillometry, which relies on measuring pupil size as an index of cognitive effort (Van Engen and McLaughlin 2018). Fluctuations in pupil size are known to happen as a result of light adaptation, but these have also been shown to reflect changes in task demands from attention and memory (Kahneman 1973). Thus, pupillometry provides an online physiological measure of demands incurred while listening.

In a study of middle-aged adults with normal or impaired hearing, Zekveld et al. (2011) found less of a change in pupil dilation when moving from easy to difficult levels of background masking, replicating a prior study conducted with young adults (Zekveld et al. 2010). In both studies, the authors argued this change in pupil dilation reflected a diminished release of effort when in less adverse listening conditions. Such release of effort is anticipated in participants when moving from difficult to less difficult listening conditions and has also been observed when participants give up on a difficult listening task (Zekveld and Kramer 2014).

Kuchinsky et al. (2012) tested older adults’ ability to identify words in background noise and found that pupil size increased as listening became more difficult. Pupil size was also found to increase as a function of the number of phonological competitors of the target word, indicating that participants were experiencing more cognitive demand as a result of lexical competition (McLaughlin et al. 2021). Piquado et al. (2010a) found that linguistic variables such as sentence length and syntactic complexity impact pupil dilation in a task testing memory for spoken sentences. Interestingly, while Piquado et al. tested both young and older adults, only young adults revealed an effect of syntactic complexity on pupil dilation. The authors concluded this finding indicated that older adults were likely processing the syntactic complexity of the sentences to a poorer extent than the young adults, supporting a "good enough" approach to listening in older adults (Amichetti et al. 2016).

9.4.2 Episodic Memory

In a classic study, Rabbitt (1968) showed a remarkable dissociation between identification and memory for spoken words. Participants listened to and repeated spoken digits. In one experiment, the first half of a list was presented in quiet, but the second half could be either in clear or in noise. Items from early in the list (which were always presented clearly, and thus with full intelligibility) were remembered less well when the latter part of the list was in noise. This finding cannot be explained by differences in first-half item acoustics or intelligibility, which were identical in the two conditions. Rabbitt proposed that the effort used to identify words in noise prevented sufficient rehearsal and encoding of *prior* words into long-term memory, and thereby negatively impacted free recall. This mechanism was later confirmed in a study by Piquado et al. (2010b), who found that that acoustic degradation of a single word disrupts memory for not only the degraded word, but also the word presented immediately prior to it. That is, the acoustic degradation interrupted cognitive processes required for memory encoding. Such a finding is not explained by an "auditory-only" framework, but instead supports a role for non-auditory cognitive processes in understanding the degraded speech.

Nearly 25 years after his original experiment, Rabbitt (1991) performed a similar experiment with older adults, and, rather than manipulating the background noise within-subjects, compared groups of older adults with and without hearing loss. He found that those with hearing loss showed poorer free recall than those with good hearing for their age, even when both groups had displayed perfect identification accuracy. Surprenant (2007) found a similar pattern when manipulating noise within-subjects for older and young adults and found that even small levels of auditory degradation that do not show changes in identification accuracy can nevertheless decrease free recall.

To more directly investigate interactions between acoustic and linguistic factors, Koeritzer et al. (2018) presented young and older listeners with lists of spoken sentences in different levels of background noise (multitalker babble). The sentences varied in their linguistic challenge, with half containing one or more ambiguous words ("bark" could refer to the sound a dog makes, or the outer covering of a tree). These high-ambiguity sentences have been shown to rely on domain-general cognitive resources (Rodd et al. 2010, 2012) and may thus potentially interact with acoustic challenge drawing on these same resources. Following an aurally presented list of sentences, listeners performed a visual recognition memory test for those sentences, which revealed that memory was poorer for high-ambiguity sentences, poorer for sentences in more challenging noise conditions, and that the two factors interacted to challenge memory. Perhaps most telling, for the older adults tested, pure-tone hearing sensitivity and measures of verbal working memory both significantly correlated with memory performance in the most challenging condition.

The key takeaway from these studies is that breakdowns in sensory processing, either via hearing loss or acoustic degradation of the stimulus, have a cascading impact upon the cognitive systems required for understanding spoken language (Gordon-Salant and Fitzgibbons 1997). These experimental findings are consistent with the fact that participants with hearing loss report that certain noisy environments require more effort or concentration while listening (Xia et al. 2015). Even if immediate perceptual identification of the stimulus is not impacted, the additional demand on cognitive processing disrupts important functions for language understanding and memory.

9.4.3 The Modulatory Effects of Context

Additional cognitive demands incurred while listening to degraded speech can, in part, be mitigated by the supportive context that frequently occurs in natural speech. McCoy et al. (2005) used a free recall approach and found that hearing-impaired older adults did not have impaired free recall relative to older adults with good hearing when the words shared a semantic context. One reason this may happen is because supportive context may reduce the need for bottom-up sensory fidelity. To this point, Sheldon et al. (2008) found that older and young adults' perceptual thresholds for words were improved when preceded by facilitative priming, sentential context, or a combination of both.

The findings of McCoy et al. (2005) and Sheldon et al. (2008) complement a wider literature that has shown that older adults greatly benefit from the addition of facilitative context as a way to compensate for age-related hearing loss (Pichora-Fuller 2008). In this sense, context provides a basis for expectation and reduces the amount of bottom-up acoustic information needed to achieve successful identification of a stimulus. For example, in a study using a word-onset gating methodology (Grosjean 1980) where listeners were given incrementing 50 ms segments of target words until recognition was achieved, Lash et al. (2013) found that listeners required fewer segments when identifying words preceded by strongly constraining sentences (e.g., "He mailed the letter without a STAMP") compared to weakly constraining sentences (e.g., "He did not say anything about the STAMP"). Lash et al. (2013) also found that this benefit of context was larger for older adults compared to young adults. An initial explanation for older adults' use of context was provided by Sommers and Danielson (1999), who held that semantic and linguistic context improved word identification by reducing the set of potential competitors in the lexicon, reducing the requirement for inhibition of phonological competitors (especially useful for older adults, who have a well-documented inhibition deficit), and thereby facilitating lexical discrimination.

