UNIVERSITY OF CALIFORNIA,
IRVINE

Psychophysical and neurophysiological investigations from three approaches to understanding human speech processing

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Psychology

by

Jonathan Henry Venezia

Dissertation Committee:
Professor Gregory Hickok, Chair
Professor Kourosh Saberi
Assistant Professor Alyssa Brewer

2014

DEDICATION

To

Megan, Nancy and Jesus

each of whom carried me through

graduate school at one time or another.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Page

ACKNOWLEDGMENTS

Thank you to Greg. Your method of critical analysis is what will stay with me the longest. You don't overthink, you don't get swept away, and you don't deal with garbage. You seem to home in effortlessly on the most reasonable interpretation of data. I believe this comes from a measured use of the interpretive mechanism – for you, reason, logic, and the synthesis of facts are its obligatory antecedents. Your practical advice to a young investigator has been indispensable as well.

Thank you also to my committee members, Alyssa Brewer and Kourosh Saberi. Through our collaborations I have benefited from a wealth of technical knowledge and theoretical insight.

Special thanks to the following:

Collaborators who contributed extensively to the work herein: Charlie Chubb, Julius Fridriksson, Steven Thurman, Paul Fillmore, and Brian Barton.

UC Irvine Faculty who gave more than a few moments of their free time to dispense advice: Ginny Richards, Ted Wright, Emily Grossman, Joachim Vandekerckhove, and Mimi Liljeholm.

Senior members of the lab who were instrumental in my technical training: Kai Okada, Corianne Rogalsky, Kenny Vaden, and Lisette Isenberg.

Two lab members who were incredibly supportive friends through different phases of graduate school: Dale Maddox and William Matchin.

Former mentors: Robert Sommer, Tamara Swaab, and Andy Yonelinas. You gave me my start, for better or worse.

Jeff Krichmar for allowing me to attend his lab's social functions.

Eunice Huang for looking after the health of my own brain while I tried to understand everyone else's.

I thank John Wiley and Sons for permission to include Chapter One of my dissertation, which originally appeared in *Language and Linguistics Compass*. Chapter Two originally appeared in *Frontiers in Psychology* and is licensed under the Creative Commons Attribution 3.0. The co-author listed in these publications (Gregory Hickok) directed and supervised research which forms the basis for the dissertation.

Financial support was provided by the University of California, Irvine and the National Institute on Deafness and Other Communication Disorders Award DC010775 from the University of California, Irvine.

CURRICULUM VITAE

Jonathan Henry Venezia
2014

Education

2014	University of California, Irvine Ph.D., Psychology Specialization: Cognitive Neuroscience
2008	University of California, Davis B.A., Psychology, Highest Honors

Awards and Honors

2012	Graduate Merit Fellowship, Center for Cognitive Neuroscience, University of California, Irvine
2010	John I. Yellott Award, Department of Cognitive Sciences, University of California, Irvine
2009, 2010	Honorable mention, National Science Foundation Graduate Research Fellowship Program
2008	Herbert A. Young Award (most outstanding graduate of the College of Letters and Science), University of California, Davis
2008	Highest Honors (awarded for senior research project), Department of Psychology, University of California, Davis
2008	Psychology Departmental Citation for Excellence, University of California, Davis
2008	Finalist, University Medal, University of California, Davis
2007	Phi Beta Kappa Honor Society
2007	Phi Kappa Phi Honor Society
2006	Finalist, Lawrence J. Andrews Award, University of California, Davis
2003-2008	Dean’s Honors List, University of California Davis, College of Letters and Science

Research Appointments

2010-2011 Interdisciplinary Training Program in Hearing Research (supported by NIDCD Award DC010775), Center for Hearing Research, University of California, Irvine

2009-2013 Graduate Student Researcher in Auditory and Language Neuroscience, Department of Cognitive Sciences, University of California, Irvine

2007-2008 Research Assistant in Cognitive Neuroscience of Language, Department of Psychology, University of California, Davis

2007-2008 Research Assistant in Human Memory, Department of Psychology, University of California, Davis

Teaching Appointments

2009-2014 Teaching Assistant, School of Social Sciences, University of California, Irvine
Introductory Psychology
Research Methods in Psychology
Computer-based Research in Social Science
Sensation and Perception
Brain Disorders
Language and Brain

Scientific Appointments

2010-2014 Ad-hoc Reviewer, *Brain & Language, Neuroimage, Psychonomic Bulletin & Review, Human Brain Mapping, Cortex, Frontiers in Language Sciences, Language & Linguistics Compass, Annual Conference of the Society for the Neurobiology of Language*

Professional Affiliations

Society for Neuroscience
Cognitive Neuroscience Society
Society for the Neurobiology of Language

Publications

Book Chapters:

Venezia J.H., Matchin, W. & Hickok, G. (In press). Multisensory Integration and Audiovisual Speech Perception. In A.W. Toga, M.M. Mesulam & S. Kastner (Eds.), *Brain Mapping: An Encyclopedic Reference*. Oxford, U.K.: Elsevier Limited.

Peer-Reviewed Journal Articles:

Venezia, J.H., Rong, F., Maddox, D., Saberi, K., & Hickok, G. (Submitted). The visual speech stream: Facial motion processing, audiovisual integration, and speech perception in the superior temporal sulcus. *Journal of Cognitive Neuroscience.*

Okada, K., Venezia, J.H., Matchin, W., Saberi, K. & Hickok, G. (2013). An fMRI Study of Audiovisual Speech Perception Reveals Multisensory Interactions in Auditory Cortex. PloS ONE 8: e68959.

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

Venezia, J.H., Saberi, K., Chubb, C. & Hickok, G. (2012). Response bias modulates the speech motor system during syllable discrimination. *Frontiers in psychology*, 157: 1-13.

Okada, K., Rong, F., Venezia, J.H. et al (2010). Hierarchical Organization of Human Auditory Cortex: Evidence from Acoustic Invariance in the Response to Intelligible Speech. *Cerebral Cortex*, 20: 2486-2495.

Venezia, J.H., and Hickok, G. (2009). Mirror neurons the motor system and language: From the motor theory to embodied cognition and beyond. *Language and Linguistics Compass*, 3: 1-14.

Conference Abstracts:

Venezia, J.H., Fillmore, P., Hickok, G. & Fridriksson, J. (2014). Cortical network for sensorimotor integration of audio-visual speech. Poster Presentation at the Annual Meeting of the Society for the Neurobiology of Language.

Venezia, J.H., Barton, B., Saberi, K., Brewer, A. & Hickok, G. (2013). The distribution of cortical surface area dedicated to auditory temporal receptive fields is symmetric between hemispheres in auditory core and belt. Poster Presentation at the Annual Meeting of the Cognitive Neuroscience Society.

Barton, B., Venezia, J.H., Saberi, K., Hickok, G., & Brewer, A.A. (2013). Orthogonal acoustic dimensions define auditory field maps in human cortex. Poster presentation at the Annual Meeting of the Cognitive Neuroscience Society.

Brewer, A.A., Barton, B., Venezia, J.H., Saberi, K. & Hickok, G. (2013). Cross sensory activation of 'clover leaf' clusters in human auditory and visual cortex. Poster presentation at the Annual Meeting of the Cognitive Neuroscience Society.

Barton, B., Venezia, J.H., Saberi, K., Hickok, G., & Brewer, A.A. (2013). Orthogonal acoustic dimensions define auditory field maps in human cortex. Poster presentation at the Annual Meeting of the Cognitive Neuroscience Society.

Venezia, J.H., Barton, B., Saberi, K., Brewer, A. & Hickok, G. (2012). The distribution of cortical surface area dedicated to auditory temporal receptive fields is symmetric between hemispheres in auditory core and belt. Poster Presentation at the Neurobiology of Language Conference.

Venezia, J.H., Matchin, W. & Hickok, G. (2012). Human “mirror system” can be trained to respond to arbitrary non-action related objects. Poster presentation at the Annual Meeting of the Cognitive Neuroscience Society.

Barton, B., Venezia, J.H., Saberi, K., Hickok, G., & Brewer, A.A. (November, 2011). Orthogonal maps of tonotopy and periodicity in the human auditory core. Poster presentation at the Annual Meeting of the Society for Neuroscience.

Venezia, J.H, Saberi, K., Hickok, G. (2011). Activation in motor speech regions tracks with experimentally induced bias. Poster presentation at the Neurobiology of Language Conference.

Venezia, J.H, Saberi, K., Hickok, G. (2011). Activation in motor speech regions tracks with experimentally induced bias. Poster presentation at the Annual Meeting of the Cognitive Neuroscience Society.

Venezia, J.H., Rong, F., Maddox, C.D., Saberi, K., Hickok, G. (2010). Integration of audiovisual speech does not rely on Broca’s area. Poster presentation at the annual meeting of the Cognitive Neuroscience Society.

Maddox, C.M., Venezia, J.H., Hickok, G. (2010). Functionally hierarchical organization of the superior temporal sulcus in speech perception. Poster presentation at the annual meeting of the Cognitive Neuroscience Society

Okada, K., Venezia, J., Matchin, W., Saberi, K., Hickok, G. (2009). Visual speech increases the gain of the response to auditory speech in human primary auditory cortex. Poster presentation at the annual meeting of the Cognitive Neuroscience Society.

Okada, K., Venezia, J., Matchin, W., Saberi, K., Serences, J., Hickok, G. (2008). Early auditory cortical regions discriminate intelligible from unintelligible speech. Poster presentation at the annual meeting of the Cognitive Neuroscience Society.

Boudewyn, M., Venezia, J., Gordon, P., Camblin, C., Polse, L., Swaab, T. (2008). The modulation of semantic priming by discourse and sentence contexts during speech comprehension: an ERP Study. Poster presented at the annual meeting of the Cognitive Neuroscience Society.

Murray, L.J., Venezia, J., Yonelinas, A.P., Parks, C.M. (2008). The effects of complexity on source memory performance. Poster session presented at the annual UC Davis Undergraduate Research, Scholarship, and Creative Activities Conference, Davis, California.

ABSTRACT OF THE DISSERTATION

Psychophysical and neurophysiological investigations from three approaches to understanding human speech processing

By

Jonathan Henry Venezia

Doctor of Philosophy in Psychology

University of California, Irvine, 2014

Professor Gregory Hickok, Chair

Human speech processing (perception and in some cases production) is approached from three levels. At the top level, I investigate the role of the motor system in top-down processing and decision-making during speech perception. At the middle level, I investigate the mechanisms underlying integration of auditory and visual speech for both perception and production of speech. At the bottom level, I investigate the organized representation of temporal modulations in sound, with an eye toward structure that may provide insight into how speech sound representations are built. The primary investigative techniques throughout are auditory and visual psychophysics and functional MRI (sometimes combined). The main findings of the investigations can be summarized briefly as follows. First, the motor system does not participate meaningfully in speech perception. Rather, speech motor activity is modulated by taxing decision-level mechanisms in laboratory speech tasks. Second, discrete visual features appear to be extracted from visual speech signals and integrated with auditory speech representations in the superior temporal sulcus (STS). Results are equivocal with respect to the level of processing at which this occurs, although speculation is provided. Also, there are dedicated sensorimotor integration networks for visual speech. Third, slow temporal modulations in sound are represented in an auditory-cortical place code that magnifies the expression of modulations within the range that is most common in natural speech (4-16 Hz).

INTRODUCTION

The chapters that follow contain a series of investigations into the computational and neural mechanisms underlying perception and production of speech. The typical motivation for studying these mechanisms is that speech circuits are presumed to be highly specialized components of the human language system (Lenneberg, Chomsky, & Marx, 1967). Language itself is of interest to anyone who wishes to understand the human mind and brain. Namely, language is a uniquely human capacity – it allows the user of any specific language to produce and understand an unbounded number of expressions, whereas other organisms are simply not equipped with this level of flexibility (Petitto, 2005). However, although speech and language are straightforwardly connected (speech is fundamentally involved in the externalization of language for communication), the nature of this connection is not clear. Speech is not necessary for normally-functioning language, yet speech systems appear to be embedded within neural circuitry that is specialized for linguistic processing (Hickok, Bellugi, & Klima, 1998).

At this point, it may be useful to clarify exactly what is meant by 'language.' From the perspective of cognitive science, language is an internal computational system that operates somewhere between transduction of external sensory signals (or production of actions) and construction of (or operation over) internal representations – namely, thoughts, ideas, beliefs, or concepts. This definition, broad and somewhat vague as it is, contains a strong implication that there exists some degree of modularity in cognitive systems (Fodor, 1983) and the neural machinery unpinning them (Gall, 1835), a notion that I accept as a natural fact about the organization of the mind. Having accepted this, it perhaps follows straightforwardly that the language system evolved as a means of externalizing thoughts, ideas, etc., for the purpose of communication (including, of course, the reverse process – i.e., mapping external communicative messages back to internal representations). Clearly language enables such functions. To give an informal example, when I attend an academic talk I essentially extract the main idea of the talk from the speaker's words, represent it internally, and I can then report it back (perhaps to other colleagues) in my own words. This is not simply a matter of capturing and regenerating the message verbatim. Rather, the message is transformed – the particular sequence of sounds is transduced, parsed, categorized, and ultimately abstracted to an efficient internal code (linguistic and conceptual structure are imposed); the message can then be recapitulated by applying the reverse transformation.

However, it need not be the case that language evolved with communication as its key function. According to Chomsky and colleagues, the Faculty of Language (Hauser, Chomsky, & Fitch, 2002) amounts in the narrowest sense to a specialized combinatorial operation for merging two elements (roughly word-like at the lowest level) into an *unordered* set containing the two original elements in unmodified form. Further, while physical constraints on the sensorimotor systems that interface with the language faculty impose a sequential linear order on sensorimotor computations (e.g., the words composing a sentence must be pronounced one after the other rather than simultaneously), order does not enter into computations that construct internal conceptual representations (Chomsky, 2007). As such, language is said to function primarily as an 'instrument of thought' (R. Berwick & Chomsky, 2011), i.e., linguistic computations are naturally structured to support internal mental operations., while externalization of language for communication is merely an ancillary function (R. C. Berwick, Friederici, Chomsky, & Bolhuis, 2013).

What is the significance of all this? I have spent the preceding paragraphs fleshing out a definition of language and briefly exploring its function in order to emphasize the (potential) distinction between language (in the narrow sense) and the sensorimotor systems that support the externalization of language for communication. This is because, as mentioned above, I study language at the sensorimotor interface (speech, in my case). For many who study language at this level, there is (at least) an implicit belief that sensorimotor speech systems are no less a part of language than the highly specialized computations described by Chomsky. This belief has merit. As mentioned above, sensorimotor systems impose real computational constraints on the use of language for communication. Following from this, we might learn something about language as a whole by first describing these constraints – including how and why they are imposed to solve computational problems – and later understanding how they govern interactions between external and internal linguistic representations at the sensorimotor interface. Indeed, a great deal of what we know about the relationship between language and the brain has been inferred from language dysfunction caused by damage to sensorimotor speech systems (Goodglass, 1993). Moreover, the first and perhaps most influential cognitive-neuroscientific model of language essentially describes a sensorimotor speech circuit specialized for interfacing with lexical-conceptual systems (Wernicke, 1969). Recent incarnations of this model suggest that sensorimotor speech networks themselves have an intrinsic organization that is specialized to accommodate different linguistic levels of processing (Hickok, 2014).

I am sympathetic to the notion that, to state the program specifically, understanding sensorimotor speech circuits will contribute to a comprehensive understanding of language, and, to restate the program more broadly, that we can understand elements of higher-level cognitive systems by understanding their inputs and outputs. However, as I asserted previously, cognitive systems are modular, so at some point this program will run out of room – specifically, if we assume that cognitive systems can be individuated on the basis of intrinsic specializations, it would not make sense to begin to study language, for instance, by studying the structure of the inner ear, even though inner ear functions clearly interact with the language system via auditory speech. A modular system is perhaps best understood using an approach in which the goal is to understand the fundamental structure and principles of the system in its own right. How, then, should one choose a level at which to study language (or the brain, or cognition generally for that matter)? Quite honestly, I am not strongly motivated to draw any hard lines when it comes to answering this particular question. Whether my investigations on speech perception, for example, will ultimately reveal something about the organization of language or the organization of auditory perception is rather inconsequential to me. Each system belongs equally to the matrix we call cognition, and to understand either of these systems, so distinguished, would be significant.

At the end of the day, I conceive of speech perception and production as processes embedded within a 'sensorimotor speech' system, specialized in its own right (at least at the interface with linguistic representations) (Hickok & Poeppel, 2007) and more or less part of the language system depending on the particular definition chosen. At the very least, understanding the sensorimotor speech system should lead to tangible, real-world benefits in terms of diagnosis and treatment of clinical disorders ranging from peripheral auditory dysfunction (Frisina & Frisina, 1997) to high-level language disorders (Goodglass, 1993). This practical thrust along with a loose faith in the bottom-up approach to studying language (i.e., through speech) and a strong conviction that I can study the sensorimotor speech system in its own right serve as the primary motivations for my research. Having established this, I will now move on to describe my approach to studying sensorimotor speech.

Essentially, I have taken a three-pronged approach to understanding the sensorimotor speech system. There is an objective associated with each "prong," described as follows: (1) clarify what speech perception *is* by establishing what speech perception *is not*; (2) establish how speech systems interface with signals from multiple sensory modalities; (3) establish the organization of central representations of auditory signals. Chapters 1 and 2 focus on objective (1), and in particular attempt to distinguish computations related to speech perception from computations related to speech production and decision-making. Chapters 3, 4, and 5 focus on objective (2) in the context of audiovisual speech – namely, these chapters identify some computational properties for integration of auditory and visual speech, describe where precisely in the brain this occurs, and determine whether and how visual speech signals interface with speech production systems. Chapter 6 works toward objective (3) by mapping the intrinsic organization of cortical auditory systems, including evaluation of potential specializations that may benefit speech perception. The included studies are a combination of literature review (Ch.1) and original investigations using psychophysical (Ch. 2 & 3) and neuroimaging (Ch. 2, 4, 5, 6) techniques. Each chapter is itself a standalone manuscript in publication format. Chs. 1 and 2 are published[1], and the remaining chapters are prepared for submission. I will provide a brief primer at the beginning of each chapter that links the chapter back to the overall approach, but only broadly. In truth, the individual studies are not strongly related to each other at a detailed level (at least not across objectives), but I have chosen to include each investigation because they unite under the broad objective of understanding the sensorimotor speech system.

To conclude, I will briefly expand upon the motivation behind each objective in the three-pronged approach described immediately above. Regarding objective (1), there has been

[1] Previously published work may be slightly altered from its published form to accommodate formatting changes, etc.

[1]Note these studies also found activation in superior temporal sulcus (STS). See below for an alternative discussion

considerable debate over the form of the so-called ‘objects of speech perception’ (Diehl, Lotto, & Holt, 2004). One set of theories asserts that articulatory gestures are the objects of speech perception (Fowler, 1986; Liberman & Mattingly, 1985; Liberman & Whalen, 2000), while another asserts that the objects of speech perception are fundamentally auditory in nature (Diehl & Kluender, 1989; Massaro, 1987; Stevens, 1989). While gesture-based theories disagree over whether a perceiver’s own speech motor system is recruited in speech perception (Fowler, 1996), and even over whether speech perception requires any deep understanding of gestural sources (Fowler & Magnuson, 2012), these theories are frequently understood in terms of the dominant Motor Theory of Speech Perception (MT). Under MT, speech perception proceeds by mapping incoming acoustic signals to internal motor representations via a vocal-tract synthesizer (Liberman & Mattingly, 1985). When I began my own investigations in 2008, MT, which had previously been debunked in the eyes of many speech scientists (Pardo & Remez, 2006), was experiencing a resurgence due in part to the discovery of mirror neurons (Fadiga & Craighero, 2006; Rizzolatti & Craighero, 2004). In particular, while certain claims of MT appeared to be untestable using behavioral techniques (MacNeilage, 1991), modern neuroimaging provided at least circumstantial evidence for MT by demonstrating that brain regions involved in speech production are also active during perception of speech (Pulvermuller et al., 2006; Watkins, Strafella, & Paus, 2003; Wilson, Saygin, Sereno, & Iacoboni, 2004). Chapter 1 examines these results in detail and establishes an alternative viewpoint – namely, perceiving speech activates the motor system because sensory speech representations are used to guide speech production. Chapter 2 is original work demonstrating that the speech motor system may be involved in top down components of speech perception engaged during typical laboratory tasks.

Regarding objective (2), interest in multimodal (audiovisual) speech perception can be related back to the debate over the objects of speech perception. First, let me establish that there is clear evidence that auditory and visual speech signals interact in speech perception (McGurk & MacDonald, 1976; Ross, Saint-Amour, Leavitt, Javitt, & Foxe, 2007; Sumby & Pollack, 1954). Given this fact, it would be informative to know precisely how these interactions occur and at what level of processing. If the objects of speech perception are gestural, then the level of interaction between visual and auditory speech is straightforward – gestural information present in the visual signal (in the form of observable actions) is combined with gestural information recovered from the auditory signal. If the objects of speech perception are auditory (the position I support), then the nature of the interaction between visual and auditory speech is less clear. While some information such as dynamic temporal patterns may be isomorphic across visual and auditory signals (Jiang, Auer, Alwan, Keating, & Bernstein, 2007), other complementary cues must somehow be extracted from the visual speech and combined with auditory speech representations (Summerfield, 1987). The nature of these visual speech cues, including the information they carry and the level at which they interact with auditory speech cues, should provide general insight into the organization of speech processing. Chapter 3 develops a new technique for identifying the visual speech cues extracted during perception, while Ch. 4 examines where in the brain these cues are extracted and at which stage in the multisensory speech processing stream. Chapter 5 examines whether visual speech information is integrated with the speech motor system during production and, if so, whether the circuits involved mirror auditory-motor speech circuits.

Finally, objective (3) is based on the assumption that, if speech perception is carried out by general auditory mechanisms (Holt & Lotto, 2008), there should be a mechanism by which complex speech objects are constructed out of lower-level auditory features, similar to hierarchical object processing in vision (Riesenhuber & Poggio, 1999). Recent evidence from neuroimaging suggests that visual cortical areas are organized into clusters of visual field maps, which may help to provide a common reference frame between nodes in the visual processing hierarchy (Brewer & Barton, 2012). I have recently been involve in research demonstrating that auditory cortical areas are also organized into clusters of auditory field maps (Barton, Venezia, Saberi, Hickok, & Brewer, 2012). Chapter 6 examines the detailed organization of these auditory field maps with respect to low-level temporal auditory features that may be crucial extraction of speech objects (Poeppel, 2003).

References

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

Berwick, Robert C, Friederici, Angela D, Chomsky, Noam, & Bolhuis, Johan J. (2013). Evolution, brain, and the nature of language. *Trends in cognitive sciences, 17*(2), 89-98.

Berwick, Robert, & Chomsky, Noam. (2011). The biolinguistic program: The current state of its evolution and development. *The biolinguistic enterprise: New perspectives on the evolution and nature of the human language faculty*, 19-41.

Brewer, A.A., & Barton, B. (2012). Visual field map organization in human visual cortex. *Visual Cortex, InTech*.

Chomsky, Noam. (2007). Approaching UG from below. *Interfaces+ recursion= language*, 1-29.

Diehl, Randy L, & Kluender, Keith R. (1989). On the objects of speech perception. *Ecological Psychology, 1*(2), 121-144.

Diehl, Randy L, Lotto, Andrew J, & Holt, Lori L. (2004). Speech perception. *Annu. Rev. Psychol., 55*, 149-179.

Fadiga, Luciano, & Craighero, Laila. (2006). Hand actions and speech representation in Broca's area. *Cortex, 42*(4), 486-490.

Fodor, Jerry A. (1983). *The modularity of mind: An essay on faculty psychology*: MIT press.

Fowler, Carol A. (1986). An event approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics, 14*(1), 3-28.

Fowler, Carol A. (1996). Listeners do hear sounds, not tongues. *The Journal of the Acoustical Society of America, 99*(3), 1730-1741.

Fowler, Carol A, & Magnuson, James S. (2012). Speech perception. *The Cambridge Handbook of Psycholinguistics*, 3-20.

Frisina, D Robert, & Frisina, Robert D. (1997). Speech recognition in noise and presbycusis: relations to possible neural mechanisms. *Hearing research, 106*(1-2), 95-104.

Gall, Franz Joseph. (1835). *On the Function of the Brain and of Each of Its Parts*: Marsh, Capen & Lyon.

Goodglass, Harold. (1993). *Understanding aphasia*: Academic Press.

Hauser, Marc D, Chomsky, Noam, & Fitch, W Tecumseh. (2002). The faculty of language: What is it, who has it, and how did it evolve? *science, 298*(5598), 1569-1579.

Hickok, Gregory. (2014). Towards an integrated psycholinguistic, neurolinguistic, sensorimotor framework for speech production. *Language, Cognition and Neuroscience, 29*(1), 52-59.

Hickok, Gregory, Bellugi, Ursula, & Klima, Edward S. (1998). The neural organization of language: evidence from sign language aphasia. *Trends in cognitive sciences, 2*(4), 129-136.

Hickok, Gregory, & Poeppel, David. (2007). The cortical organization of speech processing. *Nat Rev Neurosci, 8*(5), 393-402.

Holt, Lori L, & Lotto, Andrew J. (2008). Speech perception within an auditory cognitive science framework. *Current Directions in Psychological Science, 17*(1), 42-46.

Jiang, Jintao, Auer, Edward T, Alwan, Abeer, Keating, Patricia A, & Bernstein, Lynne E. (2007). Similarity structure in visual speech perception and optical phonetic signals. *Perception & psychophysics, 69*(7), 1070-1083.

Lenneberg, Eric H, Chomsky, Noam, & Marx, Otto. (1967). *Biological foundations of language* (Vol. 68): Wiley New York.

Liberman, Alvin M, & Mattingly, Ignatius G. (1985). The motor theory of speech perception revised. *Cognition, 21*(1), 1-36.

Liberman, Alvin M, & Whalen, Doug H. (2000). On the relation of speech to language. *Trends in cognitive sciences, 4*(5), 187-196.

MacNeilage, Peter F. (1991). *Comment: The gesture as a unit in speech perception theories.* Paper presented at the Modularity and the Motor Theory of Speech Perception: Proceedings of a Conference to Honor Alvin M. Liberman.

Massaro, D.W. (1987). *Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry*: Erlbaum Associates.

McGurk, Harry, & MacDonald, John. (1976). Hearing lips and seeing voices.

Pardo, Jennifer S, & Remez, Robert E. (2006). The perception of speech. *The handbook of psycholinguistics, 2*, 201-248.

Petitto, Laura-Ann. (2005). How the brain begets language. *The cambridge companion to chomsky*, 84-101.

Poeppel, D. (2003). The analysis of speech in different temporal integration windows: cerebral lateralization as 'asymmetric sampling in time'. *Speech Communication, 41*(1), 245-255.

Pulvermuller, F., Huss, M., Kherif, F., Moscoso del Prado Martin, F., Hauk, O., & Shtyrov, Y. (2006). Motor cortex maps articulatory features of speech sounds. *Proc Natl Acad Sci U S A, 103*(20), 7865-7870. doi: 10.1073/pnas.0509989103

Riesenhuber, Maximilian, & Poggio, Tomaso. (1999). Hierarchical models of object recognition in cortex. *Nature neuroscience, 2*(11), 1019-1025.

Rizzolatti, Giacomo, & Craighero, Laila. (2004). The mirror-neuron system. *Annu. Rev. Neurosci., 27*, 169-192.

Ross, Lars A, Saint-Amour, Dave, Leavitt, Victoria M, Javitt, Daniel C, & Foxe, John J. (2007). Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments. *Cerebral Cortex, 17*(5), 1147-1153.

Stevens, Kenneth N. (1989). On the quanta! nature of speech. *Journal of phonetics, 17*, 3-45.

Sumby, William H, & Pollack, Irwin. (1954). Visual contribution to speech intelligibility in noise. *The Journal of the Acoustical Society of America, 26*(2), 212-215.

Summerfield, Q. (1987). Some preliminaries to a comprehensive account of audio-visual speech perception. In Dodd (Ed.), *Hearing by eye: The psychology of lip-reading*: Lawrence Erlbaum Associates.

Watkins, K. E., Strafella, A. P., & Paus, T. (2003). Seeing and hearing speech excites the motor system involved in speech production. *Neuropsychologia, 41*(8), 989-994. doi: 10.1016/s0028-3932(02)00316-0

Wernicke, Carl. (1969). *The symptom complex of aphasia.* Paper presented at the Proceedings of the Boston Colloquium for the Philosophy of Science 1966/1968.

Wilson, S. M., Saygin, A. P., Sereno, M. I., & Iacoboni, M. (2004). Listening to speech activates motor areas involved in speech production. *Nat Neurosci, 7*(7), 701-702. doi: 10.1038/nn1263

CHAPTER 1

Primer

As mentioned in the Introduction, there is a long-standing debate in speech science concerning the so-called ‘objects of speech perception.’ One camp holds that the objects of speech perception are articulatory gestures, while another camp holds that the objects of speech perception are fundamentally auditory. Of the gestural accounts, The Direct Realist Theory (DRT) asserts that articulatory gestures produce specifiers or invariants in the acoustic signal that allow gestures to be perceived directly from the speech signal (Fowler, 1994). This seems to render the DRT essentially an auditory theory. The most influential gestural account, the Motor Theory of Speech Perception (MT) (Liberman & Mattingly, 1985), asserts that gestures must be recovered from the acoustic signal using the listener’s own motor system. However, MT has failed, throughout its various iterations, to account for well-known neuropsychological evidence demonstrating that speech perception is unaffected by damage to the speech motor system (see below). Despite this, gestural theories (MT in particular) have seen a recent resurgence in popularity thanks to the recent discovery of mirror neurons (among other factors reviewed below). In short, gestural theories have become a nuisance to serious (auditory-based) research programs that aim to understand speech perception. To return to the objectives listed in the main Introduction, the current chapter aims to establish definitively what speech perception *is not*. In short, speech perception *is not* perception of articulatory gestures.

Mirror neurons, the motor system and language: from the motor theory to embodied cognition and beyond

Jonathan H. Venezia and Gregory Hickok

Introduction – The Motor Theory of Speech Perception

A major problem in speech perception research is the lack of invariance in the relation between acoustic patterns and the speech sound percepts they generate; that is, the same phoneme may have a very different acoustic pattern in one context compared to another (Liberman, Delattre, & Cooper, 1952; Liberman, Delattre, Cooper, & Gestman, 1954). This results from coarticulation of speech gestures: the vocal tract gestures for successive speech sounds overlap temporally (Browman & Goldstein, 1986). Liberman (1957) noticed, however, that the gestures that produced a given phonemic percept were always similar even if the resulting acoustic pattern wasn't. In other words, perception tracks articulation. This observation lead to the development of the motor theory of speech perception (Liberman, 1957; Liberman, Cooper, Shankweiler, and Studdert-Kennedy, 1967; Liberman & Mattingly, 1985). There are three central tenets of the classic motor theory spanning its several iterations (outlined by Galantucci, Fowler, & Turvey, 2006): (1) that speech processing is special, (2) that perceiving speech is perceiving vocal tract gestures, and (3) that speech perception involves access to the motor system. According to the most recent version of the motor theory, we perceive phonetic segments in terms of motor commands in the brain that control speech production (Liberman & Mattingly, 1985).

In the decades following its introduction, intensive investigation of the major tenets of the motor theory led to empirical challenges and the theory fell out of favor among speech scientists. (Thorough reviews including summaries of the arguments for and against the motor theory can be found elsewhere in the literature–see Galantucci et al., 2006; Massaro & Chen, 2008; Lotto, Hickok, & Holt, 2009). However, the discovery of mirror neurons has thrust the motor theory once again to the forefront of the discussion on the neural implementation of speech perception. Therefore, our focus will be on mirror neurons and the studies they have inspired, and more specifically on the contribution of these studies to the current understanding of speech comprehension. Such inspiration is welcome as, indeed, descriptions of the neural computations underlying speech comprehension are often ambiguous and sorely underspecified. Nevertheless, we will show that human studies stemming from the mirror neuron literature have done little to resolve the problems in speech comprehension research, particularly within the framework of the motor theory of speech perception.

From here, we will review the proposed function of mirror neurons and their possible role in language processing. This will lead us to examine the motor system's role in speech comprehension, along with the contribution of recent evidence to a motor-based interpretation of speech perception. Such an interpretation stands in contrast to alternative views, which focus more heavily on the role of sensory representations in speech recognition. For example, the fuzzy logical model of speech perception (Massaro, 1987; Massaro, 1998) describes speech perception as a feed-forward process of pattern recognition where multiple sources of information (e.g., auditory, visual, tactile) influence speech perception directly – signals from bottom-up perceptual processes are evaluated and integrated to provide an overall degree of support for possible speech alternatives, and feedback after perception can be used to tune the prototypical values used by the evaluation process. The dual stream model of speech perception (Hickok & Poeppel, 2000/2004/2007) postulates the existence of a dorsal processing stream that maps acoustic speech information onto motor speech representations to guide speech production (see discussion below). We will argue that tight sensory-motor coupling – the most common observation in research that seeks to link the motor system with speech perception – is consistent with these "sensory first" theories of speech comprehension, and that such theories provide the best explanation of evidence concerning the role of the motor system in speech comprehension.

Mirror Neurons

The discovery of mirror neurons in the monkey frontal cortex has lead to renewed interest in the motor theory of speech perception and indeed has been taken as evidence in support of the model (di Pellegrino, Fadiga, Fogassi, Gallese, & Rizzolatti, 1992; Rizzolatti & Arbib, 1998; Rizzolatti & Craighero, 2004; Iacoboni, 2008). In the following sections, we discuss the proposed function of mirror neurons, their response properties, and a potential role for the human mirror system in action understanding.

The Discovery of Mirror Neurons

Mirror neurons were discovered during single-cell recording from macaque monkey (*macaca nemestrina*) area F5 in the inferior prefrontal cortex (di Pellegrino et al., 1992).

Previous studies in area F5 revealed sub-populations of cells with sensory, motor, and sensory-motor properties – most cells in the region respond during execution of motor acts such as grasping, holding, and tearing, and a fraction of these also respond to passive somatosensory (~40%) or visual (~17%) stimulation in the absence of action (Rizzolatti et al., 1988). The key finding in seminal work on mirror neurons was that a number of the cells responsive during both sensory stimulation and motor production were congruent with respect to object-directed action; that is, these cells respond to production and observation of the *same* action (e.g., when the monkey grasps a raisin from a tray or watches another individual grasp a raisin from the same tray; see Figure 1.1). Such congruency is the defining aspect of mirror neurons, and the property for which they are named. Mirror neurons are often cited as playing a role in action understanding (e.g., Gallese et al., 1996, Rizzolatti & Craighero, 2004), likely due to the aforementioned congruity in sensory and motor activation for the same action. Interestingly, canonical (non-mirror) F5 neurons, which are active during action production and perception of objects, are not typically implicated in object identification or understanding, but rather in visuomotor transformation for sensory access to motor acts (Nelissen, Luppino, Vanduffel, Rizzolatti, & Orban, 2005).

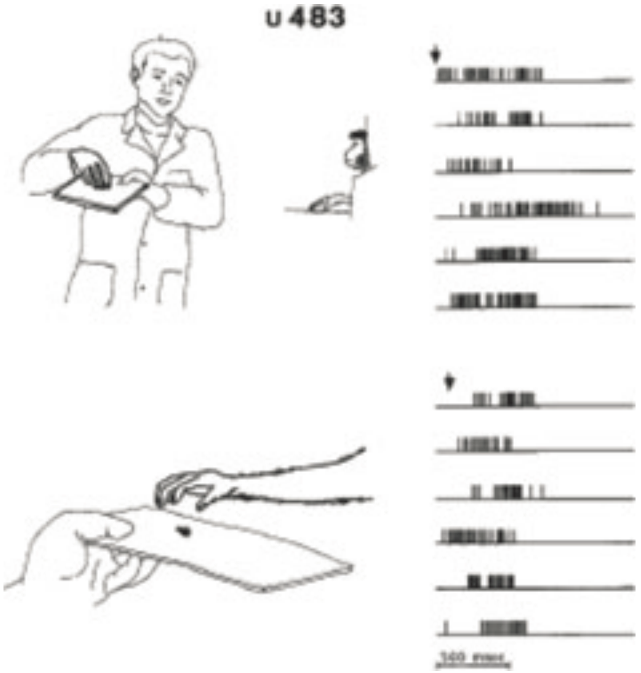


Figure 1.1. An example of a single unit selectively discharging to observation of the experimenter grasping an object (a) and the monkey grasping the same object (b). Arrowheads indicate the approximate onset of the grasping motion (from di Pellegrino et al. 1992).

Properties of Mirror Neurons

Mirror neurons do not respond to visual stimulation other than that generated by observation of goal-directed action, and do not exhibit movement preparation activity: they discharge when the monkey observes an action, stop firing when the action terminates, and remain quiet even if the object is moved toward the monkey, firing again only when the monkey initiates its own action (Gallese et al., 1996). This is an important fact as this property distinguishes mirror neurons from "set-related" neurons in monkey area 6 that discharge before movement onset (Weinrich, Wise, & Mauritz, 1984; Wise & Mauritz, 1985). As important controls for the possibility that "mirror activity" reflected some form of covert movement, Gallese, Fadiga, Fogassi, and Rizzolatti (1996) recorded from the hand area of primary motor cortex (F1 or M1), and recorded muscle activity (via electromyography or EMG) from several hand and mouth muscles during action observation. No M1 cells fired, and no EMG activity was elicited in response to action observation. Mirror neurons have subsequently been discovered in the rostral part of the inferior parietal cortex (Gallese, Fogassi, Fadiga & Rizzolatti, 2002) and, contrary to initial studies, in area M1 (Tkach, Reimer, & Hatsopoulos, 2007). Critical to the discussion of mirror neurons in language processing is the observation that monkey area F5 is considered by some to be the homolog of Broca's region in humans (Nishitani, Martin, Schürmann, Amunts & Hari, 2005), and that auditory mirror neurons – neurons that respond to producing an action and hearing the sound associated with that action – have been discovered in this area (Kohler et al., 2002).

Though work with monkeys reveals no direct evidence concerning the role of mirror neurons in action understanding (Hickok, 2009), there has been great interest in probing for a human analog to mirror neurons that might serve such a function. Several studies have identified mirror-like responses in the human central nervous system (e.g., Fadiga, Fogassi, Pavesi, & Rizzolatti, 1995; Iacoboni et al., 1999), and subsequent studies have attempted to establish a more direct connection between motor systems and action processing (Urgesi, Candidi, Ionta & Aglioti, 2007; Urgesi, Calvo-Merino, Haggard & Aglioti, 2007; Pazzaglia et al., 2008). In all, however, evidence of mirror activity in human primary and peripheral motor cortices is at best weakly consistent with a mirror theory of action understanding, and in no way does the available evidence indicate that the human mirror system is directly involved in coding the meaning of actions (we refer the interested reader to the thorough set of arguments presented by Hickok, 2009). Nonetheless, the notion of embodied semantics has inspired a host of studies seeking to investigate the role of the putative human mirror system in language processing. Many such studies seek to investigate a proposed connection between the motor system and action-semantics in language tasks (e.g., Buccino, Riggio, Melli, Binkofski, Gallese, & Rizzolatti, 2005; Pulvermuller, Hauk, Nikulin, & Ilmoniemi, 2005; Boulenger, Roy, Paulignan, Deprez, Jeannerod, & Nazir, 2006; Glenberg, Sato, Cattaneo, Riggio, Palumbo, & Buccino, 2008), though these studies will not be our focus as they are subject to similar problems as the mirror system literature reviewed above (Hickok, in press). Instead, we will highlight studies that bear directly on the discussion of motor involvement in speech processing, and we will discuss this evidence within the framework of the motor theory of speech perception. Thus, in the upcoming sections we review behavioral and neurophysiological evidence concerning the role of the motor system in speech processing.

The Motor System and Language

Let us return now to the motor theory of speech perception. If the resolution of discrete phonological units and the mapping of those units onto semantic representations were critically associated with activation of invariant speech production codes, then we would expect to see activation in motor regions during speech comprehension. We would further expect to see activation related to more than just preparatory motor activity since, as established above, speech comprehension is often paired closely in time with speech production, and such a pairing should clearly lead to motor activation simply on the basis of learned association. We will show that the evidence for motor activation specific to speech perception itself is ambiguous at best. In addition, it will be informative to review behavioral evidence for the involvement of motor systems in speech perception. If the motor system is indeed active during speech perception, and the motor theory of speech perception is accurate, we should expect to find behavioral evidence that confirms a functional role for motor activation during speech perception. Such behavioral evidence should be the foundation on which neurophysiological investigation of motor involvement in speech is built, though we will demonstrate that the behavioral evidence is sparse and at best weakly supportive of the motor theory. Finally, if speech perception is indeed critically dependent on the motor system, we should expect to see neuropsychological evidence demonstrating that damage to frontal speech production networks impairs speech comprehension ability. In fact, we will show that the evidence supports the opposite conclusion – specifically, damage to frontal motor regions does not impair speech comprehension, and speech comprehension can be impaired with frontal motor regions intact.

Motor Involvement in Speech Perception – Behavioral Evidence

Most behavioral evidence for involvement of the motor system in speech perception focuses on the identification of perceptual-motor links, i.e., evidence that acoustic stimuli affect performance in a speech production task or vice versa. However, evidence of this sort is not a clear indication that the motor system is directly involved in speech perception. Take as an example the finding of selective adaptation in speech production (Cooper, 1979). In a perceptual selective adaptation paradigm (Eimas & Corbit, 1973), repeated presentations of a syllable, e.g. /pa/, lead to fewer identifications of that syllable along an ambiguous continuum, e.g. /ba/-to-/pa/. In the speech production version of the paradigm, repeated auditory presentations of a syllable lead to reductions in voice onset times when subjects are asked to produce the same syllable. Thus, the motor system is being "primed" for production of a given speech sound when that sound is repeatedly heard. Such perceptual “priming” of the motor system is to be expected if we (reasonably) assume the presence associative links between perceptual and motor systems, but this does not necessarily indicate that speech perception is carried out by the motor system. A similar finding indicates that individual differences in a vowel production task replicate when subjects are asked to discriminate the same vowels (Bell-Berti, Raphael, Pisoni, & Sawusch, 1979). In this study, subjects first produced a series of vowels that differed in height. Consistent with the phonetic distinction in height, 4 of 10 speakers showed a gradual decrease in activity of a tongue muscle (the genioglossus muscle, which affects tongue height) as the height of vowels decreased. In a later perception test of the same vowels, the 4 speakers who had shown a height distinction in production showed larger anchoring effects when asked to discriminate the vowels on a continuum (decreased identifications of the vowel sound at the end of the continuum). The other 6 subjects did not show such large anchoring effects, demonstrating that subjects grouped in the perceptual task as they had in the production task. A possible explanation of such a finding is that the way an individual produces speech influences the way speech is perceived by the same individual. However, the opposite position is equally tenable, namely that the way an individual perceives speech influences the way speech is produced by that individual. The direction of the relationship is ambiguous given the result and, again, we find no conclusive evidence of motor involvement in speech perception.

Perhaps a more compelling finding comes in a variant of the McGurk effect (McGurk & MacDonald, 1976). The typical effect occurs when audio and visual speech information are mismatched (hearing “ba” while watching someone articulate “ga”), which can induce a perceptual illusion (“da”). A recent study found that a McGurk-like effect can be induced not only by *viewing* incongruent speech gestures, but by the listener’s own incongruent speech gestures (Sams et al., 2005). Listeners silently articulated speech sounds that were either congruent or incongruent with the syllables they were listening to. The incongruent condition led to significantly more misperceptions of the heard speech (32% correct) than the congruent condition (95% correct) suggesting that motor representations of speech can influence sensory perception of speech sounds (Sams, Mottonen, & Sihvonen, 2005). It is important to note that such a finding is not inconsistent with sensory-first theories of speech perception. In fact, it has been suggested that the source of this influence is via efferent copies of motor commands that are transmitted to auditory regions, and that this process may form a kind of predictive (forward model) mechanism that modulates the analysis of sensory input (Sams et al., 2005; Poeppel, Idsardi, & van Wassenhove, 2008; Okada & Hickok, 2009). Thus, there is behavioral evidence of motor involvement in speech perception, but we should not conclude that speech perception per se is a charge of the motor system. The neural evidence for motor involvement in speech perception likewise should not lead us to draw such a conclusion.

Mirror Neurons, Broca's Region, and Premotor Cortex – A New Motor Theory?

As mentioned above, the resurgence of interest in the motor theory of speech perception stems in large part from the discovery of mirror neurons and the analogous human mirror system. In the realm of language, it has been proposed that the human mirror system is at the center of an evolutionary development from a primitive gestural communication system to a more advanced system capable of supporting language production and comprehension. Rizzolatti and Arbib (1998) thoroughly outline the theory behind this proposed evolutionary connection. Noting activation of Broca's area during action observation (Grafton, Arbib, Fadiga, Rizzolatti, 1996;

Rizzolatti et al., 1996)[1], they suggest a tight coupling between a pre-human gestural communication system, mediated by mirror system activity, and speech production in humans, mediated by Broca's area. However, the theory makes no claim as to the coevolution of a speech perception system in frontal motor networks. Nonetheless, this theory, along with the discovery of auditory mirror neurons in monkey area F5 (Kohler et al., 2002), has motivated researchers to investigate the neurophysiological role of frontal motor systems in speech perception. Can we find any direct evidence that the motor system is active during speech perception in humans?

Fadiga, Craighero, Buccino and Rizzolatti (2002) used TMS to test the excitability of tongue muscles during speech comprehension. They reasoned that, according to the motor theory of speech perception, muscles involved in the articulation of speech sounds should be more excited when those sounds are more taxing to produce. Fadiga and colleagues used TMS to stimulate the sector of motor cortex that controls tongue muscles while recording MEPs directly from the tongue. Subjects listened to words through headphones, where one group of words required strong tongue movements and the other group required only slight tongue movements. Motor-evoked potentials revealed increased activity in the tongue muscles when words were the type that required strong tongue movements. Thus, while listening to others speak, the comprehender tracks incoming speech sounds with movements of the tongue. Combining positron emission tomography (PET) and TMS, Watkins and Paus (2004) were able to show that increased excitability in speech production areas of motor cortex is correlated with activity in the posterior part of the LIFG (Broca's region). The authors proposed that activity in Broca's region "primes" the motor system for language production in response to spoken speech, whether or not output is required. Such an explanation is consistent with the account given

[1] Note these studies also found activation in superior temporal sulcus (STS). See below for an alternative discussion of STS is speech perception.

above – that perceptual priming of the motor system or, likewise, motoric “priming” (e.g., forward models) of perceptual systems is what we should expect given the tight coupling of speech perception with speech production. The directionality of this relationship is ambiguous, and thus there is no evidence here to suggest a direct role for the motor system in the perception of speech sounds or the decoding of their meaning.