To assess the role of semantic context on speech perception, Rogers et al. (2012) measured young and older adults' word identification for masked target words preceded by clearly presented primes that created facilitative semantic context (e.g., “row-BOAT”), misleading semantic context (e.g., “row-GOAT”), and neutral context conditions (e.g., “cloud-BOAT”). The authors found that older adults had better performance than young adults on facilitative context conditions, but were more likely than young adults to falsely hear the word predicted by the misleading semantic context (e.g., reporting “BOAT” when given “row-GOAT”). This pattern was also reflected in the pattern of young and older adults’ metacognitive monitoring (i.e., confidence in their responses), where older adults were much more confident in making responses that matched the semantic context, even when their responses were incorrect. Such a pattern indicates that context use by older adults does not improve hearing per se, but rather provides a basis for older adults to respond that could be either helpful, or misleading. In the real world, where context is much more likely to be helpful than misleading, this could be of real benefit. However, older adults’ confidence in their responses indicates that this happens without their awareness, and may not be aware of when context is misleading (Rogers and Wingfield 2015; Rogers 2016). Exactly because this reflexive use of context may be useful in the world, it may reflect a “good enough” linguistic processing strategy, where older adults have learned that the potential drawbacks resulting from misleading utterances are not worth the effort needed to detect them (Ferreira et al. 2002; Christianson et al. 2006).

9.5 Neuroimaging Evidence for Interactions Between Hearing and Cognition

Complementing evidence from behavior and pupillometry is a growing literature of functional brain imaging studies that speaks to cognitive processes required to make sense of degraded speech. Only a small number of neuroimaging studies directly examine how hearing loss affects patterns of brain activity, but studies examining responses to a variety of acoustic challenges in listeners with good hearing help provide some context for the types of brain recruitment that might be expected.

9.5.1 Neuroanatomical Frameworks for Spoken Word and Sentence Comprehension

Before considering how aging and hearing loss might affect the brain networks used to understand speech, it is first useful to consider what a “core” language processing network in the brain looks like in the absence of these additional challenges. Fortunately, many decades of research on language processing in patients with brain damage, complemented by functional brain imaging in healthy listeners, have provided a relatively clear picture on what this network might look like.

Acoustic information first reaches the brain bilaterally in primary auditory cortex, and speech perception pathways flow from this initial point. As a general rule, "lower-level" speech features are processed bilaterally and in regions that are anatomically neighboring auditory cortex. Phoneme processing, for example, differentially modulates posterior superior temporal sulcus (STS) (Liebenthal et al. 2005), and single-word comprehension further activates regions of middle temporal gyrus (Binder et al. 2000). Evidence for word processing being supported to at least some degree by both left and right hemisphere comes from findings that listeners who have had their left or right hemisphere inactivated using a Wada procedure (in which a barbiturate is selectively administered to a single hemisphere) are still able to understand words (Hickok et al. 2008).

As the linguistic demands of speech become more complex, additional brain regions are engaged. For example, combinatorial processes that integrate information across multiple words ("plaid jacket" or "red boat" vs. "jacket" or "boat") engage the angular gyrus (Price et al. 2015, 2016) and anterior temporal lobe (Bemis and Pylkkänen 2013; Ziegler and Pylkkänen 2016). The brain networks active during sentence comprehension frequently involve left anterior temporal lobe (Evans et al. 2014) and left inferior frontal gyrus (IFG) (Rodd et al. 2005; Davis et al. 2011), regions that frequently show additional increases when syntactic demands are increased (Peelle et al. 2010b). Thus, regions for speech processing radiate from auditory cortex along dorsal and ventral streams that process increasingly complex aspects of the speech signal (Hickok and Poeppel 2007; Peelle et al. 2010a), and the degree of lateralization depends (at the very least) on the level of linguistic processing required (Peelle 2012).

9.5.2 Executive Attention Networks Respond to Errors in Speech Recognition: The Cingulo-opercular Network

One of the most repeated findings in neuroimaging studies of speech comprehension is elevated activity in the cingulo-opercular network when speech is acoustically challenging enough to result in word recognition errors. The cingulo-opercular network is an executive attention network comprised of the anterior cingulate and bilateral frontal opercula (or perhaps the nearby anterior insulae). These regions can be thought of as a functional network because of their frequent co-activation during various cognitive tasks, and because of the strong correlation of their time courses during rest (Dosenbach et al. 2008; Power and Petersen 2013). The anatomical location of the cingulo-opercular network and its involvement in speech tasks is shown in Fig. 9.3.

The time course of cingulo-opercular activity can provide some indication of its function during cognitive tasks (Neta et al. 2015). Relative to rest, the cingulo-opercular network shows increased activation at the beginning of a task block. However, it shows further punctate increases following errors. Thus, although it appears to have a role in error-monitoring, its function seems better defined as broadly concerned with task engagement, which is needed at the outset of a task, and needs to be revisited following errors.

Activity in the cingulo-opercular network is seldom seen when listeners process unchallenging speech (e.g., speech in quiet). However, when speech is acoustically challenging enough that listeners make mistakes in comprehension or word recognition, the cingulo-opercular network is often engaged (Eckert et al. 2009; Lee et al. 2018). Cingulo-opercular activity has been seen in young adults with good hearing listening to noise-vocoded speech (Wild et al. 2012; Erb et al. 2013), older adults listening to noise-vocoded sentences (Erb and Obleser 2013), and older adults listening to single words in noise (Vaden Jr. et al. 2016).

A particularly informative study in this context was conducted by Vaden and colleagues (2013). They conducted an fMRI study of word perception, with single words presented in background noise at an SNR difficult enough that participants made errors in word recognition. As in prior studies, Vaden et al. found elevated activity in the cingulo-opercular network following these error trials. However, they went one step further and conducted a general linear mixed model (GLMM) analysis to see whether this elevated activity was related to accuracy on the *following* trial. In other words, was cingulo-opercular activity "merely" a response to an error, or did it actually relate to participants' future performance? Their analysis showed that increased activity in the cingulo-opercular network following a word recognition error was indeed correlated with improved accuracy on the next trial. This finding is consistent with a role for the cingulo-opercular network in task engagement and suggests that following an error, participants were able to re-engage with the task (and thus perform more accurately) in proportion to activity in their cingulo-opercular network. Although initially demonstrated in young adults, this finding has also been shown in older adults with age-related hearing loss (Vaden Jr. et al. 2015).

Activity in the cingulo-opercular network also relates to which words are remembered on a subsequent memory test. Vaden and colleagues (2017) conducted an fMRI study in which they played words embedded in background noise for young

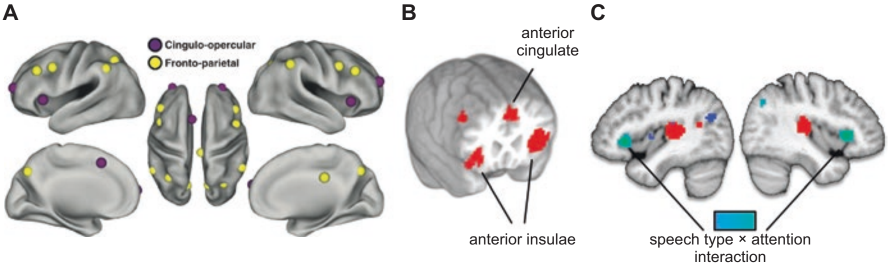

Fig. 9.3 The cingulo-opercular network. (**a**) Nodes of the cingulo-opercular and frontoparietal attention networks, defined by analysis of their time course during resting state fMRI scans (Power and Petersen 2013). (**b**) Increased activity in the cingulo-opercular network when listeners hear speech in noise (Vaden Jr. et al. 2013). (**c**) Interactions between speech clarity and attention during sentence comprehension in the cingulo-opercular network (Wild et al. 2012)

adult listeners with good hearing. Following the listening portion, listeners completed a recognition memory test on the presented words. They found that memory encoding in difficult listening conditions was poorer when cingulo-opercular activity was not sustained, suggesting a role for this network not only in perception but also in memory.

9.5.3 Responses in Prefrontal Cortex and the Successful Perception of Acoustically Challenging Speech

Although there is converging evidence regarding the role of the cingulo-opercular network when listeners make recognition errors, there is less agreement on what other neural and cognitive systems might be involved in supporting successful comprehension. Some additional anatomical evidence comes from studies showing increased activity in regions of prefrontal and premotor cortex when speech is acoustically challenging.

Davis and Johnsrude (2003) presented listeners with spoken sentences that parametrically varied in intelligibility as a result of three different acoustic manipulations: noise vocoding, background noise, or temporal interruption. Varying intelligibility in similar ways but using different signal processing approaches allowed the authors to examine whether responses to changes in intelligibility depended on the specific acoustic form of the signal. For speech that was acoustically degraded, the authors found a large swath of increased activity in left prefrontal cortex. This activity did not depend on the acoustic manipulation used, suggesting it reflects a higher-level response to a decrease of intelligibility.

Although activity in prefrontal cortex is frequently seen when speech is acoustically challenging, there is still a debate about what role this activity may be playing in perception. Some of these regions appear to overlap with portions of the frontoparietal attention network (Power and Petersen 2013), part of a set of regions that respond to a variety of general task demands (Duncan 2010; Jackson et al. 2017).

During acoustically challenging listening situations, activity is also seen in premotor cortex. This observation has led to the suggestion that motor representations may be engaged during speech perception (Watkins et al. 2003; Skipper et al. 2005). That is, when the acoustic signal is unclear, listeners may engage their own motor speech representations to help make sense of the degraded signal. However, it is important to note that the role of motor representations in speech perception is far from clear (Lotto et al. 2009). Outstanding questions remain regarding whether motor activity is obligatory or necessary during speech perception, and the degree to which its role may be influenced by the acoustic clarity of the signal (e.g., whether motor representations may be relied upon differently in quiet compared to in the presence of background noise).

9.6 A Framework for Considering Auditory and Cognitive Interactions in Speech Perception

Although this chapter has focused on auditory and cognitive interactions in listeners with age-related hearing loss, it has also emerged that principles learned from studying this group should generalize to other populations. This section presents a general framework for thinking about auditory-cognitive interactions during speech perception.

The framework, shown in Fig. 9.4a, focuses on speech perception at the level of the individual listener. In a given listening situation, the cognitive demand placed on a listener depends, minimally, on both the acoustic and linguistic challenge of the speech signal. The acoustic challenge reflects contributions of the listener (e.g., hearing sensitivity), the speech signal (e.g., clarity of articulation), and the environment (e.g., background noise) (Denes and Pinson 1993). The linguistic challenge reflects demands of speech processing (single words vs. sentences). For a given level of cognitive demand, the cognitive resources actually engaged by a listener depend on the available resources (e.g., a listener's verbal working memory capacity) and how motivated they are to engage resources to accomplish a task (Eckert et al. 2016; Richter 2016). The term "listening effort" is often applied to this act of cognitive engagement in service of speech comprehension (Pichora-Fuller et al. 2016; Peelle 2018).

An important aspect of this framework, illustrated in Fig. 9.4b, is that cognitive resources are not monolithic. That is, although it is a convenient shorthand,

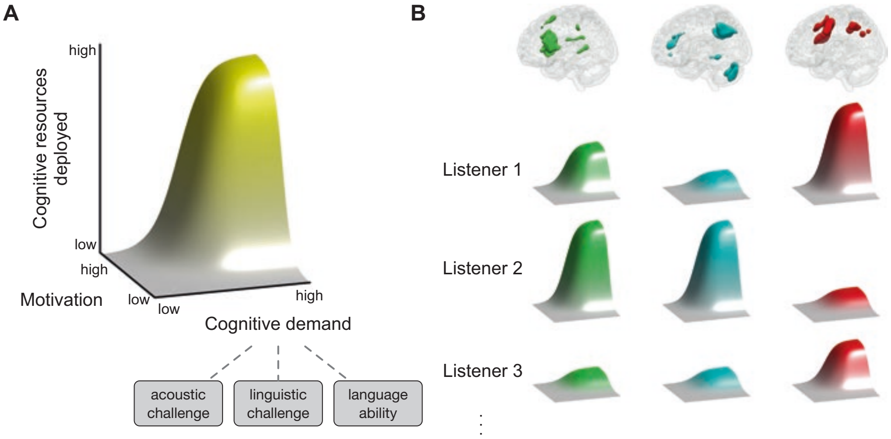

Fig. 9.4 Framework for cognitive resource allocation during speech understanding. (**a**) The cognitive resources engaged during speech understanding vary as a function of the cognitive demands of the task and a listener's motivation to understand what is being said. (**b**) Rather than monolithic "cognitive resources," different listeners may engage dissociable brain networks to various degrees in order to understand what they are hearing. (Figure available via a CC-BY4.0 license from https://osf.io/mv95h/)

speaking of a listener increasing "cognitive resources" grossly oversimplifies what listeners actually do (Wingfield 2016). Rather, each listener has a number of dissociable brain networks that support various cognitive functions, each with a biologically constrained capacity. These various networks can thus be engaged to differing degrees in a particular listening situation.