Following the lead of these studies, two fMRI experiments showed activation in similar premotor areas during both speech perception and speech production (Wilson, Saygin, Sereno, & Iacoboni, 2004; Pulvermuller et al., 2006). Later, Wilson, Molnar-Szakacs, and Iacoboni (2008) used inter-subject correlational analyses in fMRI to test whether premotor areas were responsive to time-varying characteristics of linguistic input. They argued that such an analysis would identify active voxels across subjects that were sensitive to neural activity that varies in time with stimulus properties (cf. Hasson, Nir, Levy, Fuhrmann, & Malach, 2004). Results indicated that a premotor area responded to time-varying characteristics of the input during continuous narrative speech, which may be construed as evidence that the motor system directly tracks the incoming acoustic input. Again, however, these results are not inconsistent with feed-forward sensory models of speech perception where sensory speech representations are fed to motor areas in order to guide speech production. Indeed, if the processes that subserve speech perception and production are critically dependent on proper timing, as they are during natural verbal exchanges, we should expect processing in one domain to track processing in the other, either directly or via executive function capable of influencing processing in either domain (or both). This is not evidence that speech representations are fundamentally motoric. Such a necessary role for the motor system can only be established by demonstrating a disruption of speech perception abilities when the motor system is disabled.

Just such a demonstration was attempted by Meister and colleagues (Meister, Wilson, Debleck, Wu, & Iacoboni, 2007) in a study entitled *The Essential Role of Premotor Cortex in Speech Perception*. Subjects were asked to discriminate voiceless stop consonants in single syllables presented in noise (i.e., the task is relatively hard) while repetitive TMS was applied to premotor cortex. Relative to tone and color discrimination controls, TMS caused a small but significant disruption in discrimination performance in the speech task (discrimination performance dropped by 8% from 78.9% correct in a baseline task). A possible mechanism for the observed effect was proposed: "premotor cortex generates forward models... that are compared within the superior temporal cortex with the results from initial acoustic-speech analysis... Premotor cortex provides top-down information that facilitates speech perception in circumstances such as when the acoustic signal is degraded...." (p. 1694). Again, this explanation is not inconsistent with sensory-first theories of speech perception. Though it is clear that frontal systems can play a role in speech recognition, the results do not support the conclusion that these systems are the primary mechanism or 'essential' for perceiving speech.

A more recent study acknowledged the failure of Meister et al. (2007) to provide convincing proof of a causal role for the motor system in speech perception (D'Ausilio, Pulvermuller, Salmas, Bufalari, Begliomini, & Fadiga, 2009, p. 381), and reported a stronger finding demonstrating that stimulation of human motor cortex via TMS directly affects the perception of speech sounds. Stimulation pulses were applied to lip and tongue areas of primary motor cortex while participants were asked to identify speech sounds involved in prominent lip articulation, [b] and [p], or prominent tongue articulation, [d] or [t]. A double-dissociation was observed in the reaction time data: relative to a non-stimulation baseline, participants were faster to identify tongue-related sounds when the tongue area was stimulated, and faster to identify lip-related sounds when lip areas were stimulated. The authors claimed this as evidence that speech perception is grounded in motor circuits. However, there are several alternative explanations that describe the data equally well. First, it may be that stimulation of primary motor cortex results in motor-to-sensory feedback (see previous discussion of forward models), effectively priming perceptual phonemic categories. Indeed, the error pattern observed in the data confirms a perceptual bias in favor of speech sounds concordant with the stimulation site (D'Ausilio et al., 2009, p.383). In addition, the task used in this experiment involved a difficult phoneme identification task (stimuli were degraded in noise to hold baseline performance at 75%), which is known engage a set of processes not necessarily involved in natural speech comprehension (see discussion below; Blumstein, 1995; Miceli, Gainotti, Caltagirone, & Masullo, 1980). Thus, there appears to be a role for the motor system in facilitating the perception of individual speech sounds under degraded perceptual conditions, but we have yet to see convincing evidence that the motor system is the seat of speech processing generally.

Overall, we see clear evidence that the motor system can *influence* speech perception, perhaps in a top-down fashion, but there is nothing to indicate that speech perception is a motor process by nature. As such, now is a good time to review evidence that runs strictly counter to the motor theory of speech perception.

Why Motor Theory Cannot Explain Speech Perception

As mentioned above, we can only make statements about the necessity of motor processing in speech comprehension if disabling motor regions results in some deficit in speech comprehension. This sort of evidence would be essential in confirming the major tenets of the motor theory of speech perception, and, by the same token, evidence to the contrary would cripple it.   Thus we turn to the lesion literature, the major source of evidence concerning impairment to productive regions in speech processing – evidence that is almost unanimously in opposition to the motor theory.

Damage to speech production areas is often characterized by large frontal lesions typical of those seen in Broca's aphasia, which involve most of the lateral frontal lobe, motor cortex, and anterior insula, but often also extend posteriorly to include the parietal lobe (A. R. Damasio, 1992; H. Damasio, 1991; Dronkers, Redfern, & Knight, 2000). If the motor theory of speech perception holds, such lesions should result in severe disruption of speech comprehension abilities. However, this prediction is not borne out, as little if any comprehension deficits are seen at the single word level in Broca's aphasia (H Goodglass, 1993; H. Goodglass, Kaplan, & Barresi, 2001). For example, a recent study reported that Broca's aphasics (n=9) were indistinguishable from control subjects on an auditory word comprehension test involving 236 items (Moineau, Dronkers, & Bates, 2005). It is true that Broca's aphasics can be impaired on syllable discrimination tasks, i.e., the ability to judge whether pairs of non-sense syllables are the same (/ba/ - /ba/) or different (/ba/ - /da/) (Blumstein, 1995) but these tasks double-dissociate from more ecologically valid auditory comprehension task, even when contextual information that might cue word meaning are removed (Miceli, Gainotti, Caltagirone, & Masullo, 1980; see Hickok & Poeppel, 2000,2004, 2007 for extensive discussion of this issue). It has been suggested that deficits on syllable discrimination tasks result from damage to frontal lobe-dependent working memory and/or executive systems rather than to systems supporting speech recognition. The important observation, though, is that even extensive disruption to motor speech systems does not result in commensurate disruption of speech recognition abilities in contrast to the prediction of the motor theory.

In addition to the fact that lesions to motor speech areas do not disrupt speech comprehension, we find that speech comprehension can be impaired while motor regions are intact. This is evidenced by patients who present with mixed transcortical aphasia (Bogousslavsky, Regli, & Assal, 1988; Geschwind, Quadfasel, & Segarra, 1968). This syndrome is characterized by a severe deficit in comprehension of speech and is associated with damage to left frontal and posterior parietal regions, sparing perisylvian speech-related areas such as Broca's area, superior temporal gyrus, and the tissue in between. Sensory-motor functions of speech are left intact, as evidenced by the fact the patients are able to repeat heard speech. Thus, the motor system appears to have no direct involvement in speech comprehension, as damage to the motor system does not impair speech comprehension abilities, and damage that results in speech comprehension deficits spares the motor system.

An Alternative Perspective – A Sensory Theory of Speech Production

We have repeatedly noted a tight coupling of sensory and motor activity in the evidence reviewed thus far. According to proponents of the motor theory, such a coupling is indicative of motor involvement in speech perception. While we will not argue the fact that the motor system *can be* involved in speech perception, we propose that speech perception is fundamentally a sensory process, and that sensory representations of speech in the superior temporal lobe are projected to frontal motor networks in order to guide speech production. This is the reverse relation of the one proposed in motor theories of speech perception, and as such accounts equally well for close sensory-motor links in speech (Guenther, Hampson, & Johnson, 1998; Guenther, Ghosh, Tourville, 2006; Hickok & Poeppel, 2000, 2004, 2007).

The context of this proposal is the dual stream model of speech processing set forth by Hickok and Poeppel (2000, 2004, 2007), which holds that there are two speech processing pathways, a ventral pathway that maps acoustic speech information onto conceptual semantic representations for speech comprehension, and a dorsal pathway that maps acoustic speech information onto motor speech representations. Independent evidence from developmental considerations, articulatory decline in late-onset deafness, delayed or altered speech feedback, and other sources (reviewed in Hickok & Poeppel 2000, 2004, 2007) suggests that auditory speech information is critically involved in speech production behaviors. It is proposed that the dorsal, sensory-motor stream is the pathway responsible for this interaction. On this view, sensory representations are used to guide motor-articulatory processes, and this can be accomplished via feed-forward or feedback mechanisms (cf., Guenther, 2006). Thus, the dual stream model accommodates bidirectional interaction between sensory and motor systems, and has no trouble accounting for motor involvement in speech perception (e.g., via forward models and efferent copies of motor commands).

Lesion studies (Damasio, 1991; 1992) further support the role of sensory areas in speech production, evidenced by the fact that damage to auditory-related areas in the left temporal lobe result in speech production deficits. Conduction aphasia is particularly interesting in this respect because the deficit appears to be specific to phonological-level aspects of speech production (Wilshire & McCarthy, 1996), and has been interpreted as a breakdown of the speech-related sensory-motor integration system (Hickok et al. 2003, Hickok & Poeppel, 2004). Conduction aphasia is associated with lesions to left superior temporal gyrus and the temporoparietal junction (Damasio & Damasio, 1980; Goodglass, 1992; Baldo, Klostermann, & Dronkers, 2008), that is, classic auditory-related cortical areas (see Figure 1.2).

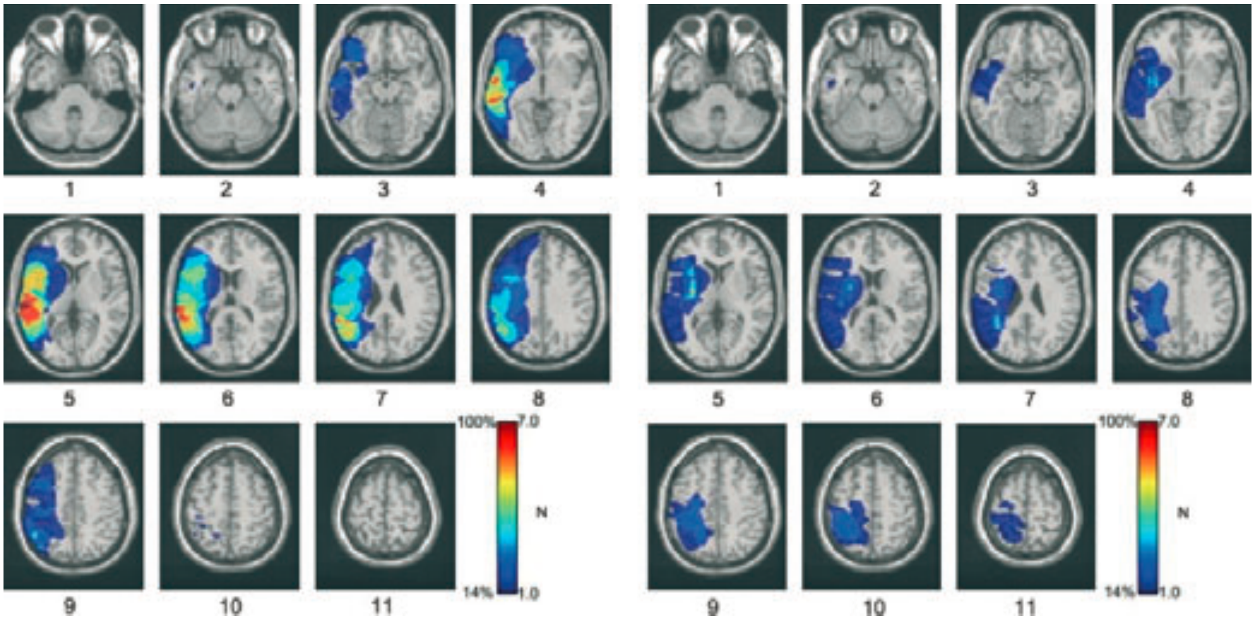


Figure 1.2. Lesion overlap of patients with conduction aphasia (left group) versus left-hemisphere damaged controls (right group). Note the location of overlap in the aphasic patients. Damage to these classic auditory-related cortical areas results in disruption of speech production. From Baldo et al. (2008).

In addition, functional imaging studies have revealed a network of regions thought to support these sensory-motor interactions. This network of cortical regions show sensory-motor response properties, and includes area Spt at the temporal-parietal boundary, which is squarely within the lesion distribution of conduction aphasia. Area Spt appears to be connected both to sensory speech areas in the superior temporal sulcus (STS), and motor speech areas in the left inferior frontal gyrus and left premotor cortex (Buchsbaum, Hickok & Humphries, 2001; Hickok, Buchsbaum, Humphries & Muftuler, 2003). Hickok and Poeppel (2000; 2004; 2007) argue that STS (bilaterally) is responsible for sensory coding of speech, while area Spt is responsible for sensory-motor integration (its activation is tightly coupled to activation in inferior frontal gyrus). By this logic, activity in Broca's region or motor cortex associated with language comprehension can be explained in terms of sensory-motor associations projecting from STS to Spt and finally to LIFG (Figure 1.3).

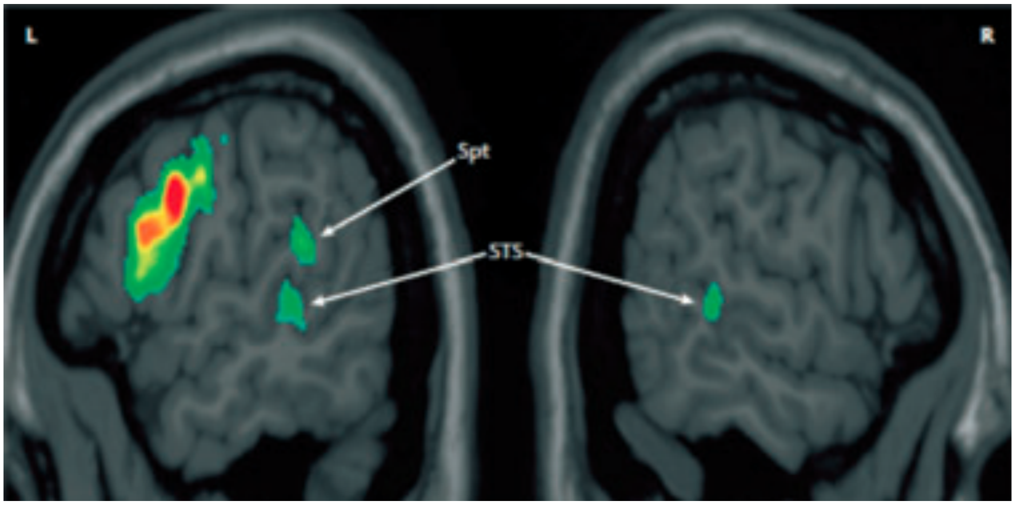


Figure 1.3. Colored regions demonstrated sensorimotor responses: they were active during both the perception of speech and the sub-vocal production of speech. Note the bilateral activation in STS and left-lateralized activation in Spt and frontal regions. According to the dual stream model, bilateral sensory/phonological representations in STS project through Spt to frontal production networks. From Hickok and Buchsbaum (2003).

Given that the dual steam model is consistent with both lesion and functional imaging evidence, we propose that it gives a much stronger account of the current empirical literature. However, motor theories are attractive and often easy to accept at first glance, and the intermediate temporal resolution of functional imaging studies often leads to results that are ambiguous with respect to directionality of interactions between sensory and motor brain regions. An example will provide clarification in this case. A recent study (Wilson & Iacoboni, 2006) had subjects rate the producibility of non-native phonemes. In a subsequent fMRI scan, the same subjects were presented native phonemes along with the non-native phonemes from the earlier behavioral session. Brain regions were identified with activations that a) were more active during perception of phonemes than at rest, b) discriminated native from non-native phonemes, and c) were correlated with the producibility of nonnative phonemes. Bilateral superior temporal, premotor, and primary motor cortices were active to the perception of phonemes compared to rest. In addition, both motor areas and superior temporal areas discriminated native from non-native phonemes (though motor areas demonstrated this distinction only in an ROI analysis, while temporal regions did so in the group and ROI analyses). Interestingly, a functional connectivity analysis demonstrated equivalent functional connectivity between premotor and superior temporal areas regardless of which area was used as the 'seed' area. At first glance, this appears to be strong evidence for motor involvement in speech perception. However, further examination of the results demonstrates that the dual stream model provides a more accurate account of the data. Namely, only superior temporal areas displayed activation that correlated with difficulty of producibility of the non-native phonemes, in addition to one other focus of such activation – area Spt (Figure 1.4). According to the authors, the "…findings suggest that superior temporal auditory areas bilaterally are crucial for the transformation of acoustic speech input to a phonetic code, since only in these areas, and not in motor areas, did signal change correlate with producibility." (p.322). Further, the observed activation in Spt is a clear indication of its function in integration between phonetic (and sensory) speech representations in superior temporal regions and motor representations in productive networks. It is likely that Spt plays some role in the *transformation* of phonological code to motor code (Goodglass, 1992), which is why activation in Spt varied with *producibility*, while activation in frontal production regions did not. These observations are exactly what the dual stream model would predict. Again, the authors focused on the role of motor activation in producing forward models of phonemic categorization that are, at some point, compared with *sensory representations* in superior temporal cortex. Indeed, it will be important for future research to resolve the role of the motor system in speech comprehension, but for now we can remain confident that the representation of speech in cortex is fundamentally a sensory process.

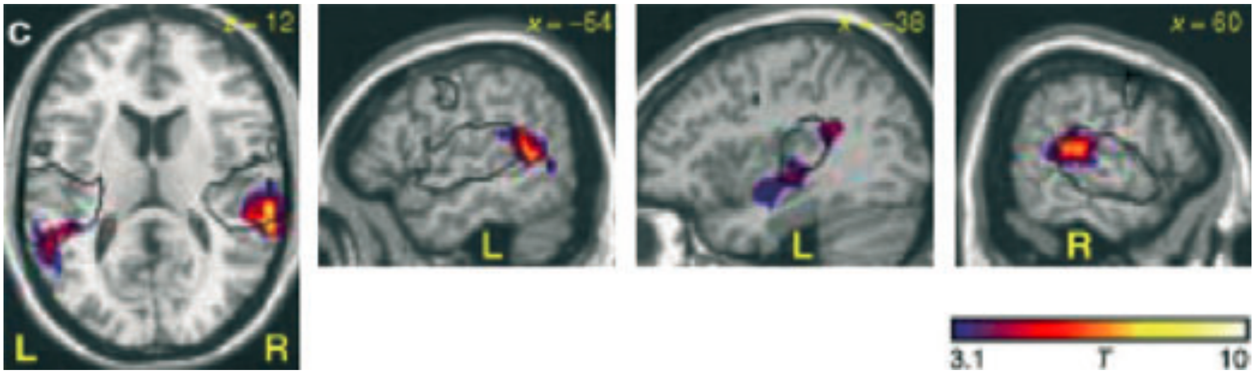


Figure 1.4. Colored regions were negatively correlated with producibility of non-native phonemes. Note the bilateral activation of superior temporal regions and unilateral activation of area Spt, and compare with Figure 1.3 above. This is an identical sensory-motor integration network (note that activation was observed in frontal regions during listening relative to rest, completing the network). We would not expect to see activation in frontal regions in the correlation analysis, where negative correlation with producibility indicates a role in transformation from sensory to phonological and motor codes. From Wilson and Iacoboni (2006).

Conclusion

There is a bevy of evidence that supports a link between motor processing and language perception. This connection is often interpreted as evidence for a motor theory of speech perception. However, there is strong evidence against this view and there is an alternative explanation for the association between sensory and motor speech systems, namely the reverse relation, a sensory theory of speech production, that more accurately describes the role of the motor system in speech processing. The discovery of mirror neurons in the macaque, as well as the human studies they have inspired, do nothing to change this conclusion.

References

Baldo, J.V., Klostermann, E.C., Dronkers, N.F. 2008. It's either a cook or a baker: Patients with conduction aphasia get the gist but lose the trace. *Brain and Language* 105.134-140.

Bell-Berti, F., Raphael, L.J., Pisoni, D.B., & Sawusch, J.R. 1979. Some relationships between speech production and perception. *Phoenetica* 36.373-383.

Blumstein, S. 1995. The neurobiology of the sound structure of language. In M. S. Gazzaniga (Ed.), *The cognitive neurosciences* (pp. 913-929). Cambridge, MA: MIT Press.

Bogousslavsky, J., Regli, F., & Assal, G. 1988. Acute transcortical mixed aphasia. A carotid occlusion syndrome with pial and watershed infarcts. *Brain* 111.631-641.

Boulenger, V., Roy, P., et al. 2006. Cross-talk between Language Processes and Overt Motor Behavior in the First 200 msec of Processing. *Journal of Cognitive Neuroscience,* 18.1607-1615.

Browman, C.P., and Goldstein, L. 1986. Towards an articulatory phonology. *Phonology Yearbook* 3.219-252.

Buccino, G., Riggio, M., et al. 2005. Listening to action-related sentences modulates the activity of the motor system: A combined TMS and behavioral study. *Brain Research,* 24.355-363.

Buchsbaum, B., Hickok, G. & Humphries, C. 2001. Role of left posterior superior temporal gyrus in phonological processing for speech perception and production. *Cognitive Science* 25.663–678.

Cooper, W. 1979. *Speech perception and production: Studies in selective adaptation.* Norwood, NJ: Ablex.

D'Ausilio, A., Pulvermuller, F., Salmas, P., Bufalari, I., Begliomini, C., Fadiga, L. 2009. The motor somatotopy of speech perception. *Current Biology* 19.381-385.

Damasio, A. R. 1992. Aphasia. *New England Journal of Medicine* 326.531-539.

Damasio, H. 1991. Neuroanatomical correlates of the aphasias. In M. Sarno (Ed.), *Acquired aphasia* (2nd ed., pp. 45-71). San Diego: Academic Press.

Damasio, H., & Damasio, A.R. 1980. The anatomical basis of conduction aphasia. *Brain* 103.337-50.

Dronkers, N. F., Redfern, B. B., & Knight, R. T. 2000. The neural architecture of language disorders. In M. S. Gazzaniga (Ed.), *The new cognitive neurosciences* (pp. 949-958). Cambridge, MA: MIT Press.

Eimas, P.D. and Corbit, J.D. 1973. Selective adaptation of linguistic feature detectors. *Cognitive Psychology* 4.99-109.

Fadiga, L, Craighero, L, Buccino, G, & Rizzolatti, G. 2002. Speech listening specifically modulates the excitability of tongue muscles: A TMS study. The *European journal of neuroscience* 15.399-402.

Fadiga, L., Fogassi, L., Pavesi, G., & Rizzolatti, G. 1995. Motor facilitation during action observation: a magnetic stimulation study. *Journal of Neurophysiology* 73.2608-2611.

Fowler, Carol A. 1994. Invariants, specifiers, cues: An investigation of locus equations as information for place of articulation. Perception & Psychophysics, 55(6), 597-610.

Galantucci, B., Fowler, C.A., & Turvey, M.T. 2006. The motor theory of speech perception reviewed. *Psychonomic Bulletin & Review* 13.361-377.

Gallese, V., Fadiga, L., Fogassi, L., & Rizzolatti, G. 1996. Action recognition in the premotor cortex. *Brain* 119.593-609.

Gallese, V., Fogassi, L., Fadiga, L., & Rizzolati, G. 2002. Action representation and the inferior parietal lobule. In W. Prinz & B. Hommel (Eds.), *Attention & Performance XIX. Common Mechanisms in Perception and Action*. Oxford: Oxford University Press.

Geschwind, N., Quadfasel, F. A., & Segarra, J. M. 1968. Isolation of the speech area. *Neuropsychologia* 6.327-340.

Glenberg, A.M., Sato, M.M., Cattaneo, L., et al. 2008. Processing abstract language modulates motor system activity. Quarterly Journal of Experimental Psychology, 61.905-919.

Goodglass, H. 1992. In *Conduction Aphasia (ed. Kohn, S. E.) 39–49*. Lawrence Erlbaum Associates: Hillsdale, New Jersey.

Goodglass, H. 1993. *Understanding aphasia*. San Diego: Academic Press.

Goodglass, H., Kaplan, E., & Barresi, B. 2001. *The assessment of aphasia and related disorders, 3rd ed*. Philadelphia: Lippincott Williams & Wilkins.

Grafton, S.T., Arbib, M.A., Fadiga, L., & Rizzolatti, G. 1996. Localization of grasp representations in humans by positron emission tomography. 2. Observation compared with imagination. *Experimental Brain Research* 112.103-111.

Guenther, F.H. 2006. Cortical interactions underlying production of speech sounds. *Journal of Communication Disorders* 39.350-365.

Guenther, F.H., Ghosh, S.S., Tourville, J.A. 2006. Neural modeling and imaging of the cortical interactions underlying syllable production. *Brain and Language* 96.280-301.

Guenther, F.H., Hampson, M., & Johnson, D. 1998. Atheoretical investigation of reference frames for the planning of speech movements. *Psychological Review* 105.611-633.

Hasson, U., Nir, Y., Levy, I., Fuhrmann, G., Malach, R. 2004. Intersubject synchronization of cortical activity during natural vision. *Science* 303.1634-1640.

Hickok, G. 2009. Eight problems for the mirror neuron theory of action understanding in monkeys and humans. *Journal of Cognitive Neuroscience* 21.1229-1243.

Hickok, G. In press. Mirror neurons, speech perception, and action word semantics: Is there any connection? *Brain and Language*

Hickok, G., Buchsbaum, B., Humphries, C. & Muftuler, T. 2003. Auditory–motor interaction revealed by fMRI: Speech, music, and working memory in area Spt. *Journal of Cognitive Neuroscience* 15.673–682.

Hickok, G. and Poeppel, D. 2004. Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language. *Cognition* 92.67-99.

Hickok, G. and Poeppel, D. 2000. Towards a functional neuroanatomy of speech perception. *Trends in Cognitive Sciences* 4.131-138.

Hickok, G., & Poeppel, D. 2007. The cortical organization of speech processing. *Nature Reviews Neuroscience* 8.393-402.

Iacoboni, 2008. The role of premotor cortex in speech perception: Evidence from fMRI and rTMS. *Journal of Physiology – Paris* 102.31-34.

Iacoboni, M., Woods, R. P., Brass, M., Bekkering, H., Mazziotta, J. C., & Rizzolatti, G. 1999. Cortical mechanisms of human imitation. *Science* 286.2526-2528.

Kohler, E., Keysers, C., Umilta, M.A., Fogassi, L., Gallese, V., Rizzolatti, G. 2002. Hearing sounds, understanding actions: Action representation in mirror neurons. *Science* 297.846-848.

Liberman, A.M. 1957. Some results of research on speech perception. *Journal of the Acoustical Society of America* 29.117-123.

Liberman, A.M., Cooper, F.S., Shankweiler, D.P., & Studdert-Kennedy, M. 1967. Perception of speech code. *Psychological Review* 74.431-461.

Liberman, A.M., Delattre, P., & Cooper, F.S. 1952. The role of selected stimulus-variables in the perception of the unvoiced stop consonants. *Americal Journal of Psychology* 65.497-516.

Liberman, A.M., Delattre, P., Cooper, F.S., & Gerstman, L. 1954. The role of consonant-vowel transitions in the perception of the stop and nasal consonants. *Psychological Monographs: General & Applied* 68.1-13.

Liberman, A.M. and Mattingly, I.G. 1985. The motor theory of speech perception revised. *Cognition* 21.1-36.

Lotto, A.J., Hickok, G. & Holt, L.L. 2009. Reflections on Mirror Neurons and Speech Perception, *Trends in Cognitive Science* 13.110-114.

Massaro, D.W. 1987. *Speech perception by ear and by eye: A paradigm for psychological inquiry.* Hillsdale, NJ: Erlbaum.

Massaro, D.W. 1998. *Perceiving talking faces: From speech perception to a behavioral principle*. Cambridge, MA: MIT press.

Massaro, D.W. & Chen, T.H., 2008. The motor theory of speech revisited. *Psychonomic Bulletin &*

Review 15.453-457.

McGurk, H. & MacDonald, J. 1976. Hearing lips and seeing voices. *Nature* 264.746-748.

Meister, I.G., Wilson, S.M., Deblieck, C., Wu, A.D., Iacoboni, M. 2007. The essential role of premotor cortex in speech perception. *Current Biology* 17.1692-1696.

Miceli, G., Gainotti, G., Caltagirone, C., & Masullo, C. 1980. Some aspects of phonological impairment in aphasia. *Brain and Language* 11.159-169.

Moineau, S., Dronkers, N. F., & Bates, E. 2005. Exploring the processing continuum of single-word comprehension in aphasia. *Journal of Speech, Language and Hearing Research* 48.884-896.

Nelissen, K., Luppino, G., Vanduffel, W., Rizzolatti, G., & Orban, G. A. 2005. Observing others: multiple action representation in the frontal lobe. *Science* 310.332-336.

Nishitani, N., Schurmann, M., Amunts, K., & Hari, R. 2005. Broca's region: from action to language. *Physiology* 20.60-69.

Okada, K., & Hickok, G. 2009. Two cortical mechanisms support the integration of visual and auditory speech. A hypothesis and preliminary data. *Neuroscience Letters* 452.219-223.

Pazzaglia, M., Smania, N., Corato, E., & Aglioti, S. M. 2008. Neural underpinnings of gesture discrimination in patients with limb apraxia. *Journal of Neuroscience* 28.3030-3041.

Di Pellegrino, G., Fadiga, L., Fogassi, L., Gallese, V., & Rizzolatti, G. 1992. Understanding motor events: A neurophysiological study. *Experimental Brain Research* 91.176-180.

Poeppel D., Idsardi W.J., van Wassenhove V. 2008. Speech perception at the interface of neurobiology and linguistics. *Philosophical Transactions of the Royal Society of Lond B Biological Sciences* 363.1071-1086.

Pulvermüller, F., Hauk, O., Nikulin, V., & Ilmoniemi. 2005. Functional links between motor and language systems. *The European Journal of Neuroscience* 21.793-797.

Pulvermuller, F., Huss, M., Kherif, F., Moscoso del Prado Martin, F., Hauk, O., & Shtyrov, Y. 2006. Motor cortex maps articulatory features of sounds. In Proceedings of the National Academies of Science USA, 103.7865-7870.

Rizzolatti, G. and Arbib, M.A. 1998. Language within our grasp. *Trends in Neurosciences* 21.188-194.

Rizzolatti, G., Camarda, R., Fogassi, L., Gentilucci, M., Luppino, G., & Matelli, M. 1988. Functional organization of inferior area 6 in the macaque monkey. II. Area F5 and the control of distal movements. *Experimental Brain Research* 71.491-507.

Rizzolatti, G., and Craighero, L. 2004. The mirror-neuron system. *Annual Reviews in Neuroscience* 27.169-192.

Rizzolatti, G., Fadiga, L., Matelli, M., Bettinardi, V., Paulesu, E., Perani, D., & Fazio, F. 1996. Localization of grasp representations in humans by PET: 1. Observation versus execution. *Experimental Brain Research* 111.246-252.

Sams M, Mottonen R, Sihvonen T. 2005. Seeing and hearing others and oneself talk. *Cognitive Brain Research* 23.429-435.

Tkach, D., Reimer, J., & Hatsopoulos, N. G. 2007. Congruent activity during action and action observation in motor cortex. *Journal of Neuroscience,* 27.13241-13250.

Urgesi, C., Calvo-Merino, B., Haggard, P., Aglioti, S.M. 2007. Transcranial Magnetic Stimulation Reveals Two Cortical Pathways for Visual Body Processing. *The Journal of Neuroscience* 27.8023-8030.

Urgesi, C., Candidi, M., Ionta, S., & Aglioti, S.M. 2007. Representation of body identity and body actions in extrastriate body area and ventral premotor cortex. *Nature Neuroscience* 10.30-31.

Watkins, K. and Paus, T. 2004. Modulation of motor excitability during speech perception: The role of broca's area. *Journal of Fognitive Neuroscience* 16.978-987.

Weinrich, M., Wise, S. P., & Mauritz, K. H. 1984. A neurophysiological study of the premotor cortex in the rhesus monkey. *Brain* 107.385-414.

Wilshire, C. E. & McCarthy, R. A. 1996. Experimental investigations of an impairment in phonological encoding. *Cognitive Neuropsychology* 13.1059–1098.

Wilson, S.M., and Iacoboni, M. 2006. Neural responses to non-native phonemes varying in producibility: Evidence for the sensorimotor nature of speech perception. *NeuroImage* 33.316-325

Wilson, S.M., Molnar-Szakacs, I., Iacoboni, M. 2008. Beyond superior temporal cortex: Intersubject correlations in narrative speech comprehension. *Cerebral Cortex* 18.230-242.

Wilson, S.M., Saygin, A.P., Sereno, M.I., Iacoboni, M. (2004). Listening to speech activates motor areas involved in speech production. *Nature Neuroscience* 7.701-702.

Wise, S. P., & Mauritz, K. H. 1985. Set-related neuronal activity in the premotor cortex of rhesus monkeys: effects of changes in motor set. *Proceedings of the Royal Society of London B Biological Sciences* 223.331-354.

CHAPTER 2

Primer

The previous chapter established (or reiterated, at least) that a functioning speech motor system is not necessary for normal speech perception. However, it also described an emerging body of evidence demonstrating that speech-motor brain regions are active during passive perception of speech sounds. This pattern was easily explained once a fundamental link between perception and production was accounted for – specifically, sensory speech representations function to guide speech production (for an expansion on this see (Hickok, 2012)). Chapter 1 also reviewed evidence suggesting that the speech motor system may play a top-down role in perception, especially when the signal is degraded or during difficult (and unnatural) laboratory tasks. In assessing top-down contributions to speech perception, it is important to distinguish the contributions of domain-general processes (attention, working memory, decision-making, etc.) from speech-specific mechanisms (e.g., motor predictions as described in Ch. 1). The current chapter contains an original investigation designed to: (a) manipulate domain-general top-down processes (decision-making) in the context of a typical laboratory speech task, and (b) assess whether (and in which) speech-related brain regions activity varies in accordance with the manipulation. Not surprisingly, activity in speech motor brain regions (but not perceptual regions) was modulated by shifts in the decision criterion (see below for details). While this result does not rule out motor contributions to perception via a domain-specific top-down mechanism (e.g., motor prediction), it does provide positive evidence that speech motor brain regions are recruited by domain-general processes. This calls into question many laboratory findings that implicate the speech motor system in perception, and reinforces the conclusion that speech perception *is not* a motoric process.

Response bias modulates the speech motor system during syllable discrimination

Jonathan H. Venezia, Kourosh Saberi, Charles Chubb & Gregory Hickok

Introduction

Recent research and theoretical discussion concerning speech perception have focused considerably on the language production system and its role in perceiving speech sounds. More specifically, it has been suggested – in varying forms and with claims of variable strength – that the cortical speech motor system supports speech perception directly (Galantucci, Fowler & Turvey, 2006; Hasson, Skipper, Nusbaum & Small, 2007; Pulvermuller & Fadiga, 2010; Rizzolatti & Craighero, 2004). Classic support for this position comes from studies involving patients with large frontal brain lesions (i.e., Broca's aphasics), which demonstrate that these patients are impaired on syllable discrimination tasks (Blumstein, 1995; Miceli, Gainotti, Caltagirone & Masullo, 1980), including worse performance when discriminating place of articulation versus voicing (Baker, Blumstein & Goodglass, 1981), and perhaps mildly impaired (and significantly slowed) in auditory word comprehension (Moineau, Dronkers & Bates, 2005). Indeed, more recent evidence demonstrates unequivocally that the cortical motor system is active during speech perception (Fadiga, Craighero, Buccino & Rizzolatti, 2002; Hickok, Buchsbaum, Humphries & Muftuler, 2003; Pulvermuller et al., 2006; Skipper, Nusbaum & Small, 2005;

Watkins, Strafella & Paus, 2003; Wilson, Saygin, Sereno & Iacoboni, 2004). Perhaps the strongest evidence for motor system involvement comes from recent transcranial magnetic stimulation (TMS) studies. For example, one study demonstrated that repetitive TMS applied to a speech region of premotor cortex impaired syllable identification but not color discrimination (Meister et al. 2007), while another found that TMS of primary motor areas for different vocal tract articulators selectively facilitated identification of phonemes relying on those articulators (D'Ausilio et al., 2009). Additional TMS studies demonstrated that disruptive TMS applied to motor/premotor cortex significantly altered performance in discrimination of synthesized syllables (Mottonen & Watkins, 2009) and phoneme discrimination (Sato, Tremblay & Gracco, 2009).

Despite the evidence listed above, neuropsychological data seem to dispel the notion that the speech motor system is critically involved in speech perception (Hickok, 2009; Venezia & Hickok, 2009). In short, Broca's aphasics generally have preserved word-level comprehension (Damasio, 1992; Goodglass, 1993; Goodglass, Kaplan & Barresi, 2001; Hillis, 2007), as do patients with bilateral lesions to motor speech regions (Levine & Mohr, 1979; Weller, 1993). Further, two recent studies of patients with radiologically confirmed lesions to motor speech areas including Broca's region and surrounds, failed to replicate earlier findings that Broca's aphasics have substantial speech discrimination deficits (Hickok, Costanzo, Capasso & Miceli, 2011; Rogalsky, Love, Driscoll, Anderson & Hickok, 2011). Additionally, children that fail to develop motor speech ability (as a result of congenital or acquired anarthria) are able to develop normal receptive speech (Bishop, Brown & Robson, 1990; Christen et al., 2000; Lenneberg, 1962). Lastly, anesthesia of the entire left hemisphere, producing complete speech arrest (mutism), leaves speech sound perception proportionately intact (Hickok et al., 2008).

Nonetheless, it remains to explain why acute disruption and/or facilitation of speech motor cortex significantly performance on speech perception tasks. First, the aforementioned TMS studies either utilized degraded or unusual (synthesized) speech stimuli (D'Ausilio et al., 2009; Meister et al., 2007; Mottonen & Watkins, 2009), or failed to produce an effect unless the phonological processing load was unusually high (Sato et al., 2009). Several studies indicate that speech motor areas of the inferior frontal cortex are more active with increasing degradation of the speech signal (Binder, Liebenthal, Possing, Medler & Ward, 2004; Davis & Johnsrude, 2003; Zekveld, Heslenfeld, Festen & Schoonhoven, 2006). Additionally, Broca's aphasics showed poor auditory comprehension when stimuli were low-pass filtered and temporally compressed (Moineau et al., 2005). Indeed, syllable identification (as in Meister et al., 2007 and D'Ausilio et al., 2009) was not impaired nor were reaction times facilitated in TMS studies using clear speech stimuli (D'Ausilio, Bufalari, Salmas & Fadiga, 2011; Sato et al., 2009).

Second, the effects of applying TMS to speech motor cortex are often small (Meister et al., 2007) and/or confined to reaction time measures (D'Ausilio et al., 2009; Sato et al., 2009). A recent functional magnetic resonance imaging (fMRI) study utilizing a two-alternative forced choice syllable identification task at varying signal-to-noise ratios demonstrated that hemodynamic activity correlated with identification performance (percent correct) in superior temporal cortex and decision load (reaction time) in inferior frontal cortex (Binder et al., 2004). Additionally, Broca's aphasics exhibit increased reaction times on an auditory word comprehension task relative to older controls and patients with right hemisphere damage (Moineau et al., 2005). Together these findings suggest that speech motor brain regions may be preferentially involved in decision-level components of speech perception tasks.

An important component of decision-level processes in standard syllable- or single-word-level speech perception assessments is response bias – i.e., when changes in a participant's decision criterion lead to a more liberal or conservative strategy that biases the participant toward a particular response (see discussion of signal detection theory below). Response bias is not properly accounted for in standard measures of performance on speech perception tasks (percent correct, reaction time, error rates) such as those reported in the studies above that appear to implicate speech motor cortex in speech perception ability (D'Ausilio et al., 2009; Meister et al., 2007). For example, a recent study in which disruptive TMS was applied to the lip region of motor cortex reported that cross-category discrimination of synthesized syllables was impaired for lip-tongue place of articulation continua (/ba/-/da/, /pa/-/ta/) but not for voice onset time or non-lip place of articulation continua (/ga/-/ka/ and /da/-/ga/, respectively; Mottonen & Watkins, 2009). However, the performance measure in this study was simply the change in proportion of "different" responses in the same-different discrimination task after application of TMS. This effect could simply be due to changes in response bias induced by application of TMS to speech motor cortex (there is no reason to believe that an effect on response bias, like an effect on accuracy, should not be articulator-specific). Indeed, another recent study demonstrated that use-induced motor suppression of the tongue resulted in a larger response bias toward the lip-related phoneme in a syllable identification task with lip- and tongue-related phonemes (/pa/ and /ta/, respectively; Sato et al., 2011). The opposite effect held for use-induced suppression of the lips, while suppression had no effect on identification performance (*d'*) in any condition.

Classic lesion data suggesting a speech perception deficit in Broca's aphasics may also be contaminated by response bias. A study by Miceli and colleagues (Miceli, et al., 1980) demonstrated that patients with a phonemic output disorder (POD+; fluent and nonfluent aphasics) were impaired on a same-different syllable discrimination task versus patients without disordered phonemic output (POD-). However, both groups made more false identities than false differences, and POD+ patients were more likely to make a false identity than POD- patients (see Miceli et al., 1980, their table 1), indicating the presence of a response bias toward "same." Similarly, a single word, minimal pair discrimination study conducted by Baker and colleagues (Baker et al., 1981) showed that Broca's aphasics were more impaired at discriminating place of articulation than voicing, but this effect was driven by a higher error rate for "different" trials. In other words, Broca's aphasics were again more likely to make a false identity. An informal analysis of the data (inferred from the error rates in same and different trials relative to the overall number of trials) indicates that overall performance on the discrimination task was quite good in Broca's aphasics when response bias is accounted for (*d'* = 3.78; Hickok et al., 2011).

In light of this information, we set out to determine whether changes in response bias modulate functional activity in speech motor cortex. Thus, the present functional magnetic resonance imaging experiment was designed to produce specific, measureable changes in response bias in a speech perception task using degraded speech stimuli. Minimal consonant-vowel stimulus pairs were presented between volume acquisitions for same-different discrimination. Speech stimuli were embedded in Gaussian noise at the threshold signal-to-noise ratio (SNR) as determined via 2-down, 1-up staircase. We manipulated bias by changing the ratio of same-to-different trials: 1:3, 1:2, 1:1, 2:1, 3:1. Ratios were blocked by run and subjects were cued to the upcoming ratio at the beginning of each run. In order to measure response bias, we modeled the data using a modified version of signal detection theory (SDT). Briefly, SDT attempts to disentangle a participant's decision criterion from true perceptual sensitivity (which should remain constant under unchanging stimulus conditions, regardless of shifts in criterion).

In the classic case of a "yes-no" detection experiment, the participant is tasked with identifying a signal in the presence of noise (e.g., a tone in noise, a brief flash of light, or a tumor on an x-ray). Two conditional, Gaussian probability distributions – one for noise trials and another for signal+noise trials – are used to model the likelihood of observing a particular level of internal (sensory) response on a given trial. The normalized distance between the means of the two distributions, known as *d'*, is taken to be the measure of perceptual sensitivity (i.e., ability to detect the signal), where this distance is an intrinsic (fixed) property of the sensory system. However, the participant must set a response criterion – a certain position on the internal response continuum – for which trials that exceed the criterion response are classified as "signal." The position of the criterion is referred to as *c*, and can change in response to a number of factors, both internal and/or external to the observer. The values for *d'* and *c* can be estimated from the proportion of response types. We have extended this analysis to our same-different design. In brief (see Materials and Methods below for an extended discussion), we have modeled the decision space as six separate conditional Gaussian distributions that represent each of the six possible stimulus pairs presented on a same-different trial. The internal response continuum is a single perceptual statistic (standard normal units) that represents the stimulus pair, where negative values are more likely to be a "same" pair (e.g., ba-ba) and positive values are more likely to be a "different" pair (e.g., ba-da). The listener sets a single criterion value on the internal response continuum, where trials that produce a response above the criterion yield a "different" response, while responses below the criterion yield a "same" response.

Based on the properties of our design – in particular, the maintenance of a constant SNR and otherwise identical stimulus conditions across runs – we assumed that the distances between the means (analogous to *d'*) of the six stimulus distributions were fixed across bias ratio conditions. The criterion value, here called *C*, was allowed to vary across conditions. To be explicit, *d'* would not be expected to change because the sensory properties of the stimuli remained constant across conditions, while *C* would be expected to change because the same-different ratio was manipulated directly in each condition. We expected changes in response bias to correlate with changes in the blood-oxygen level dependent (BOLD) signal in motor (i.e., frontal) brain regions, but not sensory (i.e., temporal) brain regions. This is precisely what we observed – response bias, *C*, varied significantly across conditions and a group-level regression of overall bias on percent signal change revealed a network of motor brain regions correlated with response bias. To the best of our knowledge, this is the first study to demonstrate a direct relationship between response bias and functional brain activity in a speech perception task. The significance of this relationship is discussed below along with details of the particular network of brain areas that correlated with response bias.

Materials and Methods

Participants

Eighteen (9 female) right-handed, native-English speakers between 18 and 32 years of age participated in the study. All volunteers had normal or corrected-to-normal vision, no known history of neurological disease, and no other contraindications for MRI. Informed consent was obtained from each participant in accordance with UCI Institutional Review Board guidelines.

Stimuli and Procedure

Participants were presented with same-different discrimination trials involving comparison of 250ms-duration consonant-vowel syllables (/ba/, /da/, or /ga/) embedded in a broadband Gaussian noise masker (independently sampled) of equal duration. Auditory stimuli were recorded in an anechoic chamber (Industrial Acoustics Company, Inc). During recording, a male, native-English speaker produced approximately 20 samples of each syllable using natural timing and intonation, pausing briefly between each sample over the course of a single continuous session. A set of four tokens was chosen for each syllable based on informal evaluation of loudness, clarity and quality of the audio recording (see Figure 2.1) for a representative member from each speech sound category). Syllables were digitally recorded at a sampling rate of 44.1 kHz and normalized to equal root-mean-square amplitude. The average A-weighted level of the syllables was 66.3 dB SPL (sd = 0.5 dB SPL). Since natural recordings were used, several tokens (i.e., from separate recordings) of each syllable were created so that no artifact of the recording process could be used to distinguish between speech sound categories. This also increased the difficulty of the task such that discrimination relied on subjects' ability to distinguish between speech sound categories rather than identify purely acoustic differences in the stimuli (i.e., both within- and between-category tokens differed acoustically). Throughout the experiment, the level of the noise masker was held constant at approximately 62 dB(A). All sounds were presented over MR-compatible, insert-style headphones (Sensimetrics model S14) powered by a 15 watt-per-channel stereo amplifier (Dayton model DTA-1). This style of headphone utilizes a disposable "earbud" insert that serves as both an earplug and sound delivery apparatus, allowing sounds to be presented directly to participants' ear canals. During scanning, a secondary protective ear cover (Pro Ears Ultra 26) was placed over the earbuds for additional attenuation of scanner noise. Stimulus delivery and timing were controlled using Cogent software (http://www.vislab.ucl.ac.uk/cogent 2000.php) implemented in Matlab R12 (Mathworks, Inc, USA) running on a dual-core IBM Thinkpad laptop.