Another critical point, not necessarily obvious from Fig. 9.4, is that different listeners might achieve a similar level of performance through different patterns of neural engagement. That is, when listening to a talker in a noisy restaurant (causing cognitive demand), one listener may increase activity in the cingulo-opercular network whereas a second listener may increase activity in prefrontal cortex. One would expect this based in part on electrophysiological studies in other animals that suggest multiple combinations of neural activity can result in identical behavior (Prinz et al. 2004; Gutierrez et al. 2013). Thus, even when performance is equated across listeners (including when performance is essentially perfect), listeners may be engaging in different neural "strategies" to achieve this level of performance.

How might this framework be applied in the context of a group of young adults, all of whom have clinically normal hearing? One would expect that by measuring their hearing ability and cognitive ability it would be possible to predict the degree to which they would need to recruit cognitive resources in order to understand speech, assuming they were motivated to do so. And, in fact, if an individual's cognitive ability were lower than the demand, one would expect performance to suffer (e.g., speech might become less intelligible compared to its intelligibility for listeners with greater cognitive ability).

9.7 Summary

Understanding spoken language relies not only on the auditory system, but on linguistic and cognitive processing that acts on the acoustic signal. Individual differences in any of these abilities can affect a listener's success at understanding speech, and the cognitive and neural systems required to achieve this level of success. Because adult aging is associated with changes in both hearing and cognition, it provides an informative window into how these domains interact in all listeners.

One clear area for future growth is that of individual difference analyses, which are important for both theoretical and clinical reasons. Theoretically, contemporary theories (such as the framework outlined in Sect. 9.6) predict that individual differences in the amount of auditory challenge will relate to cognitive demand in individual listeners. Thus, accurate estimates of ability and challenge for an individual listener are required to test this prediction. From a clinical perspective, it is necessary to make judgments about the difficulties and interventions at the level of individual listeners, and thus accurate estimates are required without pooling data across a group. In this context, it will also be critical to ensure that any measures of brain structure or function are reliable at the individual level, which may require more data per individual than is typically collected for group studies (Gordon et al. 2017).

A second area where there is ample room for improvement is moving toward the use of more sophisticated measures of hearing and cognitive ability in the context of brain and cognitive measures. Many of the published studies—particularly fMRI studies—have relied heavily on pure-tone averages as a summary measure of hearing ability. Expanding measures of hearing ability to include multiple frequencies, indications of "hidden" hearing loss, temporal processing, and auditory filter bandwidth, is likely to prove more useful in estimating the auditory challenge faced by individual listeners. Similarly, age-related cognitive decline is a multifaceted concept, and will similarly benefit from more complex measurement approaches. Finally, these increased amounts of data will need more complex theories to constrain their interpretation. These theories need to reflect a more sophisticated understanding which cognitive processes are at play in speech perception and how accurately they can be assessed, so that the conditions under which they are engaged can be determined.

Acknowledgments This work was supported by grants R01DC014281, R21DC015884, and R56AG018029 from the US National Institutes of Health.

Compliance with Ethics Requirements Jonathan Peelle declares no conflicts of interest.

Chad Rogers declares no conflicts of interest.

References

Alain C, Woods DL (1999) Age-related changes in processing auditory stimuli during visual attention: evidence for deficits in inhibitory control and sensory memory. Psychol Aging 14(3):507–519

Alain C, Roye A, Salloum C (2014) Effects of age-related hearing loss and background noise on neuromagnetic activity from auditory cortex. Front Syst Neurosci 8:8

Amichetti NM, White AG, Wingfield A (2016) Multiple solutions to the same problem: utilization of plausibility and syntax in sentence comprehension by older adults with impaired hearing. Front Psychol 7:789

Anderson S, Parbery-Clark A, White-Schwoch T, Kraus N (2012) Aging affects neural precision of speech encoding. J Neurosci 32(41):14156–14164

Anderson S, Parbery-Clark A, White-Schwoch T, Drehobl S, Kraus N (2013) Effects of hearing loss on the subcortical representation of speech cues. J Acoust Soc Am 133(5):3030–3038

Baddeley AD (1986) Working memory. Clarendon Press, Oxford

Balota DA, Dolan PO, Duchek JM (2000) Memory changes in healthy older adults. In: Tulving E, Craik FIM (eds) The Oxford handbook of memory. Oxford University Press, New York, pp 395–409

Bao J, Ohlemiller KK (2010) Age-related loss of spiral ganglion neurons. Hear Res 264:93–97

Belleville S, Rouleau N, Caza N (1998) Effect of normal aging on the manipulation of information in working memory. Mem Cognit 26(3):572–583

Bemis DK, Pylkkänen L (2013) Basic linguistic composition recruits the left anterior temporal lobe and left angular gyrus during both listening and reading. Cereb Cortex 23(8):1859–1873

Bharadwaj HM, Masud S, Mehraei G, Verhulst S, Shinn-Cunningham BG (2015) Individual differences reveal correlates of hidden hearing deficits. J Neurosci 35:2161–2172

Bilger RC, Wang MD (1976) Consonant confusions in patients with sensorineural hearing loss. J Speech Hear Res 19(4):718–748

Binder JR, Frost JA, Hammeke TA, Bellgowan PS, Springer JA, Kaufman JN, Possing ET (2000) Human temporal lobe activation by speech and nonspeech sounds. Cereb Cortex 10(5):512–528

Bugg JM, DeLosh EL, Davalos DB, Davis HP (2007) Age differences in Stroop interference: contributions of general slowing and task-specific deficits. Aging Neuropsychol Cogn 14(2):155–167

Burianova J, Ouda L, Profant O, Syka J (2009) Age-related changes in GAD levels in the central auditory system of the rat. Exp Gerontol 44:161–169

Burke DM, White H, Diaz DL (1987) Semantic priming in young and older adults: evidence for age constancy in automatic and attentional processes. J Exp Psychol Hum Percept Perform 13(1):79–88

Caspary DM, Ling L, Turner JG, Hughes LF (2008) Inhibitory neurotransmission, plasticity and aging in the mammalian central auditory system. J Exp Biol 211:1781–1791

Cerella J (1985) Information processing rates in the elderly. Psychol Bull 98(1):67–83

Christianson K, Williams CC, Zacks RT, Ferreira F (2006) Younger and older adults' "good-enough" interpretations of garden-path sentences. Discourse Process 42(2):205–238

Cowan N (1984) On short and long auditory stores. Psychol Bull 96(2):341–370

Craik FIM (1977) Age differences in human memory. In: Birren S (ed) Handbook of the psychology of aging. Van Nostrand Reinhold, New York, pp 384–420

Craik FI (2008) Memory changes in normal and pathological aging. Can J Psychiatr 53(6):343–345

Craik FIM, McDowd JM (1987) Age differences in recall and recognition. J Exp Psychol Learn 13:474–479

Craik FI, Rabinowitz JC (1985) The effects of presentation rate and encoding task on age-related memory deficits. J Gerontol 40(3):309–315

Craik FI, Rose NS (2012) Memory encoding and aging: a neurocognitive perspective. Neurosci Biobehav Rev 36(7):1729–1739

Cruickshanks KJ, Wiley TL, Tweed TS, Klein BE, Klein R, Mares-Perlman JA, Nondahl DM (1998) Prevalence of hearing loss in older adults in Beaver Dam, Wisconsin: the epidemiology of hearing loss study. Am J Epidemiol 148:879–886

Davis MH, Johnsrude IS (2003) Hierarchical processing in spoken language comprehension. J Neurosci 23(8):3423–3431

Davis MH, Ford MA, Kherif F, Johnsrude IS (2011) Does semantic context benefit speech understanding through "top-down" processes? Evidence from time-resolved sparse fMRI. J Cogn Neurosci 23:3914–3932

de Villers-Sidani E, Alzghoul L, Zhou X, Simpson KL, Lin RCS, Merzenich MM (2010) Recovery of functional and structural age-related changes in the rat primary auditory cortex with operant training. Proc Natl Acad Sci USA 107:13900–13905

del Campo HNM, Measor KR, Razak KA (2012) Parvalbumin immunoreactivity in the auditory cortex of a mouse model of presbycusis. Hear Res 294:31–39

Denes PB, Pinson EN (1993) The speech chain: the physics and biology of spoken language. Waveland Press, Long Grove

Dosenbach NUF, Fair DA, Cohen AL, Schlaggar BL, Petersen SE (2008) A dual-networks architecture of top-down control. Trends Cogn Sci 12:99–105

Duncan J (2010) The multiple-demand (MD) system of the primate brain: mental programs for intelligent behaviour. Trends Cogn Sci 14:172–179

Eckert MA (2011) Slowing down: age-related neurobiological predictors of processing speed. Front Neurosci 5:25

Eckert MA, Menon V, Walczak A, Ahlstrom J, Denslow S, Horwitz A, Dubno JR (2009) At the heart of the ventral attention system: the right anterior insula. Hum Brain Mapp 30:2530–2541

Eckert MA, Cute SL, Vaden KI Jr, Kuchinsky SE, Dubno JR (2012) Auditory cortex signs of age-related hearing loss. J Assoc Res Otolaryngol 13:703–713

Eckert MA, Teubner-Rhodes S, & Vaden KI (2016). Is Listening in Noise Worth It? The Neurobiology of Speech Recognition in Challenging Listening Conditions. Ear and hearing, 37 Suppl 1(Suppl 1), 101S–10S. https://doi.org/10.1097/AUD.0000000000000300

Engle JR, Gray DT, Turner H, Udell JB, Recanzone GH (2014) Age-related neurochemical changes in the rhesus macaque inferior colliculus. Front Aging Neurosci 6:73

Erb J, Obleser J (2013) Upregulation of cognitive control networks in older adults' speech comprehension. Front Syst Neurosci 7:116

Erb J, Henry MJ, Eisner F, Obleser J (2013) The brain dynamics of rapid perceptual adaptation to adverse listening conditions. J Neurosci 33:10688–10697

Evans S, Kyong JS, Rosen S, Golestani N, Warren JE, McGettigan C, Mourão-Miranda J, Wise RJS, Scott SK (2014) The pathways for intelligible speech: multivariate and univariate perspectives. Cereb Cortex 24:2350–2361

Ferreira F, Bailey KGD, Ferraro V (2002) Good-enough representations in language comprehension. Curr Dir Psychol Sci 11:11–15

Fleischman DA, Gabrieli JD (1998) Repetition priming in normal aging and Alzheimer's disease: a review of findings and theories. Psychol Aging 13(1):88–119

Gao F, Wang G, Ma W, Ren F, Li M, Dong Y, Liu C, Liu B, Bai X, Zhao B, Edden RAE (2015) Decreased auditory GABA+ concentrations in presbycusis demonstrated by edited magnetic resonance spectroscopy. NeuroImage 106:311–316

Gazzaley A, Cooney JW, Rissman J, D'Esposito M (2005) Top-down suppression deficit underlies working memory impairment in normal aging. Nat Neurosci 8:1298–1300

Golomb JD, Peelle JE, Addis KM, Kahana MJ, Wingfield A (2008) Age differences in temporal and semantic associations in free and serial recall. Mem Cognit 36:947–956

Goman AM, Lin FR (2016) Prevalence of hearing loss by severity in the United States. Am J Public Health 106(10):1820–1822