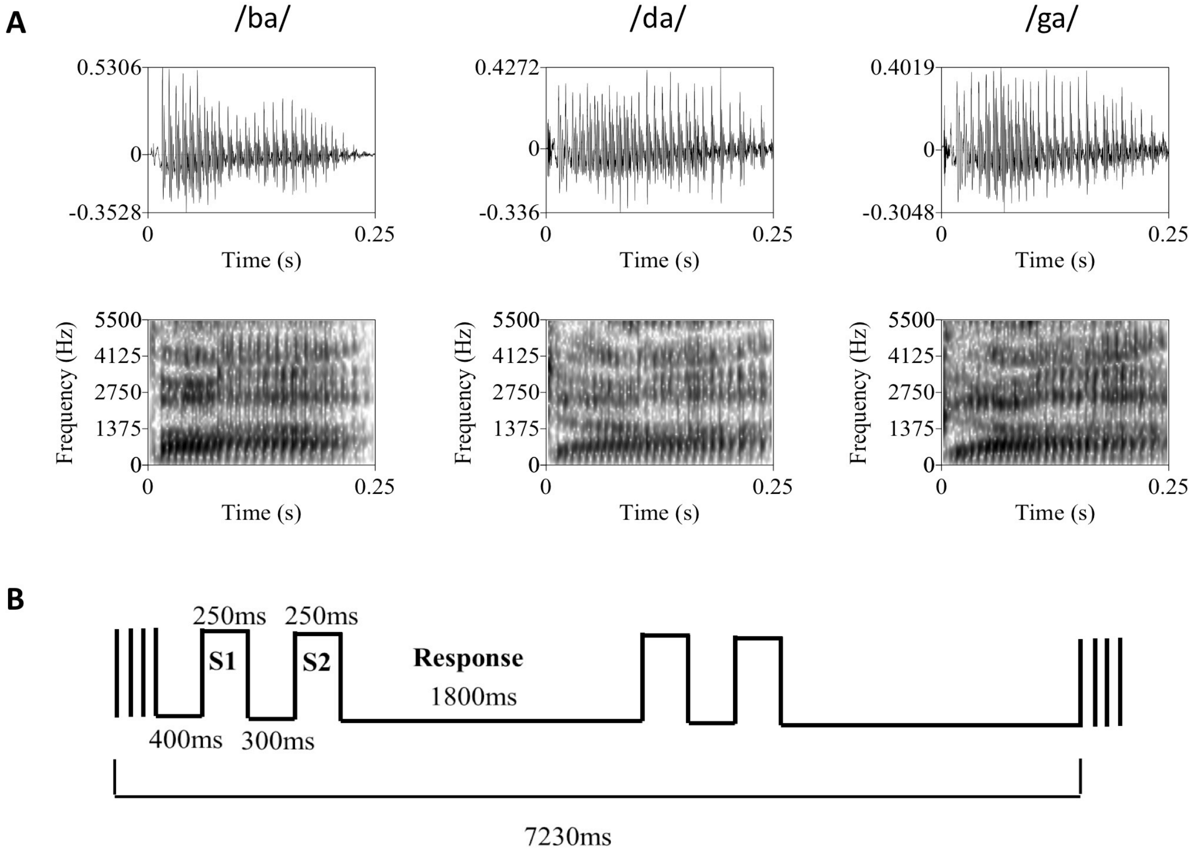


Figure 2.1. (A) Representative tokens from each of the three syllable categories used in the discrimination task. The top row contains the raw waveforms for each token with amplitude (y-axis) plotted against time (x-axis). The bottom row contains the associated spectrograms for each token with frequency in Hz (y-axis) plotted against time (x-axis), where darker sections indicate greater sound energy. During the experiment, auditory syllable stimuli were presented in broadband Gaussian noise at a constant level of 62 dB SPL. Syllable amplitude was held constant at the psychophysically determined threshold level (mean SNR = -13.1 dB). (B) The basic trial and block structure of the fMRI experiment. Each block consisted of two same-different discrimination trials with 250ms auditory syllable presentations separated by a 300ms ISI. Subjects had 1800ms to respond. Trials occurred in the silent period between 1630ms volume acquisitions, with a 400ms silent period at the beginning of each block of two trials.

Trials followed a two-interval same-different discrimination procedure. Two response keys were operated by the index finger of the left hand. Trials consisted of presentation of one of the syllables, followed by a 300ms interstimulus interval, then presentation of a second syllable. Participants pressed key 1 if the two syllables were from the same category (e.g., /ba/-/ba/) and key 2 if the syllables were from different categories (e.g., /ba/-/da/). During the period between responses, participants rested their index finger at a neutral center-point spaced equidistantly from each response key. Participants were instructed to respond as quickly and accurately as possible. Each participant took part in a behavioral practice session in a quiet room outside the scanner. During this practice session, participants were asked to perform 24 practice trials, followed by 72 trials in a "2-down, 1-up" staircase procedure that tracks the participant's 71% threshold (Levitt, 1971). During the staircase procedure, syllable amplitude was varied with 4 dB step size. Participants then performed a second block of 72 trials following the same staircase procedure with a 2 dB step size. Threshold level was determined by eliminating the first four reversals and averaging the amplitude of the remaining reversals (four minimum). Once the threshold level was determined participants performed additional blocks of 72 trials at threshold until behavioral performance stabilized between 65 and 75 percent correct. Many subjects continued to improve over several runs, therefore the experimenter was instructed to make 1-2 dB adjustments to the syllable level in between runs in order to keep performance in the target range. All trials prior to scanning were presented back-to-back with a 500ms intertrial interval and a constant same-different ratio of 1:1. Practice trials were self-paced (i.e., the next trial did not begin until a response was entered).

During the scan session, participants were placed inside the MRI scanner and, following initial survey scans, the scanner was set to "standby" in order to minimize the presence of external noise due to cooling fans and pumps in the scan room. Participants were then required to repeat the staircase procedure described above. It was necessary to set threshold performance inside the scanner because the level of ambient noise in the scan room could not be kept equivalent to our behavioral testing room. The threshold level determined inside the scanner was used throughout the remainder of the experiment (mean = 48.9 dB, sd = 5.8 dB; mean SNR = -13.1 dB). After a short rest following threshold determination, the MRI scanner was set to "start" for fMRI data collection. Volumes were acquired using a traditional sparse scanning sequence with a volume acquisition time of 1630ms and an interscan interval of 5600ms. Blocks of two trials, each with a 1.8s response period, occurred in the silent period between single volume acquisitions (Figure 2.1b). Rest blocks (no task) were included at random at a rate of one in every six blocks. Each scanning run contained a total of 36 task blocks (72 trials) and six rest blocks. Subjects performed a total of 10 runs (720 trials). In order to manipulate response bias, subjects were cued to a particular ratio of same-different trials – 1:3, 1:2, 1:1, 2:1, or 3:1 – at the beginning of each run. Each ratio appeared twice per subject and the order of ratios was randomized across subjects. Crucially, during the practice, adaptive (pre-scan) and fMRI portions of the experiment, each run of 72 trials was designed so that the four tokens for each syllable (/ba/, /da/, /ga/) were presented 12 times. For example, in the 1:1 ratio condition (all pre-scan runs were of this type) there were 36 "same" trials (12 ba-ba, 12 da-da, 12 ga-ga) and 36 "different" trials (6 ba-da, 6 ba-ga, 6 da-ba, 6 da-ga, 6 ga-ba, 6 ga-da). So, there were 48 presentations of /ba/, 48 presentations of /da/, and 48 presentations of /ga/, where each group of 48 was divided evenly between each of the four tokens available for that speech sound. The order of trial type (same or different), speech sound identity of the stimulus pair (e.g., ba-ba, ba-da, da-ga, etc), and token identity were drawn pseudorandomly to fit the run structure. Thus, the actual physical stimuli presented in each run were identical and this was true of each bias ratio condition.

Scanning Parameters

MR images were obtained in a Philips Achieva 3T (Philips Medical Systems, Andover, MA) fitted with an 8-channel SENSE receiver/head coil, at the John Tu and Thomas Yuen Center for Functional Onco-Imaging facility at the University of California, Irvine. We collected a total of 430 echo planar imaging (EPI) volumes over 10 runs using single pulse Gradient Echo EPI (matrix = 76 x 76, repetition time [TR] = 7.23 s, acquisition time [TA] = 1630ms, echo time [TE] = 25 ms, size = 2.875 x 2.875 x 3.5 mm, flip angle = 90). Thirty axial slices provided whole brain coverage. Slices were acquired sequentially with a 0.5mm gap. After the functional scans, a T1-weighted structural image was acquired (140 axial slices; slice thickness = 1 mm; field of view = 240 mm; matrix 240 × 240; repetition time = 11 ms, echo time = 3.55 ms; flip angle = 18°; SENSE factor reduction 1.5 × 1.5)

Data analysis – Behavior

We assume the participant extracts from the stimulus pair presented on a given trial a statistic (a random variable) that reflects the strength of the difference between the two substimuli in the pair. We further assume that the distribution of this difference-strength statistic is invariant with respect to the order of substimuli in a pair. Thus, for example, the difference-strength statistic characterizing a ba-da pair is assumed to be identically distributed to the statistic characterizing a da-ba pair. Under these assumptions, there are six classes of stimuli, S_k, k=1,2,…,6, three in which the two substimuli are drawn from the same category and three in which the two substimuli are drawn from different categories. We assume that the difference-strength statistic the participant extracts from a given presentation of S_k, k=1,2,…,6, is a normally distributed random variable X_k with standard deviation 1 and mean μ_k.. In a bias condition with proportion q of “different” trials, we assume the participant judges stimulus S_k to be “different” just if $X_k > C_q$, where C_q is the criterion adopted by the participant in the given bias condition. Under this model, the probability of a correct response in the bias condition with proportion q of “different” responses given a stimulus S_k is

$$P_{k,q} = \begin{cases} \Phi(C_q - \mu_k) & \text{if } S_k \text{ comprises substimuli from the same category} \\ \Phi(\mu_k - C_q) & \text{if } S_k \text{ comprises substimuli from different categories} \end{cases} \qquad (1)$$

where Φ is the standard normal cumulative distribution function.

This model has 11 parameters: μ_k, k = 1,2,…, 6, and C_q, for q ranging across the five ratios (1:3, 1:2, 1:1, 2:1, or 3:1) of same to different trials. However, the model is underconstrained if all 11 parameters are free to vary as can be seen by considering Eq. (1). Note, in particular, that for any real number α, the probabilities $P_{k,q}$ remain the same if we substitute $C_q+\alpha$ for each of the C_q’s and $\mu_k+\alpha$ for each of the μ_k’s. For current purposes it is convenient to insure that the model parameters are uniquely determined by imposing the additional constraint that the μ_k’s sum to 0. Thus, the model actually has only 10 degrees of freedom.

For any values μ_k, k=1,2,…,6 and C_1, C_2,…, C_5 (with the μ_k's constrained to sum to 0), the log likelihood function is

$$L(\mu_1,...,\mu_6,C_1,...,C_5) = \sum_{k=1}^{6}\sum_{q=1}^{5}\left(H_{k,q}\log\left(P_{k,q}\right) + M_{k,q}\log\left(1-P_{k,q}\right)\right)$$

Where $P_{k,q}$ is given by Eq. (1), $H_{k,q}$ ($M_{q,k}$) is the number of correct (incorrect) responses given to stimulus S_k in the bias condition with proportion q of "different" trials. Data from trials for which no response was recorded were excluded from analysis (mean proportion dropped = 0.01, max = 0.057).

In short, the values μ_k can be thought of as six "perceptual distances" (analogous to *d'*) that characterize the sensory representation of the stimulus pairs, where the model estimates of these distances are assumed to be constant across bias conditions. The model is constrained such that the mean of the values μ_k is set to zero. The values C_Q are the five criterion values – one for each bias condition – and serve as a measure of response bias where negative values indicate a bias to respond "different," positive values indicate a bias to respond "same," and larger values indicate a stronger bias (standard normal units). See Figure 2.2 for a visual representation of the parameter space based on a representative subject's actual data.

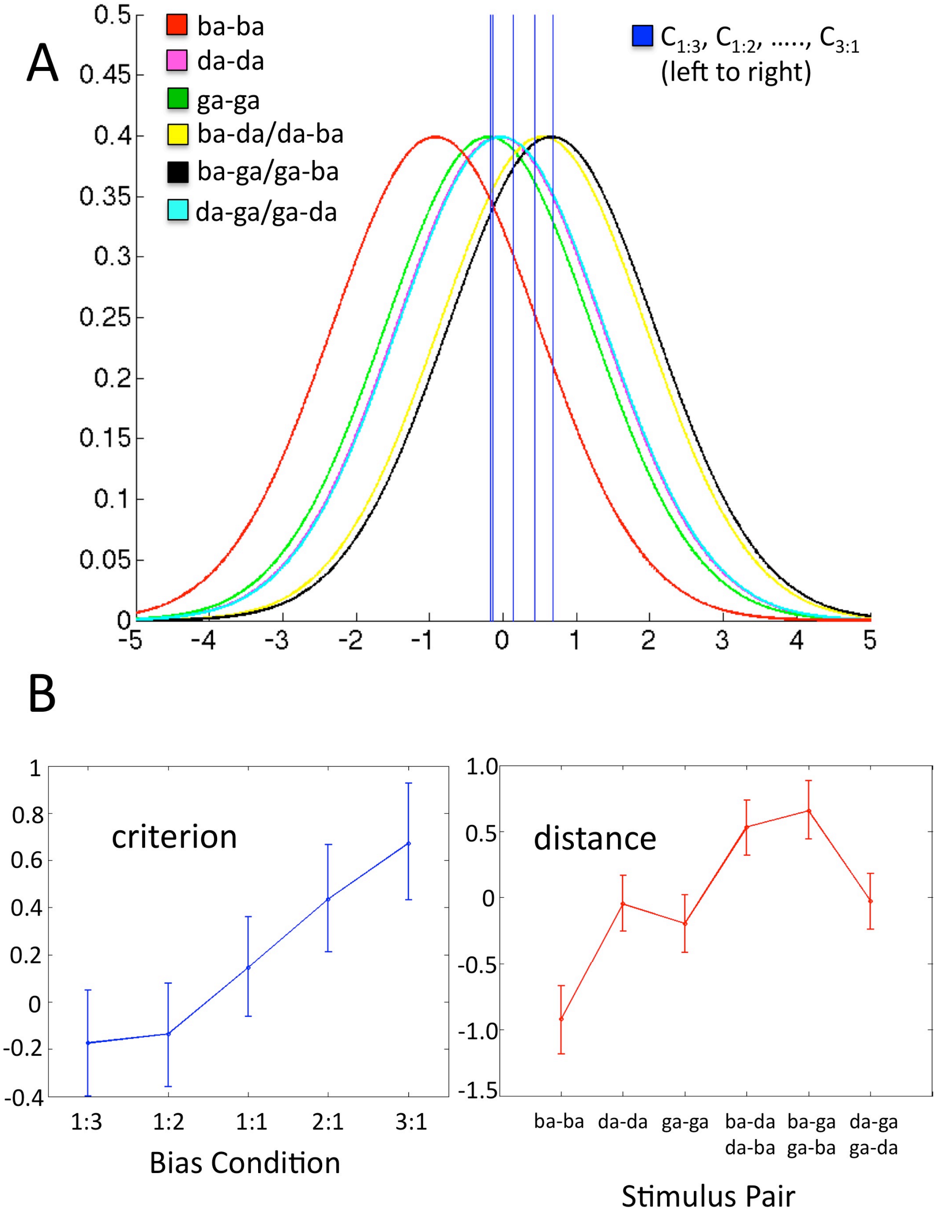


Figure 2.2. Data in each panel are from a representative subject. (A) Schematic of the decision space including six conditional Gaussian distributions (one for each stimulus pair) representing the likelihood of observing a given sensory response, and five criterion values (one for each bias ratio condition), where C0.333, C0.5, …., C3, fall left to right on the graph. The x-axis is the value of a difference-strength statistic representing the perceptual difference between two substimuli in a pair. The y-axis is the probability of observing a given value of the difference-strength statistic. For a given bias ratio condition, values of the difference-strength stastistic that fall above (right of) the criterion line yield a "different" response, while values that fall below (left of) the criterion yield a "same" response. The positions of the six stimulus-pair distributions remain fixed across conditions. Note that the magenta (da-da) and teal (da-ga/ga-da) distributions fall roughly on top of one another, as do the criteria for the 1:3 and 1:2 bias ratio conditions. (B) Model estimates (over all q = proportion of different trials, k = stimulus pair) for Cq (criterion, left) and µk (distance, right) plotted as line graphs with the Bayesian 95% credible intervals plotted as error bars. These values reflect the five criterion lines and the means of the six stimulus-pair distributions plotted in (A) above, respectively.

To estimate the parameter values, we used a Bayesian modeling procedure to fit the data for each participant. This fitting procedure employs a Markov chain Monte Carlo (MCMC) algorithm that yields a sample of size 100,000 from the posterior density characterizing the joint distribution of the model parameters. The prior density for each parameter was taken to be uniform on the interval (-10,10). Model parameters were estimated from an initial run of 100,000 with starting values of μ_k= 0 and C_Q= 0 over all values of k and Q, where the first 20,000 samples were discarded as burn-in and the remaining 80,000 were utilized for posterior estimation. A second run of 100,000 was then executed with starting values of μ_k and C_Q equal to the mean parameter estimates from the initial run. Final parameter estimates and 95% credible intervals were derived from all 100,000 samples of the second run. A given parameter was estimated by the mean sample value for that parameter, and the 95% credible intervals were estimated by taking the 0.025 and 0.975 quantiles of the sample for that parameter. Data from a representative subject are plotted in Figure 2.2b: five criterion values and six distance values are displayed as line graphs with the 95% credible interval as error bars.

Since we were only interested in parameter differences induced by our same-different ratio manipulation, only the vector of estimated criterion values, C=($C_{0.333}$, $C_{0.5}$, C_1, C_2, C_3) were entered into second-level analyses. The mean 95% credible interval across C_q, q = 0.333, 0.5,1,2,3, was calculated for each subject as a means to determine the precision of the model fit. Three subjects with a mean 95% credible interval greater than 1 (and negligible variation in the C_k's) were excluded from further analysis. Individual subject parameter estimates for C were then entered in a multivariate analysis of variance (mANOVA) to test for differences in the group means across bias ratio conditions.

Data analysis – MRI

Study-specific Template Construction and Normalization of Functional Images

Group-level localization of function in fMRI, including identification of task-related changes in activation, can be highly dependent on accurate normalization to a group template. Surface-based (Argall, Saad & Beauchamp, 2005; Desai, Liebenthal, Possing, Waldron & Binder, 2005) and non-linear (Klein et al., 2009; Klein et al., 2010) warping techniques have recently been utilized to improve normalization by accounting for individual anatomical variability. Here, we used a diffeomorphic registration method implemented within the Advanced Normalization Tools software (ANTS; Avants et al., 2008; Avants & Gee, 2004). Symmetric diffeomorphic registration (SyN) uses diffeomorphisms (differentiable and invertible maps with a differentiable inverse) to capture both large deformations and small shape changes. We constructed a study-specific group template using a diffeomorphic shape and intensity averaging technique and a cross-correlation similarity metric (Avants, Anderson, Grossman & Gee, 2007; Avants et al., 2010). The resulting template was then normalized using SyN to the MNI space ICBM template (http://www.bic.mni.mcgill.ca/ServicesAtlases/ICBM152NLin2009; ICBM 2009a Nonlinear Symmetric). A low-resolution (2x2x2mm) version of the study-specific template was constructed for alignment of the functional images. Functional images for each individual subject were first motion-corrected, slice timing corrected and aligned to the individual subject anatomy in native space using AFNI software (http://afni.nimh.nih.gov/afni). Following this step, the series of diffeomorphic and affine transformations mapping each individual subject's anatomy to the MNI-space, study-specific template was applied to the aligned functional images, using the low-resolution template as a reference image. The resulting functional images were resampled to 2x2x2mm voxels and registered to the study-specific group template in MNI space.

fMRI Analysis

Preprocessing of the data was performed using AFNI software. For each run, motion correction, slice timing correction, and coregistration of the EPI images to the high resolution anatomical were performed in a single interpolation step. Normalization of the functional images to the group template was performed as described above. Images were then high pass filtered at .008 Hz and spatially smoothed with an isotropic 6-mm full-width half-maximum (FWHM) Gaussian kernel. Each run was then mean scaled in the temporal domain. The global mean signal was calculated at each time point and entered as a regressor of no interest in the individual subject analysis along with motion parameter estimates.

A Generalized Least Squares Regression analysis was performed in individual subjects in AFNI (3dREMLfit). To create the regressors of interest, a stimulus-timing vector was created for each bias ratio condition by modeling each sparse image timepoint as “on” or “off” for that condition. The resulting regression coefficients for each bias ratio represented the mean percent signal change (PSC) from rest. A linear contrast representing the average activation (versus rest) across all bias ratio conditions was also calculated for each subject.

Group Analysis

First, a mask of "active" voxels was created by entering the individual subject contrast coefficients for average activation across bias ratio conditions (relative to rest) in a Mixed-Effects Meta-Analysis (AFNI 3dMEMA) at the group level. This procedure is similar to a standard group level t-test but also takes into account the level of intra-subject variation by accepting t-scores from each individual subject analysis. A voxel-wise threshold was applied using the false discovery rate (FDR) procedure at $q < 0.05$. Voxels surviving this analysis demonstrated a mean level of activity that was significantly greater than baseline across all bias conditions and all subjects at the chosen threshold. All further analyses were restricted to this set of voxels.

To evaluate whether voxels were sensitive to changes in behaviorally measured response bias, we performed an orthogonal linear regression with absolute value of the bias score as the predictor variable and percent signal change as the dependent variable (orthogonal regression accounts for measurement error in both the predictor and dependent variables). We chose to use the absolute value of our bias measure because the sign of C_q reflects the direction of response bias (toward "same" or "different"), and we wanted to assess the effect of overall bias magnitude on percent signal change, without respect to direction. As such, we will subsequently refer to the vector of bias values entered in the group fMRI analysis as $|C|=(|C_{0.333}|, |C_{0.5}|, |C_1|, |C_2|, |C_3|)$. Individual subject vectors $|C|$ and $\mathrm{PSC}=(\mathrm{PSC}_{0.333}, \mathrm{PSC}_{0.5}, \mathrm{PSC}_1, \mathrm{PSC}_2, \mathrm{PSC}_3)$ were concatenated across subjects and entered in the group regression. To account for between subject variability in $|C|$ and PSC, measures in each individual subject vector were converted to z-scores prior to regression. Thus, the ratio of error variances in the orthogonal regression was assumed to be 1, such that the equation for the slope of the regression line ($y = mx + b$) took the form

$$m = \frac{\left(\left(\sum_{i}^{N} V_i^2 - \sum_{i}^{N} U_i^2 \right) + \sqrt{\left\{ \sum_{i}^{N} V_i^2 - \sum_{i}^{N} U_i^2 \right\}^2 + 4 \left\{ \sum_{i}^{N} U_i V_i \right\}^2} \right)}{2 \sum_{i}^{N} U_i V_i},$$

where $U_i = x_i - \text{mean}(x)$ and $V_i = y_i - \text{mean}(y)$. A one-out jackknife procedure was used to estimate the standard error of the slope estimator. Jackknife t-statistics were constructed in the form

$$t = \frac{N * m - (N-1) * \overline{m}}{(N-1) * (\hat{\sigma} / \sqrt{N})}$$

where m is the slope estimator, $\overline{m}$ is the mean of the jackknife distribution of slope estimators, and $\hat{\sigma}$ is the standard deviation of the jackknife distribution of slope estimators. Hypothesis testing was performed against values from a student’s t distribution with N-2 degrees of freedom.

In sum, our group analysis consisted of orthogonal linear regression of 75 bias scores on their corresponding 75 PSC measures (five bias ratio conditions, 15 subjects; three subjects were excluded on the basis of our criterion on the maximum allowable mean 95% Bayesian credible interval for *C*). Voxels were deemed to be significant at an FDR-corrected threshold of $q < 0.01$ with a minimum cluster size of 20 voxels.

Results

Behavioral Results

During the fMRI session, subjects performed a total of 144 same-different discrimination trials in each of five bias ratio conditions: 1:3, 1:2, 1:1, 2:1, 3:1. Consonant vowel pairs were presented in a background of Gaussian noise at a constant SNR based on a behaviorally determined threshold performance (see methods). Subjects were expected to make use of same-different ratio information to bias their response patterns – e.g., in the 1:3 ratio condition, subjects would be expected to respond "different" more often on uncertain trials, and in the 3:1 ratio condition subjects would be expected to respond "same" more often. As such, our behavioral measure of response bias, *C*, was expected to vary significantly across bias ratio conditions. The results bear out this expectation: group mean *C* varied significantly ($\Lambda = 0.170$, $F(4,11) = 13.461$, $p < 0.001$) and in the expected direction (larger negative values for ratio conditions with a greater number of different trials and larger positive values for ratio conditions with a greater number of same trials (Figure 2.3). This result confirms that our treatment succeeded in manipulating response bias while holding the physical stimuli constant across conditions. For completeness, we also calculated a summary *d'* measure for each condition by tabulating the overall hit rate and false alarm rate across all trial types and entering these values in the standard signal detection formula for same-different designs (Independent Observation model; see Macmillan & Creelman, 2005). Indeed, group mean *d'* values did not vary significantly across conditions ($\Lambda = 0.757$, $F(4,14) = 1.125$, $p = 0.384$).

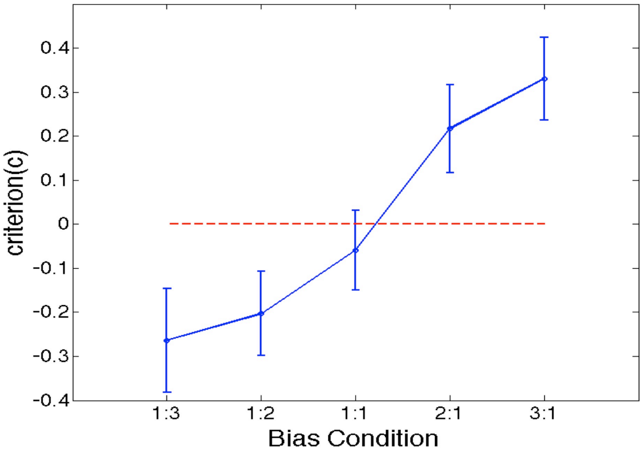


Figure 2.3. Group behavioral results for our bias measure, C. The data are plotted as a line graph where the x-axis is the same-different ratio and the y-axis is the group mean value of C. Error bars reflect ± one standard error of the mean. The zero criterion value (no bias) is plotted as a dotted line in red. Clearly, C varies significantly in the expected direction (negative values indicate bias toward responding "different" and positive values indicate bias toward responding "same"). Also, the mean criterion value in the 1:1 bias ratio condition is closest to zero (and contains zero within ± 1SE), as expected.

fMRI Results

Overall activation to speech discrimination versus rest was measured on the basis of a linear contrast modeling mean activation across bias ratio conditions (see Methods). The group result for this contrast (Figure 2.4) reveals a typical perisylvian language network including activation in bilateral auditory cortex, anterior and posterior superior temporal gyrus (STG), posterior superior temporal sulcus (STS), and planum temporale. Activation was also observed in speech motor brain regions including left inferior frontal gyrus (IFG) and insula, bilateral motor/premotor cortex and bilateral supplementary motor area (SMA). Other active areas include bilateral parietal lobe (including somatosensory cortex), thalamus and basal ganglia, cerebellum, prefrontal cortex, and visual cortex. All activations are reported at FDR-corrected P

< 0.05. Subsequent analyses were restricted to suprathreshold voxels in this task versus rest analysis.

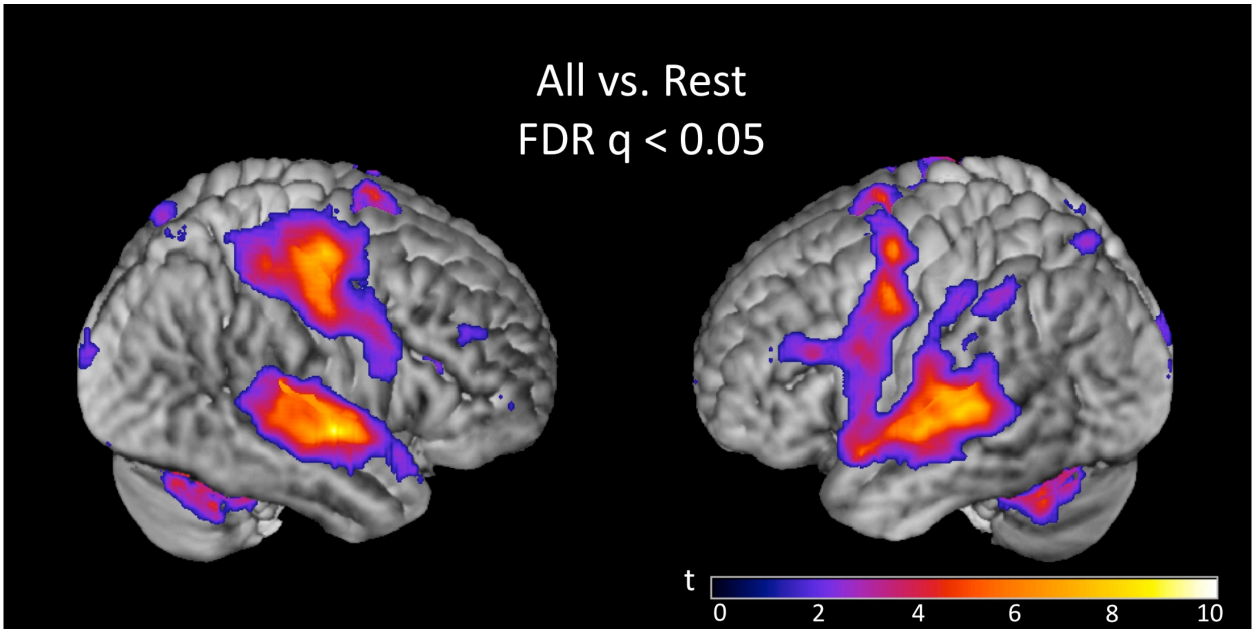


Figure 2.4. Group (n=15) t-map for the contrast corresponding to mean activation in the syllable discrimination task (versus rest) across all five bias ratio conditions. The statistical image was thresholded at FDR-corrected q < 0.05. The set of voxels identified in this contrast were used as a mask for all subsequent analyses.

In order to isolate active voxels for which activity was modulated significantly by changes in response bias, we carried out a group level regression of |*C*| against measured percent signal change (PSC). Each measure was converted to a z-score prior to regression in order to account for between-subject variability in |*C*| and PSC. In other words, we wanted to identify voxels for which changes in response bias (regardless of direction) were associated with changes in PSC in individual subjects. We hypothesized that voxels in speech motor brain regions would be most strongly modulated by changes in response bias. Indeed, significant voxels were almost exclusively restricted to motor and/or frontoparietal sensory-motor brain regions. Clusters significantly modulated by response bias were identified in left ventral precentral gyrus bordering on IFG, a more dorsal aspect of the left ventral precentral gyrus, left insula, bilateral SMA, left ventral postcentral gyrus extending into frontosylvian cortex, right superior parietal lobule, left inferior parietal lobule, and bilateral superior frontal cortex including middle frontal gyrus, superior frontal gyrus, and dorsal precentral gyrus (see Table 2.1 for MNI coordinates). One additional cluster was identified in right peri-calcarine visual cortex. Each of these clusters demonstrated a strong negative relationship between $|C|$ and PSC (i.e., signal was generally stronger when participants exhibited less response bias (Figure 2.5). The significance of this relationship is discussed at length below but, briefly, we believe that signal increases were produced by more effortful processing when, 1) probabilistic information was not available to subjects (in the 1:1 condition), or 2) subjects chose to ignore available probabilistic information (low measured response bias in the 1:3, 1:2, 2:1, or 3:1 conditions), leading to increased recruitment of the sensory-motor network elaborated previously. Indeed, no clusters were found to demonstrate a positive relationship between response bias and PSC, and activity in temporal lobe structures was not correlated with response bias.

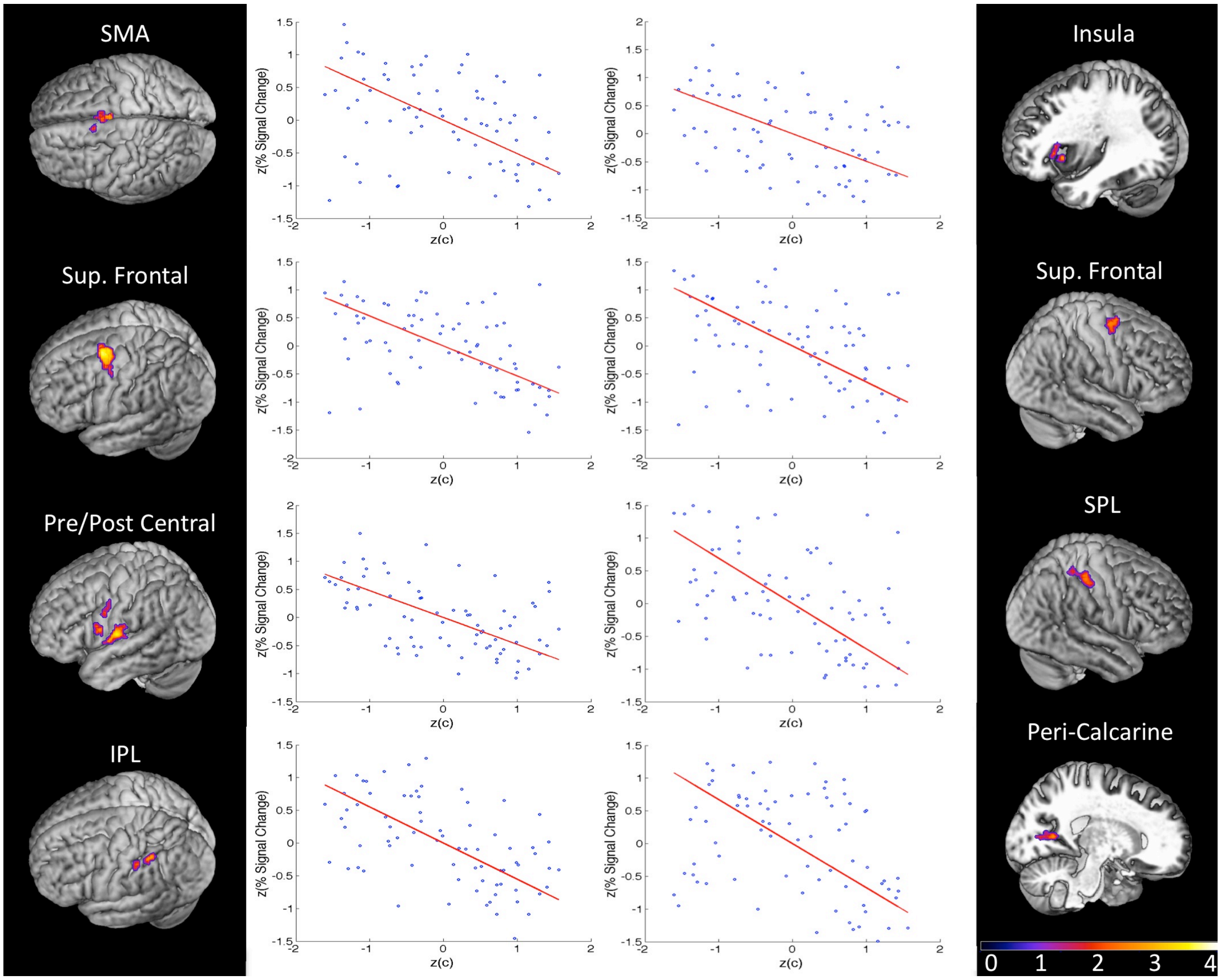


Figure 2.5. Group results (z-maps, where FDR q values are converted to z-scores) for the orthogonal regression of response bias (|C|) on BOLD percent signal change (PSC). Each region pictured was significant at FDR-corrected q < 0.01 with a minimum cluster size of 20 voxels. Next to each significant cluster is a scatter plot of normalized bias score (x-axis) against normalized percent signal change (averaged across all voxels in the region). Each plot contains five data points from each of the 15 subjects (blue) corresponding to the five same-different ratio conditions in the syllable discrimination experiment. The orthogonal least squares fit is plotted in red. There is a strong negative relationship between response bias and percent signal change in every region of this fronto-parietal network.

Table 2.1

MNI coordinates of the center of mass in activated cluster (thresholded FDR $q < 0.01$, minimum 20 voxels, group analysis)

	Number of Voxels	Hemisphere			
			x	y	z
Correlated with Bias (|C|)					
Superior Frontal Gyrus	434	Left	-30	-1	60
Supplementary Motor Area	287	Bilateral	-1	2	56
Ventral Postcentral Gyrus	255	Left	-62	-10	14
Superior Parietal Lobule/Post-Central Gyrus	125	Right	51	-31	59
Superior Frontal Gyrus	95	Right	34	-4	65
Inferior Parietal Lobule	48	Left	-48	-43	51
Peri-Calcarine Cortex	47	Right	28	-65	6
Insula	37	Left	-35	22	9
Ventral Pre-central Gyrus	37	Left	-58	7	18
Dorsal Pre-Central Gyrus	26	Left	-62	-1	40
Insula	24	Left	-42	16	2
Inferior Parietal Lobule	23	Left	-59	-30	54

Discussion

Performance on a speech sound discrimination task involves at least two processes, perceptual analysis and response selection. In the present experiment we effectively held perceptual analysis (signal-to-noise ratio) constant while biasing response selection (probability of same vs. different trials). Our behavioral findings confirmed that this manipulation was successful, as our measure of bias, *C*, changed significantly across conditions as expected. This allowed us a means for identifying the brain regions involved in the response selection component of the task. When the discrimination task was compared against baseline (rest) we found a broad area of activation including superior temporal, frontal, and parietal regions bilaterally, implicating auditory as well as motor and sensory-motor regions in the performance of the task. However, when we tested for activation that was correlated with changes in response bias, only the motor and sensory-motor areas were found to be significantly modulated; no temporal lobe regions were identified. This finding is consistent with a model in which auditory-related regions in the temporal lobe are performing perceptual analysis of speech, while the motor-related regions support (some aspect of) the response selection component of the task.

A similar dissociation has been observed in monkeys performing vibrotactile frequency discrimination (VTF). In a VTF experiment, monkeys must compare the frequency of vibration of two tactile stimuli, f1 and f2, separated by a time gap. The monkey must decide whether the frequency of vibration was greater for f1 or f2, communicating its answer by pressing a button with the nonstimulated hand. Single-unit recordings taken from neurons in primary somatosensory cortex (S1) indicate that, for many S1 neurons, the average firing rate increases monotonically with increasing stimulus frequency (Hernandez, Zainos & Romo, 2000). It has been argued that this rate code serves as sensory evidence for the discrimination decision (Romo, Hernandez, Zainos & Brody, 2000; Romo, Hernandez, Zainos & Salinas, 1998; Salinas, Hernandez, Zainos & Romo, 2000). However, S1 firing rates do not dissociate on the basis of trial type – that is, there is no difference in the mean firing rate (across frequencies) during the comparison (f2) period for trials of type f1>f2 versus f2>f1, so S1 responses strictly reflect f1 frequency in the first interval and f2 frequency in the second interval, not their relation (cf., Gold & Shadlen, 2005). However, neurons in several areas exhibit activity patterns that do reflect a comparison between f1 and f2. In particular, neurons in the ventral and medial premotor cortices persist in firing during the delay period between f1 and f2, and discriminate trial type on the basis of mean firing rate during the comparison period (Romo, Hernandez & Zainos, 2004). This indicates that these neurons are likely participating in maintenance and comparison of sensory representations (Hernandez, Zainos & Romo, 2002; Romo et al., 2004).

In the case of our syllable discrimination paradigm, it is unclear exactly what aspect(s) of response selection is (are) being performed by the motor and sensory-motor brain regions identified in our fronto-parietal network. Beyond perceptual analysis of the stimuli, a discrimination task involves short-term maintenance of the pair of stimuli and some evaluation and decision process. Given that our bias measure was negatively correlated with BOLD signal in the fronto-parietal network, i.e., that activation was higher with less bias, the following account is suggested. Bias can simplify a response decision by providing a viable response option in the absence of strong evidence from a perceptual analysis. In our case, knowledge of the same versus different trial ratio provides a probabilistically determined response option in the case of uncertainty. So when subjects were in doubt based on the perceptual analysis, a simple decision, go with the most likely response, was available and if used would tend to decrease activation in a neural network involved in response selection.

Although our experiment cannot determine which aspects of the response selection process are driving activation in the fronto-parietal network, the location of some of the activations suggest some likely possibilities. For example, the involvement of lower-level motor speech areas such as ventral and dorsal premotor cortex, regions previously implicated in phonological working memory (Buchsbaum et al., 2011; Buchsbaum, Olsen Koch et al., 2005; Buchsbaum, Olsen, Koch, Kohn, et al., 2005; Hickok et al., 2003), suggest that these regions support response selection via short-term maintenance of phonological information. One possibility, therefore, is that in the absence of either decisive perceptual information or a strong decision bias, subjects will work harder to come to a decision and as part of this effort will maintain short-term activation of the stimuli in working memory for a longer period of time resulting in more activation in regions supporting articulatory rehearsal.

An alternative view of the role of the motor system is that it is particularly involved in speech perception during degraded listening conditions, such as in the current experiment in which stimuli were presented in noise and near psychophysical threshold (Binder et al., 2004; Callan, Jones, Callan & Akahane-Yamada, 2004; D'Ausilio et al., 2011; Shahin, Bishop & Miller, 2009; Zekveld et al., 2006). This view is not inconsistent with the present findings, at least broadly. For example, the motor system may assist in perception via its role in phonological working memory. In this case, the explanation of our findings proposed above would hold equally well as this "alternative." However, most motor-oriented theorists promote a more powerful role for the motor system in speech perception, holding that it contributes substantively to the perceptual analysis. Some of these authors have argued for a strong version of the motor theory of speech perception by which motor representations themselves must be activated in order for perception to occur (Fadiga & Craighero, 2006; Gallese, Fadiga, Fogassi & Rizzolatti, 1996). Others hold a more moderate view in which the motor system is able to modulate sensory analysis of speech, for example, via predictive coding (Bever & Poeppel, 2010; Callan et al., 2004; Skipper, van Wassenhove, Nusbaum & Small 2007; Wilson & Iacoboni, 2006). The present data do not provide strong support for either of these possibilities because if the motor system were contributing to perceptual analysis, then modulating motor activity, as we successfully achieved in our study would have been expected to result in a modulation of perceptual discrimination, which it did not.

The above argument, that modulation of the motor system did not result in a corresponding modulation of perceptual discrimination and therefore the motor system is not contributing to perception, is dependent on whether we modulated the relevant components of the motor system. To assess this, we examined the relation between our biased-induced modulation and two prominent previous TMS studies that have targeted the motor speech system and found modulatory effects on (potentially biased) measures of performance. Figure 2.6 shows that the regions that are correlated with our measure of bias overlap those regions that were stimulated in previous TMS studies of motor involvement in speech perception. This suggests that we were successfully able to modulate activity in these same regions and yet still failed to observe an effect on perceptual discrimination.

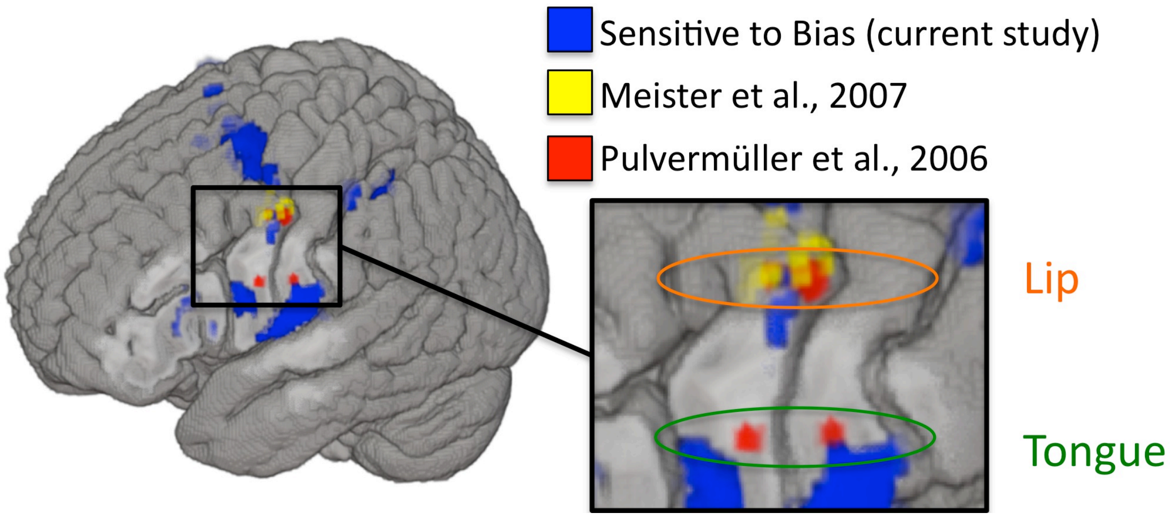

🟦 Sensitive to Bias (current study)
🟨 Meister et al., 2007
🟥 Pulvermüller et al., 2006
Lip
Tongue

Figure 2.6. Left hemisphere activations that were sensitive to changes in response bias are plotted in blue (current study). MNI coordinates used to target rTMS in Meister et al. (2007; listed in their table 1) are plotted as 2mm radius spheres in yellow. Motor ROIs from Pulvermüller et al. (2006) are plotted in red as 3mm radius spheres around the MNI peak activation foci (8mm radius ROIs were used in the original study). Depth is represented faithfully – activations nearest to the displayed cortical surface are increasingly bold in color, while activations farthest from the displayed surface are increasingly transparent. Targets in Meister et al. were peak activations from an fMRI localizer experiment involving perception of auditory speech. Disruptive stimulation of these targets resulted in a slight decrement in syllable identification performance. The motor ROIs from Pulvermüller et al. were identified on the basis of activation to lip and tongue movements in a motor localizer experiment. The motor somatotopy established therein is also shown (lip foci are circled in orange and tongue foci are circled in green). The two posterior foci (one in the lip region and one in the tongue region) were chosen to target TMS stimulation in d'Ausilio et al. (2009), which demonstrated that excitatory stimulation to lip and tongue motor cortex selectively facilitated identification of phonemes relying on those articulators.

What might explain the discrepancy between our finding, that the motor speech system is modulated by manipulations of bias (and not perceptual discrimination), and TMS findings, which show that stimulation of portions of the motor speech system affect measures of speech perception? Given that none of the previous TMS studies used an unbiased dependent measure, it is a strong possibility that what is being affected by TMS to motor speech areas is not speech perception ability but response bias. An alternative possibility is consistent with our proposed interpretation of what is driving the bias correlation in our experiment, namely that motor speech regions support speech perception tasks only indirectly via articulatory rehearsal. This could explain Meister et al.'s (2007) result that stimulation to premotor cortex caused a decline in performance on the assumption that active maintenance of the stimulus provides some benefit to performance. It could also explain the somatotopic-specificity result of d'Ausilio et al. (2009) on the assumption that stimulating lip or tongue areas biased the contents of phonological working memory. Further TMS studies using unbiased measures will be needed to sort out these possibilities.

To review, we specifically modulated response bias in a same-different syllable discrimination task by cueing participants to the same-different ratio, which was varied over five values (1:3, 1:2, 1:1, 2:1, 3:1). We used the measure *C* from our modified signal detection analysis to evaluate performance, where this measure is taken to represent behavioral response bias. Our experimental manipulation was successful in that response bias varied significantly across conditions while the physical stimuli remained constant, and we demonstrated that overall magnitude of response bias ($|C|$) correlated significantly with BOLD percent signal change in a frontoparietal network of motor and sensory-motor brain regions. We had predicted that we would observe a significant relationship between response bias and BOLD activity in frontal but not temporal brain structures. Indeed, eight of twelve clusters demonstrating a significant relationship between bias and percent signal change were entirely confined to frontal cortex or contained voxels in frontal cortex. None of the clusters contained voxels in temporal cortex. In each region there was a strong negative relationship between response bias and BOLD activation level, which we argued was due to our sensory-motor network participating in response selection components of the syllable discrimination task. In particular, we argued that the load on response selection was reduced when a probabilistically determined response option was available, resulting in lower activation levels. We also demonstrated that several of our clusters in the vicinity of the left premotor cortex overlap quite well with premotor foci previously implicated in modulation of speech perception ability. However, our results undermine claims that speech perception is supported directly by premotor cortex since our manipulation of response bias successfully modulated activity in these regions without modulating speech discrimination performance.

References

Argall, B. D., Z. S. Saad, and M. S. Beauchamp. 2006. “Simplified intersubject averaging on the cortical surface using SUMA.” *Hum Brain Mapp* no. 27 (1):14-27.

Avants, B., C. Anderson, M. Grossman, and J. C. Gee. 2007. “Spatiotemporal normalization for longitudinal analysis of gray matter atrophy in frontotemporal dementia.” *Medical image computing and computer-assisted intervention : MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention* no. 10 (Pt):303-10.

Avants, B. B., P. Yushkevich, J. Pluta, D. Minkoff, M. Korczykowski, J. Detre, and J. C. Gee. 2010. “The optimal template effect in hippocampus studies of diseased populations.” *Neuroimage* no. 49 (3):2457-2466.

Avants, B., J. T. Duda, J. Kim, H. Zhang, J. Pluta, J. C. Gee, and J. Whyte. 2008. “Multivariate Analysis of Structural and Diffusion Imaging in Traumatic Brain Injury.” *Academic Radiology* no. 15 (11):1360-1375.

Avants, B., and J. C. Gee. 2004. “Geodesic estimation for large deformation anatomical shape averaging and interpolation.” *Neuroimage* no. 23:139-50.

Baker, Errol, Sheila E. Blumstein, and Harold Goodglass. 1981. “Interaction between phonological and semantic factors in auditory comprehension.” *Neuropsychologia* no. 19 (1):1-15. Doi: 10.1016/0028-3932(81)90039-7.

Bever, T.G.; Poeppel, D. 2010. “Analysis by synthesis: A (re-) emerging program of research for language and vision.” *Biolinguistics* no. 4 (2-3):174-200.

Binder, J. R., E. Liebenthal, E. T. Possing, D. A. Medler, and B. D. Ward. 2004. “Neural correlates of sensory and decision processes in auditory object identification.” *Nat Neurosci* no. 7 (3):295-301.

Bishop, D. V., B. B. Brown, and J. Robson. 1990. “The relationship between phoneme discrimination, speech production, and language comprehension in cerebral-palsied individuals.” *Journal of speech and hearing research* no. 33 (2):210-9.

Blumstein, Sheila E. 1995. “The neurobiology of the sound structure of language.” In *The cognitive neurosciences.*, edited by Michael S. Gazzaniga, 915-929. Cambridge, MA, US: The MIT Press.

Buchsbaum, B. R., J. Baldo, K. Okada, K. F. Berman, N. Dronkers, M. D’Esposito, and G. Hickok. 2011. “Conduction aphasia, sensory-motor integration, and phonological short-term memory – An aggregate analysis of lesion and fMRI data.” *Brain Lang* no. 119 (3):119-28. Doi: 10.1016/j.bandl.2010.12.001.

Buchsbaum, B. R., R. K. Olsen, P. Koch, and K. F. Berman. 2005. “Human dorsal and ventral auditory streams subserve rehearsal-based and echoic processes during verbal working memory.” *Neuron* no. 48 (4):687-97.

Buchsbaum, B. R., R. K. Olsen, P. F. Koch, P. Kohn, J. S. Kippenhan, and K. F. Berman. 2005. “Reading, hearing, and the planum temporale.” *Neuroimage* no. 24 (2):444-54.

Callan, D. E., J. A. Jones, A. M. Callan, and R. Akahane-Yamada. 2004. “Phonetic perceptual identification by native- and second-language speakers differentially activates brain regions involved with acoustic phonetic processing and those involved with articulatory-auditory/orosensory internal models." *Neuroimage* no. 22 (3):1182-94.

Christen, H. J., F. Hanefeld, E. Kruse, S. Imh‰user, J. P. Ernst, and M. Finkenstaedt. 2000. "Foix-Chavany-Marie (anterior operculum) syndrome in childhood: a reappraisal of Worster-Drought syndrome." *Developmental medicine and child neurology* no. 42 (2):122-32.

D'Ausilio, A., F. Pulvermuller, P. Salmas, I. Bufalari, C. Begliomini, and L. Fadiga. 2009. "The motor somatotopy of speech perception." *Curr Biol* no. 19 (5):381-5. Doi: 10.1016/j.cub.2009.01.017.

Damasio, Antonio R. 1992. "Aphasia." *New England Journal of Medicine* no. 326 (8):531-539. Doi: doi:10.1056/NEJM199202203260806.

Davis, M. H., and I. S. Johnsrude. 2003. "Hierarchical processing in spoken language comprehension." *J Neurosci* no. 23 (8):3423-31.

Desai, R., E. Liebenthal, E. T. Possing, E. Waldron, and J. R. Binder. 2005. "Volumetric vs. surface-based alignment for localization of auditory cortex activation." *Neuroimage* no. 26 (4):1019-29.

Fadiga, L., and L. Craighero. 2006. "Hand Actions and Speech Representation in Broca's Area." *Cortex* no. 42 (4):486-490.

Fadiga, Luciano, Laila Craighero, Giovanni Buccino, and Giacomo Rizzolatti. 2002. "Speech listening specifically modulates the excitability of tongue muscles: a TMS study." *European Journal of Neuroscience* no. 15 (2):399-402. Doi: 10.1046/j.0953-816x.2001.01874.x.

Galantucci, B., C. A. Fowler, and L. Goldstein. 2009. "Perceptuomotor compatibility effects in speech." *Atten Percept Psychophys* no. 71 (5):1138-49. Doi: 10.3758/APP.71.5.1138.

Gallese, V., L. Fadiga, L. Fogassi, and G. Rizzolatti. 1996. "Action recognition in the premotor cortex." *Brain* no. 119:593-609.

Gold, J. I., and M. N. Shadlen. 2007. "The neural basis of decision making." *Annu Rev Neurosci* no. 30:535-74.

Goodglass, Harold. 1993. *Understanding aphasia*. San Diego: Academic Press.

Goodglass, Harold, Edith Kaplan, and Barbara Barresi. 2001. *The assessment of aphasia and related disorders*. Philadelphia: Lippincott Williams & Wilkins.

Hasson, U., J. I. Skipper, H. C. Nusbaum, and S. L. Small. 2007. "Abstract coding of audiovisual speech: beyond sensory representation." *Neuron* no. 56 (6):1116-26. Doi: 10.1016/j.neuron.2007.09.037.

Hernández, A., A. Zainos, and R. Romo. 2002. "Temporal evolution of a decision-making process in medial premotor cortex." *Neuron* no. 33 (6):959-72.

Hernández, A., A. Zainos, and R. Romo. 2000. "Neuronal Correlates of Sensory Discrimination in the Somatosensory Cortex." *Proc Natl Acad Sci U S A* no. 97 (11):6191-6196.

Hickok, G., M. Costanzo, R. Capasso, and G. Miceli. 2011. "The role of Broca's area in speech perception: Evidence from aphasia revisited." *Brain Lang* no. 119 (3):214-20. Doi: 10.1016/j.bandl.2011.08.001.

Hickok, G., K. Okada, W. Barr, J. Pa, C. Rogalsky, K. Donnelly, L. Barde, and A. Grant. 2008. "Bilateral capacity for speech sound processing in auditory comprehension: Evidence from Wada procedures." *Brain and Language* no. 107 (3):179-184.

Hickok, Gregory. 2009. "Eight Problems for the Mirror Neuron Theory of Action Understanding in Monkeys and Humans." *Journal of Cognitive Neuroscience* no. 21 (7):1229-1243. Doi: 10.1162/jocn.2009.21189.

Hickok, Gregory. (2012). Computational neuroanatomy of speech production. *Nature Reviews Neuroscience, 13*(2), 135-145.

Hickok, Gregory, Bradley Buchsbaum, Colin Humphries, and Tugan Muftuler. 2003. "Auditory–Motor Interaction Revealed by fMRI: Speech, Music, and Working Memory in Area Spt." *Journal of Cognitive Neuroscience* no. 15 (5):673-682. Doi: 10.1162/jocn.2003.15.5.673.

Hillis, A. E. 2007. "Aphasia: progress in the last quarter of a century." *Neurology* no. 69 (2):200-13. Doi: 10.1212/01.wnl.0000265600.69385.6f.

Klein, A., J. Andersson, B. A. Ardekani, J. Ashburner, B. Avants, M. C. Chiang, G. E. Christensen, D. L. Collins, J. Gee, P. Hellier, J. H. Song, M. Jenkinson, C. Lepage, D. Rueckert, P. Thompson, T.

Vercauteren, R. P. Woods, J. J. Mann, and R. V. Parsey. 2009. “Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration.” *Neuroimage* no. 46 (3):786-802.

Klein, A., S. S. Ghosh, B. Avants, B. T. T. Yeo, B. Fischl, B. Ardekani, J. C. Gee, J. J. Mann, and R. V. Parsey. 2010. “Evaluation of volume-based and surface-based brain image registration methods.” *Neuroimage* no. 51 (1):214-220.

Lenneberg, E. H. 1962. “Understanding language without ability to speak: a case report.” *Journal of abnormal and social psychology* no. 65:419-25.

Levine, D. N., and J. P. Mohr. 1979. “Language after bilateral cerebral infarctions: role of the minor hemisphere in speech.” *Neurology* no. 29 (7):927-38.

Levitt, H. (1971). “Transformed up-down methods in psychoacoustics.” *J. Acoust. Soc. Am.* No. 49 (2): 467-477.

Macmillan, Neil A., and C. Douglas Creelman. 2005. *Detection theory : a user’s guide*. Mahwah, NJ [u.a.]: Erlbaum.

Meister, I. G., S. M. Wilson, C. Deblieck, A. D. Wu, and M. Iacoboni. 2007. “The essential role of premotor cortex in speech perception.” *Curr Biol* no. 17 (19):1692-6. Doi: 10.1016/j.cub.2007.08.064.

Miceli, Gabriele, Guido Gainotti, Carlo Caltagirone, and Carlo Masullo. 1980. “Some aspects of phonological impairment in aphasia.” *Brain and Language* no. 11 (1):159-169. Doi: 10.1016/0093-934x(80)90117-0.

Moineau, Suzanne, Nina F. Dronkers, and Elizabeth Bates. 2005. “Exploring the Processing Continuum of Single-Word Comprehension in Aphasia.” *J Speech Lang Hear Res* no. 48 (4):884-896. Doi: 10.1044/1092-4388(2005/061).

Möttönen, Riikka, and Kate E. Watkins. 2009. “Motor Representations of Articulators Contribute to Categorical Perception of Speech Sounds.” *The Journal of Neuroscience* no. 29 (31):9819-9825. Doi: 10.1523/jneurosci.6018-08.2009.

Pulvermuller, F., and L. Fadiga. 2010. “Active perception: sensorimotor circuits as a cortical basis for language.” *Nat Rev Neurosci* no. 11 (5):351-60. Doi: 10.1038/nrn2811.

Pulvermuller, F., M. Huss, F. Kherif, F. Moscoso del Prado Martin, O. Hauk, and Y. Shtyrov. 2006. “Motor cortex maps articulatory features of speech sounds.” *Proc Natl Acad Sci U S A* no. 103 (20):7865-70. Doi: 10.1073/pnas.0509989103.

Rizzolatti, G., and L. Craighero. 2004. “The mirror-neuron system.” *Annu Rev Neurosci* no. 27:169-92. Doi: 10.1146/annurev.neuro.27.070203.144230.

Rogalsky, C., T. Love, D. Driscoll, S. W. Anderson, and G. Hickok. 2011. “Are mirror neurons the basis of speech perception? Evidence from five cases with damage to the purported human mirror system.” *Neurocase* no. 17 (2):178-87. Doi: 10.1080/13554794.2010.509318.

Romo, R., A. Hern·ndez, and A. Zainos. 2004. “Neuronal correlates of a perceptual decision in ventral premotor cortex.” *Neuron* no. 41 (1):165-73.

Romo, R., A. Hern·ndez, A. Zainos, C. D. Brody, and L. Lemus. 2000. “Sensing without touching: psychophysical performance based on cortical microstimulation.” *Neuron* no. 26 (1):273-8.

Romo, R., A. Hernandez, A. Zainos, and E. Salinas. 1998. “Somatosensory discrimination based on cortical microstimulation.” *Nature.* No. 392 (6674):387.

Salinas, E., A. Hernandez, A. Zainos, and R. Romo. 2000. “Periodicity and firing rate as candidate neural codes for the frequency of vibrotactile stimuli.” *J Neurosci* no. 20 (14):5503-15.

Sato, M., P. Tremblay, and V. L. Gracco. 2009. “A mediating role of the premotor cortex in phoneme segmentation.” *Brain Lang* no. 111 (1):1-7. Doi: 10.1016/j.bandl.2009.03.002.

Sato, Marc, Krystyna Grabski, Arthur M. Glenberg, Amélie Brisebois, Anahita Basirat, Lucie Ménard, and Luigi Cattaneo. 2011. “Articulatory bias in speech categorization: Evidence from use-induced motor plasticity.” *Cortex* no. 47 (8):1001-1003. Doi: 10.1016/j.cortex.2011.03.009.

Shahin, A. J., C. W. Bishop, and L. M. Miller. 2009. “Neural mechanisms for illusory filling-in of degraded speech.” *Neuroimage* no. 44 (3):1133-1143.

Skipper, J. I., H. C. Nusbaum, and S. L. Small. 2005. “Listening to talking faces: motor cortical activation during speech perception.” *Neuroimage* no. 25 (1):76-89. Doi: 10.1016/j.neuroimage.2004.11.006.

Skipper, J. I., V. van Wassenhove, H. C. Nusbaum, and S. L. Small. 2007. “Hearing lips and seeing voices: how cortical areas supporting speech production mediate audiovisual speech perception.” *Cereb Cortex* no. 17 (10):2387-99. Doi: 10.1093/cercor/bhl147.

Venezia, Jonathan H., and Gregory Hickok. 2009. “Mirror Neurons, the Motor System and Language: From the Motor Theory to Embodied Cognition and Beyond.” *Language and Linguistics Compass* no. 3 (6):1403-1416. Doi: 10.1111/j.1749-818X.2009.00169.x.

Watkins, K. E., A. P. Strafella, and T. Paus. 2003. “Seeing and hearing speech excites the motor system involved in speech production.” *Neuropsychologia* no. 41 (8):989-994. Doi: 10.1016/s0028-3932(02)00316-0.

Weller, M. 1993. “Anterior opercular cortex lesions cause dissociated lower cranial nerve palsies and anarthria but no aphasia: Foix-Chavany-Marie syndrome and “automatic voluntary dissociation” revisited.” *Journal of neurology* no. 240 (4):199-208.

Wilson, S. M., and M. Iacoboni. 2006. “Neural responses to non-native phonemes varying in producibility: evidence for the sensorimotor nature of speech perception.” *Neuroimage* no. 33 (1):316-25. Doi: 10.1016/j.neuroimage.2006.05.032.

Wilson, S. M., A. P. Saygin, M. I. Sereno, and M. Iacoboni. 2004. “Listening to speech activates motor areas involved in speech production.” *Nat Neurosci* no. 7 (7):701-2. Doi: 10.1038/nn1263.

Zekveld, A. A., D. J. Heslenfeld, J. M. Festen, and R. Schoonhoven. 2006. “Top-down and bottom-up processes in speech comprehension.” *Neuroimage* no. 32 (4):1826-36. Doi: 10.1016/j.neuroimage.2006.04.199.

CHAPTER 3

Primer

The preceding chapters worked toward the conclusion that speech perception is fundamentally carried out within sensory (as opposed to motor) brain regions. This supports a broader conclusion, albeit indirectly, that the objects of speech perception are auditory as opposed to gestural. However, speech perception is inherently multisensory. While sound is the primary medium for speech, different parameters of the speech signal can be estimated from visual or even haptic (Boothroyd, Kishon-Rabin, & Waldstein, 1995; Fowler & Dekle, 1991) signals. Chapters 3 and 4 will explore the computational and neural mechanisms underlying audiovisual integration of speech. In particular, the focus will shift from describing what speech perception *is not* as in Chs. 1 and 2, to what speech perception *is*: a set of computations capable of integrating information across multiple sensory channels. The broad assumption is that we can learn something about the speech perceptual mechanism by understanding how speech information is differentially encoded within, and subsequently integrated across, these sensory channels. This may be relatively simple for certain low-level temporal parameters (e.g., onset and offset of voicing (K. G. Munhall, Jones, Callan, Kuratate, & Vatikiotis-Bateson, 2004)) that can be estimated from temporal co-modulation between auditory and visual speech signals (and thus processed by general mechanisms). The case is more complex for integration of visual cues that specify phonetic information – i.e., information about the shape of the vocal tract conveyed through a variety of facial configuration and motion cues over time. These cues are often complementary to (rather than redundant with) the auditory signal (Q. Summerfield, 1987). The current chapter begins to work on this problem by developing a method to identify the particular visual cues that are integrated with the auditory signal for a given audiovisual speech stimulus. The technique allows a high degree of temporal precision and I exploit this precision to reveal exactly which visual cues are integrated in different temporal contexts. The results inform current models of audiovisual integration in speech.

The temporal dynamics of audiovisual fusion in speech

Jonathan H. Venezia, Steven Thurman, William Matchin, Sahara George, and Gregory Hickok

Introduction

It has been argued that processing of visual speech gestures proceeds in two distinct modes – a correlated mode in which dynamic visual features partially duplicate the spectro-temporal patterns in heard speech, and a complementary mode in which visual speech provides reliable cues that can disambiguate underspecified parts of the auditory speech stream (Campbell, 2008; Q. Summerfield, 1987). Recent evidence suggests these modes are related to two stages of processing in audiovisual integration of speech – an early *binding* stage in which coherent auditory and visual speech signals are bound (or not) to a single processing stream, and a late *fusion* stage in which auditory and visual speech signals are integrated, *per se*, into a unified percept (Berthommier, 2004; Nahorna, Berthommier, & Schwartz, 2012). More generally, it has been pointed out that perceptual systems must solve a problem of causal inference in order to determine whether multiple signals originate from a single source, and once this problem has been solved the nervous system must further determine how to integrate these signals (Shams & Kim, 2010). Binding and fusion of auditory and visual speech, as described above, likely correspond to mechanisms for solving these general problems of multisensory integration.

Fusion is essentially conditional on binding – fusion should occur only when auditory and visual speech signals are bound, and fusion strength may be partially a function of some continuous measure of audiovisual coherence, perhaps expressed as a probability (Magnotti, Ma, & Beauchamp, 2013). Recent evidence indicates that binding is sensitive (at least) to temporal and phonetic coherence (Nahorna et al., 2012), the latter of which rules out temporal comodulation between the visual and auditory speech signals (Chandrasekaran, Trubanova, Stillittano, Caplier, & Ghazanfar, 2009; Grant, 2001; Grant & Seitz, 2000) as the primary or sole cue for binding. Nonetheless, it is instructive to examine the effects of temporal coherence on audiovisual speech binding, if for no other reason than the fact that it is simple to examine the limits of binding with respect to temporal synchrony – one can simply ask subjects to make explicit synchrony judgments for a range of audiovisual temporal offsets. A line of research utilizing this technique has yielded consistent results: auditory and visual speech signals are judged to be synchronous over a relatively long temporal window ranging from ~50ms audio-lead to ~200ms visual-lead (where "lead" is relative to natural timing), and this holds for connected speech (Grant, Wassenhove, & Poeppel, 2004), words (Conrey & Pisoni, 2006), and syllables (V. van Wassenhove, Grant, & Poeppel, 2007).

The marked asymmetry in this temporal window favoring visual-lead is rather striking. A plausible explanation based on the natural dynamics of audiovisual speech has been generated to explain this effect (Grant et al., 2004; V. van Wassenhove et al., 2007). As it goes, visual speech is relatively coarse with respect to both time and informational content – that is, the information conveyed by speechreading is limited primarily to place of articulation (Grant & Walden, 1996; Massaro, 1987; Q. Summerfield, 1987; Quentin Summerfield, 1992), which evolves over a syllabic interval of ~200ms (Greenberg, 1999). Conversely, auditory speech contains robust cues that can be processed on a fine phonemic time scale of ~20-40 ms (Poeppel, 2003). The argument concerns the primacy of auditory speech signals: when relatively robust auditory information is processed before visual speech cues have been realized (i.e, even at rather short audio-lead asynchronies), there is no need to "wait around" for slowly-unfolding visual speech information, and so the auditory and visual speech streams are unbound. The opposite is true for situations in which visual speech information is processed before auditory-phonemic cues have been realized (visual-lead) – it pays to wait as long as possible for auditory information to disambiguate among candidate representations activated by visual speech (auditory and visual streams are bound). Under this characterization, visual speech plays a predictive role in bimodal speech perception (Virginie van Wassenhove, Grant, & Poeppel, 2005). Indeed, mouth movements appear to lead the voice within a range of 100-300ms for audiovisual speech with natural timing (Chandrasekaran et al., 2009), but see also (Schwartz & Savariaux, 2014).

It might be argued that the large temporal window for binding of audiovisual speech is an artifact of the somewhat unnatural synchrony judgment task. However, the same window – in terms of duration and visual-lead asymmetry – has been measured for audiovisual sentences using speech recognition as the task (Grant & Greenberg, 2001). In this study, the auditory signal was filtered into two 1/3-octave bands (298-375 Hz, 4762-6000 Hz), rendering auditory sentences largely unintelligible (18.6% recognition). Adding visual speech with natural timing boosted intelligibility to 62.6% and this boost remained for visual-lead asynchronies up to 200ms. A drop in intelligibility (46.5%) was observed for auditory-lead asynchronies beginning at 40ms. Interestingly, this suggests that the audiovisual boost to intelligibility, presumably derived from fusion computations that integrate complementary visual speech cues with the auditory signal, are essentially uniform within the audiovisual binding window. Similar evidence can be drawn from the McGurk effect (McGurk & MacDonald, 1976) – an illusion in which an auditory syllable (e.g., /pa/) dubbed onto video of an incongruent visual syllable (e.g., /ka/) yields a perceived syllable that matches neither the auditory or visual component (e.g., /ta/). Here, fusion is measured explicitly (e.g., % illusory /ta/ responses), and McGurk fusion is demonstrably robust to temporal asynchrony over the typical binding window (<50ms audio-lead to ~150ms visual-lead) (V. van Wassenhove et al., 2007).

This syllable-length window has been dubbed the 'temporal window of integration' for audiovisual speech (V. van Wassenhove et al., 2007), meaning that auditory and visual speech signals that align anywhere within this temporal window will be bound (and effectively integrated when useful information is present, as evidence by the temporal window observed for audiovisual intelligibility gains and the McGurk effect). This suggests that the set of visual cues that are integrated with the auditory signal does not depend on the temporal relationship between visual and auditory speech, so long as the two signals are within the temporal window of integration. In other words, the fusion mechanism should operate uniformly for a given set of visual cues in a bound audiovisual speech signal. However, this need not be the case. Multisensory neurons in animal models are modulated by changes in audiovisual synchrony even when these changes are within the window in which auditory and visual signals are perceptually bound (King & Palmer, 1985; Meredith, Nemitz, & Stein, 1987; Stein, Meredith, & Wallace, 1993). The same effect is observed in humans (as measured in fMRI) when audiovisual speech is the stimulus, and seems to be related to a fairly high level or processing – the effect is localized to association cortex in the superior temporal lobe (Stevenson, Altieri, Kim, Pisoni, & James, 2010).

As such, it should be informative to examine precisely the visual speech cues that contribute to fusion and their temporal relationship to the auditory speech signal. Here, we present a novel technique based on the McGurk effect that allows for such specification by examining, frame by frame (of a digital video), the visual speech information that leads to fusion. To implement the technique, we overlaid a McGurk stimulus with a spatiotemporally correlated visual masker that revealed different components of the visual speech signal at random on different trials, such that fusion was achieved on some trials but not on others based on the masking pattern. Visual information crucial to fusion was identified by comparing the making patterns on fusion trials to the patterns on non-fusion trials. This produced a high resolution spatiotemporal map of the visual features that contributed to fusion. Further, we took advantage of the fact that audiovisual speech is bound over a large range of temporal asynchronies – namely, we implemented a temporal synchrony manipulation that allowed us to examine changes in the map of fusion-related visual features relative to changes in the temporal relationship between visual and auditory speech signals. We specifically chose asynchrony values that fell within the temporal window for perceptual binding. Finally, we extracted the temporal dynamics of lip-related movements in the McGurk stimulus (Chandrasekaran et al., 2009) and compared these dynamics to the temporal dynamics of the fusion process, estimated using our masking technique. We posed the following questions: (1) What precisely are the visual cues that contribute to fusion? (2) When do these cues unfold relative to the auditory signal (i.e., is there any preference for visual information that precedes onset of the auditory signal)? (3) Are these cues related to any features in the temporal dynamics of lip movements? (4) Do the particular cues that contribute to fusion vary depending on audiovisual synchrony, even when stimuli are behaviorally indistinguishable?

Methods

Participants

A total of 34 (6 male) participants were recruited to take part in two experiments. All participants were right-handed, native speakers of English with normal hearing and normal or corrected-normal vision (self-report). Participants were students enrolled at UC Irvine and received course credit for their participation. Informed consent was obtained from each participant in accordance with the UC Irvine Institutional Review Board guidelines. Of the 34 participants, 20 were recruited for the main experiment and 14 for an additional calibration study. Three participants (all female) did not complete the main experiment and were excluded from analysis.

Stimuli

Digital videos of a single male actor producing a sequence of vowel-consonant-vowel (VCV) non-words were recorded on a high-speed camera at a native resolution of 1080p at 60 frames per second. Videos captured the head and neck of the actor against a green screen. In post-processing, the videos were cropped to 500x500 pixels and the green screen was replaced with a uniform gray background. Individual clips of each VCV were extracted such that each contained 78 frames (duration 1.3s). Audio was simultaneously recorded on separate device, digitized (44.1 kHz, 16-bit), and synced to the main video sequence in post-processing (a deliberately-produced acoustic transient in the waveform was manually aligned to the same feature in the camera audio).

In each VCV, the consonant was preceded and followed by the vowel /ɑ/ (as in 'father'). At least nine VCV clips were produced for each of the English voiceless stops – i.e, APA, AKA, ATA. Of these clips, five each of APA and ATA and one clip of AKA were selected for use in the study. To create a McGurk stimulus, audio from one APA clip was dubbed onto the video of the AKA clip. The APA audio waveform was manually aligned to the original AKA audio waveform such that the differences at offset of the initial vowel and onset of the consonant burst were minimized. This resulted in the onset of the consonant burst in the McGurk-aligned APA leading the onset of the consonant burst in the original AKA by 6ms. This McGurk stimulus will henceforth be referred to as 'Lag0' to reflect the natural timing in the alignment of the auditory and visual speech signals. Two additional McGurk stimuli were created by altering the temporal alignment of the Lag0 stimulus. Specifically, two clips with visual-lead asynchronies within the canonical temporal window of integration were created by lagging the auditory signal by 50ms (Lag50) and 100ms (Lag100), respectively (Figure 3.1). A buffer was added to the beginning of the Lag50 and Lag100 audio files to maintain duration at 1.3s. In a calibration experiment, the Lag0 McGurk clip was presented in a 4-AFC design along with congruent clips of APA, AKA, and ATA (experimental conditions were the same as for the main experiment; see below). Participants were asked to indicate what they perceived (APA, AKA, ATA, OTHER). The Lag0 stimulus was perceived as ATA 92% (± 3% SEM) of the time on average, indicating a high degree of fusion. All congruent stimuli were perceived accurately >90% of the time.

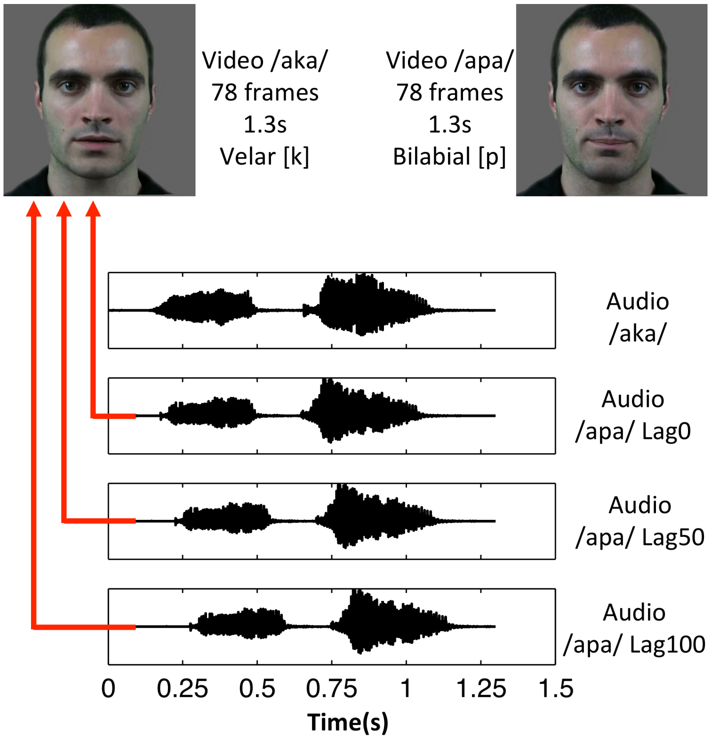


Figure 3.1. Construction of McGurk stimuli. The audio signal from an APA video clip was extracted and dubbed over an AKA video clip. This was done with three separate versions of the APA audio clip: APA audio synchronized with AKA audio (Lag0), APA audio lagged 50ms behind AKA audio (Lag50), and APA audio lagged 100ms behind AKA audio (Lag100). Audio waveforms for AKA, Lag0, Lag50, and Lag100 are displayed. Red lines show stimulus combinations used to produce the McGurk stimuli.

Procedure

For all experimental sessions, stimulus presentation and response collection was implemented in Psychtoolbox-3 (Kleiner et al., 2007) on an IBM ThinkPad running Ubuntu Linux. Auditory stimuli were presented over Sennheiser HD 280 Pro headphones and responses were collected on a DirectIN keyboard (Empirisoft). Participants were seated ~20 inches in front of the testing laptop inside an anechoic chamber (IAC Acoustics). All auditory stimuli (including those in audiovisual video clips) were presented at an average of 68 dB SPL (A) against a background of white noise at 62 dB SPL (A). This auditory signal-to-noise ratio (+6 dB) was chosen to increase the likelihood of McGurk fusion (Magnotti et al., 2013) without significantly disrupting identification of the auditory signal in isolation.

Each participant completed three days of testing spread over no more than a week. The task was phoneme identification on a six-point confidence scale: on each trial of the experiment, participants were asked to indicate whether or not they perceived the non-word APA using the 1-6 keys on the response keyboard. Participants were told they would be presented some VCV non-word of the form /ɑ/-X-/ɑ/. The '1' key indicated highest confidence for APA and the '6' key indicated highest confidence for Not-APA, with the boundary between '3' and '4' corresponding to a categorical decision boundary. The response key as displayed to participants follows:

Sure apa 1 2 3 4 5 6 Sure Not apa

Phoneme identification was performed in three conditions: audio-only, clear audiovisual (Clear-AV), and masked audiovisual (masked-AV). In the audio-only condition, participants completed two blocks of 100 trials of auditory phoneme identification. The 100 trials comprised 50 trials in which the stimulus was APA, and 50 trials in which the stimulus was ATA. There five separate APA tokens (including the APA audio used to create McGurk stimuli) and five separate ATA tokens (10 trials per token), presented in random order. In each trial, a black fixation cross was presented against a gray background over a jittered inter-trial interval (0.5-1.5s); at onset of the auditory signal, the fixation cross was replaced by the response key which remained on screen until participants made their response.

In the clear-AV condition, participants completed six blocks of 24 trials of audiovisual phoneme identification. In each block, 16 trials contained a McGurk stimulus, 4 trials contained one of four congruent APA videos, and 4 trials contained one of four congruent ATA videos. The congruent videos served as perceptual "anchoring" stimuli for the McGurk stimulus (participants were not explicitly aware that the McGurk stimulus was incongruent). The McGurk stimulus in each block was Lag0, Lag50, or Lag100 (2 blocks each, random order). In each trial, a blank gray background appeared during a variable inter-trial interval (based on video loading times), followed by presentation of the video (1.3s); at the end of the video, the response key was flashed on screen and remained until participants made their response.

The crucial condition was the masked-AV condition. Here, participants completed 24 blocks of 40 trials of audiovisual syllable identification. In each block, there were 32 McGurk trials and 8 "anchoring" trials with congruent APA or ATA videos. The trial structure was the same as for clear-AV. The major difference was that a visual masker was placed over the mouth of the speaker on each trial. The masker disrupted McGurk fusion on some trials but not on others (see section immediately following).

Visual masking technique

We developed a novel visual masking technique designed to reveal the temporal dynamics of audiovisual fusion in speech. This technique is based on "bubbles" techniques applied in some of our previous work (Thurman, Giese, & Grossman, 2010; Thurman & Grossman, 2011). Masking was applied to VCV video clips in the masked-AV condition. To apply our technique, we first down-sampled the clip to 120x120 pixels, and from this down-sampled clip we selected a 30x35 pixel region covering the mouth and part of the lower jaw of the speaker. This pixel values in this region were demeaned and a 30x35 mouth-region masker was applied as follows: (1) a random image (uniform(0,1)) was created for each frame; (2) a Gaussian blur was applied to the random image sequence in the temporal domain (sigma = 10 frames); (3) a Gaussian blur was applied to the random image sequence in the spatial domain (sigma = 14 pixels); (4) the blurred image sequence was scaled back to a range of [0 1] and a power transform (^4) was applied; (5) the processed 30x35 image sequence was multiplied to the 30x35 mouth region of the original video separately in each RGB color frame; (6) the contrast variance in the masked mouth region was adjusted to 4 and mean intensity was set to 128; (7) the fully processed sequence was up-sampled to 480x480 pixels for display. In the resultant video, a masker with spatiotemporally correlated alpha (transparency) values covered the mouth. Specifically, the mouth was (at least partially) visible in certain frames of the video, but not in other frames (Figure 3.2). Maskers were generated online and at random for each trial, such that no masker had the same pattern of transparent pixels. The crucial manipulation was masking of McGurk stimuli, where the logic of the masking process is as follows: when transparent components of the masker reveal critical visual features (i.e., of the mouth during articulation), McGurk fusion will be observed; on the other hand, when critical visual features are blocked by the masker, McGurk fusion will not be observed. The set of visual features that contribute reliably to fusion can be estimated from the relation between behavioral response patterns and masker patterns over many trials. The specific masker created for each trial was saved for later analysis.

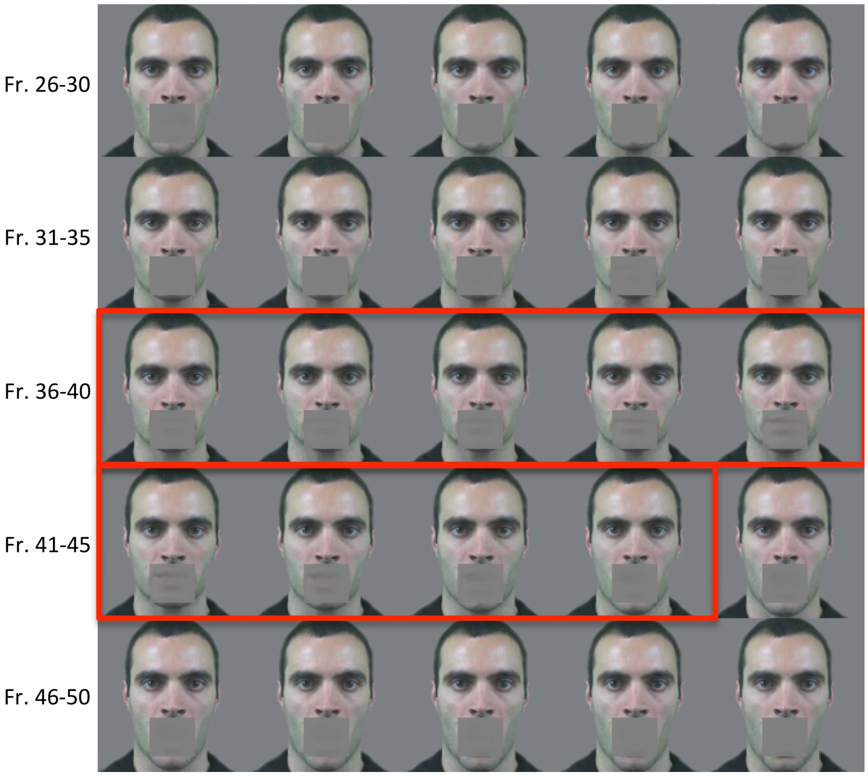


Figure 3.2. Twenty-five frames from an example masked-AV stimulus. Masker alpha (transparency) values were spatiotemporally correlated such that only certain frames would be revealed on a given trial. These frames are outlined in red on the example stimulus here. If you look closely, you will see that the mouth is visible in these frames but not in others. This effect was more natural in live videos.

Data Analysis

Performance in the audio-only and clear-AV conditions was evaluated simply in terms of % APA responses and mean confidence rating. The same measures were tabulated for congruent stimuli in the masked-AV condition. The main analysis involved constructing 'classification movies' (CMs) for the Lag0, Lag50, and Lag100 McGurk stimuli based on behavior collected in the masked-AV condition. Data from McGurk trials were grouped by stimulus: Lag0, Lag50, Lag100 (256 trials each). Separately for each stimulus, reverse correlation of trial-by-trial behavior with trial-by-trial masker alpha values was employed to construct CMs. Alpha values ranged from 0 (most opaque) to 0.5 (most transparent), and a separate alpha movie (the masker) was stored for each trial (30x35 pixels, 78 frames). Behavior was converted from six-point confidence ratings to binary weights: on trials for which participants responded APA (1-3 on the confidence scale), the response was assigned a value of -1; on trials for which participants responded Not-APA (4-6 on the confidence scale), the response was assigned a value of 1. APA responses were correct with respect to the auditory signal (no fusion occurred) while Not-APA responses were incorrect with respect to the auditory signal (fusion occurred). In other words, fusion responses were weighted positively and non-fusion responses were weighted negatively. In the reverse correlation, we performed a trial-by-trial sum of the masker alpha movies weighted by the behavioral response. The weights were scaled by the overall fusion rate (proportion Not-APA responses). Specifically, we performed a weighted sum of the alpha values (a) over all trials (*t*) for each pixel location (*x,y*) in each frame (*f*) of the masker alpha movies. For a given fusion rate, FR:

$$CM_{x,y,f} = \sum_{t=1}^{256} \begin{cases} (1 - FR) * a_t & \text{if response is fusion} \\ -FR * a_t & \text{if response is \sim fusion} \end{cases} \quad \text{for } x = 1, \dots ,30; y = 1, \dots ,35; f = 1, \dots ,78$$

The result was a CM in which large positive values appeared at pixels that were frequently transparent (in the masker) when fusion occurred, while large negative values appeared at pixels that were frequently transparent when fusion did not occur. A CM for Lag0, Lag50, and Lag100 was created separately for each participant.

In order to identify CM pixels that were reliably positive across participants, we constructed group CMs for each stimulus. Individual participant CMs were first z-scored (globally across all pixels and time points). The normalized CMs were then summed across participants and divided by sqrt(n), creating a group CM with a standard normal test statistic (group z-score) at each pixel. Group CMs were thresholded using the False Discovery Rate (FDR; $q < 0.05$) to control for multiple comparisons (only the pixels in frames 10-65 were included in statistical testing and multiple comparison correction). Visual features that contributed significantly to fusion were identified by overlaying the thresholded group CMs on the McGurk video. In order to chart the temporal dynamics of fusion, we created group classification time-courses by averaging across all pixels in each frame of the group CMs. We marked time-points that contributed significantly to fusion by identifying frames in which >25% of pixels survived the FDR-corrected threshold (considering only positive-valued z-scores).

Temporal dynamics of lip movements in McGurk stimuli

For comparison with the group classification time-courses, we measured and plotted the temporal dynamics of lip movements in the McGurk video following established methods (Chandrasekaran et al., 2009). The interlip distance, which tracks the time-varying amplitude of the mouth opening, was measured frame-by-frame by manually. For plotting, the resulting time course was smoothed using a Savitzky-Golay filter (order 3, window = 9 frames). The "velocity" of the lip opening was calculated by approximating the derivative of the interlip distance (Matlab 'diff'). The velocity time course was smoothed for plotting in the same way as the interilp distance time course. Two features related to production of the stop consonant (/k/)

were identified from the interlip distance and velocity curves. Consonants typically involve a rapid closing of the mouth before opening to produce the subsequent sound. To identify the temporal signature of this closing phase, we looked backward in time from the onset of the consonant burst to find the point at which the interlip distance just started to decrease. This was marked by a trough in the velocity curve, and corresponded to initiation of the closure movement. We then looked forward in time to find the next peak in the velocity curve, which marked the point at which the mouth was half-closed and beginning to decelerate. The time between this half-closure point and the onset of the consonant burst, known as 'time-to-voice,' was calculated to be 300ms for our McGurk stimulus. The interlip distance time course and velocity time course of the McGurk visual AKA are plotted together with the Lag0 audio signal (APA) in Figure 3.3. The various phases of consonant production are marked. The purple shaded region corresponds to production of the initial vowel. The yellow shaded region corresponds to consonant-related lip movements that take place prior to onset of the auditory signal. The green shaded region corresponds to the time period during which consonant-related auditory signal is present (consonant burst through onset of vowel steady state).

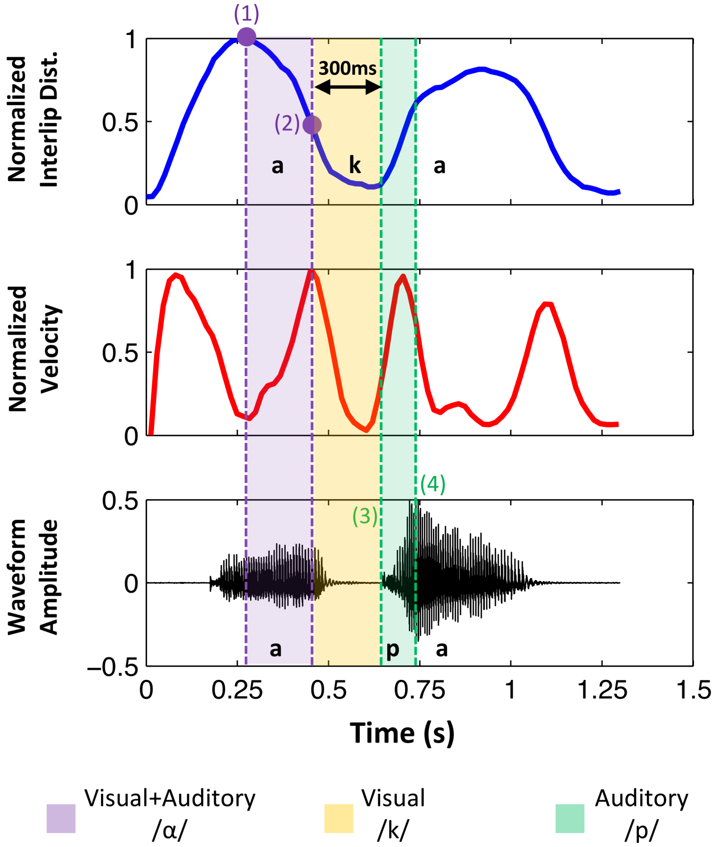


Figure 3.3. McGurk visual stimulus parameters. Pictured are normalized curves (max = 1) showing the visual interlip distance (blue, top) and lip velocity (red, middle) over time. These curves describe the evolution of the visual AKA signal in our McGurk stimuli. Also pictured is the auditory APA waveform used in the Lag0 (synchronized) McGurk stimulus. Several features are marked by numbers on the graphs: (1) corresponds to the onset of lip closure during the initial vowel production; (2) corresponds to the point at which the lips were half-way closed toward the offset of initial vowel production; (3) corresponds to onset of the consonant burst in the auditory waveform; (4) corresponds to onset of vowel steady state in the auditory waveform. The time between (2) and (3) is the so-called ‘time to voice.’ The purple shaded region corresponds to visual and auditory information that both specify the initial vowel /a/. The yellow shaded region corresponds to visual information that (presumably) specifies the consonant /k/. The green shaded region corresponds to auditory information that specifies the consonant /p/. There could also be visual information that specifies the consonant /k/ during the green-shaded period.

Results

Audio-only and clear-AV

Auditory APA stimuli were perceived as APA 90% (± 1% SEM) of the time on average, and the mean confidence rating was 1.78 (± 0.07 SEM). Auditory ATA stimuli were perceived as APA 9% (± 2% SEM) of the time on average, and the mean confidence rating was 5.22 (± 0.14 SEM). The APA audio used to create the McGurk stimuli was perceived as APA 89% (± 2% SEM) of the time on average, and the mean confidence rating was 1.82 (± 0.11 SEM). Overall, this indicates that some perceptual uncertainty was introduced for auditory stimuli at the +6dB SNR chosen for auditory presentation, but overall auditory-only perception was quite accurate.

For reporting the results of the clear-AV condition, we will focus on the McGurk stimuli (performance for congruent AV stimuli was at ceiling). Recall that in McGurk stimuli, an auditory APA was dubbed on a visual AKA. Responses that did not conform to the identity of the auditory signal were considered fusion responses. The Lag0 stimulus was perceived as APA 5% (± 3% SEM) of the time on average, with a mean confidence rating of 5.34 (± 0.16). The Lag50 stimulus was perceived as APA 6% (± 3% SEM) of the time on average, with a mean confidence rating of 5.33 (± 0.15). The Lag100 stimulus was perceived as APA 6% (± 3% SEM) of the time on average, with a mean confidence rating of 5.34 (± 0.17). Three conclusions are clear from these data. First, a very large proportion of responses (>90%) deviated from the identity of the auditory signal, indicating a high rate of fusion. Second, this rate of fusion did not differ across the McGurk stimuli (and nor did confidence ratings), suggesting that the McGurk stimuli were equally well bound despite the asynchrony manipulation. Third, McGurk stimuli were judged as Not-APA with roughly the same frequency and confidence as for auditory ATA stimuli, suggesting a high reliance on visual information (this was the intended effect of adding low-intensity white noise to the auditory signal).