Gordon EM, Laumann TO, Gilmore AW, Newbold DJ, Greene DJ, Berg JJ, Ortega M, Hoyt-Drazen C, Gratton C, Sun H, Hampton JM, Coalson RS, Nguyen AL, McDermott KB, Shimony JS, Snyder AZ, Schlaggar BL, Petersen SE, Nelson SM, Dosenbach NUF (2017) Precision functional mapping of individual human brains. Neuron 95(4):791–807

Gordon-Salant S, Fitzgibbons PJ (1997) Selected cognitive factors and speech recognition performance among young and elderly listeners. J Speech Lang Hear Res 40(2):423–431

Gordon-Salant S, Shader MJ, Wingfield A (2020) Age-related changes in speech understanding: peripheral versus cognitive influences. In: Helfer KS, Bartlett EL, Popper AN, Fay RR (eds) Aging and hearing, Springer handbook of auditory research, vol 72. Springer, Cham, pp 199–230

Grosjean F (1980) Spoken word recognition processes and the gating paradigm. Percept Psychophys 28:267–283

Gutierrez GJ, O'Leary T, Marder E (2013) Multiple mechanisms switch an electrically coupled, synaptically inhibited neuron between competing rhythmic oscillators. Neuron 77:845–858

Hasher L, Zacks RT (1988) Working memory, comprehension, and aging: a review and a new view. In: Bower GA (ed) The psychology of learning and motivation: advances in research and theory, vol Vol. 22. Academic Press, San Diego, pp 193–225

Herrmann B, Henry MJ, Johnsrude IS, Obleser J (2016) Altered temporal dynamics of neural adaptation in the aging human auditory cortex. Neurobiol Aging 45:10–22

Hertzog C, Dunlosky J (2011) Metacognition in later adulthood: spared monitoring can benefit older adults' self-regulation. Curr Dir Psychol Sci 20(3):167–173

Hickok G, Poeppel D (2007) The cortical organization of speech processing. Nat Rev Neurosci 8:393–402

Hickok G, Okada K, Barr W, Pa J, Rogalsky C, Donnelly K, Barde L, Grant A (2008) Bilateral capacity for speech sound processing in auditory comprehension: evidence from Wada procedures. Brain Lang 107:179–184

Humes LE, Kidd GR, Lentz JJ (2013) Auditory and cognitive factors underlying individual differences in aided speech-understanding among older adults. Front Syst Neurosci 7:55

Jackson J, Rich AN, Williams MA, Woolgar A (2017) Feature-selective attention in frontoparietal cortex: multivoxel codes adjust to prioritize task-relevant information. J Cogn Neurosci 29(2):310–321

Kahneman D (1973) Attention and effort. Prentice Hall, Englewood Cliffs

Kelley CM, Sahakyan L (2003) Memory, monitoring, and control in the attainment of memory accuracy. J Mem Lang 48:704–721

Koeritzer MA, Rogers CS, Van Engen KJ, Peelle JE (2018) The impact of age, background noise, semantic ambiguity, and hearing loss on recognition memory for speech. J Speech Lang Hear Res 61:740–751

Kuchinsky SE, Vaden KI Jr, Keren NI, Harris KC, Ahlstrom JB, Dubno JR, Eckert MA (2012) Word intelligibility and age predict visual cortex activity during word listening. Cereb Cortex 22:1360–1371

Kujawa SG, Liberman MC (2009) Adding insult to injury: Cochlear nerve degeneration after "temporary" noise-induced hearing loss. J Neurosci 29:14077–14085

Lash A, Rogers CS, Zoller A, Wingfield A (2013) Expectation and entropy in spoken word recognition: effects of age and hearing acuity. Exp Aging Res 39:235–253

Laver GD, Burke DM (1993) Why do semantic priming effects increase in old age? A meta-analysis. Psychol Aging 8(1):34–43

Lee YS, Wingfield A, Min NE, Kotloff E, Grossman M, Peelle JE (2018) Differences in hearing acuity among "normal-hearing" young adults modulate the neural basis for speech comprehension. eNeuro 5:e0263–0217.2018. https://doi.org/10.1523/ENEURO.0263-17.2018

Liebenthal E, Binder JR, Spitzer SM, Possing ET, Medler DA (2005) Neural substrates of phonemic perception. Cereb Cortex 15:1621–1631

Ling LL, Hughes LF, Caspary DM (2005) Age-related loss of the GABA synthetic enzyme glutamic acid decarboxylase in rat primary auditory cortex. Neuroscience 132:1103–1113

Lotto AJ, Hickok GS, Holt LL (2009) Reflections on mirror neurons and speech perception. Trends Cogn Sci 13:110–114

McCabe DP, Roediger HLI, McDaniel MA, Balota DA, Hambrick DZ (2010) The relationship between working memory capacity and executive functioning: evidence for a common executive attention construct. Neuropsychology 24:222–243

McCoy SL, Tun PA, Cox LC, Colangelo M, Stewart R, Wingfield A (2005) Hearing loss and perceptual effort: downstream effects on older adults' memory for speech. Q J Exp Psychol 58(1):22–33

McLaughlin DJ, Zink M, Gaunt L, Spehar B, Van Engen KJ, Sommers MS, Peelle JE (2021) Pupillometry reveals cognitive demands of lexical competition during spoken word recognition in young and older adults. PsyArXiv. https://doi.org/10.31234/osf.io/6pa3g

Melton AW (1963) Implications of short-term memory for a general theory of memory. J Verb Learn Verb Behav 2:1–21

Merchant SN, Nadol JB (2010) Schuknecht's pathology of the inner ear, 3rd edn. People's Publishing House, Shelton

Mitchell P, Gopinath B, Wang JJ, McMahon CM, Schneider J, Rochtchina E, Leeder SR (2011) Five-year incidence and progression of hearing impairment in an older population. Ear Hear 32:251–257

Morrell CH, Gordon-Salant S, Pearson JD, Brant LJ, Fozard JL (1996) Age- and gender-specific reference ranges for hearing level and longitudinal changes in hearing level. J Acoust Soc Am 100(4):1949–1967

Multhaup KS, Balota DA, Cowan N (1996) Implications of aging, lexicality, and item length for the mechanisms underlying memory span. Psychon Bull Rev 3(1):112–120

Näätänen R, Paavilainen P, Rinne T, Alho K (2007) The mismatch negativity (MMN) in basic research of central auditory processing: a review. Clin Neurophysiol 118(12):2544–2590