Masked-AV

Congruent APA videos were perceived as APA 95% of the time on average, while congruent ATA videos were perceived as APA 4% of the time on average, indicating that perception of congruent videos was largely unaffected by the masker. The Lag0 McGurk stimulus was perceived as APA 40% (± 4% SEM) on average, with a mean confidence rating of 3.87 (± 0.80). The Lag50 McGurk stimulus was perceived as APA 37% (± 4% SEM) on average, with a mean confidence rating of 3.97 (± 0.71). The Lag100 McGurk stimulus was perceived as APA 33% (± 4% SEM) on average, with a mean confidence rating of 4.13 (± 0.65). Thus, we observed a net increase of APA responses equal to 35% for Lag0, 31% for Lag50, and 27% for Lag100, indicating a significant reduction of fusion responses due to the masker. This reduction was the basis for classification of the visual features that contribute to fusion.

Example frames from the FDR-corrected classification movie (CM) for the Lag0 stimulus are presented in Figure 3.4. Some comments are warranted. First, there are several frames in which significant negative-valued pixels can be identified (i.e., pixels that were reliably transparent when fusion was not observed). We have plotted the negative values for completeness, but we were not interested in the negative patterns and so will not address them further. Second, since the masker region was rather small (i.e., confined to the mouth), and because a high spatial correlation was induced in the maskers, it is difficult to make meaningful conclusions about the spatial patterns of significant pixels in the CMs. We were primarily interested in the temporal dynamics of fusion, so from this point forward we fill focus on the classification time-courses.

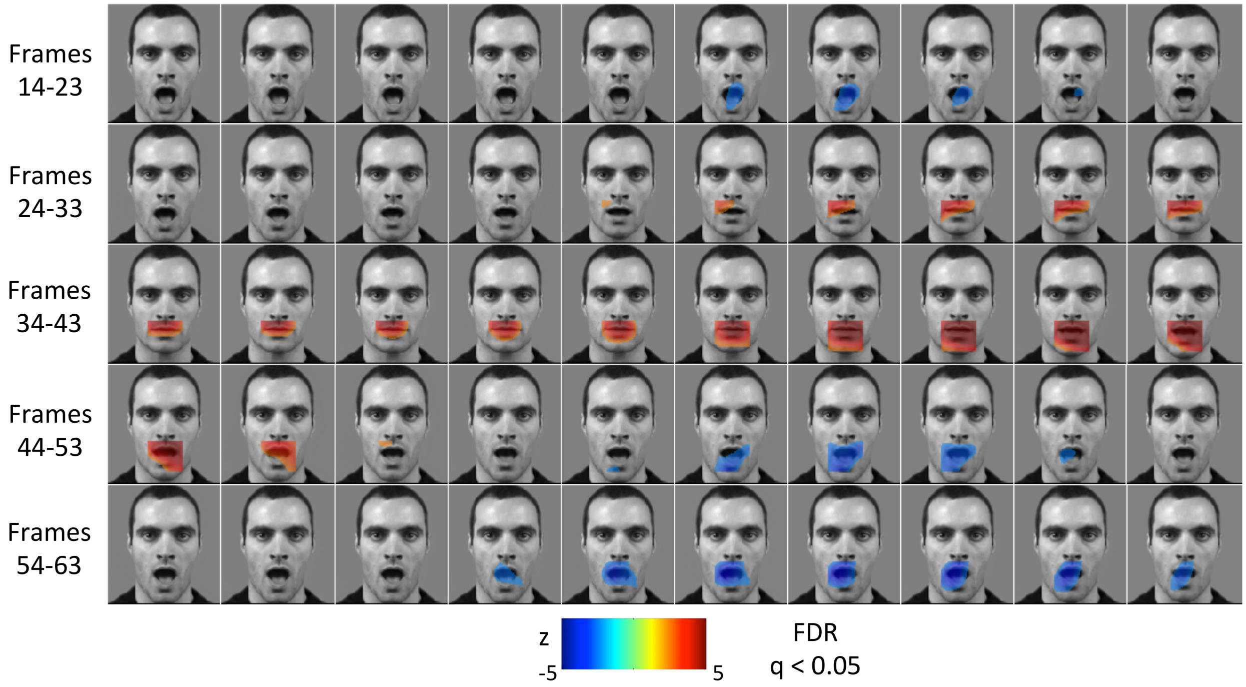


Figure 3.4. Results: Group classification movie for Lag0. Fifty example frames from the classification movie for the Lag0 McGurk stimulus are displayed. Warm colors mark pixels that contributed significantly to fusion. When these pixels were transparent, fusion was reliably observed. Cool colors mark pixels that showed the opposite effect. When these pixels were transparent, fusion was reliably blocked. Only pixels that survived multiple comparison correction at FDR q < 0.05 are assigned a color.

Classification time-courses for the Lag0, Lag50, and Lag100 stimuli are plotted in Figure 3.5 along with a trace of the auditory waveform for each stimulus. Recall that these time-courses were created by simply averaging the group classification z-scores at each frame of the CM. As such, large positive values correspond to frames that contributed most to fusion. Frames for which >25% of pixels were positive-valued and exceeded the FDR-corrected threshold are labeled with red circles. We will henceforth refer to these as 'significant frames.' In Figure 3.5, several results are immediately apparent: (1) each of the classification time-courses reaches its peak in exactly the same region; (2) the morphology of the Lag0 time-course differs dramatically from the Lag50 and Lag100 time-courses; (3) there are more significant frames in the Lag0 time-course than the Lag50 and Lag100 time-courses. Regarding (1), the exact location of the peak in each time-course was frame 42, and this pattern was rather stable across subjects. For the Lag0 stimulus, 11 of 17 subjects had their classification peak within ± 2 frames of the group peak and 14 of 17 subjects had a local maximum within ± frames of the group peak. For the Lag50 stimulus, these proportions were 12/17 and 15/17, respectively; and for the Lag100 stimulus, 13/17 and 16/17, respectively. Regarding (3), the range of significant frames for the Lag0 stimulus was 30 through 45 (266.7ms). The range of significant frames was 37 through 45 (150ms) and 39 through 46 (133.4ms) for the Lag50 and Lag100 stimuli, respectively.

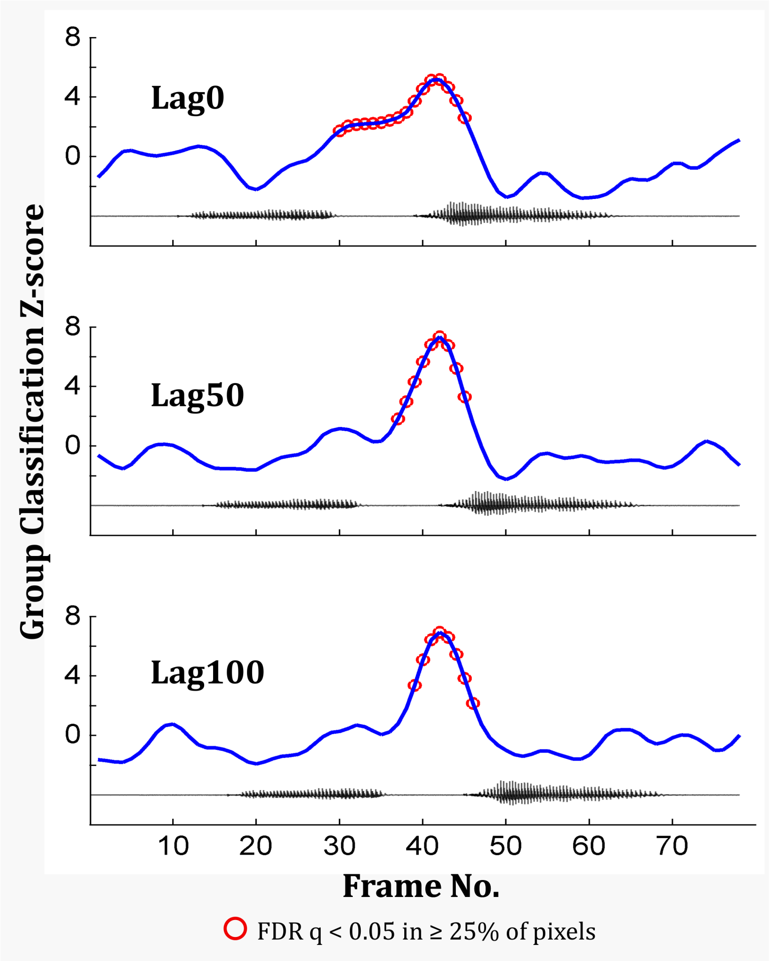


Figure 3.5. Results: Group classification time-courses for each McGurk stimulus. The group-mean classification z-scores are shown for each frame in the Lag0 (top), Lag50 (middle), and Lag100 (bottom) McGurk stimuli. Significant frames – i.e., those for which at least 25% of pixels in the group classification movie were positive-valued and exceeded the FDR threshold – are labeled by red circles. These frames contributed significantly to McGurk fusion. There are differences in both overall morphology and patterns of significant frames across the McGurk stimuli. In particular, earlier frames tend to play a larger role in fusion for the Lag0 stimulus. However, the peak is identical across each curve. The waveform of the auditory signal (black) for each stimulus is plotted beneath the classification time course (blue).

In Figure 3.6, we zoom on the portion of classification time-courses containing significant frames, and we plot these portions aligned to the lip velocity curve over the same time period. Phases of the lip movement related to consonant production are labeled on the velocity curve. The shaded regions from Figure 3.3 are reproduced, accounting for shifts in the audio for the Lag50 and Lag100 stimuli. Two features on this plot are immediately clear. First, the peak region on each classification time-course clearly corresponds to the region of the lip velocity curve describing acceleration of the lips to peak velocity approaching and following the release of airflow in visual /k/. For the Lag0 stimulus, this part of the curve falls squarely in the time period containing auditory cues related to the /p/ sound. In other words, the most salient visual cues for fusion strongly overlap the auditory signal when audiovisual timing is natural. Second, 10 significant frames in the Lag0 time-course fall in the time period (shaded yellow in Figure 3.6) corresponding to pre-release lip movements – i.e., lip movements that precede the onset of the consonant-related auditory signal when the audiovisual timing is natural. This number falls to just 3 significant frames for Lag50 and 1 significant frame for Lag100. This shift constitutes the major difference in classification of fusion-related cues across the McGurk stimuli. The fact that such a difference exists excludes the possibility that visual cues are integrated uniformly within the temporal window of integration.

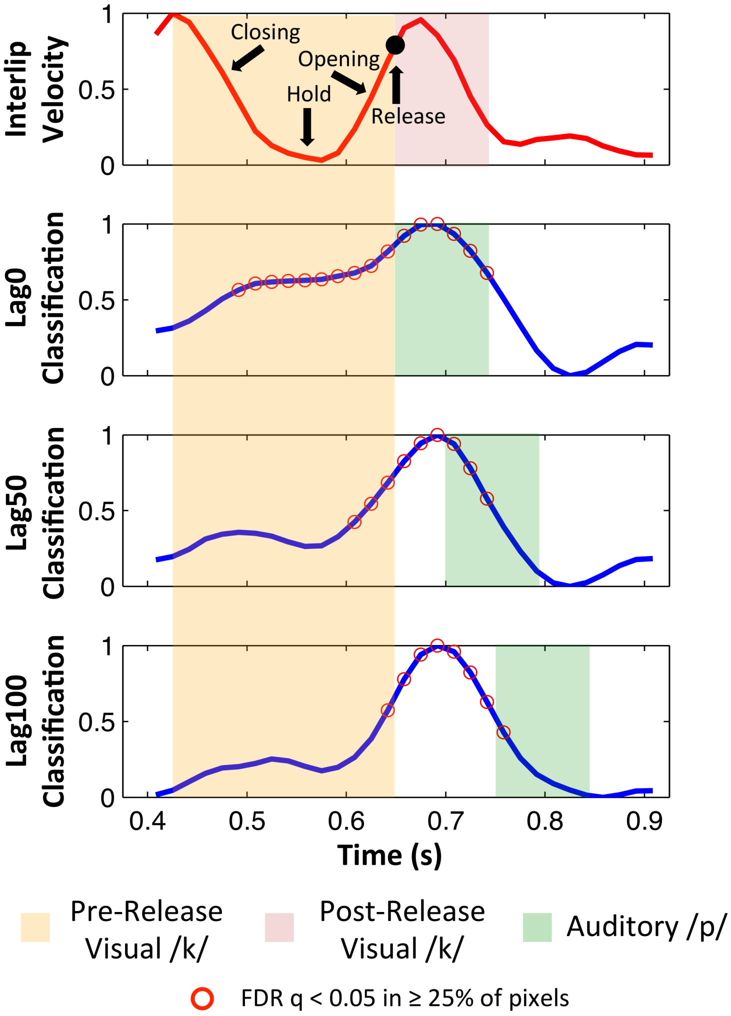


Figure 3.6. Classification time-courses for the Lag0, Lag50, and Lag100 McGurk stimuli (blue) are plotted along with the lip velocity function (red). The figure is "zoomed in" on the time period containing frames that contributed significantly to fusion (marked as red circles). Classification time-courses have been normalized (max = 1). The yellow-shaded period from Figure 3.3 is duplicated here. The onset of this period corresponds to lip closure following the initial vowel, and the offset corresponds to release of airflow at the consonant burst. We have labeled this the 'pre-release' visual /k/. Shaded in green is the period containing the auditory consonant /p/ from initial burst to onset of vowel steady state. The green shaded region is shifted appropriately for each McGurk stimulus (to account for auditory lags in Lag50 and Lag100). A region on the lip velocity curve is shaded pink. This region corresponds to 'post-release' visual /k/, as estimated from the classification time-courses. Different phases of articulation are labeled (black) on the lip velocity function.

Discussion

We have developed a novel technique for mapping the temporal dynamics of audiovisual fusion in speech. To implement this technique, we employed a phoneme identification task in which we overlaid McGurk stimuli with a spatiotemporally correlated visual masker that revealed critical visual cues on some trials but not on others. As a result, McGurk fusion was observed only on trials for which critical visual cues were available. Behavioral patterns in phoneme identification (fusion or no fusion) were reverse correlated with masker patterns over many trials, yielding a classification time-course of the visual cues that contributed significantly to fusion. We performed this classification for three McGurk stimuli with different temporal offsets between the auditory and visual signals – natural timing (Lag0), 50ms visual-lead (Lag50), and 100ms visual-lead (Lag100) – in order to test whether this temporal relationship altered the set of visual cues that contributed to fusion. Three significant findings sum up the results of the study. First, the Lag0, Lag50, and Lag100 McGurk stimuli were rated identically in a phoneme identification task with no visual masker. Specifically, each stimulus elicited a high degree of fusion, suggesting that all of the stimuli fell within the canonical temporal window of integration. Second, the primary visual cue contributing to fusion, identified using our masking technique, was identical across the McGurk stimuli (i.e., regardless of the temporal offset between the auditory and visual signals). Third, despite this fact, there were significant differences in the set of visual cues that contributed to fusion across the McGurk stimuli. Namely, early visual cues – that is, lip movements that precede the onset of the relevant auditory signal when audiovisual timing is natural – contributed significantly to fusion for the Lag0 stimulus, but not for the other McGurk stimuli. This finding is significant because it reveals details of the audiovisual fusion computation that are not available using traditional behavioral measurements. We discuss this and other findings in light of models of audiovisual integration below.

As mentioned in the Introduction, previous work on audiovisual integration in speech suggests that, so long as the auditory and visual speech signals are bound to a single perceptual stream, the fusion computation should operate rather uniformly. The theoretical background for this claim is analysis-by-synthesis (Halle, 2003; Skipper, van Wassenhove, Nusbaum, & Small, 2007; Virginie van Wassenhove et al., 2005). According to this approach, visual speech influences auditory speech by establishing a predictive context – namely, visual speech gestures, which tend to originate before the onset of the auditory speech signal, generate an abstract hypothesis for the identity of the incoming speech sound. This hypothesis serves as the context for subsequent auditory-phonemic analysis (incoming auditory signals are evaluated against the visual hypothesis). This type of model neatly explains why audiovisual integration of speech is disrupted when the auditory signal is made to artificially lead the visual signal (even at short asynchronies) – if visual cues have not been fully processed upon arrival and processing of the auditory signal, then there is no predictive context against which to evaluate the incoming auditory information. As such, there is no need to bind the auditory and visual signals, and visual speech information is not integrated. The opposite is true when the auditory signal is made to lag behind the visual signal. In this case, any relevant visual cues can be fully processed prior to processing of the auditory signal, leading to generation of a visual hypothesis that is maintained in memory for as long as possible, until finally it is disambiguated by the incoming auditory signal. The result is that visual speech information is integrated even at fairly long visual-lead asynchronies (Conrey & Pisoni, 2006; Grant & Greenberg, 2001; Grant et al., 2004; K. G. Munhall, Gribble, Sacco, & Ward, 1996; Virginie van Wassenhove et al., 2005).

In our study, a baseline measurement of the visual cues that contribute to fusion (i.e., cues that would lead to generation of a visual hypothesis under analysis-by-synthesis) is given by the classification time-course for the Lag0 McGurk stimulus (natural audiovisual timing). Inspection of this time course reveals that 16 video frames (30-45) contributed significantly to fusion. If analysis-by-synthesis is correct, this pattern should be largely unchanged for the Lag50 and Lag100 time-courses. Specifically, the Lag50 and Lag100 stimuli were constructed with relatively short visual-lead asynchronies (50ms and 100ms, respectively) that produced no behavioral differences in McGurk fusion. In other words, the 'visual hypothesis' for each stimulus remained the same in spite of the temporal synchrony manipulation. However, the set of visual cues that contributed to fusion for Lag50 and Lag100 was drastically different than the set for Lag0. In particular, all of the early significant frames dropped out – there were only 9 video frames (37-45) that contributed to fusion for Lag50, and only 8 video frames (39-46) contributed to fusion for Lag100. Overall, early video frames had progressively less influence on fusion as the auditory signal was lagged further in time. This provides evidence that the fusion computation is not uniform with respect to the temporal relationship between the auditory and visual speech signals, even when those signals are within the canonical temporal window of integration. It also suggests that analysis-by-synthesis must be adapted, at the very least, to account for the current data. In particular, there was a nonlinear dropout of significant frames moving from Lag0 to Lag50 with respect to the synchrony manipulation. In particular, a 50ms shift in the auditory signal, which should correspond to a three-frame shift with respect to the visual signal, caused *seven* early frames (116.7ms) to drop out from the classification moving from Lag0 to Lag50. This suggests that discrete visual events contributed to "hypotheses" of varying strength, such that a relatively low-strength hypothesis related to visual events in the early frames (those labeled ‘pre-release’ in Figure 3.6) was no longer influential when the auditory signal was lagged by 50ms.

Thus, we suggest an alternative explanation. In our account, dynamic (perhaps kinematic) visual features enter into the fusion computation. These features correspond to different phases of articulation but need not have any particular level of phonological specificity (Chandrasekaran et al., 2009; K. G. Munhall & Vatikiotis-Bateson, 2004; Q. Summerfield, 1987; H. C. Yehia, Kuratate, & Vatikiotis-Bateson, 2002; H. Yehia, Rubin, & Vatikiotis-Bateson, 1998). Several findings in the current study support the existence of such features. Immediately above, we described a nonlinear dropout with respect to the contribution of early visual frames in the Lag50 classification relative to Lag0. This suggests that a discrete visual feature (likely related to the catch+hold phase of articulation, i.e., cutting off airflow in the vocal tract) no longer contributed to fusion when the auditory signal was lagged by 50ms. A linear shift in the classification time-course would be expected otherwise. Further, the peak of the classification time-courses was identical across all McGurk stimuli, regardless of the temporal offset between the auditory and visual speech signals. We believe this peak corresponds to a visual feature related to the release of air in consonant production (see Figure 3.6, which compares the classification time-courses to the lip velocity profile). We suggest that visual features are weighted in the fusion computation according to three factors: (1) visual salience, (2) information content, and (3) temporal proximity to the auditory signal (closer = greater weight). To be precise, representations of visual features are activated with strength proportional to visual salience and information content (high for the ‘release’ feature here), and this activation decays over time such that visual features farther in time from the auditory signal are weighted less heavily (pre-release visual-articulatory features here). This allows the auditory system to “look back" in time for informative visual information. The 'release' feature in our McGurk stimuli remained influential even when it was temporally distanced from the auditory signal (e.g., Lag100) because of its high salience and because it was the only informative feature that remained activated upon arrival and processing of the auditory signal. Qualitative neurophysiological evidence (dynamic source reconstructions form MEG recordings) suggests that cortical activity loops between auditory cortex, visual motion cortex, and heteromodal superior temporal cortex when audiovisual convergence has not been reached, e.g. during lipreading (Arnal, Morillon, Kell, & Giraud, 2009). This may reflect maintenance of visual features in memory over time.

Other research supports the notion of a dynamic-feature-based fusion computation. It is evident from experiments with the McGurk effect that fusion depends on the time-varying dynamics of visual speech. In particular, the McGurk effect is observed in situations when detailed configuration cues (the exact postures of the face and mouth at any point in time) are obscured or unavailable. This holds when the speaker's mouth is at too great a distance from the observer to provide reliable configuration cues (Jordan & Sergeant, 2000), and when visual speech is conveyed as a point-light stimulus (Rosenblum & Saldaña, 1996). Moreover, the audiovisual advantage for speech intelligibility in noise is maintained when the visual speech signal is low-pass filtered in the spatial frequency domain (i.e., blurred) (K. Munhall, Kroos, Jozan, & Vatikiotis-Bateson, 2004). There is also specific evidence that visual and auditory speech information interact at a featural level (Green, 1998). For example, the presence of a visual bilabial stop (the /b/ in /ibi/) shifts perception of simultaneously-presented auditory tokens on an ambiguous /iri/-/ili/ continuum (Green & Norrix, 2001). Presumably this reflects the influence of visual features in the bilabial stop (e.g., rapid opening of the lips after closure), which specify temporal characteristics of the co-articulatory context (e.g., shortening of the vowel steady-state in /i/). These temporal characteristics are the basis for the observed perceptual shift.

Another aspect of our results deserves some further elaboration. As mentioned several times above, classification time-courses peaked over the same visual frames across all three McGurk stimuli. This peak region coincided with an acceleration of the lips immediately preceding and following the release of airflow during consonant production. Examination of the Lag0 stimulus (natural audiovisual timing) indicates that this visual-articulatory gesture unfolded over the same time period as the consonant-related portion of the auditory signal. As such, the most influential visual information in the stimulus temporally overlapped the auditory signal. This information was equally influential in the Lag50 and Lag100 stimuli when it preceded the onset of the auditory signal. This is interesting in light of the theoretical importance placed on visual speech cues that lead the onset of the auditory signal (Arnal et al., 2009; Virginie van Wassenhove et al., 2005). In our study, the most informative visual information was related to the actual release of airflow during articulation, rather than the preparatory catch and hold (closing the vocal tract to produce the stop), and this was true whether this information preceded or temporally overlapped the auditory signal. However, this may have been an artifact of the McGurk stimulus. In particular, the visual velar /k/ is less distinct during vocal tract closure and makes a relatively weak prediction of the consonant identity (relative to, say, a bilabial /p/) (Arnal et al., 2009; Q. Summerfield, 1987; Quentin Summerfield, 1992; Virginie van Wassenhove et al., 2005).

Finally, we should address several of the design choices in the current study. First, regarding our visual masking technique, we chose to mask only the part of the visual stimulus containing the mouth and part of the lower jaw. This choice obviously limits our conclusions to mouth-related visual features. This is a potential shortcoming since it is well known that other aspects of face and head movement are correlated with the acoustic speech signature (Jiang, Alwan, Keating, Auer, & Bernstein, 2002; Jiang, Auer, Alwan, Keating, & Bernstein, 2007; K. G. Munhall et al., 2004; H. C. Yehia et al., 2002; H. Yehia et al., 1998). However, restricting the masker to the mouth region reduced computing time and thus experiment duration since maskers were generated in real time. Moreover, we were quite confident that masking the mouth region alone would produce a considerable reduction in McGurk fusion and we confirmed this in pilot testing (which was also used to tune the parameters of the masker). Second, we added 62 dB SPL of noise to auditory speech signals (+6 dB SNR) throughout the experiment. As mentioned above, this was done to increase the likelihood of fusion by increasing perceptual reliance on the visual signal (Alais & Burr, 2004; Shams & Kim, 2010). We needed the highest possible fusion rates in order to conclude that any observed reduction in fusion was due primarily (if not entirely) to the presence of the visual masker. A potential criticism is that adding ambiguity to the auditory signal changed the nature of the fusion computation. Specifically, one could argue that the perceptual process shifted from fusion *per se* to visual capture. This is unlikely for the following reasons: auditory-only identification performance for the McGurk stimulus was at 90%, the results of our classifications depended on the position of the auditory signal, and the percept in our task was phenomenologically auditory. Third, we chose to collect responses on a 6-point confidence scale that emphasized identification of the nonword APA (i.e., the choices were between APA and Not-APA). The major drawback of this choice is that we do not know precisely what subjects perceived on fusion (Not-APA) trials. A 4-AFC calibration study shows that our McGurk stimulus was perceived overwhelmingly as ATA (92%) in a different group of subjects. Nevertheless, some ambiguity remains regarding the interpretation of behavior in the masked-AV condition of the current experiment. This was driven by necessity because we needed to use a task with a binary response variable in order to implement the classification analysis. A simple choice would have been to force subjects to choose between APA and ATA, but any subjects who perceived, for example, AKA on a significant number of trials would have been forced to arbitrarily assign this to APA or ATA. We chose to use an identification task with APA as the target so that any response involving some visual interference (AKA, ATA, AKTA, etc.) would be attributed to the Not-APA category. There is some debate regarding whether responses such as AKA or AKTA represent true fusion, but in these cases it is clear that some complementary visual information has influenced auditory perception. This is how we define fusion, so we were comfortable grouping these responses into a single category. A final issue concerns the generalizability of our results. In the present study, we have presented classification data based on a single voiceless McGurk token, spoken by just one individual. This was done to facilitate collection of the large number of trials needed for a reliable classification. As a result, generalizability may be low for some aspects of the data. However, at least one conclusion with broad theoretical implications can be made quite strongly from the current data – the particular visual speech information that is integrated during fusion depends on the temporal relationship between the visual and auditory speech signals. Moreover, we have provided here a method for classification of fusion-related visual features that can now be extended or modified in future research.

In conclusion, our visual masking technique successfully classified the mouth-related visual cues that contribute to audiovisual fusion in speech. We were able to chart the temporal dynamics of fusion at a high resolution – our classifications were based on visual stimuli that were recorded at a high frame rate (60 Hz). The results of this procedure revealed details of the fusion computation that were not available in typical behavioral measurements. In particular, a different set of visual cues influenced McGurk fusion depending on the temporal offset between the auditory and visual speech signals, even though the rate of perceived McGurk fusion was identical at each offset. We interpreted this result in terms a model of audiovisual fusion in which dynamic visual features are extracted and integrated proportional to their salience, informational content, and temporal proximity to the auditory speech signal. This model is potentially at odds with the influential analysis-by-synthesis account of audiovisual fusion.

References

Alais, David, & Burr, David. (2004). The ventriloquist effect results from near-optimal bimodal integration. *Current biology, 14*(3), 257-262.

Arnal, L. H., Morillon, B., Kell, C. A., & Giraud, A. L. (2009). Dual neural routing of visual facilitation in speech processing. *J Neurosci, 29*(43), 13445-13453. doi: 10.1523/JNEUROSCI.3194-09.2009

Berthommier, Frédéric. (2004). A phonetically neutral model of the low-level audio-visual interaction. *Speech Communication, 44*(1), 31-41.

Boothroyd, Arthur, Kishon-Rabin, Liat, & Waldstein, Robin. (1995). *Studies of tactile speechreading enhancement in deaf adults.* Paper presented at the Seminars in Hearing.

Campbell, R. (2008). The processing of audio-visual speech: empirical and neural bases. *Philos Trans R Soc Lond B Biol Sci, 363*(1493), 1001-1010. doi: 10.1098/rstb.2007.2155

Chandrasekaran, Chandramouli, Trubanova, Andrea, Stillittano, Sébastien, Caplier, Alice, & Ghazanfar, Asif A. (2009). The natural statistics of audiovisual speech. *PLoS computational biology, 5*(7), e1000436.

Conrey, Brianna, & Pisoni, David B. (2006). Auditory-visual speech perception and synchrony detection for speech and nonspeech signals. *The Journal of the Acoustical Society of America, 119*(6), 4065-4073.

Fowler, Carol A, & Dekle, Dawn J. (1991). Listening with eye and hand: Cross-modal contributions to speech perception. *Journal of Experimental Psychology: Human Perception and Performance, 17*(3), 816.

Grant, Ken W. (2001). The effect of speechreading on masked detection thresholds for filtered speech. *The Journal of the Acoustical Society of America, 109*(5), 2272-2275.

Grant, Ken W, & Greenberg, Steven. (2001). *Speech intelligibility derived from asynchronous processing of auditory-visual information.* Paper presented at the AVSP 2001-International Conference on Auditory-Visual Speech Processing.

Grant, Ken W, & Seitz, Philip-Franz. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. *The Journal of the Acoustical Society of America, 108*(3), 1197-1208.

Grant, Ken W, & Walden, Brian E. (1996). Evaluating the articulation index for auditory–visual consonant recognition. *The Journal of the Acoustical Society of America, 100*(4), 2415-2424.

Grant, Ken W, Wassenhove, Virginie van, & Poeppel, David. (2004). Detection of auditory (cross-spectral) and auditory–visual (cross-modal) synchrony. *Speech Communication, 44*(1), 43-53.

Green, Kerry P. (1998). The use of auditory and visual information during phonetic processing: Implications for theories of speech perception. *Hearing by eye II*, 3-26.

Green, Kerry P, & Norrix, Linda W. (2001). Perception of/r/and/l/in a stop cluster: Evidence of cross-modal context effects. *Journal of Experimental Psychology: Human Perception and Performance, 27*(1), 166.

Greenberg, Steven. (1999). Speaking in shorthand–A syllable-centric perspective for understanding pronunciation variation. *Speech Communication, 29*(2), 159-176.

Halle, Morris. (2003). *From memory to speech and back: Papers on phonetics and phonology 1954-2002* (Vol. 3): Walter de Gruyter.

Jiang, Jintao, Alwan, Abeer, Keating, Patricia A, Auer, Edward T, & Bernstein, Lynne E. (2002). On the relationship between face movements, tongue movements, and speech acoustics. *EURASIP Journal on Applied Signal Processing, 11*, 1174-1188.

Jiang, Jintao, Auer, Edward T, Alwan, Abeer, Keating, Patricia A, & Bernstein, Lynne E. (2007). Similarity structure in visual speech perception and optical phonetic signals. *Perception & psychophysics, 69*(7), 1070-1083.

Jordan, Timothy R, & Sergeant, Paul. (2000). Effects of distance on visual and audiovisual speech recognition. *Language and Speech, 43*(1), 107-124.

King, AJ, & Palmer, AR. (1985). Integration of visual and auditory information in bimodal neurones in the guinea-pig superior colliculus. *Experimental Brain Research, 60*(3), 492-500.

Kleiner, Mario, Brainard, David, Pelli, Denis, Ingling, Allen, Murray, Richard, & Broussard, Christopher. (2007). What's new in Psychtoolbox-3. *Perception, 36*(14), 1.1-16.

Magnotti, John F, Ma, Wei Ji, & Beauchamp, Michael S. (2013). Causal inference of asynchronous audiovisual speech. *Frontiers in psychology, 4*.

Massaro, D.W. (1987). *Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry*: Erlbaum Associates.

McGurk, Harry, & MacDonald, John. (1976). Hearing lips and seeing voices.

Meredith, M Alex, Nemitz, James W, & Stein, Barry E. (1987). Determinants of multisensory integration in superior colliculus neurons. I. Temporal factors. *The Journal of Neuroscience, 7*(10), 3215-3229.

Munhall, Kevin G, Gribble, P, Sacco, L, & Ward, M. (1996). Temporal constraints on the McGurk effect. *Perception & Psychophysics, 58*(3), 351-362.

Munhall, Kevin G, Jones, Jeffery A, Callan, Daniel E, Kuratate, Takaaki, & Vatikiotis-Bateson, Eric. (2004). Visual prosody and speech intelligibility head movement improves auditory speech perception. *Psychological science, 15*(2), 133-137.

Munhall, Kevin G, & Vatikiotis-Bateson, ERIC. (2004). Spatial and temporal constraints on audiovisual speech perception. *The handbook of multisensory processes*, 177-188.

Munhall, KG, Kroos, C, Jozan, G, & Vatikiotis-Bateson, E. (2004). Spatial frequency requirements for audiovisual speech perception. *Perception & Psychophysics, 66*(4), 574-583.

Nahorna, Olha, Berthommier, Frédéric, & Schwartz, Jean-Luc. (2012). Binding and unbinding the auditory and visual streams in the McGurk effect. *The Journal of the Acoustical Society of America, 132*(2), 1061-1077.

Poeppel, D. (2003). The analysis of speech in different temporal integration windows: cerebral lateralization as 'asymmetric sampling in time'. *Speech Communication, 41*(1), 245-255.

Rosenblum, Lawrence D, & Saldaña, Helena M. (1996). An audiovisual test of kinematic primitives for visual speech perception. *Journal of Experimental Psychology: Human Perception and Performance, 22*(2), 318.

Schwartz, Jean-Luc, & Savariaux, Christophe. (2014). No, There Is No 150 ms Lead of Visual Speech on Auditory Speech, but a Range of Audiovisual Asynchronies Varying from Small Audio Lead to Large Audio Lag. *PLoS Comput Biol, 10*(7), e1003743. doi: citeulike-article-id:13320829

doi: 10.1371/journal.pcbi.1003743

Shams, Ladan, & Kim, Robyn. (2010). Crossmodal influences on visual perception. *Physics of life reviews, 7*(3), 269-284.

Skipper, J. I., van Wassenhove, V., Nusbaum, H. C., & Small, S. L. (2007). Hearing lips and seeing voices: how cortical areas supporting speech production mediate audiovisual speech perception. *Cereb Cortex, 17*(10), 2387-2399. doi: 10.1093/cercor/bhl147

Stein, Barry E, Meredith, M Alex, & Wallace, Mark T. (1993). The visually responsive neuron and beyond: multisensory integration in cat and monkey. *Progress in brain research, 95*, 79-90.

Stevenson, Ryan A, Altieri, Nicholas A, Kim, Sunah, Pisoni, David B, & James, Thomas W. (2010). Neural processing of asynchronous audiovisual speech perception. *Neuroimage, 49*(4), 3308-3318.

Summerfield, Q. (1987). Some preliminaries to a comprehensive account of audio-visual speech perception. In Dodd (Ed.), *Hearing by eye: The psychology of lip-reading*: Lawrence Erlbaum Associates.

Summerfield, Quentin. (1992). Lipreading and audio-visual speech perception. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 335*(1273), 71-78.

Thurman, Steven M, Giese, Martin A, & Grossman, Emily D. (2010). Perceptual and computational analysis of critical features for biological motion. *Journal of Vision, 10*(12), 15.

Thurman, Steven M, & Grossman, Emily D. (2011). Diagnostic spatial frequencies and human efficiency for discriminating actions. *Attention, Perception, & Psychophysics, 73*(2), 572-580.

van Wassenhove, V., Grant, K. W., & Poeppel, D. (2007). Temporal window of integration in auditory-visual speech perception. *Neuropsychologia, 45*(3), 598-607. doi: 10.1016/j.neuropsychologia.2006.01.001

van Wassenhove, Virginie, Grant, Ken W, & Poeppel, David. (2005). Visual speech speeds up the neural processing of auditory speech. *Proceedings of the National Academy of Sciences of the United States of America, 102*(4), 1181-1186.

Yehia, Hani C, Kuratate, Takaaki, & Vatikiotis-Bateson, Eric. (2002). Linking facial animation, head motion and speech acoustics. *Journal of Phonetics, 30*(3), 555-568.

Yehia, Hani, Rubin, Philip, & Vatikiotis-Bateson, Eric. (1998). Quantitative association of vocal-tract and facial behavior. *Speech Communication, 26*(1), 23-43.

CHAPTER 4

Primer

The results of Ch. 3 suggested that discrete visual features are extracted from the visual speech signal and integrated with auditory speech at some (unspecified) level of representation. This question – i.e., ‘At what processing stage are visual and auditory speech integrated?’ – has plagued speech scientists for some time (Bernstein, 2005; K. P. Green, 1998; Massaro, 1987; Rosenblum, Pisoni, & Remez, 2005; Schwartz, Robert-Ribes, & Escudier, 1998; Summerfield, 1987). Recent neurophysiological evidence suggests that many cognitive processes relevant to audiovisual speech perception – including audiovisual integration, biological motion processing, and phonological speech processing – converge in one particular brain area: the superior temporal sulcus (STS; see below for details). However, it has heretofore been unclear whether these computations converge to a single region of the STS or, instead, spread over multiple STS subregions. The answer to this question may shed light on the issue of processing stages in audiovisual speech perception. In the current chapter, I present the results of an fMRI study designed to examine: (a) the relation between unimodal (i.e., visual feature extraction) and bimodal (i.e., audiovisual integration) regions of the STS, and (b) the relation between speech-specific (perhaps phonological) and general sensory processing regions of the STS. The results allow an update to current neuroanatomically-informed models of audiovisual speech perception.

The visual speech stream: Facial motion processing, audiovisual integration, and speech perception in the superior temporal sulcus

Jonathan H. Venezia, Feng Rong, Dale Maddox, Kourosh Saberi, and Gregory Hickok

Introduction

The superior temporal sulcus (STS) has been implicated as a crucial processing center in a wide variety of human perceptual abilities. Among these are audiovisual integration (Amedi, Kriegstein, Atteveldt, Beauchamp, & Naumer, 2005; Michael S Beauchamp, Lee, Argall, & Martin, 2004), auditory speech perception (J. R. Binder, Swanson, Hammeke, & Sabsevitz, 2008; J. Binder et al., 2000; G. Hickok & Poeppel, 2004; Gregory Hickok & Poeppel, 2007; Price, 2010), and biological motion perception (Allison, Puce, & McCarthy, 2000; Michael S Beauchamp, Lee, Haxby, & Martin, 2003; E. D. Grossman, Battelli, & Pascual-Leone, 2005; E. D. Grossman & Blake, 2002; E. Grossman et al., 2000; Puce & Perrett, 2003). Where speech perception is concerned, it is widely established that visual speech information influences auditory speech perception (Callan et al., 2003; Dodd, 1977; McGurk & MacDonald, 1976; Reisberg, Mclean, & Goldfield, 1987; Sumby & Pollack, 1954), and thus it seems likely that biological motion information – specifically, information about the configuration of the various articulators that specify vocal tract shape – should interact with auditory speech representations in the STS or elsewhere, as others have suggested (Callan et al., 2003). In fact, visual speech strongly engages the STS (Callan et al., 2003; Campbell et al., 2001; Okada & Hickok, 2009).

Such an interaction implies that some form of integration of auditory and visual speech information must take place. The STS is well positioned for this task as it lies between visual association cortex in the posterior lateral temporal region (Michael S Beauchamp, Lee, Haxby, & Martin, 2002) and auditory association cortex in the superior temporal gyrus (Kaas & Hackett, 2000; Rauschecker, Tian, & Hauser, 1995; Wessinger et al., 2001). In nonhuman primates it has been demonstrated that polysensory fields in STS receive convergent input from unimodal auditory and visual cortical regions (Lewis & Van Essen, 2000; Seltzer & Pandya, 1978, 1994) and that these fields contain auditory, visual and bimodal neurons (Benevento, Fallon, Davis, & Rezak, 1977; Bruce, Desimone, & Gross, 1981; Dahl, Logothetis, & Kayser, 2009). Indeed, a host of studies have identified the STS as a multisensory convergence zone for speech (M. S. Beauchamp, Argall, Bodurka, Duyn, & Martin, 2004; Michael S Beauchamp, Nath, & Pasalar, 2010; Calvert, Campbell, & Brammer, 2000; Nath & Beauchamp, 2011, 2012; Stevenson, Altieri, Kim, Pisoni, & James, 2010; Stevenson & James, 2009; Stevenson, VanDerKlok, Pisoni, & James, 2011; Szycik, Tausche, & Münte, 2008; Wright, Pelphrey, Allison, McKeown, & McCarthy, 2003).

Two classes of models have been proposed for audiovisual integration in the STS (Figure 4.1). In what we call "direct integration" models, auditory speech representations housed in auditory association cortex (e.g., the superior temporal gyrus) and visual speech representations housed in visual association cortex (e.g., motion-sensitive/lateral occipital extrastriate visual cortex) converge on bimodal speech representations in multisensory STS; these bimodal representations are intrinsically-abstract, high-level representations of individual speech sounds (Michael S Beauchamp et al., 2010). Thus, according to direct integration models, auditory and visual speech signals are integrated at the phonological level and the output of this integration process is essentially a categorized speech sound.

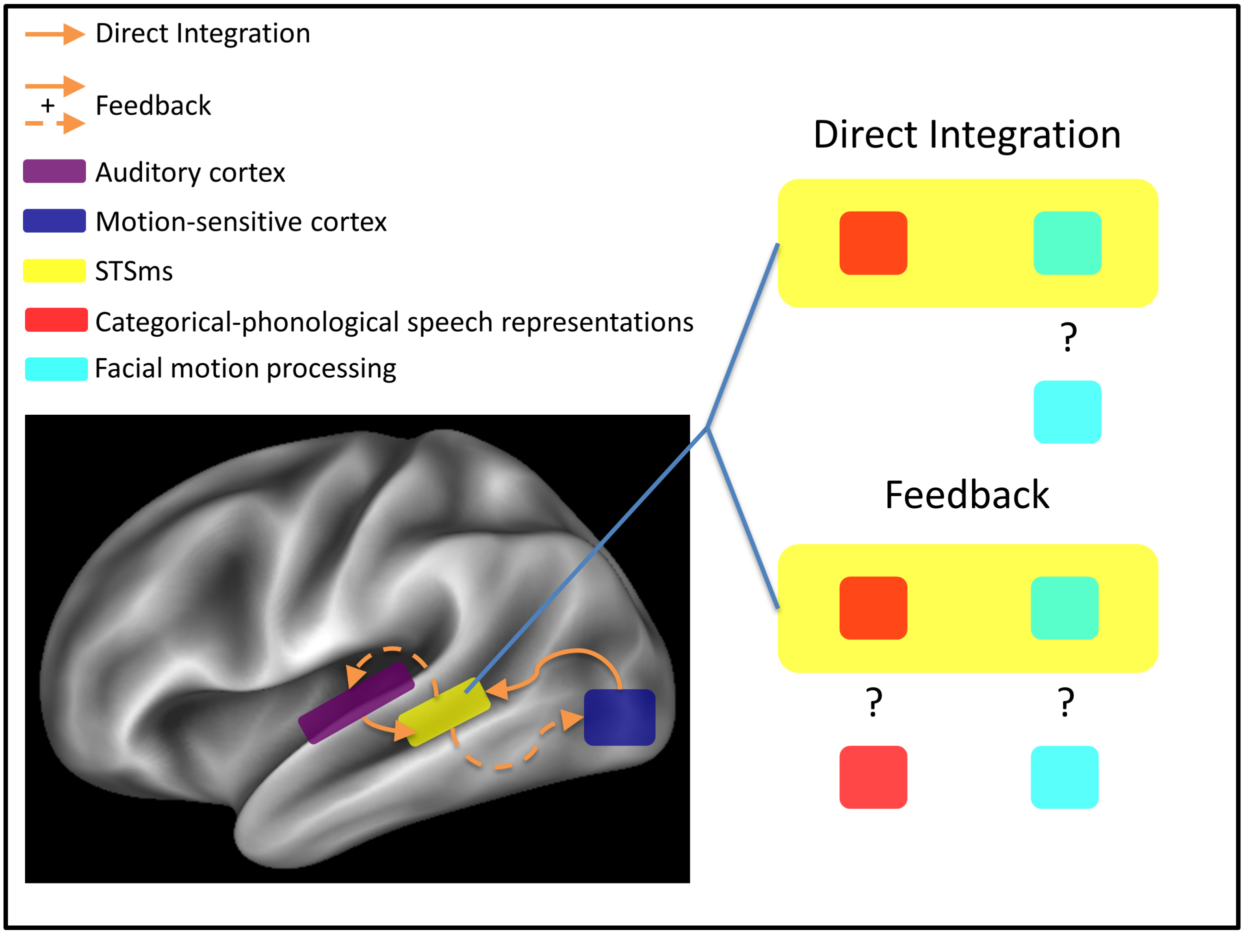


Figure 4.1. Schematic of current models of multisensory speech processing. Brain diagram at left: approximate locations of processing stages in multisensory speech processing. Arrows show directional connectivity. A color-coded legend describing the labeled regions is located top left. In Direct Integration models (solid arrows) the auditory speech signal (processed first in auditory cortex/superior temporal gyrus) is combined with the visual speech signal (processed first in the lateral occipital/posterior medial temporal lobe) in the multisensory region of the STS. In Feedback models (solid and dashed arrows), auditory and visual speech signals are compared in STSms and fed back to unimodal cortices. Schematic at right: expansion of STSms (yellow) to show different predictions made by Direct Integration and Feedback models. In Direct Integration models, auditory and visual speech signals converge on abstract, bimodal speech sound representations at the phonological level (red square inside STSms). Feedback models do not specify whether auditory and visual speech signals converge on phonological representations in STSms or elsewhere (red square both inside and outside STSms). Neither Direct Integration nor Feedback models characterize the relation between STSms and facial motion processing regions of the STS (teal square both inside and outside STSms).

In what we term “feedback” models, the STS serves as a multimodal intermediary that processes or combines cross-modal speech signals and provides feedback to unimodal cortices (Calvert et al., 1999; Calvert et al., 1997; Calvert et al., 2000; Driver & Spence, 2000; Ghazanfar, Maier, Hoffman, & Logothetis, 2005; Okada, Venezia, Matchin, Saberi, & Hickok, 2013; van Wassenhove, Grant, & Poeppel, 2005). In general, these feedback models are rather nonspecific regarding the level at which auditory and visual speech signals are processed or combined in the STS, and also with respect to the cortical location at which signals ultimately converge onto high-level speech sound representations (i.e., does this process happen directly in the multisensory STS, in the auditory cortex following feedback from the STS, or elsewhere?).

Both direct integration and feedback models are largely silent with respect to the role of biological-motion-processing regions of the STS in multimodal speech perception. By and large, activations to facial motion in particular – including natural facial motion (Puce, Allison, Bentin, Gore, & McCarthy, 1998), movements of facial line drawings (Puce et al., 2003), and point-light facial motion (Bernstein, Jiang, Pantazis, Lu, & Joshi, 2011) – yield activation quite posteriorly in the STS, a location potentially distinct from auditory and visual speech-related activations (Hein & Knight, 2008; Okada & Hickok, 2009). However, these posterior STS regions may be crucial for extracting high-level properties of biological movements (e.g., action class or action goal) that are invariant with respect to particular motion kinematics, image size, or viewpoint (E. D. Grossman, Jardine, & Pyles, 2010; Lestou, Pollick, & Kourtzi, 2008). This computation, applied specifically to facial motion, is likely to contribute to the formation of visual speech representations with any level of phonological specificity, yet the relation between biological motion systems and audiovisual speech integration systems has not been fully elucidated.

In short, current models of multimodal speech processing in the STS, taken as a whole, are equivocal with respect to the following questions: (1) Are biological/facial motion processing regions of the STS involved in visual speech processing and do these regions overlap with multisensory STS? And (2) at what representational level are auditory and visual speech signals processed or combined in multisensory STS?