Neely JH (1977) Semantic priming and retrieval from lexical memory: roles of inhibitionless spreading activation and limited-capacity attention. J Exp Psychol Gen 106:226–254

Nelson TO, Narens L (1990) Metamemory: a theoretical framework and some new findings. In: Bower GH (ed) The psychology of learning and motivation, vol 26. Academic Press, New York, pp 125–173

Neta M, Miezin FM, Nelson SM, Dubis JW, Dosenbach NUF, Schlaggar BL, Petersen SE (2015) Spatial and temporal characteristics of error-related activity in the human brain. J Neurosci 35:253–266

Park DC, Royal D, Dudley W, Morrell R (1988) Forgetting of pictures over a long retention interval in young and older adults. Psychol Aging 3(1):94–95

Park DC, Smith AD, Lautenschlager G, Earles JL, Frieske D, Zwahr M, Gaines CL (1996) Mediators of long-term memory performance across the life span. Psychol Aging 11(4):621–637

Parkinson SR, Perey A (1980) Aging, digit span, and the stimulus suffix effect. J Gerontol 35(5):736–742

Paul BT, Bruce IC, Roberts LE (2017) Evidence that hidden hearing loss underlies amplitude modulation encoding deficits in individuals with and without tinnitus. Hear Res 344:170–182

Peelle JE (2012) The hemispheric lateralization of speech processing depends on what "speech" is: a hierarchical perspective. Front Hum Neurosci 6:309

Peelle JE (2018) Listening effort: how the cognitive consequences of acoustic challenge are reflected in brain and behavior. Ear Hear 39:204–214

Peelle JE, Wingfield A (2016) The neural consequences of age-related hearing loss. Trends Neurosci 39:486–497

Peelle JE, Johnsrude IS, Davis MH (2010a) Hierarchical processing for speech in human auditory cortex and beyond. Front Hum Neurosci 4:51

Peelle JE, Troiani V, Wingfield A, Grossman M (2010b) Neural processing during older adults' comprehension of spoken sentences: age differences in resource allocation and connectivity. Cereb Cortex 20:773–782

Peelle JE, Troiani V, Grossman M, Wingfield A (2011) Hearing loss in older adults affects neural systems supporting speech comprehension. J Neurosci 31:12638–12643

Pichora-Fuller MK (2008) Use of supportive context by younger and older adult listeners: balancing bottom-up and top-down information processing. Int J Audiol 47:S72–S82

Pichora-Fuller MK, Kramer SE, Eckert MA, Edwards B, Hornsby BWY, Humes LE, Lemke U, Lunner T, Matthen M, Mackersie CL, Naylor G, Phillips NA, Richter M, Rudner M, Sommers MS, Tremblay KL, Wingfield A (2016) Hearing impairment and cognitive energy: the framework for understanding effortful listening (FUEL). Ear Hear 37:5S–27S

Piquado T, Isaacowitz D, Wingfield A (2010a) Pupillometry as a measure of cognitive effort in younger and older adults. Psychophysiology 47:560–569

Piquado T, Cousins KAQ, Wingfield A, Miller P (2010b) Effects of degraded sensory input on memory for speech: Behavioral data and a test of biologically constrained computational models. Brain Res Bull 1365:48–65

Plomp R, Mimpen AM (1979) Speech-reception threshold for sentences as a function of age and noise level. J Acoust Soc Am 66:1333–1342

Power JD, Petersen SE (2013) Control-related systems in the human brain. Curr Opin Neurobiol 23:223–228

Price AR, Bonner MF, Peelle JE, Grossman M (2015) Converging evidence for the neuroanatomic basis of combinatorial semantics in the angular gyrus. J Neurosci 35:3276–3284

Price AR, Peelle JE, Bonner MF, Grossman M, Hamilton RH (2016) Causal evidence for a mechanism of semantic integration in the angular gyrus as revealed by high-definition transcranial direct current stimulation. J Neurosci 36(13):3829–3838

Prinz AA, Bucher D, Marder E (2004) Similar network activity from disparate circuit parameters. Nat Neurosci 7(12):1345–1352

Profant O, Balogová Z, Dezortová M, Wagnerová D, Hájek M, Syka J (2013) Metabolic changes in the auditory cortex in presbycusis demonstrated by MR spectroscopy. Exp Gerontol 48:795–800

Rabbitt PMA (1968) Channel capacity, intelligibility and immediate memory. Q J Exp Psychol 20:241–248

Rabbitt PMA (1991) Mild hearing loss can cause apparent memory failures which increase with age and reduce with IQ. Acta Otolaryngol Suppl 476:167–176

Richter, Michael (2016) The Moderating Effect of Success Importance on the Relationship Between Listening Demand and Listening Effort, Ear and Hearing: July/August 2016 - Volume 37 - Issue - p 111S–117S. https://doi.org/10.1097/AUD.0000000000000295

Rodd JM, Davis MH, Johnsrude IS (2005) The neural mechanisms of speech comprehension: fMRI studies of semantic ambiguity. Cereb Cortex 15:1261–1269

Rodd JM, Johnsrude IS, Davis MH (2010) The role of domain-general frontal systems in language comprehension: evidence from dual-task interference and semantic ambiguity. Brain Lang 115:182–188

Rodd JM, Johnsrude IS, Davis MH (2012) Dissociating frontotemporal contributions to semantic ambiguity resolution in spoken sentences. Cereb Cortex 22:1761–1773

Rogers CS (2016) Semantic priming, not repetition priming, is to blame for false hearing. Psychon Bull Rev 24:1194–1204

Rogers CS, Wingfield A (2015) Stimulus-independent semantic bias misdirects word recognition in older adults. J Acoust Soc Am 138:EL26

Rogers CS, Jacoby LL, Sommers MS (2012) Frequent false hearing by older adults: the role of age differences in metacognition. Psychol Aging 27:33–45

Rogers CS, Payne L, Maharjan S, Wingfield A, Sekuler R (2018) Older adults show impaired modulation of attentional alpha oscillations: evidence from dichotic listening. Psychol Aging 33(2):246–258