Here, we present an fMRI experiment designed to answer at least the first and possibly the second question. Participants were presented with blocks of auditory and visual speech (CV syllables) and nonspeech (spectrally rotated speech, facial gurning) stimuli. To identify multisensory speech regions (STSms) we performed a conjunction of activation to auditory and visual speech. To identify facial motion processing regions (STSfm) we performed the conjunction of activation to visual speech and nonspeech facial gestures. Critically, we included as a baseline condition blocks with a *stationary* face to account for activation related to face processing generally. We also identified regions of the STS that were preferentially involved in processing speech versus nonspeech by contrasting activations to visual speech with activations to nonspeech facial gestures, and activations to auditory speech with activations to spectrally rotated (unintelligible) speech. We performed multivariate pattern analysis (MVPA) based on regions of interest defined in individual subjects in order to test the *representational* properties of different subregions of the STS (Mur, Bandettini, & Kriegeskorte, 2009; Okada et al., 2010). The broad motivation, in light of the previously stated questions, was to test whether and to what extent multisensory speech processing and facial motion processing draw upon the same neural resources in the STS, and whether these neural resources preferentially activate to (or distinguish in their representational profile) auditory and visual speech compared to nonspeech.

Materials and Methods

Participants

Twenty (three female) right-handed native English speakers between 18 and 30 years of age participated in the study. All volunteers had normal or corrected-to-normal vision, normal hearing by self-report, no known history of neurological disease, and no other contraindications for MRI. Informed consent was obtained from each participant in accordance with UC Irvine Institutional Review Board guidelines. Five subjects were excluded from MRI analysis leaving N = 15 for the imaging analysis (see below).

Stimuli and Procedure

Six two-second video clips were recorded in each of five experimental conditions featuring a single male actor shown from the neck up (Figure 4.2). In three speech conditions – auditory speech (A), visual speech (V), and audiovisual speech (AV) – the stimuli were six visually distinguishable consonant-vowel (CV) syllables (\ba\, \tha\, \va\, \bi\, \thi\, \vi\). In the A condition, clips consisted of a still frame of the actor's face paired with auditory recordings of the syllables (44.1 kHz, 16-bit resolution). In the V condition, videos of the actor producing the syllables were presented without sound (30 frames/s). In the AV condition, videos of the actor producing the syllables were presented simultaneously with congruent auditory recordings. There were also two non-speech conditions – spectrally rotated speech I and nonspeech facial gurning (G). In the R condition, spectrally inverted (Blesser, 1972) versions of the syllable recordings were presented along with a still frame of the actor. Auditory speech stimuli were first bandpass filtered (100-3900Hz) and then spectrally inverted about the center frequency (2000Hz). Rotation of the signal preserves the spectrotemporal complexity of speech and rotated speech is acoustically similar to clear speech, but the rotation process renders sentence-length utterances unintelligible (Narain et al., 2003; Okada et al., 2010; Scott, Blank, Rosen, & Wise, 2000) and significantly reduces discrimination accuracy for individual speech sounds separated by a category boundary (E. Liebenthal, Binder, Spitzer, Possing, & Medler, 2005). In the G condition, the actor produced the following series of nonspeech, lower-face gestures (without sound): partial opening of the mouth with leftward deviation, opening of mouth with rightward deviation, opening of mouth with lip protrusion, tongue protrusion, lower lip biting, and lip retraction. These gestures contain movements of a similar extent and duration as those used to produce the syllables in the speech conditions, but cannot be construed as speech (Campbell et al., 2001). A rest condition consisted of a still frame of the actor with no sound. Auditory speech stimuli were bandpass filtered to match the bandwidth of the rotated speech stimuli and all auditory stimuli were normalized to equal root-mean-square amplitude.

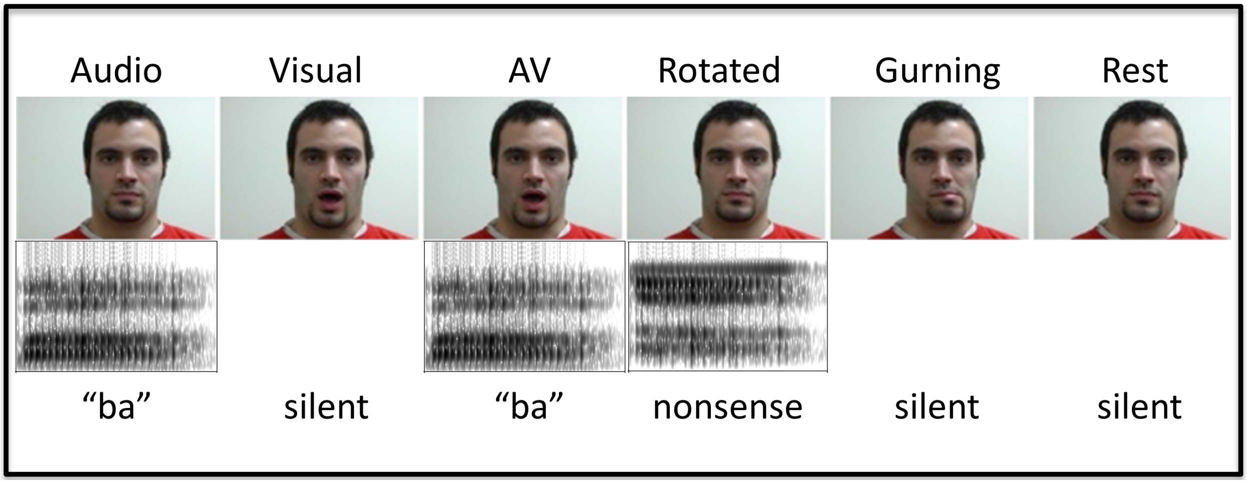


Figure 4.2. Example stimuli from each experimental condition.

Participants were presented with 12-second blocks in each of the experimental conditions. Each block contained all six of the clips composing that condition (i.e., all six CVs, all six rotated CVs, or all six gurning gestures). The clips were concatenated in random permutations of stimulus order to form 35 distinct blocks in each condition. Five additional "oddball" blocks were constructed for each condition (including rest), consisting of five within-condition clips and a single oddball clip from one of the other conditions (e.g., an oddball block might contain five A clips and a single AV clip). Oddball clips were randomly placed in the second through sixth positions in a block. An oddball could deviate either visually (e.g., a V clip in a G block), acoustically (e.g., an A clip in an R block), or both (e.g., an AV clip in an R block or an A clip in a V block). Each of these types of deviation occurred with equal frequency so that participants would attend equally to auditory and visual components of the stimuli (see task below).

Functional imaging runs consisted of pseudo-random presentation of 21 blocks, three from each condition along with three rest blocks and three oddball blocks. Blocks were by a 500ms period during which a black fixation cross was presented against a grey background. Participants were instructed to press a button each time an oddball was detected, and oddball blocks were modeled as a regressor of no interest and excluded from further analysis. The experiment started with a short practice session inside the scanner during which participants were exposed to a single block from each condition including a rest block and an oddball block. Participants were then scanned for ten functional runs immediately followed by acquisition of a high-resolution T1 anatomical volume. Auditory stimuli were presented through an MR compatible headset (ResTech) and stimulus delivery and timing were controlled using Cogent software (http://www.vislab.ucl.ac.uk/cogent_2000.php) implemented in Matlab 6 (Mathworks

Inc., USA).

Scanning Parameters

MR images were obtained on a Philips Achieva 3T (Philips Medical Systems, Andover, MA) fitted with an 8-channel SENSE receiver/head coil, at the Research Imaging Center facility at the University of California, Irvine. We collected a total of 1090 echo planar imaging (EPI) volumes over 10 runs using single pulse Gradient Echo EPI (matrix = 112 x 110, repetition time [TR] = 2.5 s, echo time [TE] = 25 ms, size = 1.957 x 1.957 x 1.5 mm, flip angle = 90). Forty-Four axial slices provided whole brain coverage. Slices were acquired sequentially with a 0.5mm gap. After the functional scans, a high-resolution anatomical image was acquired with a magnetization prepared rapid acquisition gradient echo [Mprage] pulse sequence in the axial plane (matrix = 240 x 240, TR = 11 ms, TE = 3.54 ms, size = 1 x 1 x 1 mm).

Behavioral Data Analysis

The Signal Detection Theory measure of sensitivity, *d'*, was calculated to determine performance on the oddball detection task (D. M. Green & Swets, 1966). A hit was defined as a positive response (button press) to an oddball block, while a false alarm was defined as a positive response to a non-oddball block. The hit rate (H) was calculated as the number of hits divided by the total number of oddball blocks, while the false alarm rate (F) was calculated as the number of false alarms divided by the number of non-oddball trials, and *d'* was calculated as:

$$d' = \Phi^{-1}[H] - \Phi^{-1}[F],$$

where Φ is the standard normal cumulative distribution function. Participants with a *d'* more than 1.5 standard deviation beneath the mean were excluded from further analysis (N = 2; see Results). We also calculated H separately for each experimental condition (based on the type of clips that composed the standard condition in oddball blocks). Hit rates by condition were entered in a repeated measures ANOVA with Greenhouse/Geisser correction.

fMRI Analysis

In addition to the two participants excluded for poor behavioral performance, three participants were excluded from the MRI analysis for poor raw image quality (visible artifacts). Thus, N = 15 for the MRI analysis. Preprocessing of the data was performed using AFNI software (http://afni.nimh.nih.gov/afni). For each run, slice timing correction was performed followed by realignment (motion correction) and coregistration of the EPI images to the high resolution anatomical image in a single interpolation step. Images were spatially smoothed with an isotropic 6-mm full-width half-maximum (FWHM) Gaussian kernel. Each run was then scaled to have a mean of 100 across time at each voxel.

First level regression analysis (AFNI 3dDeconvolve) was performed in individual subjects. To create the regressors of interest, a stimulus timing vector was created for each experimental condition was convolved with a model hemodynamic response function. Five such regressors were used in estimation of the model corresponding to the five experimental conditions: A, V, AV, R, G. The ‘still face’ rest condition was not modeled explicitly and was thus included in the baseline term. An additional twelve regressors corresponding to motion parameters determined during the realignment stage of preprocessing along with their temporal derivatives were entered into the model. Oddball blocks were modeled as a single regressor of no interest. Individual time points were censored from analysis when more than 10% of in-brain voxels were identified as outliers (AFNI 3dToutcount) or when the Euclidean norm of the motion derivatives exceeded 0.4.

Group Analysis

For the group analysis only (as opposed to the ROI analysis, below), functional images were registered to a common space prior to smoothing, scaling, and first-level regression. A study-specific anatomical template image was created using symmetric diffeomorphic registration (SyN) in the Advanced Normalization Tools (ANTS) software (B. Avants et al., 2008; B. Avants & Gee, 2004). Each participant’s T1 anatomical image was skull stripped in AFNI and submitted to the template-construction processing stream in ANTS (buildtemplateparallel.sh), which comprises rigid and SyN registration steps. For SyN, we used a cross correlation similarity metric (B. B. Avants et al., 2011) with a three-level multi-resolution registration with 50x70x10 iterations. A low-resolution version of the study-specific anatomical template was created to match the resolution of native-space functional images. The set of affine and diffeomorphic transformations mapping each participant’s T1 anatomical to the study-specific template were applied to the corresponding coregistered functional images using the low-resolution template as a reference image. The resulting functional images were aligned in the common study-specific template space.

A second-level analysis of variance was performed on the parameter estimates from each participant, treating ‘participant’ as a random effect. Statistical parametric maps (t-statistics) were created for each individual condition and all contrasts of interest. Active voxels were defined as those for which t-statistics exceeded the $p < 0.005$ level with a cluster extent threshold of 159 voxels. This cluster threshold was determined by Monte Carlo simulation (AFNI 3dClustSim) to hold the family-wise error rate (FWER) less than 0.05 (i.e., corrected for multiple comparisons). Estimates of smoothness in the data were drawn from the residual error time series for each participant after first-level analysis (AFNI 3dFWHMx). These estimates were averaged across participants separately in each voxel dimension for input to 3dClustSim. Simulations were restricted to in-brain voxels.

To identify STSms at the group level, we performed the conjunction $A \cap V$, and to identify STSfm at the group level, we performed the conjunction $V \cap G$. Conjunctions were performed by constructing minimum *t*-maps (e.g., minimum T score from [A,V] at each voxel) and these maps were thresholded at $p < 0.005$ with a cluster extent threshold of 159 voxels (FWER < 0.05, as for individual condition maps). This tests the ‘conjunction null’ hypothesis (Nichols, Brett, Andersson, Wager, & Poline, 2005). We also performed contrasts for activations greater for speech than nonspeech, matched for input modality: A>R and V>G. Finally, we tested for activations greater for bimodal speech processing than unimodal speech processing: AV>A and AV>V.

For convenience in reporting and displaying group results, the study-specific anatomical template was aligned to the MNI-space ICBM template (Vladimir Fonov et al., 2011; VS Fonov, Evans, McKinstry, Almli, & Collins, 2009) using a 12-parameter affine registration in ANTS. This affine transformation was applied to second-level estimates of mean PSC. Thus, group plots reflect thresholded PSC maps in MNI space. Plots are visualized on the Conte69 atlas in MNI152 space in CARET v5.65 (http://brainvis.wustl.edu/wiki/index.php/Caret:Download).

ROI Selection and Analysis

We also aimed to define STSms and STSfm as regions of interest in individual subjects (native space). To identify STSms, we performed the conjunction (minimum t-stat method, as above) A∩V thresholded at p < 0.005 (uncorrected); these will henceforth be referred to as the A∩V ROIs. To identify STSfm, we performed the conjunction V∩G thresholded at p < 0.005, (uncorrected); these will henceforth be referred to as the V∩G ROIs. To localize conjunctions to STS, a local peak t-score was identified on the minimum t-maps. A 10mm-radius sphere was formed around the peak and only voxels significant in the conjunction and located within the sphere were included in the ROI. The logic for this selection procedure is that including voxels from a restricted region around the conjunction peak is likely to identify voxels that were maximally active to the two conditions entered in the conjunction. In addition, it is well established that there is large degree of intersubject variability in the position of functionally-localized STSms (Michael S Beauchamp, 2005a, 2005b), making it advantageous to identify this region precisely in individual subjects (Arnal, Morillon, Kell, & Giraud, 2009; Stevenson et al., 2010; Stevenson & James, 2009; Stevenson et al., 2011).

Significant voxels were largely restricted to the STS, sometimes including voxels in nearby superior temporal or middle temporal gyri. ROIs had a minimum of 35 significant voxels, although the vast majority had over 100 significant voxels (**Error! Reference source not found.**). Separate A∩V ROIs were defined for left and right hemisphere (N=15 each); correspondingly, separate V∩G ROIs were defined for left and right hemisphere (N=15 each). Regions of interest were determined using only the odd-numbered runs, and subsequent analyses were then performed on data extracted from the ROI over even-numbered runs (separate linear models for odd and even runs). For the purpose of reporting statistics, using this split-plot approach ensures that the voxel selection procedure and subsequent analyses are independent (Kriegeskorte, Simmons, Bellgowan, & Baker, 2009; Poldrack & Mumford, 2009).

Table 4.1. MNI coordinates of individual subject ROIs

	A∩V								**V∩G**							
	LH_STS			**Vox**	**RH_STS**			**Vox**	**LH_STS**			**Vox**	**RH_STS**			**Vox**
Sub	x	y	z		x	y	z		x	y	z		x	y	z	
1	-62.6	-22.1	-1.9	*255*	57.7	-27.4	-3.2	*226*	-57.7	-42.3	8.2	*186*	53.5	-42.7	8.4	*346*
2	-53.1	-43.1	7	*54*	60.8	-9.6	-4.2	*35*	-59.7	-63.8	13.2	*253*	52.2	-37	5.3	*47*
3	-62.7	-38	6.7	*169*	53	-30.3	1.3	*183*	-49.8	-47.7	9.8	*173*	52.6	-17.2	-8	*136*
4	-61.6	-30.3	4.9	*417*	62.5	-30.9	3	*300*	-61.5	-31.8	5.3	*260*	59.3	-34.8	5.8	*319*
5	-58.9	-34.7	3.8	*247*	65.5	-29.7	2.7	*347*	-51.9	-46.9	6.1	*222*	52.4	-37.9	7.9	*254*
7	-62.9	-19	-7.9	*225*	61	-33.6	4.3	*167*	-50.6	-50.8	8.6	*143*	48.9	-32.1	1.8	*64*
9	-62.6	-35.7	3.8	*256*	51.3	-41.2	9.8	*369*	-53.4	-54.7	12.3	*406*	56.1	-45.5	10.3	*382*
10	-67.7	-28.1	3.4	*202*	66.9	-32.1	5.8	*151*	-58.8	-59	12.2	*232*	51.9	-44.3	12	*124*
11	-52	-47.6	17.1	*234*	57.4	-30.8	-0.7	*111*	-45.4	-52.4	16.8	*133*	57.2	-53.5	7.4	*227*
12	-55.8	-49.5	11	*372*	63.2	-23.2	1.3	*191*	-56	-54.4	9.1	*152*	54.6	-48.3	6.3	*486*
14	-65.9	-42.2	-2.9	*348*	63.7	-49.2	12.5	*271*	-55	-68.3	17.2	*391*	64.8	-49.5	13.8	*212*
16	-54.6	-45.2	13.1	*195*	57.7	-31.1	1.2	*117*	-54.4	-51.6	9.9	*201*	54.8	-49.5	10	*138*
18	-59.9	-21.5	2.7	*145*	68.2	-32.6	4.5	*232*	-59.6	-59.4	14.8	*272*	57.3	-51	9.1	*330*
19	-62	-38.3	4.8	*349*	63.6	-34.3	3.7	*366*	-55.6	-41.7	9	*97*	62.5	-35.5	1.1	*416*
20	-51.5	-43.3	14.8	*193*	57.2	-43.8	8.6	*115*	-49.4	-49	13.2	*265*	54.3	-45.2	13.3	*47*
Mean	-59.6	-35.9	5.4	244	60.6	-32	3.4	212	-54.6	-51.6	11	226	55.5	-41.6	7	235

To assess the response properties of these ROIs, we extracted the mean PSC from baseline (still face) in each of the experimental conditions for the even-numbered runs, based on the regions defined using the odd-numbered runs. The mean PSCs for each subject in each condition were entered into separate repeated-measures ANOVAs for each region. The Greenhouse/Geisser correction was applied and F tests were thresholded at $p < 0.05$. Pairwise contrasts of interest were assessed using paired samples *t* tests. Five pairwise contrasts – AvsV, AvvsA, AvvsV, AvsR, VvsG – were assessed for each ROI at a significance level of $p < 0.05$ (two-sided). To correct for multiple comparisons we applied a Bonferroni correction within each family of tests (i.e., for each ROI), making the corrected significance level $p < 0.01$.

Pattern Classification

MVPA was implemented in all four ROIs (A∩V: left and right STS; V∩G: left and right STS), identified in individual subjects. All analyses described below were performed on even-numbered runs only, that is, runs that were independent from the ROI selection runs. MVPA was achieved using a support vector machine (SVM) (MATLAB Bioinformatics Toolbox v3.1, The MathWorks, Inc., Natick, MA) as the pattern classification method. The logic behind this approach is that if a trained SVM classifier is able to successfully classify one condition from another based on the pattern of response in an ROI, then the distribution of activity among the voxels within the ROI must contain information that distinguishes the two conditions. In each ROI, five different pairwise classifications were performed: (i) AvsV, (ii) AvvsA, (iii) AvvsV, (iv) AvsR, (v) VvsG. Note that (i)-(iii) involve classification of conditions containing identical speech information and differing only in terms of modality of presentation, while (iv) and (v) involve classification of conditions containing intelligible speech versus nonspeech.

Inputs to the classifier were estimates of activation to each block (event) based on "Least Squares – Separate" (LS-S) regression (Mumford, Turner, Ashby, & Poldrack, 2012). Specifically, we performed LS-S regressions (AFNI 3dLSS) using data from the even runs (preprocessed as described above), wherein each regression included one regressor of interest modeling a single block from a given condition (e.g., A), and five nuisance regressors modeling all other events in the condition of interest (e.g., A) and all events in the remaining conditions (e.g., V, AV, R, G), respectively (Turner, Mumford, Poldrack, & Ashby, 2012). The output of each regression was an LS-S coefficient representing activity from a single block in the condition of interest. LS-S coefficients representing all 15 blocks for each condition (3 blocks/run X 5 even runs) were calculated and stored with appropriate run labels at each voxel in all four ROIs described above. Prior to classification, LS-S coefficients for each ROI were z-scored (across voxels) for each block (time point), effectively removing mean amplitude differences across blocks (only spatial pattern information remained) (Coutanche, 2013; Mumford et al., 2012).

We performed SVM classification on the LS-S data using a leave-one-out cross validation approach (Vapnik, 1999). In each iteration, we used data from all but one even session to train an SVM classifier and then used the trained classifier to test the data from the remaining session. The SVM-estimated condition labels for the testing data set were then compared with the real labels to compute classification sensitivity. Following signal detection convention, one condition was arbitrarily defined as "signal" and the other as "noise." A classifier "hit" was counted when the SVM-estimated condition label matched the real condition label for the "signal" condition, and a "false alarm" was counted when the SVM-estimated label did not match the real condition label in the "noise" condition. Measure of sensitivity, *d'*, was calculated following the formula for a yes-no experiment as listed in the behavioral analysis above. Classification *d'* for each subject was derived by averaging the *d'* scores across all leave-one-out sessions, and an overall *d'* was computed by averaging across subjects for each pairwise classification.

Classification *d'* scores were evaluated statistically using a nonparametric bootstrap method (Lunneborg, 2000). Classification procedures were repeated 10000 times for each pairwise classification within each individual data set. The only difference from the above method is that the condition labels in the training data set for each leave-one-out session were randomly reshuffled per repetition. Therefore, we obtained a random distribution of the bootstrap classification *d'* scores that could range from -6.18 to 6.18 for each subject and pairwise classification. By examining the bootstrapped *d'* values we confirmed that the ideal mean of this distribution is at the *d'* value of 0, corresponding to chance performance. We then tested the null hypothesis that the original classification *d'* score is equal to the mean of the bootstrap distribution by computing a one-tailed accumulated percentile (*P*) of the original classification accuracy score in the distribution. If the accumulated $P > 0.95$, then we rejected the null hypothesis and concluded that for this subject, signal from the corresponding ROI can classify the two tested experimental conditions. Furthermore, a bootstrap-T approach was used to assess the significance of the classification *d'* at the group level. For each repetition of the bootstrap, a *t*-test of the *d'* scores across all subjects against the ideal chance *d'* score (0 in our case) was performed. The *t*-score from the original classification procedures across the subjects was then statistically tested against the mean value of the distributed bootstrap *t*-scores. As in the within-subject approach, an accumulated $P_t > 0.95$ guarantees rejection of the null hypothesis (*d'* is significantly greater than chance). We will report statistical significance of the classification results with canonically used "p-values" calculated as p = 1-P_t.

Results

Behavior

Two participants performed below the behavioral cutoff and were excluded from further analysis (*d'* = 1.85, hits = 14/30; and *d'* = 2.13, hits = 14/30). The remaining eighteen participants performed well on the task (mean *d'* = 3.40 ± 0.14 SEM, mean hits = 26 ± 0.56 SEM) indicating that subjects attended to both auditory and visual components of the stimuli. Among participants that made the behavioral cutoff, there was not a significant difference in hit rate across conditions [F(2.3, 38.7) = 2.07, p = 0.13].

Whole-Brain Group Analysis

Activation maps for each of the five experimental conditions relative to rest are pictured in Figure 4.3 (top; FWER < 0.05). In general, we observed activation in temporal and frontal cortices, including bilateral posterior superior temporal regions and Broca's area, for the A, V, and AV speech conditions. Activation in the nonspeech conditions (R and G) was qualitatively similar to activation in the corresponding speech conditions (A and V, respectively), with differences highlighted in the contrasts described below. In a standard subtraction analysis, we tested for voxels showing an enhanced response for audiovisual speech versus auditory speech and visual speech alone (Figure 4.3, bottom). For the AV>A contrast, we observed activation in bilateral primary and secondary visual cortices, lateral occipital-temporal visual regions, inferior and middle temporal gyri, and posterior STS. For the AV>V contrast, we observed activation in core auditory cortex extending along the superior temporal gyrus and into the STS in the right hemisphere. In line with previous work, the posterior superior temporal sulcus showed enhanced activation to audiovisual speech versus auditory or visual speech alone. However, these activations were largely non-overlapping in the STS. This may have been due to functional-anatomic variability in individual subjects, a fact that should not affect the ROI analyses reported below since they are carried out in native space.

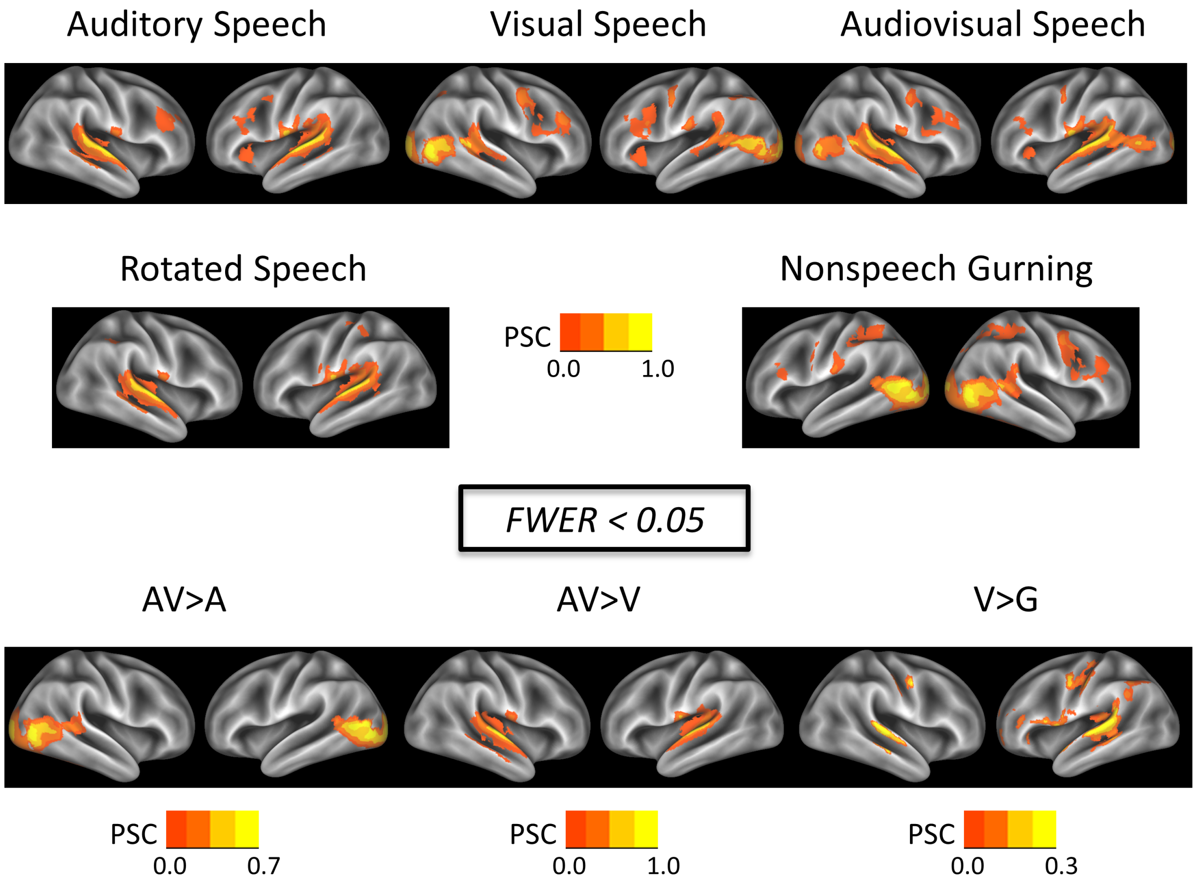


Figure 4.3. Group results: Whole-brain statistical parametric maps for individual conditions and contrasts of interest. Maps are thresholded at $p < 0.005$ with a cluster extent threshold of 159 voxels (FWER < 0.05). Color scale reflects mean percent signal change (PSC) for the individual conditions, and mean differences in PSC for the contrasts.

We also tested for voxels showing an enhanced response to intelligible speech (i.e., speech vs. nonspeech). To assess auditory speech intelligibility, we contrasted auditory speech versus rotated speech. This contrast (A>R) did not yield any significant activation at the group level. Although this is not consistent with previous imaging work, (Scott, Blank, Rosen & Wise, 2000; Narain et al., 2003; Liebenthal, Binder, Spitzer, Possing & Medler, 2005; Okada et al., 2010), we believe that our use of simpler stimuli, i.e. syllables versus sentences, may have produced this null result (see Discussion). Finally, to assess visual speech intelligibility, we contrasted visual speech versus nonspeech facial gurning (V>G). This contrast yielded an activation network consistent with previous work (Campbell et al., 2001; Okada & Hickok, 2009), including bilateral STS, left inferior frontal gyrus, and a host of inferior parietal and frontal sensory-motor brain regions (Figure 4.3 and Figure 4.4).

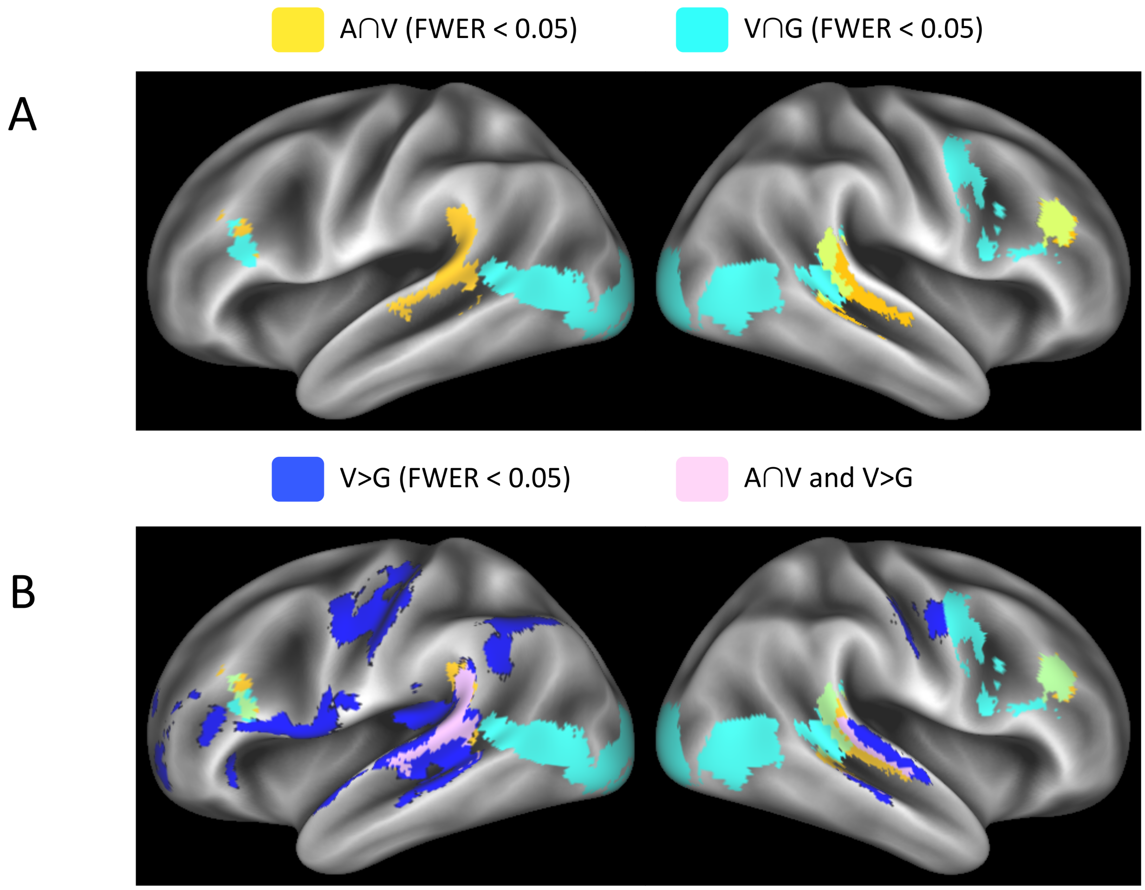


Figure 4.4. Group results: Whole-brain conjunction analyses. A: Group-level results for the two conjunctions used to define STSms (A∩V, yellow) and STSfm (V∩G, teal) are plotted together on an inflated surface rendering of the Conte69 template in MNI space. The results demonstrate a clear distinction: A∩V activations fall anterior to V∩G activations in the STS. Overlap appears green-yellow. B: Conjunction analyses are plotted together with the contrast VvsG (blue), which highlights regions that activate preferentially to visual speech versus nonspeech facial gurning. These visual speech-specific regions fall anterior to V∩G in the STS and overlap strongly with A∩V (pink). Family-wise error rate controlled < 0.05.

Finally, we conducted two conjunction analyses meant to highlight STSms (A∩V) and STSfm (V∩G). Activation maps for each conjunction appear overlaid on the same image in Figure 4.4 (FWER < 0.05). In general, A∩V activations were anterior to V∩G activations in the superior temporal lobe, including the STS (MNI coordinates of peak STS activations: A∩V LH = -65, -41, 5; A∩V RH = 62 -35 2; V∩G LH = -56, -58, 11; V∩G RH = 55, -46, 8), in both hemispheres, a fact we confirm in individual subjects below. Moreover, V∩G activations appeared to immediately abut the A∩V activations moving posterior-to-anterior in STS, with greater overlap in the right hemisphere. Both conjunctions showed activation in the left inferior and middle frontal gyri and the right middle frontal gyrus. Outside of the STS, selective activation to A∩V was present in the left temporoparietal junction, while selective activation to V∩G was present in bilateral visual cortices including hMT and surrounds, as well as right inferior frontal and premotor regions.

Of interest, visual activations specific to speech (V>G) occupied much of the same STS territory as activations to A∩V (i.e., STSms), considering especially the extent of activations along the posterior-anterior axis of the STS. The V∩G region of the STS (i.e., STSfm) did not overlap at all with speech-specific (V>G) activations (Figure 4.4). This suggests that the subregion of the STS activating preferentially to speech is anatomically comparable to STSms, a finding we explore further in the ROI analyses below. Note that much of the V>G activation on the ventral bank of the STS was due to deactivation in the G condition, rather than large activations in the V condition (see V-only maps in Figure 4.3). STS regions that activated significantly to V and were also significant in V>G were likely to be occupied by A∩V.

Region of Interest – Percent Signal Change

We defined a set of four ROIs in the STS in each individual subject (native space). Two ROIs, one in each hemisphere, were defined based on the conjunction A∩V. This conjunction was chosen to identify STSms in each subject. An additional two ROIs, again one in each hemisphere, were defined in STS based on the conjunction V∩G. This method was chosen to identify STSfm in each subject. Voxels from A∩V and V∩G ROIs were free to overlap based on the selection criteria outlined in the Materials and Methods above. In order to report the centers of mass of native-space ROIs in standard MNI space, we warped activation masks of each ROI to MNI space using the series of transformations reported in the Group Analysis subsection of the Methods. The MNI coordinates of individual subject centers of mass for each ROI in each hemisphere are listed in Table 4.1 and pictured on the group template in Figure 4.5. As in the group-level conjunctions, A∩V ROIs showed an anterior location bias relative to V∩G ROIs in both hemispheres (particularly strong in the left). In fact, ROIs followed this pattern in the majority of subjects (left: 14/15, right: 11/15), with varying degrees of overlap (generally more in the right). Overall, three sources of evidence – (1) group level conjunction maps, (2) plots of individual-subject centers of mass together in MNI space, (3) location patterns of individual subject ROIs in native space – all support the conclusion that STSfm (V∩G) transitions to STSms (A∩V) moving posterior to anterior in the STS.

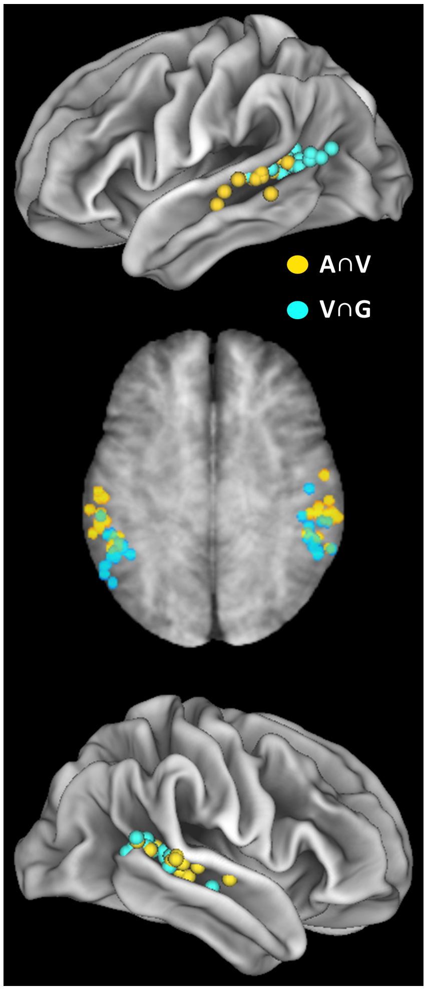


Figure 4.5. Location of individual subject ROIs visualized in MNI space. Top and Bottom: MNI coordinates of center of mass for individual subject ROIs are plotted as 3.5mm spheres on the fiducial surface of the Conte69 template (Yellow = A∩V, Teal = V∩G). Middle: The same coordinates are visualized as 3mm spheres on a smoothed volume rendering of our study-specific template warped to MNI space. The image is shown in the axial plane (top-down, neurological convention) with spheres at all depths shown (only left-right and posterior-anterior location can be discerned). On all three images, it is clear that the distribution of yellow spheres is centered anterior to the distribution of teal spheres, with less separation in the right hemisphere.

In the following section we present a standard ROI analysis of percent signal change, employed here to assess whether the ROIs just described showed an enhanced response for bimodal (AV) speech over A or V alone (AvvsA, AvvsV), or for speech versus nonspeech (AvsR, VvsG). We also tested for activation differences between the conditions used to select the ROIs (AvsV and, again, VvsG). For each ROI, we extracted the mean percent signal change for each experimental condition. These values were entered in a group-level analysis of differences in the mean percent signal change (PSC). All statistical tests were performed on independent data from those used to define the ROIs. The results of PSC-based analyses are presented graphically in Figure 4.6.

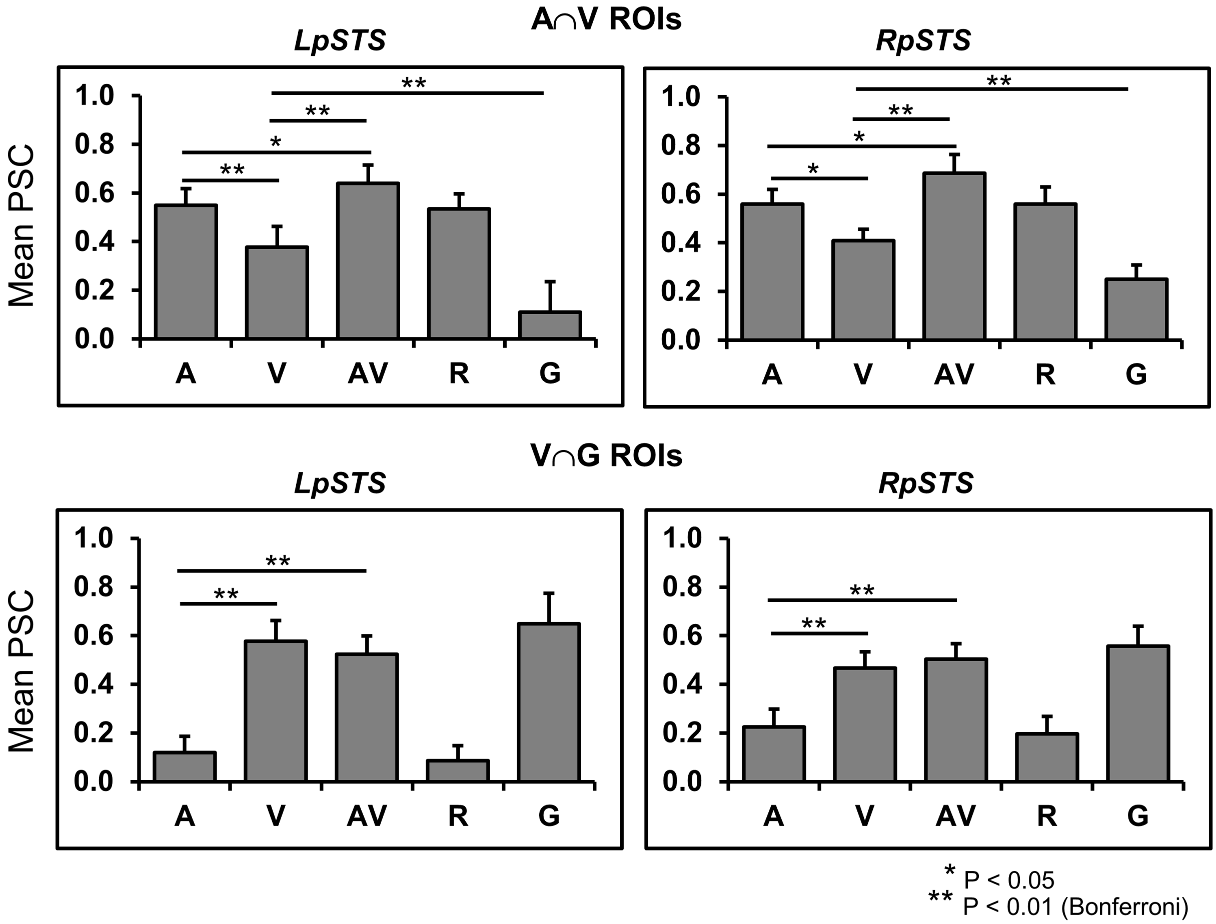


Figure 4.6. ROI results: Percent signal change. Top: Mean percent signal change in each condition for the A∩V ROIs. Bottom: Mean percent signal change in each condition for the V∩G ROIs. Trending ($p < 0.05$) and significant ($p < 0.01$) pairwise comparisons are starred. We planned five contrasts in total: AvsV, AVvsA, AVvsV, AvsR, VvsG. In general, A∩V ROIs show greater activation to audiovisual speech (AV) over auditory (A) or visual (V) speech alone, and greater activation to visual speech than nonspeech facial gurning (G). V∩G ROIs show greater activation to visual conditions (AV, V, G) than auditory-only conditions (A, R).

A∩V ROIs

The ANOVA on experimental condition was significant in both left [F(1.98, 27.73) = 23.84, p $p < 0.001$] and right [F(1.76, 24.60) = 15.70, $p < 0.001$] hemispheres. We also assessed each of the four contrasts that were tested in the whole-brain analysis in addition to the contrast AvsV. Within each ROI the contrasts were thresholded at the Bonferroni-corrected level $p < 0.01$. The results are presented graphically in Figure 4.6. In the left hemisphere, the AvvsA contrast demonstrated a strong trend [t(14) = 2.64, p = 0.019] and the AvvsV contrast was significant [t(14) = 4.30, p = 0.001]. The AvsR contrast was not significant [t(14) = 0.40, p = 0.70], while the VvsG contrast was significant [t(14) = 5.83, $p < 0.001$]. Finally, the AvsV was significant [t(14) = 3.26, p = 0.006]. In the right hemisphere there was a similar pattern. The AvvsA contrast trended strongly toward significance [t(14) = 2.90, p = 0.012] while the AvvsV contrast was significant [t(14) = 4.77, $p < 0.001$]. The AvsR contrast was not significant [t(14) = -0.01, p = 0.989], while the VvsG contrast demonstrated significance [t(14) = 3.51, p = 0.003]. The AvsV contrast trended strongly toward significance [t(14) = 2.77, p = 0.015]. Overall, the A∩V ROIs showed enhanced activation for AV relative to A and V alone, which is consistent with a role in audiovisual integration using the max criterion (AV>max[A,V]) (Michael S Beauchamp, 2005b) and supports the conclusion that these ROIs occupy STSms. In addition, there was increased activation to intelligible visual speech versus nonspeech facial movements (VvsG), which is consistent with the group results and indicates a possible role in higher-level phonological speech processing. Finally, there was greater activation to auditory speech than visual speech, suggesting a possible bias induced by the ROI selection process. The presence of such a bias would not affect our interpretation of the data.

V∩G ROIs

An ANOVA on experimental condition was significant in the left [F(3.62, 16.60) = 13.06, p = 0.002] and right [F(1.40, 19.65) = 12.28, p = 0.001] hemispheres. The pattern of results in the pairwise contrasts was identical between hemispheres (Figure 4.6). The AvsV [left: t(14) = -4.02, p = 0.001; right: t(14) = -3.90, p = 0.002] and AvvsA [left: t(14) = 4.24, p = 0.001; right: t(14) = 5.76, p < 0.001] contrasts were significant. The AvvsV [left: t(14) = -1.40, p = 0.18; right: t(14) = 1.20, p = 0.25], AvsR [left: t(14) = 1.27, p = 0.22; right: t(14) = -0.01, p = 0.99], and VvsG [left: t(14) = -1.14, p = 0.27; right: t(14) = -1.75, p = 0.10] contrasts were not significant. Overall, the V∩G ROIs responded well to stimuli containing facial motion (V, AV, and G), and poorly to auditory conditions (A, R), which supports the conclusion that these ROIs occupy STSfm. Importantly, these regions did not respond differentially to V and G, the facial motion conditions used to define the ROIs via conjunction. Additionally, there was no activation difference between the speech conditions containing facial motion (V and AV).

Region of Interest – Multivariate Pattern Analysis

Within the same ROIs discussed in the section above, we used MVPA to assess spatial patterns of activation rather than differences in mean signal strength. Again, the logic is that if our SVM algorithm is able to reliably classify two conditions within a given ROI, then this ROI must contain distinct information about each condition. A significant classification can be interpreted as support for the existence of distinct distributions of neuronal ensembles within the region that are differentially sensitive to each condition. The pattern of activity in each ROI was assessed for the following contrasts (identical to the contrasts employed in the percent signal change analysis above): (i) AvsV, (ii) AvvsA, (iii) AvvsV, (iv) AvsR, (v) VvsG. Brain regions that are sensitive to modality of presentation should discriminate at least one of contrasts (i)-(iii), while brain regions that are sensitive to intelligible (speech-specific) information should discriminate one or both of (iv)-(v).

A∩V ROIs

The results of MVPA for A∩V ROIs are pictured in Figure 4.7 (top). The AvsV contrast was discriminated successfully by our pattern classifier in the left [accuracy = 77.3%; *d'* = 3.23, t(14) = 8.08, p = 0.005] and right [accuracy = 76.2%; *d'* = 2.87; t(14) = 8.76, p < 0.001] hemispheres. This indicates that there are likely distinct neuronal ensembles that subserve auditory and visual processing in these regions, a known feature of STSms. Additionally, the AvvsA contrast was discriminated successfully in the left [accuracy = 58.4%; *d'* = 1.01; t(14) = 2.42, p = 0.03] and right [accuracy = 61.3%; *d'* = 1.31; t(14) = 3.39, p = 0.008] hemispheres, as was the case for AvvsV in the left [accuracy = 0.70%; *d'* = 2.40; t(14) = 4.66, p = 0.01] and right [accuracy = 66.2%; *d'* = 1.82; t(14) = 4.17, p = 0.006] hemispheres. It is worth noting that AV may classify versus unimodal conditions (A, V) simply due to the presence of distinct neuronal ensembles representing A and V – in other words, a significant contrast of bimodal versus unimodal speech does not necessarily support the existence of any distinct neuronal ensemble representing AV.

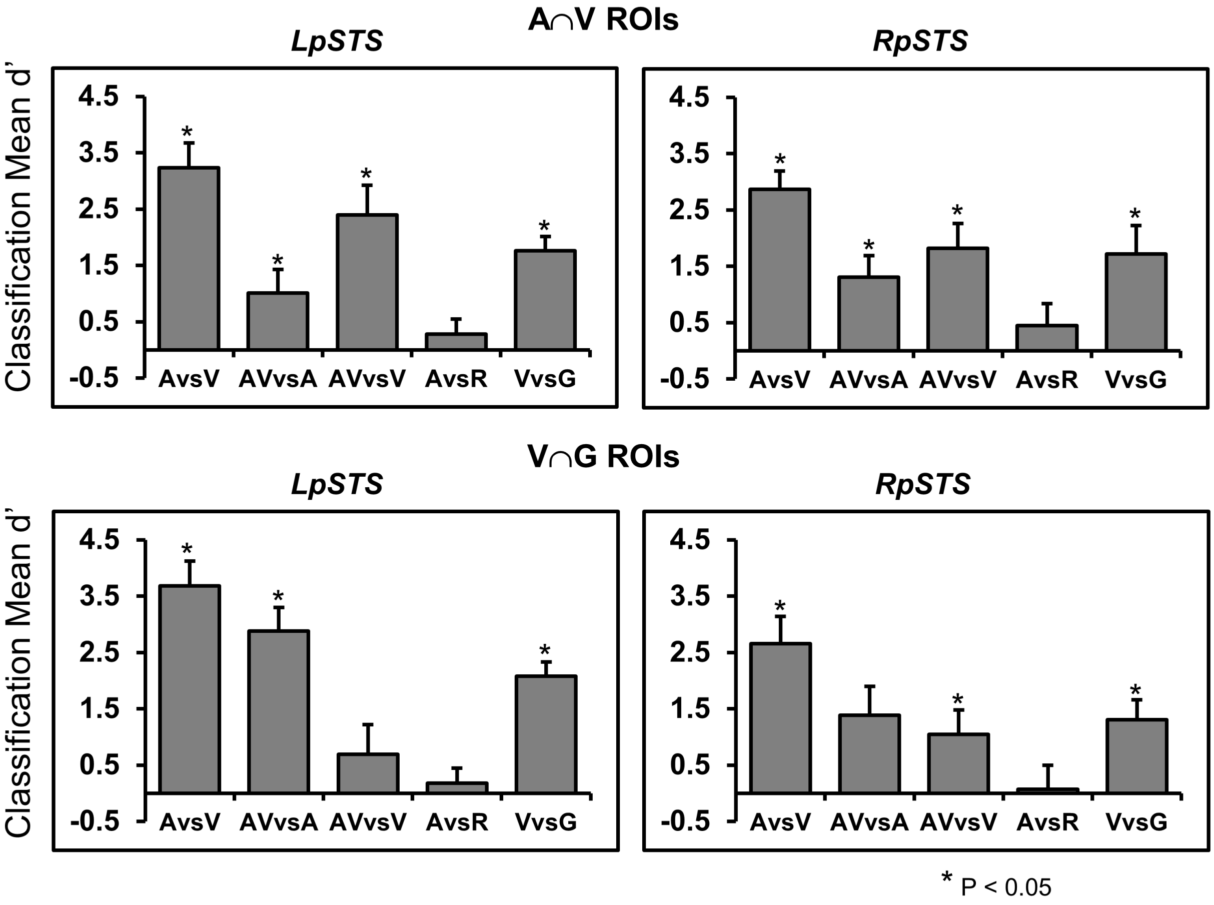


Figure 4.7. ROI results: Multivariate pattern analysis. Top: Classification accuracy (d') for the five planned pairwise comparisons for A∩V ROIs. Bottom: Classification accuracy for the same five comparisons in V∩G. Significant comparisons (bootstrap methods) are starred. The A∩V ROIs successfully distinguish contrasts of speech in different modalities (AvsV, AVvsA, AVvsV) and visual speech versus nonspeech (VvsG). V∩G ROIs distinguish conditions containing visual information from those that do not (AvsV, AVvsA) and also classify the contrast of visual speech versus nonspeech (VvsG).

Regarding the intelligibility contrasts, VvsG was classified successfully in the left [accuracy = 66.0%; *d'* = 1.76; t(14) = 4.698, p = 0.007] and right [accuracy = 65.1%; *d'* = 1.72; t(14) = 3.38, p = 0.03] hemispheres. The AvsR contrast was not classified successfully in the left [accuracy = 53.1%, *d'* = 0.28; t(14) = 0.83, p = 0.22] or the right [accuracy = 54.4%, *d'* = 0.45; t(14) = 1.14, p = 0.15] hemisphere. Overall, the MVPA results for A∩V ROIs replicate the above observations based on PSC, but in terms of spatial patterns rather than mean signal differences.

V∩G ROIs

The results of MVPA for V∩G ROIs are pictured in Figure 4.7 (bottom). The AvsV contrast was significant in the left [accuracy = 81.3%, *d'* = 3.68, t(14) = 8.40, p < 0.001] and right [accuracy = 73.1%, *d'* = 2.66, t(14) = 5.51, p = 0.005] hemispheres. The AvvsA contrast was also discriminated successfully in the left [accuracy = 76.0%, *d'* = 2.88, t(14) = 6.91, p = 0.004] but only marginally in the right [accuracy = 62.9%, *d'* = 1.39, t(14) = 2.786, p = 0.06] hemisphere. The reverse pattern was present for the AvvsV contrast, which was significant in the right [accuracy = 59.1%, *d'* = 1.05; t(14) = 2.44, p = 0.04] but not in the left [accuracy = 56.0%, *d'* = 0.69; t(14) = 1.33, p = 0.15] hemisphere.

Regarding the intelligibility contrasts, AvsR was not classified successfully in the left [accuracy = 52.0, *d'* = 0.18, t(14) = 0.66, p = 0.28] or right [accuracy = 50.0%, *d'* = 0.07, t(14) = 0.16, p = 0.43] hemisphere. However, the VvsG contrast was discriminated successfully in the left [accuracy = 69.3%, *d'* = 2.08, t(14) = 8.32, p = 0.003] and right [accuracy = 62.2%, *d'* = 0.51, t(14) = 1.166, p = 0.142] hemispheres. Overall, the MVPA results for V∩G ROIs deviated from the results based on PSC in two important ways. First, the V∩G ROIs distinguished visual speech (V) from nonspeech (G) in terms of spatial patterns, whereas this distinction could not be made on the basis of differences in mean signal strength. Second, the right hemisphere V∩G ROI distinguished bimodal (AV) from unimodal (V) speech on the basis of spatial patterns, whereas these conditions did not produce reliable differences in mean signal strength from this region. A visual region would not be expected to classify two conditions with identical visual information. As such, it is tempting to conclude that AvvsV classified because a sizable auditory population was included in the right hemisphere V∩G ROIs (there was more overlap between A∩V and V∩G in the right hemisphere). However, the 8 subjects with the highest classification accuracy in the AvvsV contrast (mean = 69.58%) showed less auditory activation (mean PSC = 0.12) than the 7 subjects with the lowest classification accuracy (mean = 47.1%, mean PSC = 0.34). Thus, we must recognize the possibility that right hemisphere facial motion regions are modulated by concurrent auditory information in AV stimuli.

Discussion

In the current study we set out to answer two questions concerning the organization of the visual speech stream in the STS: (1) Are biological/facial motion processing regions of the STS (here denoted STSfm) involved in visual speech processing and do these regions overlap with multisensory STS (here denoted STSms)?, and (2) At what representational level are auditory and visual speech signals processed or combined in STSms?

The answer to the first question falls out neatly from the current data. We identified STSfm by taking the conjunction of activation to visual speech (V) and nonspeech facial motion (G). Recall that activation in all conditions was relative to a ‘still face’ baseline. The logic behind the V∩G conjunction was that it should identify voxels that (a) respond to conditions that differ from baseline primarily in terms of facial motion, and (b) respond to visual speech, specifically. This conjunction reliably identified voxels in the pSTS bilaterally, both at the group level and in individual subjects. Our individual-subject ROI analysis demonstrated that V∩G

ROIs activated well to all conditions with facial motion (V, AV, G), but poorly to conditions without facial motion (A, R). Moreover, these regions did not distinguish between the facial motion conditions in terms of differences in mean signal strength. This supports the conclusion that the V∩G conjunction successfully identified STSfm. However, V∩G ROIs distinguished two classes of facial motion (speech vs. nonspeech, i.e., VvsG) based on spatial patterns of activation, which suggests these regions likely code motion information at a more abstract level (E. D. Grossman et al., 2010; Lestou et al., 2008). The fact that STSfm can distinguish speech from nonspeech suggests a role for this region in visual speech perception.

The second component of question (1) concerned whether STSfm and STSms share the same neural territory. We identified STSms by taking the conjunction of activation to auditory (A) and visual (V) speech. The logic here was simple: voxels in STSms should respond to speech regardless of the sensory modality of the input. The A∩V conjunction reliably identified voxels in the pSTS bilaterally, both at the group level and in individual subjects. The A∩V ROIs in our individual subject analysis showed an enhanced response to bimodal speech using the max criterion (AV > max[A,V]) and also discriminated among speech conditions in different sensory modalities based on spatial patterns of activation (AvsV, AvvsA, AvvsV). These results confirm that voxels identified by A∩V have the expected response profile for STSms. Regarding the localization of STSms with respect to STSfm, we observed that activations to A∩V were consistently positioned anterior to activations to V∩G in the STS. Three sources of evidence support this observation. First, group-level conjunction maps showed that A∩V activations were largely non-overlapping with V∩G activations, with A∩V occupying more anterior regions of STS. Second, when we warped individual subject ROIs to standard group space (MNI) and plotted centers of mass for both A∩V and V∩G ROIs, the same pattern emerged in which A∩V centers of mass were distributed anterior to V∩G centers of mass. Third, location patterns of individual subject ROIs in native space consistently showed the same anterior bias for A∩V. Recall that individual subject ROIs were defined based on voxels showing the *best* response to A∩V and V∩G, respectively. Of note, individual subject ROIs for A∩V and V∩G occasionally overlapped (to varying degrees) and group level activations to A∩V immediately abutted (and slightly overlapped in the right hemisphere) activations to V∩G moving posterior to anterior in the STS. Overall, we suggest the results reflect a processing gradient in the pSTS that transitions gradually from STSfm to STSms moving posterior to anterior.

Previous work suggests a similar posterior to anterior division along the pSTS in terms of cortex that responds preferentially to visual speech versus cortex that responds to high-level orofacial motion analysis. A recent fMRI study (Bernstein et al., 2011) employed a rather comprehensive set of visual speech and nonspeech stimuli (but no auditory stimuli), demonstrating that a more anterior region of pSTS responds preferentially to orofacial visual motion when it is speech-related, while more posterior regions of pSTS respond to orofacial motion whether or not it is speech-related. The authors dubbed the anterior speech-related area the temporal visual speech area (TVSA). Indeed, the TVSA appears to be anatomically and functionally comparable to our A∩V region (i.e, STSms). These results fall in line with our characterization of a posterior-to-anterior visual speech-processing gradient in the STS. But what does this mean in light of current models of multisensory speech perception in the STS?

The results of a recent investigation (Arnal et al., 2009) combined with our results shed light on this question. Those researchers constructed a set of audiovisual syllables that varied in terms of both visual predictability (viseme specificity) and phonological congruence (i.e., does the A syllable match the V syllable?). Using MEG, they demonstrated latency facilitation of the early auditory-cortical response (M100) that depended on visual predictability but not phonological congruence. An effect of phonological congruence was observed 20ms after the M100 latency effect, likely arising from a signal generated in the STS – the M170 response to visual syllables in the STS was delayed by exactly 20ms relative to the M170 response originating from motion-sensitive visual cortex, and the STS response peaked simultaneously with the peak response originating from auditory cortex. From this, the authors concluded that speech-related motion signals originating from visual cortex split onto two targets – auditory cortex and the STS – leading to two integration pathways: (a) an early direct pathway (with low phonological specificity) from motion sensitive cortex to auditory cortex, and (b) a later indirect pathway (with greater phonological specificity) from motion sensitive cortex to STSms. The delay in the indirect pathway may be due to further "preprocessing" of the visual signal in STSfm, performed in order to extract abstract features of the motion signal relevant to phonological classification. This dovetails with the current data, which describe a processing stream running from STSfm to STSms.

The second question posed at the beginning of this section concerned the level at which auditory and visual signals are combined in STSms. Recall from the Introduction that direct integration models support integration at a high level – that is, auditory and visual speech signals are said to converge on bimodal, abstract speech sound representations directly in STSms. Feedback models, on the other hand, are rather nonspecific regarding this question and state simply that auditory and visual speech signals are compared in STSms and fed back to unimodal sensory cortices for further evaluation. Based on previous data, we feel confident asserting that high-level phonological processing is carried out in the STS (J. R. Binder et al., 2008; J. Binder et al., 2000; G. Hickok & Poeppel, 2004; Gregory Hickok & Poeppel, 2007; Okada & Hickok, 2006; Price, 2010; Vaden Jr, Muftuler, & Hickok, 2010). As such, we would like to specifically evaluate the feasibility of direct integration models and to ask, more generally, whether STSms overlaps with auditory-phonological speech regions in the STS. To look ahead, our data are somewhat suggestive but ultimately cannot resolve among the competing positions.

One way to address the question at hand would be to plot the relation between STSms and regions of the STS that are selective for auditory speech compared to auditory nonspeech, which is a contrast that is often used to identify high-level speech processing networks. However, we failed to observe significant activation (or differences in pattern information) for the AvsR contrast in STSms or elsewhere in the STS, in contrast to previous studies (E. Liebenthal et al., 2005; Einat Liebenthal et al., 2010). We believe the manipulation may have failed due to the presence of phonetic information under spectral rotation at the syllable level (E. Liebenthal et al., 2005). In contrast, we were able to examine the relation between STSms and speech-specific activations in the visual modality. We found large clusters of activation in the pSTS bilaterally for the group V>G contrast. These clusters were positioned similarly to the A∩V conjunction (as opposed to V∩G) at the group level and, furthermore, only the individual-subject A∩V ROIs showed significant differences in mean signal for VvsG, favoring visual speech over nonspeech gurning. Thus, regions of the STS that prefer visual speech over nonspeech correspond well to STSms, providing partial support for audiovisual integration at a high level.

Co-localization of the TVSA with STSms may provide similar support, but in fact the authors who identified the TVSA assert that speech sounds are categorized downstream in the middle STS (mSTS) (Bernstein et al., 2011): "TVSA provides a linguistically relevant integration of cues that is projected for categorization by other areas…specifically, the more anterior mSTG/mSTS area previously identified as having a role in auditory speech perception" (p. 1672). Indeed, several studies have emphasized the role of the mSTS in categorical speech processing. We will review two significant studies here. First, a recent fMRI study employed dynamic sound morphing of consonant-vowel syllables to investigate temporal lobe speech processing (Specht, Osnes, & Hugdahl, 2009). This technique involves a parametric manipulation that mixes speech sounds with white noise using an increasingly large interpolation factor. The result is a continuum from white noise (morph step 1) to speech (morph step 7) that gradually reveals the spectral and temporal characteristics of the speech in a step-wise manner. Musical sounds of matched length were manipulated to create a control continuum. A significant stimulus (speech, music) by manipulation (seven-step morph continuum) interaction was observed in the left mSTS, whereby activation to speech increased linearly from step two to five of the morph continuum, after which activation leveled off, while activation to music was overall less than activation to speech and increased steadily, without leveling off. Crucially, the structure of the first and second formants of the speech stimuli were revealed in morph step five, such that the activation profile in mSTS reflected a rather categorical difference between auditory speech and nonspeech. On the other hand, a region of pSTS immediately posterior to the mSTS region showed a linearly increasing parametric response to the morphing manipulation across all seven levels for speech stimuli. This suggests a posterior-anterior distinction in which the pSTS supports high-level spectrotemporal analysis while the mSTS performs more abstract, perhaps categorical phonological processing.

A second fMRI study from a different group demonstrated a similar distinction between pSTS and mSTS (Einat Liebenthal et al., 2010). In this study, participants performed categorization (identification) of seven-step speech (/ba/-/da/) and nonspeech (rotated analogs) continua. The speech and nonspeech stimuli were identical with the exception that the spectral energy was inverted at the first formant of nonspeech stimuli, rendering those stimuli phonetically unfamiliar. Participants were scanned while performing identification trials for speech and nonspeech stimuli both before and after a two week training period that gave participants practice categorizing the two stimulus classes. Left pSTS and mSTS showed markedly different activation profiles with respect to stimulus (speech, nonspeech) and training (before, after). While the mSTS activated more to speech versus nonspeech both before and after training, the pSTS showed increased activation to speech versus nonspeech in the pretraining session but showed the reverse pattern after training. In their interpretation of these results, the authors argued that pSTS provides a short-term representation of sound features relevant to categorization, while mSTS mediates categorization of highly familiar phonemic patterns.

In light of these findings, an elegant expansion of the pattern in the current data would be to simply extend the posterior-to-anterior processing gradient running from STSfm to STSms into the mSTS. In other words, the visual speech stream in STS would run as follows: (1) extraction of high-level motion properties (vocal tract configurational information) in STSfm, (2) integration of visual speech representations with auditory speech representations in STSms, (3) categorization of speech sounds in the mSTS. However, the current data alone cannot mediate between this model (which is essentially an expanded feedback model) and direct integration models (both pictured in Figure 4.8). Future research might combine an appropriate localizer for STSms (James & Stevenson, 2012) with a design to identify categorical speech regions in the mSTS (within the same group of subjects).

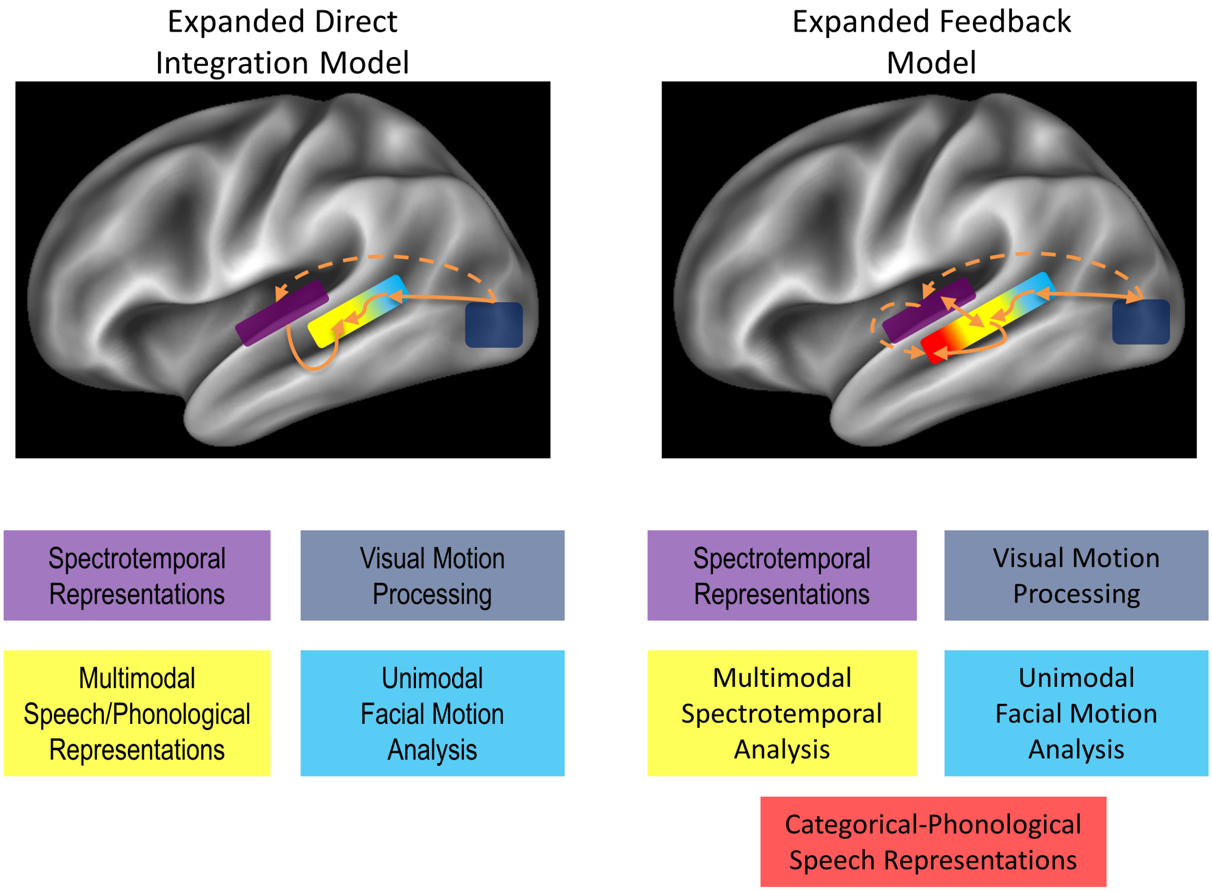

Expanded Direct Integration Model
Expanded Feedback Model
Spectrotemporal Representations
Visual Motion Processing
Multimodal Speech/Phonological Representations
Unimodal Facial Motion Analysis
Spectrotemporal Representations
Visual Motion Processing
Multimodal Spectrotemporal Analysis
Unimodal Facial Motion Analysis
Categorical-Phonological Speech Representations

Figure 4.8. Updated model schematics for multisensory speech processing in the STS. Left: An expanded Direct Integration model. Essentially, we have expanded the processing stream in the STS to reflect the posterior to anterior gradient from STSfm (teal) to STSms (yellow) identified in the current study, along with the a connection between motion-sensitive cortex (blue) and auditory cortex (purple) to account for the early influence of visual speech on auditory speech perception. Right: An expanded feedback model. Here, we have extended the posterior to anterior gradient into the mSTS (red) where categorical speech sound representations may be located. Feedback connections are present. Auditory cortex feeds directly and indirectly (via STSms) to the mSTS. Left & Right: Dotted arrows represent connections that do not run through the STSms. Color-coded boxes beneath the figures represent the computations at each stage/location in the processing stream. From the perspective of the Direct Integration model, the "red" stage of the Feedback model is subsumed by the "yellow" stage.

In sum, we have demonstrated that different subregions of the STS are involved in processing facial motion versus multisensory speech. STSms is positioned anterior to STSfm in the posterior STS, but STSfm appears to transition gradually to STSms moving posterior to anterior. Intelligible visual speech can be distinguished from nonspeech facial gestures only on the basis of spatial patterns in STSfm, but by the time the signal reaches STSms the mean signal strength to visual speech exceeds the mean signal strength to nonspeech facial gestures. Thus, visual speech representations are elaborated gradually along the posterior-to-anterior processing gradient. It remains for future research to determine whether multisensory speech interactions in STSms reflect phonological processing directly, or, as others have suggested (Arnal et al., 2009; Calvert et al., 1999; Skipper, van Wassenhove, Nusbaum, & Small, 2007), whether the outcome of multisensory integration merely informs phonological mechanisms in other brain regions such as the mSTS.

References

Allison, Truett, Puce, Aina, & McCarthy, Gregory. (2000). Social perception from visual cues: role of the STS region. *Trends in Cognitive Sciences, 4*(7), 267-278.

Amedi, A, Kriegstein, K von, Atteveldt, NM van, Beauchamp, MS, & Naumer, MJ. (2005). Functional imaging of human crossmodal identification and object recognition. *Experimental Brain Research, 166*(3), 559-571.

Arnal, L. H., Morillon, B., Kell, C. A., & Giraud, A. L. (2009). Dual neural routing of visual facilitation in speech processing. *J Neurosci, 29*(43), 13445-13453. doi: 10.1523/JNEUROSCI.3194-09.2009

Avants, B., Duda, J. T., Kim, J., Zhang, H., Pluta, J., Gee, J. C., & Whyte, J. (2008). Multivariate Analysis of Structural and Diffusion Imaging in Traumatic Brain Injury. *Academic Radiology, 15*(11), 1360-1375.

Avants, B., & Gee, J. C. (2004). Geodesic estimation for large deformation anatomical shape averaging and interpolation. *Neuroimage, 23*, 139-150.

Avants, Brian B, Tustison, Nicholas J, Song, Gang, Cook, Philip A, Klein, Arno, & Gee, James C. (2011). A reproducible evaluation of ANTs similarity metric performance in brain image registration. *Neuroimage, 54*(3), 2033-2044.

Beauchamp, M. S., Argall, B. D., Bodurka, J., Duyn, J. H., & Martin, A. (2004). Unraveling multisensory integration: patchy organization within human STS multisensory cortex. *Nat Neurosci, 7*(11), 1190-1192. doi: 10.1038/nn1333

Beauchamp, Michael S. (2005a). See me, hear me, touch me: multisensory integration in lateral occipital-temporal cortex. *Current opinion in neurobiology, 15*(2), 145-153.

Beauchamp, Michael S. (2005b). Statistical criteria in FMRI studies of multisensory integration. *Neuroinformatics, 3*(2), 93-113.

Beauchamp, Michael S, Lee, Kathryn E, Argall, Brenna D, & Martin, Alex. (2004). Integration of auditory and visual information about objects in superior temporal sulcus. *Neuron, 41*(5), 809-824.

Beauchamp, Michael S, Lee, Kathryn E, Haxby, James V, & Martin, Alex. (2002). Parallel visual motion processing streams for manipulable objects and human movements. *Neuron, 34*(1), 149-159.

Beauchamp, Michael S, Lee, Kathryn E, Haxby, James V, & Martin, Alex. (2003). FMRI responses to video and point-light displays of moving humans and manipulable objects. *Journal of Cognitive Neuroscience, 15*(7), 991-1001.

Beauchamp, Michael S, Nath, Audrey R, & Pasalar, Siavash. (2010). fMRI-guided transcranial magnetic stimulation reveals that the superior temporal sulcus is a cortical locus of the McGurk effect. *The Journal of Neuroscience, 30*(7), 2414-2417.

Benevento, Louis A, Fallon, James, Davis, BJ, & Rezak, Michael. (1977). Auditory-visual interaction in single cells in the cortex of the superior temporal sulcus and the orbital frontal cortex of the macaque monkey. *Experimental neurology, 57*(3), 849-872.

Bernstein, LYNNE E. (2005). Phonetic processing by the speech perceiving brain. *The handbook of speech perception*, 79-98.

Bernstein, Lynne E, Jiang, Jintao, Pantazis, Dimitrios, Lu, Zhong-Lin, & Joshi, Anand. (2011). Visual phonetic processing localized using speech and nonspeech face gestures in video and point-light displays. *Hum Brain Mapp, 32*(10), 1660-1676.

Binder, Jeffrey R, Swanson, Sara J, Hammeke, Thomas A, & Sabsevitz, David S. (2008). A comparison of five fMRI protocols for mapping speech comprehension systems. *Epilepsia, 49*(12), 1980-1997.

Binder, JR, Frost, JA, Hammeke, TA, Bellgowan, PSF, Springer, JA, Kaufman, JN, & Possing, ET. (2000). Human temporal lobe activation by speech and nonspeech sounds. *Cerebral Cortex, 10*(5), 512-528.

Blesser, Barry. (1972). Speech perception under conditions of spectral transformation: I. Phonetic characteristics. *Journal of Speech, Language and Hearing Research, 15*(1), 5.

Bruce, Charles, Desimone, Robert, & Gross, CHARLES G. (1981). Visual properties of neurons in a polysensory area in superior temporal sulcus of the macaque. *J Neurophysiol, 46*(2), 369-384.

Callan, D. E., Jones, J. A., Munhall, K., Callan, A. M., Kroos, C., & Vatikiotis-Bateson, E. (2003). Neural processes underlying perceptual enhancement by visual speech gestures. *Neuroreport, 14*(17), 2213-2218. doi: 10.1097/01.wnr.0000095492.38740.8f

Calvert, Gemma A, Brammer, Michael J, Bullmore, Edward T, Campbell, Ruth, Iversen, Susan D, & David, Anthony S. (1999). Response amplification in sensory-specific cortices during crossmodal binding. *Neuroreport, 10*(12), 2619.

Calvert, Gemma A, Bullmore, Edward T, Brammer, Michael J, Campbell, Ruth, Williams, Steven CR, McGuire, Philip K, . . . David, Anthony S. (1997). Activation of auditory cortex during silent lipreading. *Science, 276*(5312), 593-596.

Calvert, Gemma A, Campbell, Ruth, & Brammer, Michael J. (2000). Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. *Current Biology, 10*(11), 649-658.

Campbell, Ruth, MacSweeney, Mairead, Surguladze, Simon, Calvert, Gemma, McGuire, Philip, Suckling, John, . . . David, Anthony S. (2001). Cortical substrates for the perception of face actions: an fMRI study of the specificity of activation for seen speech and for meaningless lower-face acts (gurning). *Cognitive Brain Research, 12*(2), 233-243.

Coutanche, Marc N. (2013). Distinguishing multi-voxel patterns and mean activation: Why, how, and what does it tell us? *Cognitive, Affective, & Behavioral Neuroscience, 13*(3), 667-673.

Dahl, C. D., Logothetis, N. K., & Kayser, C. (2009). Spatial organization of multisensory responses in temporal association cortex. *J Neurosci, 29*(38), 11924-11932. doi: 10.1523/JNEUROSCI.3437-09.2009

Dodd, Barbara. (1977). The role of vision in the perception of speech. *Perception, 6*(1), 31-40.

Driver, Jon, & Spence, Charles. (2000). Multisensory perception: beyond modularity and convergence. *Current Biology, 10*(20), R731-R735.

Fonov, Vladimir, Evans, Alan C, Botteron, Kelly, Almli, C Robert, McKinstry, Robert C, & Collins, D Louis. (2011). Unbiased average age-appropriate atlases for pediatric studies. *NeuroImage, 54*(1), 313-327.

Fonov, VS, Evans, AC, McKinstry, RC, Almli, CR, & Collins, DL. (2009). Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. *NeuroImage, 47*, S102.

Ghazanfar, Asif A, Maier, Joost X, Hoffman, Kari L, & Logothetis, Nikos K. (2005). Multisensory integration of dynamic faces and voices in rhesus monkey auditory cortex. *The Journal of Neuroscience, 25*(20), 5004-5012.

Green, David Marvin, & Swets, John A. (1966). *Signal detection theory and psychophysics* (Vol. 1): Wiley New York.

Green, Kerry P. (1998). The use of auditory and visual information during phonetic processing: Implications for theories of speech perception. *Hearing by eye II*, 3-26.

Grossman, Emily D, Battelli, Lorella, & Pascual-Leone, Alvaro. (2005). Repetitive TMS over posterior STS disrupts perception of biological motion. *Vision research, 45*(22), 2847-2853.

Grossman, Emily D, & Blake, Randolph. (2002). Brain areas active during visual perception of biological motion. *Neuron, 35*(6), 1167-1175.

Grossman, Emily D, Jardine, Nicole L, & Pyles, John A. (2010). fMR-adaptation reveals invariant coding of biological motion on the human STS. *Front Hum Neurosci, 4*.

Grossman, Emily, Donnelly, M, Price, R, Pickens, D, Morgan, V, Neighbor, G, & Blake, R. (2000). Brain areas involved in perception of biological motion. *Journal of Cognitive Neuroscience, 12*(5), 711-720.

Hein, Grit, & Knight, Robert T. (2008). Superior Temporal Sulcus-It's My Area: Or Is It? *Journal of Cognitive Neuroscience, 20*(12), 2125-2136.

Hickok, G., & Poeppel, D. (2004). Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language. *Cognition, 92*(1-2), 67-99. doi: 10.1016/j.cognition.2003.10.011

Hickok, Gregory, & Poeppel, David. (2007). The cortical organization of speech processing. *Nat Rev Neurosci, 8*(5), 393-402.

James, TW, & Stevenson, RA. (2012). The use of fMRI to assess multisensory integration.

Kaas, Jon H, & Hackett, Troy A. (2000). Subdivisions of auditory cortex and processing streams in primates. *Proceedings of the National Academy of Sciences, 97*(22), 11793-11799.

Kriegeskorte, Nikolaus, Simmons, W Kyle, Bellgowan, Patrick SF, & Baker, Chris I. (2009). Circular analysis in systems neuroscience: the dangers of double dipping. *Nature neuroscience, 12*(5), 535-540.

Lestou, Vaia, Pollick, Frank E, & Kourtzi, Zoe. (2008). Neural substrates for action understanding at different description levels in the human brain. *Journal of Cognitive Neuroscience, 20*(2), 324-341.

Lewis, James W, & Van Essen, David C. (2000). Corticocortical connections of visual, sensorimotor, and multimodal processing areas in the parietal lobe of the macaque monkey. *The Journal of comparative neurology, 428*(1), 112-137.

Liebenthal, E., Binder, J. R., Spitzer, S. M., Possing, E. T., & Medler, D. A. (2005). Neural substrates of phonemic perception. *Cereb Cortex, 15*(10), 1621-1631. doi: 10.1093/cercor/bhi040

Liebenthal, Einat, Desai, Rutvik, Ellingson, Michael M, Ramachandran, Brinda, Desai, Anjali, & Binder, Jeffrey R. (2010). Specialization along the left superior temporal sulcus for auditory categorization. *Cerebral Cortex, 20*(12), 2958-2970.

Lunneborg, Clifford E. (2000). *Data analysis by resampling: Concepts and applications*: Duxbury Pacific Grove, CA.

Massaro, D.W. (1987). *Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry*: Erlbaum Associates.

McGurk, Harry, & MacDonald, John. (1976). Hearing lips and seeing voices.

Mumford, Jeanette A, Turner, Benjamin O, Ashby, F Gregory, & Poldrack, Russell A. (2012). Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses. *NeuroImage, 59*(3), 2636-2643.

Mur, Marieke, Bandettini, Peter A, & Kriegeskorte, Nikolaus. (2009). Revealing representational content with pattern-information fMRI—an introductory guide. *Social cognitive and affective neuroscience, 4*(1), 101-109.

Narain, C, Scott, Sophie K, Wise, Richard JS, Rosen, Stuart, Leff, Alexander, Iversen, SD, & Matthews, PM. (2003). Defining a left-lateralized response specific to intelligible speech using fMRI. *Cerebral Cortex, 13*(12), 1362-1368.

Nath, Audrey R, & Beauchamp, Michael S. (2011). Dynamic changes in superior temporal sulcus connectivity during perception of noisy audiovisual speech. *The Journal of Neuroscience, 31*(5), 1704-1714.

Nath, Audrey R, & Beauchamp, Michael S. (2012). A neural basis for interindividual differences in the McGurk effect, a multisensory speech illusion. *Neuroimage, 59*(1), 781-787.

Nichols, Thomas, Brett, Matthew, Andersson, Jesper, Wager, Tor, & Poline, Jean-Baptiste. (2005). Valid conjunction inference with the minimum statistic. *Neuroimage, 25*(3), 653-660.

Okada, K., & Hickok, G. (2009). Two cortical mechanisms support the integration of visual and auditory speech: a hypothesis and preliminary data. *Neurosci Lett, 452*(3), 219-223. doi: 10.1016/j.neulet.2009.01.060

Okada, K., Rong, F., Venezia, J., Matchin, W., Hsieh, I. H., Saberi, K., . . . Hickok, G. (2010). Hierarchical organization of human auditory cortex: evidence from acoustic invariance in the response to intelligible speech. *Cereb Cortex, 20*(10), 2486-2495. doi: 10.1093/cercor/bhp318

Okada, Kayoko, & Hickok, Gregory. (2006). Identification of lexical–phonological networks in the superior temporal sulcus using functional magnetic resonance imaging. *Neuroreport, 17*(12), 1293-1296.

Okada, Kayoko, Venezia, Jonathan H, Matchin, William, Saberi, Kourosh, & Hickok, Gregory. (2013). An fMRI Study of Audiovisual Speech Perception Reveals Multisensory Interactions in Auditory Cortex. *PloS one, 8*(6), e68959.

Poldrack, Russell A, & Mumford, Jeanette A. (2009). Independence in ROI analysis: where is the voodoo? *Social Cognitive and Affective Neuroscience, 4*(2), 208-213.

Price, Cathy J. (2010). The anatomy of language: a review of 100 fMRI studies published in 2009. *Annals of the New York Academy of Sciences, 1191*(1), 62-88.

Puce, Aina, Allison, Truett, Bentin, Shlomo, Gore, John C., & McCarthy, Gregory. (1998). Temporal Cortex Activation in Humans Viewing Eye and Mouth Movements. *The Journal of Neuroscience, 18*(6), 2188-2199.

Puce, Aina, & Perrett, David. (2003). Electrophysiology and brain imaging of biological motion. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 358*(1431), 435-445. doi: 10.1098/rstb.2002.1221

Puce, Aina, Syngeniotis, Ari, Thompson, James C, Abbott, David F, Wheaton, Kylie J, & Castiello, Umberto. (2003). The human temporal lobe integrates facial form and motion: evidence from fMRI and ERP studies. *Neuroimage, 19*(3), 861-869.

Rauschecker, Josef P, Tian, Biao, & Hauser, Marc. (1995). Processing of complex sounds in the macaque nonprimary auditory cortex. *Science; Science*.

Reisberg, Daniel, Mclean, John, & Goldfield, Anne. (1987). Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli.

Rosenblum, Lawrence D, Pisoni, DB, & Remez, R. (2005). Primacy of multimodal speech perception. *Handbook of speech perception*, 51-78.

Schwartz, Jean-Luc, Robert-Ribes, Jordi, & Escudier, Pierre. (1998). Ten years after Summerfield: a taxonomy of models for audio-visual fusion in speech perception. *Hearing by eye II: Advances in the psychology of speechreading and auditory-visual speech*, 85-108.

Scott, Sophie K, Blank, C Catrin, Rosen, Stuart, & Wise, Richard JS. (2000). Identification of a pathway for intelligible speech in the left temporal lobe. *Brain, 123*(12), 2400-2406.

Seltzer, Benjamin, & Pandya, Deepak N. (1978). Afferent cortical connections and architectonics of the superior temporal sulcus and surrounding cortex in the rhesus monkey. *Brain Res, 149*(1), 1.

Seltzer, Benjamin, & Pandya, Deepak N. (1994). Parietal, temporal, and occipita projections to cortex of the superior temporal sulcus in the rhesus monkey: A retrograde tracer study. *The Journal of comparative neurology, 343*(3), 445-463.

Skipper, J. I., van Wassenhove, V., Nusbaum, H. C., & Small, S. L. (2007). Hearing lips and seeing voices: how cortical areas supporting speech production mediate audiovisual speech perception. *Cereb Cortex, 17*(10), 2387-2399. doi: 10.1093/cercor/bhl147

Specht, Karsten, Osnes, Berge, & Hugdahl, Kenneth. (2009). Detection of differential speech-specific processes in the temporal lobe using fMRI and a dynamic "sound morphing" technique. *Hum Brain Mapp, 30*(10), 3436-3444.

Stevenson, Ryan A, Altieri, Nicholas A, Kim, Sunah, Pisoni, David B, & James, Thomas W. (2010). Neural processing of asynchronous audiovisual speech perception. *Neuroimage, 49*(4), 3308-3318.

Stevenson, Ryan A, & James, Thomas W. (2009). Audiovisual integration in human superior temporal sulcus: inverse effectiveness and the neural processing of speech and object recognition. *Neuroimage, 44*(3), 1210-1223.

Stevenson, Ryan A, VanDerKlok, Ross M, Pisoni, David B, & James, Thomas W. (2011). Discrete neural substrates underlie complementary audiovisual speech integration processes. *Neuroimage, 55*(3), 1339-1345.

Sumby, William H, & Pollack, Irwin. (1954). Visual contribution to speech intelligibility in noise. *The Journal of the Acoustical Society of America, 26*(2), 212-215.

Summerfield, Q. (1987). Some preliminaries to a comprehensive account of audio-visual speech perception. In Dodd (Ed.), *Hearing by eye: The psychology of lip-reading*: Lawrence Erlbaum Associates.

Szycik, Gregor Rafael, Tausche, Peggy, & Münte, Thomas F. (2008). A novel approach to study audiovisual integration in speech perception: localizer fMRI and sparse sampling. *Brain research, 1220*, 142-149.

Turner, Benjamin O, Mumford, Jeanette A, Poldrack, Russell A, & Ashby, F Gregory. (2012). Spatiotemporal activity estimation for multivoxel pattern analysis with rapid event-related designs. *NeuroImage, 62*(3), 1429-1438.

Vaden Jr, Kenneth I, Muftuler, L Tugan, & Hickok, Gregory. (2010). Phonological repetition-suppression in bilateral superior temporal sulci. *Neuroimage, 49*(1), 1018-1023.

van Wassenhove, Virginie, Grant, Ken W, & Poeppel, David. (2005). Visual speech speeds up the neural processing of auditory speech. *Proceedings of the National Academy of Sciences of the United States of America, 102*(4), 1181-1186.

Vapnik, Vladimir. (1999). *The nature of statistical learning theory*: springer.

Wessinger, CM, VanMeter, J, Tian, B, Van Lare, J, Pekar, J, & Rauschecker, JP. (2001). Hierarchical organization of the human auditory cortex revealed by functional magnetic resonance imaging. *Journal of Cognitive Neuroscience, 13*(1), 1-7.

Wright, Tarra M, Pelphrey, Kevin A, Allison, Truett, McKeown, Martin J, & McCarthy, Gregory. (2003). Polysensory interactions along lateral temporal regions evoked by audiovisual speech. *Cerebral Cortex, 13*(10), 1034-1043.

CHAPTER 5

Primer

Recall from Ch. 1 that auditory speech sound representations are engaged not only in perception of speech but also during speech production. Specifically, speech sound representations serve as the sensory "targets" for speech production. Chapters 3 and 4 were spent examining the mechanisms underlying audiovisual speech perception under the assumption that an understanding of how information from multiple modalities is combined could reveal general principles of organization with respect to perceptual speech systems. The current chapter applies the same approach to speech production. Namely, this chapter asks whether visual speech is integrated with the speech motor system in the same way as auditory speech. The motivation behind this question is a recent finding (details to follow) that some non-fluent aphasics recover a great deal of their productive speech capacity when following along with a *video* of a talker's face during an auditory repetition task. These patients often have extensive damage to motor and sensorimotor-integration systems for speech, and as a result they may struggle to produce more than a few words per utterance. Yet, some of these patients can mimic audiovisual stimuli enabling them to produce fluent speech in real time. This same effect does not hold for audio- or visual-only speech stimuli. Thus, it seems that auditory and visual speech signals can be combined to somehow access impoverished (but spared) motor speech commands in (some) non-fluent aphasics. How does this happen? Visual speech must have access to the speech motor system. Perhaps visual speech accesses the motor system by combining synergistically with auditory speech within canonical auditory-motor integration networks. But, as mentioned, these networks are often damaged in non-fluent aphasics. Another possibility is that dedicated networks exist to connect visual speech with vocal tract control mechanisms. In this chapter, I will present the results of an fMRI study that uses a covert rehearsal task to map sensorimotor integration networks for auditory, visual, and audiovisual speech inputs. Understanding the role of visual speech in production will provide insight into the general organization of vocal tract control mechanisms, and an idea of precisely how these mechanisms function to support speech production.

Perception drives production across sensory modalities: A network for sensorimotor integration of visual speech

Jonathan H. Venezia, Paul Fillmore, Lisette Isenberg, William Matchin, Gregory Hickok and Julius Fridriksson

Introduction

Visual speech refers to the motion and configuration cues associated with watching a talker's head, face and mouth during articulation. The neuro-computational role of visual speech is often couched in terms of its influence on auditory speech perception. This is not surprising given that a very large proportion of visual speech research has focused on the perceptual effects induced by adding visual speech to an auditory speech signal, which include improved intelligibility for speech in noise (Erber, 1969; MacLeod & Summerfield, 1987; McCORMICK, 1979; Neely, 1956; Ross, Saint-Amour, Leavitt, Javitt, & Foxe, 2007; Sumby & Pollack, 1954) or even for speech with undistorted acoustics (Arnold & Hill, 2001), alteration of auditory syllable identity (Massaro, 1998; McGurk & MacDonald, 1976), and improved acquisition of non-native speech sound categories (Hardison, 2003). Conversely, research on auditory speech has focused not only on its fundamental contribution to perception, but also on the crucial role of auditory speech systems in supporting speech production. Specifically, evidence suggests that auditory speech representations serve as the sensory "targets" for speech production. According to current theory, speech sound representations constitute both the initial goals and end-stage consequences of motor speech output, and they are integrated with motor systems via a dorsal sensorimotor processing stream (Guenther, 2006; Gregory Hickok, 2012, 2014; Gregory Hickok, Houde, & Rong, 2011; Gregory Hickok & Poeppel, 2007; Indefrey & Levelt, 2004; Tourville, Reilly, & Guenther, 2008). Evidence that auditory speech supports production includes articulatory decline in adult-onset deafness (Waldstein, 1990), disruption of speech output by delayed auditory feedback (Stuart, Kalinowski, Rastatter, & Lynch, 2002; Yates, 1963), and compensation for altered auditory feedback (Burnett, Freedland, Larson, & Hain, 1998; Purcell & Munhall, 2006).

There is also evidence that visual speech plays at least a complementary role in supporting speech production. A classic study (Reisberg, Mclean, & Goldfield, 1987) that is in fact frequently cited to support claims that visual speech increases auditory intelligibility actually suggests that audiovisual speech facilitates production. In this study, subjects were asked to *shadow* (listen to and immediately repeat word-by-word) spoken passages that were easy to hear but hard to understand – specifically, passages were spoken in a recently acquired second language, spoken in accented English, or drawn from semantically and syntactically complex content. The dependent variable was the tracking (speech production) rate in words per minute, and this rate significantly increased when spoken passages were accompanied by concurrent visual speech.

Circumstantial evidence that visual speech supports production can be drawn from recent neurophysiological research indicating that visual and audiovisual speech activate the speech motor system (Callan et al., 2003; Hasson, Skipper, Nusbaum, & Small, 2007; Ojanen et al., 2005; K. Okada & Hickok, 2009; Skipper, van Wassenhove, Nusbaum, & Small, 2007; Watkins, Strafella, & Paus, 2003). Although this evidence is often interpreted as supporting a role for the motor system in visual or multimodal speech *perception* (Hasson et al., 2007; Möttönen & Watkins, 2012; Schwartz, Basirat, Ménard, & Sato, 2012), the reverse relation – visual speech supports *production* – is perhaps equally plausible. This notion motivated some of our own research examining the effects of visual speech on production. Upon observing that visual and auditory speech perception activate the speech motor system (Fridriksson et al., 2008; Meister, Wilson, Deblieck, Wu, & Iacoboni, 2007; Rorden, Davis, George, Borckardt, & Fridriksson, 2008), we hypothesized that perceptual training with audiovisual speech would improve the speech output of nonfluent aphasics. Indeed, when patients were trained on a word-picture matching task, significant improvement in subsequent picture naming was observed, but only when the training phase included audiovisual words (Fridriksson et al., 2009). We later discovered a striking effect we termed "Speech Entrainment" (SE), in which shadowing of audiovisual speech allowed patients with nonfluent aphasia to increase their speech output by a factor of two or more (Fridriksson et al., 2012). This effect was not observed for shadowing of auditory- or visual-only speech, which suggests the following: motor commands for speech are relatively intact in some cases of nonfluent aphasia, and visual speech when combined with auditory speech provides crucial information allowing access to these motor commands.