Rose NS, Myerson J, Sommers MS, Hale S (2009) Are there age differences in the executive component of working memory? Evidence from domain-general interference effects. Aging Neuropsychol Cogn 16(6):633–653

Rudner M, Ronnberg J (2008) The role of the episodic buffer in working memory for language processing. Cogn Process 9(1):19–28

Salthouse TA (1991) Mediation of adult age differences in cognition by reductions in working memory and speed of processing. Psychol Sci 2(3):179–183

Salthouse TA (1996a) General and specific speed mediation of adult age differences in memory. J Gerontol B Psychol Sci Soc Sci 51(1):30–42

Salthouse TA (1996b) The processing-speed theory of adult age differences in cognition. Psychol Rev 103(3):403–428

Scharenborg O, Weber A, Janse E (2015) Age and hearing loss and the use of acoustic cues in fricative categorization. J Acoust Soc Am 138(3):1408–1417

Sheldon S, Pichora-Fuller MK, Schneider BA (2008) Priming and sentence context support listening to noise-vocoded speech by younger and older adults. J Acoust Soc Am 123(1):489–499

Skipper JI, Nusbaum H, Small SL (2005) Listening to talking faces: motor cortical activation during speech perception. NeuroImage 25:76–89

Skoe E, Krizman J, Anderson S, Kraus N (2015) Stability and plasticity of auditory brainstem function across the lifespan. Cereb Cortex 25(6):1415–1426

Smith AD (1977) Adult age differences in cued recall. Dev Psychol 13:326–331

Sommers MS, Danielson SM (1999) Inhibitory processes and spoken word recognition in young and older adults: the interaction of lexical competition and semantic context. Psychol Aging 14:458–472

Souchay C, Isingrini M (2004) Age related differences in metacognitive control: role of executive functioning. Brain Cogn 56(1):89–99

Sperling G (1960) The information available in brief visual presentations. Psychol Monogr Gen Appl 74:1–29

Stroop J (1935) Studies of interference in serial verbal reactions. J Exp Psychol 18:643–662

Surprenant AM (2007) Effects of noise on identification and serial recall of nonsense syllables in older and younger adults. Aging Neuropsychol Cogn 14(2):126–143

Tulving E, Schacter DL (1990) Priming and human memory systems. Science 247(4940):301–306

Tulving E, Szpunar KK (2009) Episodic memory. Scholarpedia 4(8):3332

Tun PA, O'Kane G, Wingfield A (2002) Distraction by competing speech in young and older adult listeners. Psychol Aging 17(3):453–467

Vaden KI Jr, Kuchinsky SE, Cute SL, Ahlstrom JB, Dubno JR, Eckert MA (2013) The cingulo-opercular network provides word-recognition benefit. J Neurosci 33:18979–18986

Vaden KI Jr, Kuchinsky SE, Ahlstrom JB, Dubno JR, Eckert MA (2015) Cortical activity predicts which older adults recognize speech in noise and when. J Neurosci 35:3929–3937

Vaden KI Jr, Kuchinsky SE, Ahlstrom JB, Teubner-Rhodes SE, Dubno JR, Eckert MA (2016) Cingulo-opercular function during word recognition in noise for older adults with hearing loss. Exp Aging Res 42:67–82

Vaden KI Jr, Teubner-Rhodes S, Ahlstrom JB, Dubno JR, Eckert MA (2017) Cingulo-opercular activity affects incidental memory encoding for speech in noise. NeuroImage 157:381–387

Van Engen KJ, McLaughlin DJ (2018) Eyes and ears: using eye tracking and pupillometry to understand challenges to speech recognition. Hear Res 369:56–66

Verhaeghen P, De Meersman L (1998) Aging and the Stroop effect: a meta-analysis. Psychol Aging 13(1):120–126

Verhaeghen P, Salthouse TA (1997) Meta-analyses of age-cognition relations in adulthood: estimates of linear and nonlinear age effects and structural models. Psychol Bull 122(3):231–249

Verhaeghen P, Marcoen A, Goossens L (1993) Facts and fiction about memory aging: a quantitative integration of research findings. J Gerontol 48(4):P157–P171

Watkins KE, Strafella AP, Paus T (2003) Seeing and hearing speech excites the motor system involved in speech production. Neuropsychologia 41(8):989–994

Waugh NC, Norman DA (1965) Primary memory. Psychol Rev 72:89–104

Weeks JC, Hasher L (2017) Older adults encode more, not less: evidence for age-related attentional broadening. Aging Neuropsychol Cogn 25:576–587

Wild CJ, Yusuf A, Wilson D, Peelle JE, Davis MH, Johnsrude IS (2012) Effortful listening: the processing of degraded speech depends critically on attention. J Neurosci 32:14010–14021

Wingfield A (2016) Evolution of models of working memory and cognitive resources. Ear Hear 37:35S–43S

Wingfield A, Kahana MJ (2002) The dynamics of memory retrieval in older adulthood. Can J Exp Psychol 56(3):187–199

Wingfield A, Tun PA, McCoy SL (2005) Hearing loss in older adulthood: what it is and how it interacts with cognitive performance. Curr Dir Psychol Sci 14:144–148

Xia J, Nooraei N, Kalluri S, Edwards B (2015) Spatial release of cognitive load measured in a dual-task paradigm in normal-hearing and hearing-impaired listeners. J Acoust Soc Am 137(4):1888–1898

Zacks RT, Hasher L, Li KZH (2000) Human memory. In: FIM C, Salthouse TA (eds) The handbook of aging and cognition, 2nd edn. Lawrence Erlbaum Associates, Mahwah, pp 293–357

Zekveld AA, Kramer SE (2014) Cognitive processing load across a wide range of listening conditions: insights from pupillometry. Psychophysiology 51:277–284

Zekveld AA, Kramer SE, Festen JM (2010) Pupil response as an indication of effortful listening: the influence of sentence intelligibility. Ear Hear 31:480–490

Zekveld AA, Kramer SE, Festen JM (2011) Cognitive load during speech perception in noise: the influence of age, hearing loss, and cognition on the pupil response. Ear Hear 32:498–510

Ziegler J, Pylkkänen L (2016) Scalar adjectives and the temporal unfolding of semantic composition: an MEG investigation. Neuropsychologia 89:161–171