As such, current evidence points to the conclusion that visual speech plays a role in speech motor control, at least in some situations. To be specific, we assume that the noted behavioral increases in speech output following exposure to audiovisual speech reflect the addition of a complementary set of visual speech “targets” that combine with auditory speech “targets” to facilitate speech motor control. Evidence from other domains demonstrates unambiguously that auditory and visual signals interact to support motor control. The canonical animal model for audiovisual integration at the cellular level, the cat superior colliculus (Meredith, Nemitz, & Stein, 1987; Meredith & Stein, 1983; Stein & Stanford, 2008), is in fact a sensorimotor structure involved extensively in oculomotor control. An extensive body of evidence demonstrates that audiovisual integration facilitates the latency and accuracy of saccades and manual movements in humans (Colonius & Arndt, 2001; Corneil, Van Wanrooij, Munoz, & Van Opstal, 2002; Diederich & Colonius, 2004; Frens, Van Opstal, & Van der Willigen, 1995; Hughes, Reuter-Lorenz, Nozawa, & Fendrich, 1994), and in nonhuman primates it has been demonstrated that audiovisual responses in the superior colliculus drive such facilitation for saccades (Bell, Meredith, Van Opstal, & Munoz, 2005). Posterior parietal visuomotor integration regions that support saccades, reaching, and grasping in primates also have multisensory properties (Andersen, 1997; Cohen & Andersen, 2002).

A straightforward hypothesis concerning the mechanism for sensorimotor integration of visual speech is that the visual signals are first translated to an auditory-phonological code. This would grant visual speech indirect access to the speech motor system via auditory dorsal stream networks. Another possibility is that visual speech activates sensorimotor speech networks directly. Indeed, speech-reading (perceiving visual speech) activates both multimodal sensory speech regions in the posterior superior temporal lobe and a well-known sensorimotor integration region for speech (Spt) in left posterior Sylvian cortex (K. Okada & Hickok, 2009). A third possibility is that dedicated sensorimotor networks exist for visual speech. Speech Entrainment provides indirect support for this position – namely, the addition of visual speech improves speech output in a population for which canonical auditory-motor integration networks are often extensively damaged (Fridriksson, Fillmore, Guo, & Rorden, 2014; Fridriksson et al., 2012). In the current study, we attempted to disambiguate among these possibilities by studying the organization of sensorimotor networks for auditory, visual, and audiovisual speech. Specifically, we used BOLD fMRI to test whether covert repetition of visual or audiovisual speech: (1) increased activation in known auditory-motor circuits (relative to repetition of auditory-only speech), (2) activated sensorimotor pathways unique to visual speech, (3) both, or (4) neither. Covert repetition is often employed to identify auditory-to-vocal-tract networks (Buchsbaum, Hickok, & Humphries, 2001; Gregory Hickok, Buchsbaum, Humphries, & Muftuler, 2003; Kayoko Okada & Hickok, 2006; Rauschecker, Pringle, & Watkins, 2008; Wildgruber, Ackermann, & Grodd, 2001), where a typical paradigm involves presenting participants with blocks of auditory non-words in each of the following conditions: perception followed by covert rehearsal (P+Reh), perception followed by rest (P+Rest), and continuous perception (CP). P+Reh is the task of interest, and regions involved the Motor phase of the task are isolated by the contrast P+Reh > P+Rest, while regions involved in the Sensory phase are isolated by the contrast CP > baseline, and the conjunction of the two contrasts identifies Sensorimotor areas. Adapting this paradigm, we asked whether using visual (V) or audiovisual (AV) stimuli as the input recruited different sensorimotor networks versus auditory-only (A) input. We computed the Sensory contrast, Motor contrast, and Sensorimotor conjunction for each of the input modalities separately. To explore the space of hypotheses listed above, i.e., (1), (2), (3), and (4), we compared the Sensorimotor networks for each modality and also identified regions demonstrating an increased Motor response for V or AV input relative to A. We also assessed whether including visual speech in the input produced effects (if any) only in the AV condition, or whether such effects could also be observed for V alone. To the best of our knowledge, this is the first time this type of sensorimotor-speech design has been applied comprehensively to multiple input modalities within the same group of participants.

Materials and Methods

Participants

Twenty (16 female) right-handed native English speakers between 20 and 30 years of age participated in the study. All volunteers had normal or corrected-to-normal vision, normal hearing by self-report, no known history of neurological disease, and no other contraindications for MRI. Informed consent was obtained from each participant in accordance with University of South Carolina Institutional Review Board guidelines.

Stimuli and Procedure

Forty-five digital video clips (3s duration, 30 fps) were produced featuring a single male actor shown from the neck up. In each clip, the actor produced a sequence of four consonant-vowel (CV) syllables drawn from a set of six visually distinguishable CVs – \ba\, \tha\, \va\, \bi\, \thi\, \vi\. The CVs were articulated as a continuous sequence with the onset of each component syllable timed to a visual metronome at 2 Hz. Each of the six CVs appeared exactly 30 times across all 45 clips and any given CV was never repeated in a sequence. Otherwise, the ordinal position of each CV within a sequence was selected at random. Videos were recorded in a single session against a 'green screen' at 720p resolution and post-processed in Final Cut Pro 7 (Apple Inc.). The green screen was replaced with a uniform gray background, individual clips were cut to 3s duration with an equal number of frames preceding and following articulation, and clips were cropped and compressed to 640x480 pixels. The concurrent auditory speech signals were recorded on a separate microphone during the video recording session, digitized (44.1 kHz, 16-bit mono), and synced manually with the video recordings using Audacity software. Auditory stimuli were normalized to equal root-mean-square amplitude.

Syllable sequences were presented to participants in each of three modalities (Figure 5.1): auditory-only (A), visual-only (V) and audiovisual (AV). In the A modality, clips consisted of a still frame of the actor's face paired with auditory recordings of CV syllable sequences. In the V modality, videos of the actor producing syllable sequences were presented without sound. In the AV modality, videos of the actor producing syllable sequences were presented along with the concurrent auditory speech signal. In addition, syllable sequences were presented in three experimental conditions: perceive and rehearse (P+Reh), perceive and rest (P+Rest), and continuous perception (CP). A single trial in each condition comprised a 10s period (Figure 5.1): a visual cue indicating the condition (1.5s) followed by a blank gray screen (uniformly jittered duration, 0.5-2s), stimulation (6s), and then a black fixation "X" on the gray background (remainder of 10s). Only the 6s stimulation period varied by condition. In the P+Reh condition, participants were asked to "perceive" (watch, listen to, or both) syllable sequences (3s) and then covertly rehearse (single repetition) the "perceived" sequence in the period immediately following (3s). In the P+Rest condition, participants were asked to perceive syllable sequences (3s) followed immediately by a period of rest (3s without covert articulation). In the CP condition, participants were asked to perceive a syllable sequence that was presented twice so as to fill the entire stimulation period (6s). There were also rest trials in which the black fixation "X" was presented for the entire 10s.

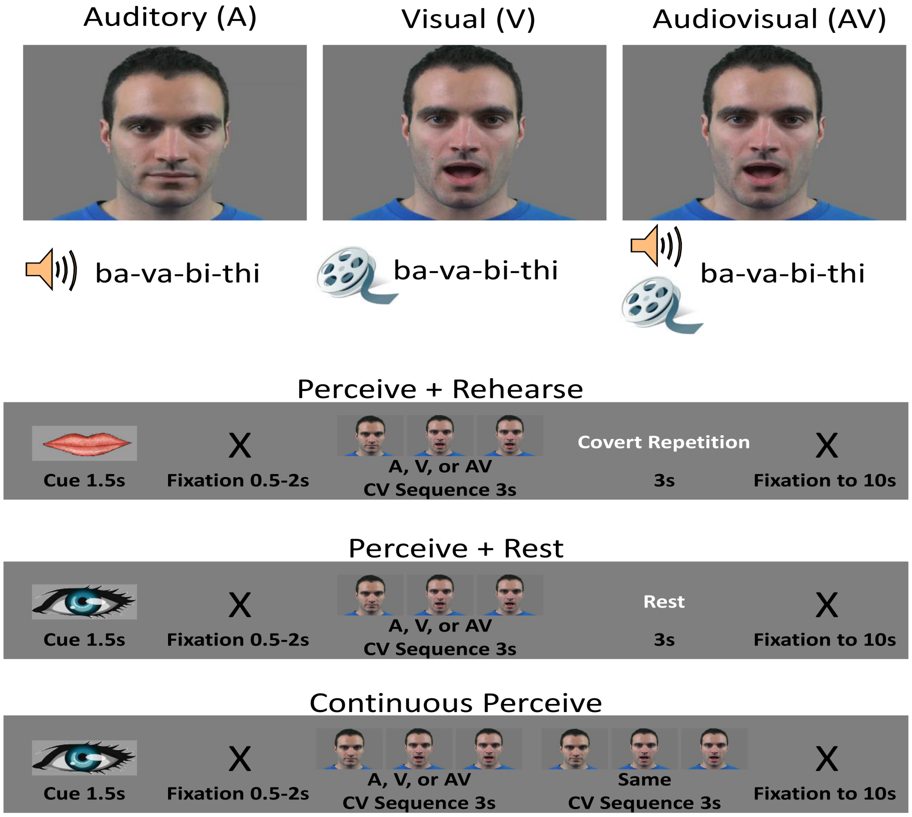


Figure 5.1. Design schematic of multimodal sensorimotor speech task. Input modality (top) was crossed with condition (bottom). Each run contained 30 trials from a given input modality – A, V or AV – and 10 rest trials (not pictured). Of the 30 trials, 10 each were perceive+rehearse, perceive+rest, and continuous perceive. The perceive+rehearse trials were cued by an image of lips at the onset of the trial, and the perceive+rest and continuous perceive trials were cued by an image of an eye at the onset of the trial. Trial structures are shown for each condition. White text indicates what subjects were actually doing, rather than instructions on screen. Stimuli were 3s CV syllable sequences drawn from the set of visually distinguishable CVs /ba/, /bi/, /tha/, /thi/, /va/, /vi/. The CV sequence shown is just one possible example.

Functional imaging runs consisted of 40 trials, 10 from each condition and 10 rest trials. Runs were blocked by modality – that is, of nine functional runs, there were three A runs, three V runs, and three AV runs, presented in pseudo-random order (the same modality was never repeated more than once and each participant encountered a different presentation order). Within each run, trials from each condition were presented in pseudo-random order (same condition never repeated more than once and rest trials were never repeated). The 45 syllable sequences in our stimulus set (see above) were presented twice in each modality (total of 90 trials, 30 in each condition). The same sequence was never presented twice in a given run and sequences were balanced across the first and second halves of the experiment: within a given modality, all 45 sequences appeared once during the initial 45 trials and once during the final 45 trials. The sequence order was random otherwise. Participants were scanned for nine functional runs immediately followed by acquisition of a high-resolution T1 anatomical volume. Stimulus delivery and timing were controlled using the Psychtoolbox-3 (Kleiner et al., 2007) implemented in Matlab (Mathworks Inc., USA).

Scanning Parameters

MR images were obtained on a Siemens 3T fitted with a 12-channel head coil and an audio-visual presentation system. We collected a total of 1872 echo planar imaging (EPI) volumes per subject over 9 runs using single pulse Gradient Echo EPI (matrix = 104 x 104, repetition time [TR] = 2s, echo time [TE] = 30ms, size = 2 x 2 x 3.75 mm, flip angle = 76). Thirty-one sequentially acquired axial slices provided whole brain coverage. After the functional scans, a high-resolution T1 anatomical image was acquired in the sagittal plane (1 mm^3).

Imaging Analysis – Study-Specific Anatomical Template Construction

A study-specific anatomical template image was created using symmetric diffeomorphic registration (SyN) in the Advanced Normalization Tools (ANTS) software (B. Avants et al., 2008; B. Avants & Gee, 2004). Each participant's whole-head T1 anatomical image was submitted to the template-construction processing stream in ANTS (buildtemplateparallel.sh), which comprises rigid and SyN registration steps. For SyN, we used a cross correlation similarity metric (B. B. Avants et al., 2011) with a three-level multi-resolution registration with 50x70x10 iterations. The output of this registration process was a whole-head T1 template approximating the group average shape and intensity. The whole-head template was skull stripped in ANTS (antsBrainExtraction.sh) via registration with the prebuilt NKI template with probabilistic brain mask (Avants, Brian; Tustison, Nick (2014): ANTs/ANTsR Brain Templates. Fig**share**. http://dx.doi.org/10.6084/m9.figshare.915436). A brain+cerebellum mask of the skull-stripped template was inverse-warped to each participant's native space and used to skull strip the individual participant T1 images. These skull-stripped images were then re-registered to the skull-stripped template using SyN (antsRegistration) in order to improve registration accuracy. Finally, the skull-stripped template was aligned to the MNI152-space ICBM template (Vladimir Fonov et al., 2011; VS Fonov, Evans, McKinstry, Almli, & Collins, 2009) using a 12-parameter affine registration in ANTS. The complete set of affine and diffeomorphic transformations mapping each participant's T1 anatomical to the study-specific T1 template, and the affine transformation mapping the study-specific T1 template to MNI space, were later used to bring each participant's functional data into alignment with the study-specific template in MNI space.

Imaging Analysis –fMRI

Preprocessing of the data was performed using AFNI software (http://afni.nimh.nih.gov/afni). For each run, slice timing correction was performed followed by realignment (motion correction) and coregistration of the EPI images to the high resolution anatomical image in a single interpolation step. Functional data were then warped to the study-specific template in MNI space using the set of transforms defined in ANTS. Finally, images were spatially smoothed with an isotropic 6-mm full-width half-maximum (FWHM) Gaussian kernel and each run was scaled to have a mean of 100 across time at each voxel.

First level regression analysis (AFNI 3dREMLfit) was performed in individual subjects. The hemodynamic response function (HRF) for events from each cell of the design was estimated using a cubic spline (CSPLIN) function expansion with 8 parameters modeling the response from 2 to 16s after stimulation onset (spacing = 1TR). The HRF was assumed to start (0s post-stimulation) and end (18s post-stimulation) at zero. Thus, a total of 72 regressors were used to model the HRF from each of the 9 event types in the experiment: A P+Reh, A P+Rest, A CP, V P+Reh, V P+Rest, V CP, AV P+Reh, AV P+Rest, AV CP. The amplitude of the response was calculated by averaging the HRF values from 6-10s post-stimulation. These amplitude estimates were fed to the 2nd level for group analysis. The "cue" periods from each trial were modeled as a single regressor of no interest corresponding to an event timing vector convolved with a canonical hemodynamic response function. Rest trials were not modeled explicitly and were thus included in the baseline term. An additional twelve regressors corresponding to motion parameters determined during the realignment stage of preprocessing along with their temporal derivatives were entered into the model. Individual time points were censored from analysis when more than 10% of in-brain voxels were identified as outliers (AFNI 3dToutcount) or when the Euclidean norm of the motion derivatives exceeded 0.4.

A second-level mixed effects analysis (Chen, Saad, Nath, Beauchamp, & Cox, 2012) was performed on the HRF amplitude estimates from each participant, treating 'participant' as a random effect. This procedure is similar to a standard group-level *t*-test but also takes into account the level of intra-subject variation by accepting *t*-scores from each individual subject analysis. Statistical parametric maps (t-statistics) were created for each contrast of interest. Active voxels were defined as those for which t-statistics exceeded the $p < 0.005$ level with a cluster extent threshold of 173 voxels. This cluster threshold was determined by Monte Carlo simulation (AFNI 3dClustSim) to hold the family-wise error rate (FWER) less than 0.05 (i.e., corrected for multiple comparisons). Estimates of smoothness in the data were drawn from the residual error time series for each participant after first-level analysis (AFNI 3dFWHMx). These estimates were averaged across participants separately in each voxel dimension for input to 3dClustSim. Simulations were restricted to in-brain voxels.

We performed two group-level contrasts to identify different components of speech-related sensorimotor brain networks. The first, which we term the 'Sensory' contrast, tested for activation greater in the CP condition than baseline (CP > Rest) and was intended to identify brain regions involved in the sensory phase of the perceive+rehearse task. The second, which we term the 'Motor' contrast, tested for activation greater in the P+Reh condition than the P+Rest condition (P+Reh > P+Rest). This contrast factored out activation to the sensory phase and was thus intended to identify brain regions involved in the (covert) motor phase of the perceive+rehearse task. Finally, Sensorimotor brain regions (i.e., those involved in both phases of the task) were identified by performing a conjunction of the Sensory and Motor contrasts (CP > rest ∩ P+Reh > P+Rest). The conjunction analysis was performed by constructing minimum *t*-maps (e.g., minimum T score from [Sensory, Motor] at each voxel) thresholded at $p < 0.005$ with a cluster extent threshold of 173 voxels (FWER < 0.05, as for individual condition maps). The Sensory, Motor, and Sensorimotor analyses were performed separately for each modality to form a total of 9 group-level SPMs: A-Sensory, A-Motor, A-Sensorimotor, V-Sensory, V-Motor, V-Sensorimotor, AV-Sensory, AV-Motor, AV-Sensorimotor. We tested directly for differences in motor activation across input modalities by performing two interaction contrasts: VvsA-Motor (V-Motor > A-Motor) and AvvsA-Motor (AV-Motor > A-Motor).

Activations were visualized on the Conte69 atlas in MNI152 space in CARET v5.65 (http://brainvis.wustl.edu/wiki/index.php/Caret:Download), or on the study-specific template in MNI152 space in AFNI. Displayed group-average time-course plots were formed by taking the average of individual subject HRF regression parameters at each time point and performing cubic spline interpolation with a 0.1s time step.

Results

Auditory-motor integration networks for the vocal tract have previously been identified quite reliably using a standard imaging paradigm in which subjects listen to and covertly rehearse sequences of auditory nonwords. In particular, auditory-motor integration regions are identified by testing for voxels that respond significantly to both the listen (Sensory) and rehearsal (Motor) phases of the task. Here, we have extended this design to multiple input modalities: in addition to auditory (A) speech (CV syllables), participants were asked to perceive and covertly rehearse visual (V) and audiovisual (AV) speech.

We hypothesized the following with respect to our multimodal perceive+rehearse task: (1) Additional sensorimotor brain regions (i.e., outside canonical auditory-motor networks as assessed in the A modality) will be recruited when the input modality is V, AV, or both; (2) Additional motor activation will be observed (either within canonical auditory-motor regions or in visual-specific regions) when the input modality is V, AV, or both (i.e., there will be motor activation over and above that observed for auditory-only input). The motivation for (1) and (2) is behavioral work (detailed above) reporting an increase in speech output when the task is to repeat a stimulus that contains visual speech – specifically, we hypothesized that these behavioral improvements are driven by recruitment of unique visual-to-motor speech pathways that lead to an increased motor response (the presumed neural correlate of behavioral improvements, although we do not test this directly). To assess (1), we simply observed differences in Sensorimotor maps across modalities qualitatively. To assess (2), we tested directly for differences in Motor (P+Reh > P+Rest) activation across modalities by performing two interaction contrasts: AvvsA-Motor (AV-Motor > A-Motor) and VvsA-Motor (V-Motor > A-Motor). These interaction contrasts were designed to identify brain regions demonstrating an enhanced motor response while factoring out activations to the sensory phase of the task in each modality.

Sensorimotor speech networks for multiple input modalities

In this section we report the results of the Sensorimotor conjunction analysis designed to identify sensorimotor integration networks in each of three input modalities: A, V, and AV. This conjunction analysis tested for voxels showing a significant response in both Sensory and Motor phases of the perceive+rehearse task. It should be noted for the purposes of viewing activation time-courses that estimates of the hemodynamic response are based on 6s of effective stimulation for the Sensory phase and only 3s of effective stimulation for the Motor phase (refer to the Sensory and Motor contrasts in the Methods).

The A-Sensorimotor map (Figure 5.2, top left) comprised significant clusters in canonical motor-speech brain regions including the left inferior frontal gyrus/frontal operculum (IFG), the left precentral gyrus (PreM), and bilateral supplementary motor area (SMA; preponderance of activation in the left hemisphere). In addition, there were significant clusters in the left posterior Sylvian region (Spt), right PreM, and the right cerebellum (cerebellar activations pictured in Figure 5.2, middle right). This network matches up quite well with previously identified auditory-motor integration networks for the vocal tract (Buchsbaum et al., 2001; Gregory Hickok et al., 2003; Isenberg, Vaden, Saberi, Muftuler, & Hickok, 2012; Kayoko Okada & Hickok, 2006).

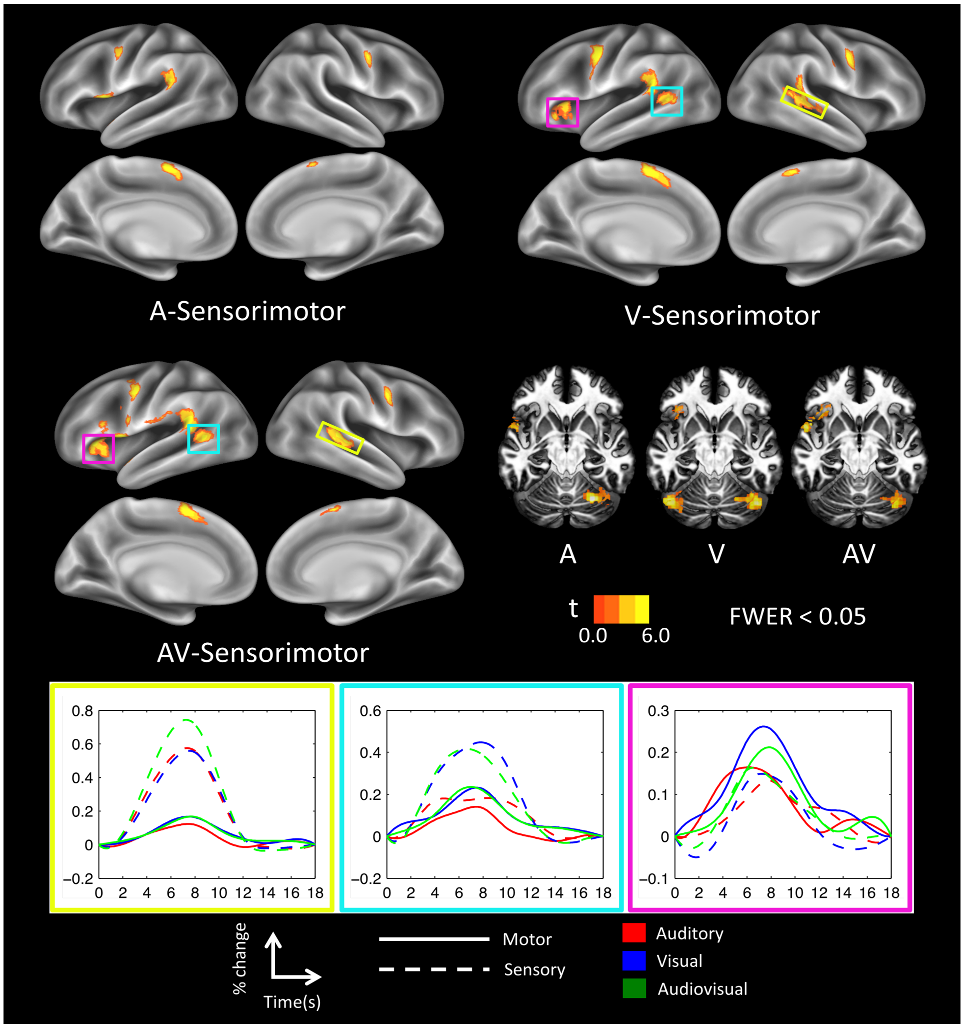


Figure 5.2. Sensorimotor conjunction SPMs. Sensorimotor brain regions were highlighted in each modality by taking the conjunction of Sensory (CP > baseline) and Motor (P+Reh > P+Rest) contrasts. These regions are displayed on separate cortical surface renderings for each input modality: A, V, AV. Also shown is a volume rendering for each modality with axial slices peeled away to allow visualization of cerebellar activation. Activation time-courses are shown for sensorimotor regions that were unique to the V and AV modalities. Yellow: Right pSTS. Teal: Left pSTS/MTG. Magenta: Left insula.

We hypothesized that Sensorimotor maps for V and/or AV would contain additional regions consistent with distinct pathways for integrating visual speech information with the speech motor system. This is precisely what we found. The V-Sensorimotor (Figure 5.2, top right) and AV-Sensorimotor maps (Figure 5.2, middle left) included the same network of brain regions as the A-Sensorimotor map, but with the following differences. First, the extent of activation in typical auditory-motor integration regions was greater for both V and AV. This was observed for Spt, SMA, and PreM (Table 5.1). Second, several new clusters emerged in both the V-Sensorimotor and AV-Sensorimotor maps. Common to both maps, additional clusters were active in the left posterior superior temporal sulcus/middle temporal gyrus (STS/MTG), the right posterior STS, and the left Insula (all highlighted in colored boxes in Figure 5.2). Examination of the activation time-courses (Figure 5.2, bottom) indicates that increased Motor activation in the V and AV modalities (relative to A) likely drove these additional clusters above threshold. This effect was subtle for the right STS (indeed, there were 114 suprathreshold voxels that did not survive cluster correction in the A-Sensorimotor conjunction). Unique to the V-Sensorimotor map, an additional cluster was present in the left cerebellum, while additional clusters unique to the AV-Sensorimotor map were observed in the left ventral PreM, the left IFG, and the left putamen (Table 5.1). Overall, sensorimotor integration of visual speech, whether V or AV, recruited a more extensive sensorimotor speech network, possibly via additional activation of posterior superior temporal regions.

Table 5.1. Centers of mass (MNI) of significant clusters in Sensorimotor conjunction maps

	Region	Hemisphere	x	y	z	Vol (voxels)	Approximate Cytoarch. Area
A-Sensorimotor	SMA	L	-2.2	6.5	64.1	450	6
	Spt	L	-59.6	-42.3	22.9	364	IPC (PF)
	Cerebellum	R	29.2	-61.7	-24.6	323	Lobule VI
	PreM	L	-55.5	-3.4	49.5	320	6
	PreM	R	57.7	0.9	44.6	271	6
	IFG	L	-52.8	9.4	-1.6	251	45
V-Sensorimotor	PreM	L	-54.2	-0.1	46.3	927	6
	Spt/STS	L	-56.3	-47.5	16.3	782	IPC (PF)
	SMA	L	-1.9	7.3	62.7	734	6
	PreM	R	56.9	0.8	45.3	458	6
	STS	R	54.8	-36.5	9.6	395	n/a
	Cerebellum	R	37.6	-64.3	-25.9	391	Lobule VI
	Cerebellum	L	-40.5	-65.7	-26.6	348	Lobule VIIa Crus I
	Insula	L	-35.7	22.6	3.2	313	n/a
AV-Sensorimotor	Spt/STS	L	-58.8	-44.3	17.1	1271	IPC(PF)
	IFG/Insula/vPreM	L	-47.9	15.2	6.2	1028	44
	SMA	L	-2	5.8	64	690	6
	PreM	L	-53.4	-2.5	49.4	525	6
	PreM	R	57.5	0.1	44	342	6
	STS	R	47.1	-37.8	6	337	n/a
	Cerebellum	R	36.1	-65.6	-25.1	214	Lobule VI
	Putamen	L	-22	4.8	6.4	192	n/a

Explicit tests for increased Motor activation in the context of visual speech

The previous section highlighted differences in sensorimotor integration networks for the vocal tract based on the input modality. However, these differences were inferred on the basis of qualitative inspection of multiple conjunction maps. It is possible that some of the observed patterns emerged on the basis of the cluster correction threshold we imposed in order to control the family-wise error rate. As such, we also wanted to test directly for differences in activation based on the input modality. In particular, we focused on differences in activation in the Motor (rehearsal) phase of the task. If the presence of visual speech information in fact recruits additional pathways to the motor system, we should observe at least one of the following in VvsA-Motor or AvvsA-Motor interaction contrast maps: (1) increased activation in canonical motor and/or sensorimotor speech areas that activate in response to covert production across input modalities; (2) increased activation in additional motor regions that come online only when the input contains visual speech; (3) increased activation in additional sensorimotor regions responsible for interfacing visual speech information with the motor system.

The VvsA-Motor and AvvsA-Motor interaction contrast maps are displayed in Figure 5.3 (warm colors), overlaid with the V-Sensorimotor and AV-Sensorimotor maps (blue) from Figure 5.2, respectively. The following is immediately apparent: extensive networks were highlighted by the interaction contrasts, but these networks did not overlap strongly (hardly at all, in fact) with the sensorimotor networks identified in the V-Sensorimotor and AV-Sensorimotor conjunction maps. The only region of considerable overlap is the left posterior MTG, and this overlap is present for both the V and AV input modalities. Thus, it seems there was not a significant "gain" on canonical sensorimotor integration regions in the presence of visual speech. However, as mentioned, large networks of *additional* brain regions were active in the interaction contrasts. In both the AvvsA-Motor and VvsA-Motor maps, large clusters in bilateral ventral occipital-temporal regions were active, as was a cluster in the right ventral pre/post-central region. Unique to the VvsA-Motor map, clusters emerged in the left posterior STS, IFG, Insula, cingulate cortex, and bilateral caudate nucleus. Unique to the AvvsA-Motor map, extensive pre- and post-central clusters emerged in addition to clusters in bilateral superior parietal lobules and paracentral lobules.

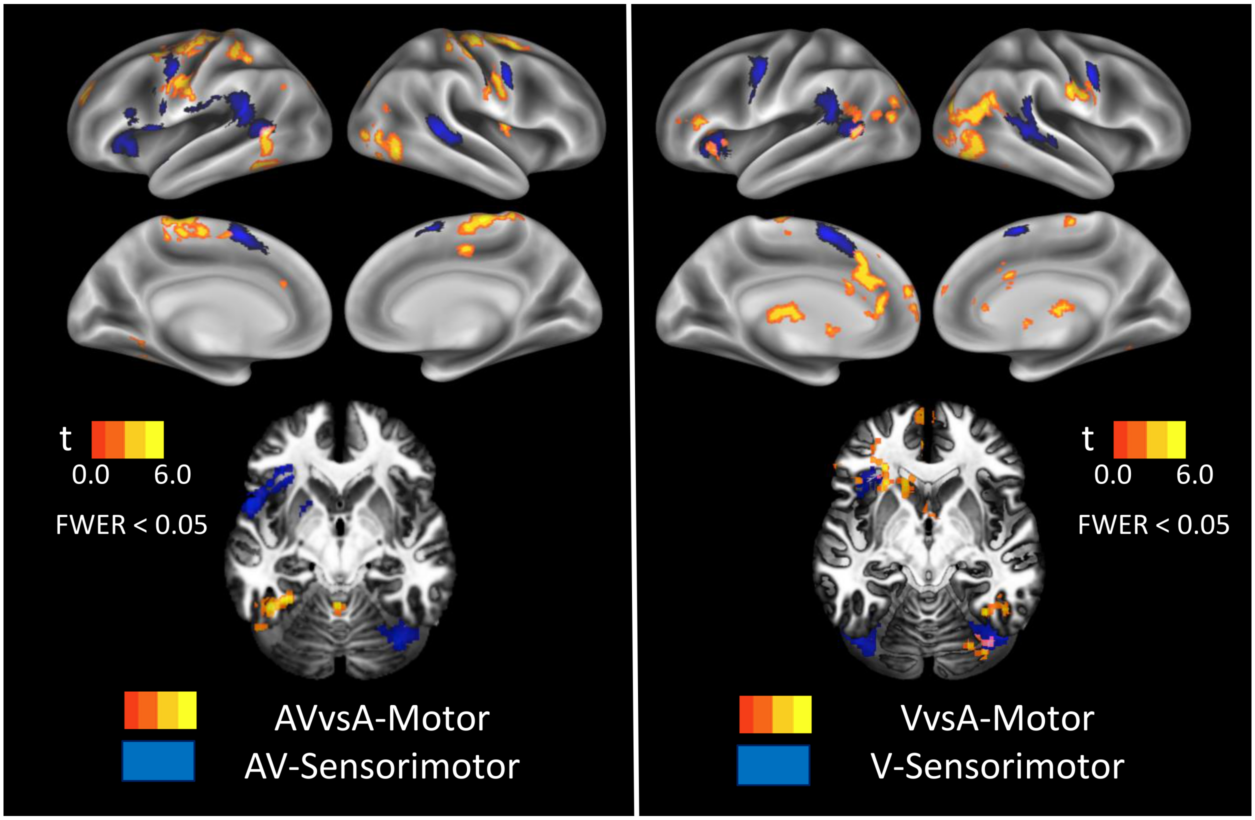


Figure 5.3. Interaction contrast SPMs (warm) overlaid with Sensorimotor conjunction SPMs (blue). Regions that displayed significantly greater Motor activation in the presence of visual speech are shown in the interaction contrast maps. Left: the AVvsA-Motor (AV-Motor > A-Motor) interaction contrast map is shown on a cortical surface rendering and a volume rendering with axial slices removed to allow cerebellar activation. Right: the VvsA-Motor (V-Motor > A-Motor) interaction contrast map is shown on a cortical surface rendering and a volume rendering with axial slices removed to allow cerebellar activation. Interaction contrast SPMs did not overlap strongly with Sensorimotor SPMs, indicating recruitment of additional motor areas.

To further examine the properties of brain regions identified in the interaction contrasts, we narrowed in on particular subregions to plot the activation time-courses of Sensory and Motor contrasts across all three input modalities. To locate interesting subregions, we restricted the search to voxels that were also significantly active in the Motor contrast alone. We did this because it was possible for voxels to reach significance in the interaction contrasts (e.g., when comparing across modalities as in AvvsA) but not in the within-modality Motor contrasts (e.g., AV-motor). In these cases, a significant interaction contrast would be due to differential patterns

of deactivation in the Motor contrast across modalities, which are difficult to interpret. As such, we identified clusters of interest using overlap analyses (logical conjunction). For significant voxels in the VvsA-Motor interaction contrast map, we identified clusters that were also significantly active to V-Motor alone (VvsA-Motor AND V-Motor). For significant voxels in the AvvsA-Motor interaction contrast map, we identified clusters that were also significantly active to AV-Motor alone (AvvsA-Motor AND AV-Motor). The idea was to highlight regions that showed an augmented Motor response in the presence of visual speech (V or AV relative to A) and also a significant Motor response relative to baseline. Overlap clusters are listed in Table 5.2 with MNI coordinates.

Table 5.2. Centers of mass (MNI) of significant clusters in interaction contrast maps overlapped with individual Motor maps (each thresholded FWER < 0.05)

	Region	**Hemisphere**	**x**	**y**	**z**	**Vol (voxels)**	**Approximate Cytoarch. Area**
VvsA-Motor AND V-Motor	PreCen Sulcus	L	-32.6	-4.6	53.8	282	6
	PreCen Sulcus	R	27.3	-9.1	59.7	251	6
	Inf Par Lobule	L	-37	-38.9	50.5	86	2
	MTG	L	-57.1	-59.9	8.8	58	n/a
	PreM	L	-58.5	1.2	33.6	55	6
	Cen Sulcus	L	-50.9	-14.2	52.2	38	1
AvvsA-Motor AND AV-Motor	ACC	L	-7.9	23.8	33.7	177	n/a
	Caudate Nucl.	L	-12.5	12.4	6.2	123	n/a
	Cerebellum	R	28.7	-71.9	-25.5	117	Lobule VIIa Crus I
	MTG	L	-57.5	-60.1	12.8	110	n/a
	Insula	L	-27	23.3	3.3	84	n/a
	IFG	L	-47.3	32.4	17	70	45

In Figure 5.4, we show clusters that were active in both the VvsA-Motor interaction contrast and the V-Motor contrast. These clusters were located in the left insula, IFG, caudate nucleus, cingulate cortex, MTG, and right cerebellum. The pattern of responses in the majority of these clusters was quite similar – Motor activation followed a graded pattern with the largest response in the V modality, followed by the AV modality and then the A modality (often little or no response), and there is very little Sensory activation across modalities. The one exception is the left MTG which responded well to both Sensory and Motor contrasts for V and AV, but poorly in both contrasts for A. In Figure 5.5, we show clusters that were active in both the AvvsA-Motor interaction contrast and the AV-Motor contrast. These clusters were located in the left PreM, pre-central sulcus, central sulcus/post-central gyrus, inferior parietal lobule, MTG, and the right pre-central sulcus. Three of these six clusters showed similar response patterns – Motor activation followed a graded pattern with the largest response in the AV Modality, followed by the V modality, and very little response in the A modality, with very little Sensory activation across modalities. The left PreM and inferior parietal lobule showed a strong Motor response in all three modalities but the response in AV and V exceeded the response in A. The left MTG cluster was in nearly the exact same region identified in the overlap between the VvsA-Motor interaction contrast and the V-Motor contrast, with nearly identical response properties. Overall, the overlap between Motor interaction contrast maps and individual-modality Motor contrast maps identified a network of motor-related brain regions that responded better during rehearsal of speech that contained a visual signal (V or AV). Some of these regions responded only to V and AV and thus correspond to additional motor regions recruited via visual-to-motor pathways, while other regions responded to all three modalities but responded best in the presence of a visual speech signal.

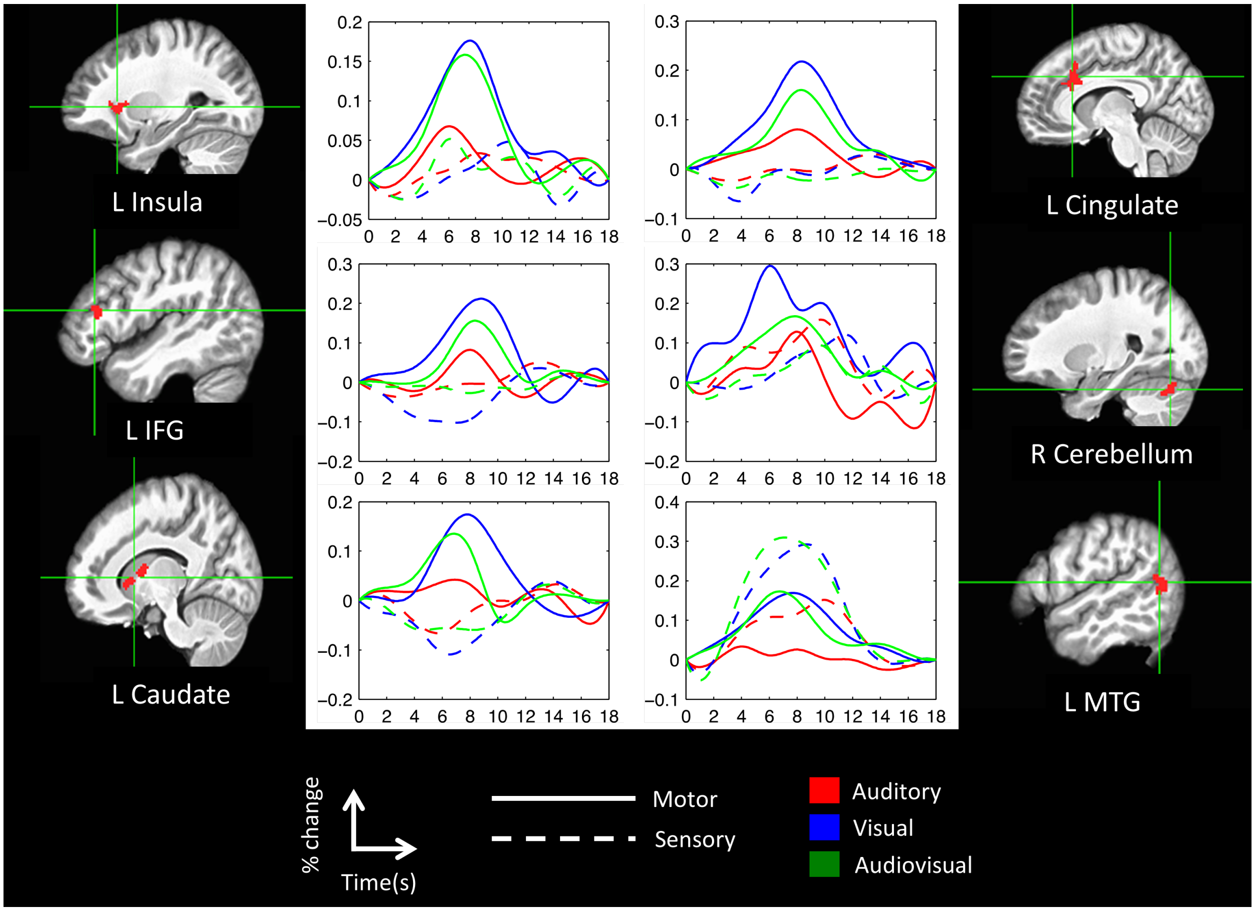


Figure 5.4. Overlap Analysis: VvsA-Motor and V-Motor. Clusters that were significant in the VvsA-Motor interaction contrast (V-Motor > A-Motor) and also to V-Motor alone are shown in volume space along with mean activation time-courses.

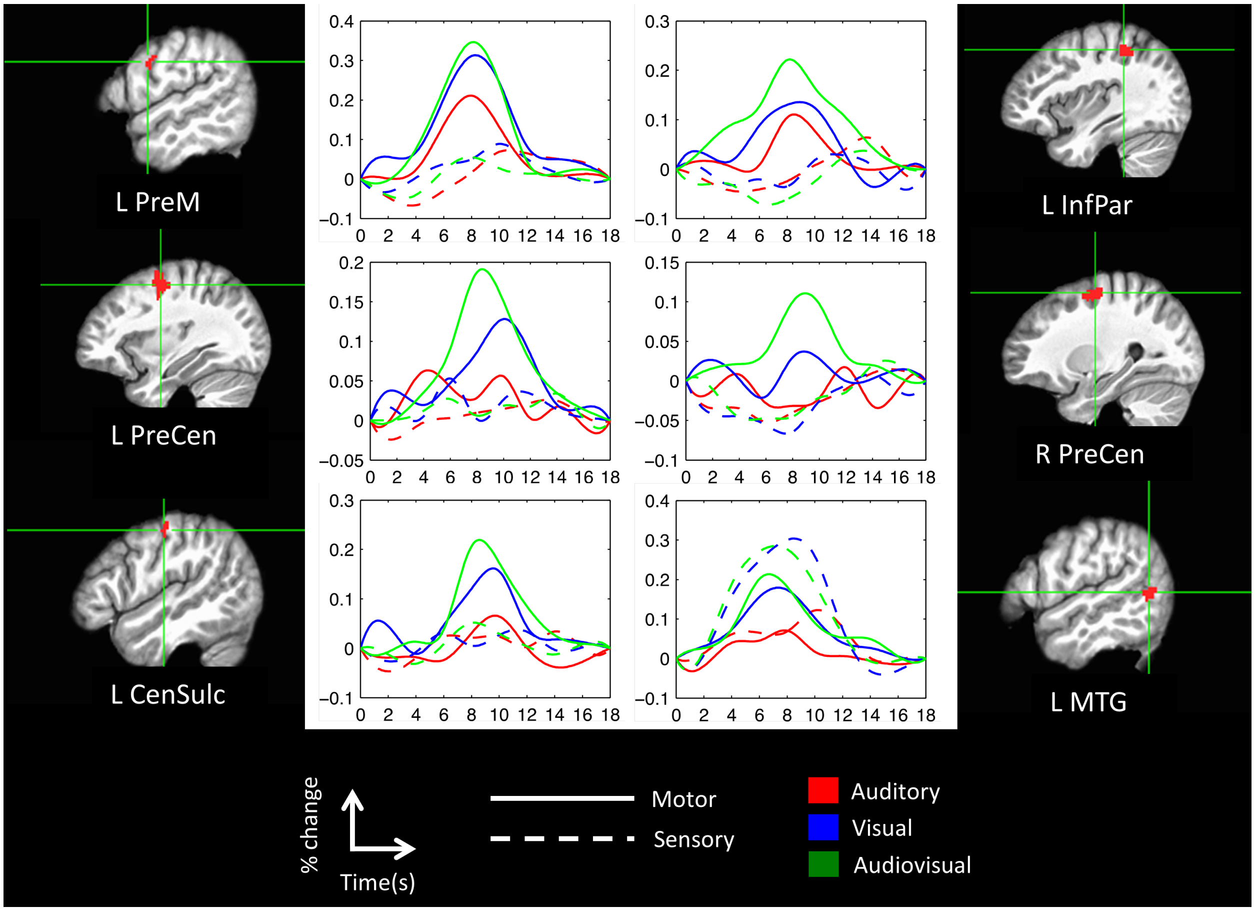


Figure 5.5. Overlap Analysis: AVvsA-Motor and AV-Motor. Clusters that were significant in the AVvsA-Motor interaction contrast (AV-Motor > A-Motor) and also to AV-Motor alone are shown in volume space along with mean activation time-courses.

Of note, the left posterior MTG has been consistently highlighted in every analysis – Sensorimotor conjunctions, Motor interaction contrasts, and overlap maps – strongly suggesting this region is a crucial sensorimotor node in a network for communicating visual speech information to the motor system.

Discussion

In the current study we asked whether sensorimotor integration networks for speech differed depending on the sensory modality of the input stimulus. In particular, we conducted an fMRI experiment in which participants were asked to perceive and immediately repeat a sequence of consonant-vowel syllables presented in one of three sensory modalities: auditory (A), visual (V), or audiovisual (AV). We measured activation to both the Sensory and Motor phases of the task, and we identified Sensorimotor brain regions by testing for voxels that activated significantly to both phases. We also tested for regions showing an increased Motor response when visual speech was included in the input (V or AV) relative to auditory-only input. We hypothesized that inclusion of visual speech in the input would either augment the activation in known auditory-motor networks (via multisensory integration of AV inputs) or recruit additional sensorimotor regions to support speech production (V or AV). Three noteworthy results will be discussed at further length below. First, speech motor regions were more activated when the input stimulus included visual speech. Second, certain motor and sensorimotor regions were only activated when the input stimulus included visual speech. Third, regions that activated preferentially for V input also tended to activate well to AV input and vice versa.

Visual speech inputs increase motor speech activation during rehearsal

Two sources of evidence support this conclusion. The first concerns differences in the V-Sensorimotor and AV-Sensorimotor networks relative to the A-Sensorimotor network. Both V and AV inputs increased the extent of sensorimotor activation in canonical motor speech regions including the ventral premotor cortex. Additionally, a significant cluster of sensorimotor activation was observed in the left insula in the V-Sensorimotor and AV-Sensorimotor maps, but not the A-Sensorimotor map. Examination of the activation time-course in this left insula region indicates that increased Motor activation in the V and AV modalities (relative to A) drove this effect. The second source of evidence comes from direct comparison of Motor activation in the V and AV modalities relative to A. The AvvsA-Motor map showed extensive activation in rolandic cortex bilaterally with additional activation in premotor regions. The VvsA-Motor map revealed activation in the right premotor cortex, left insula, and bilateral striatum.

This result concurs with behavioral evidence from normal and aphasic individuals indicating that shadowing audiovisual speech leads to increased speech output compared to shadowing of auditory-only speech (Fridriksson et al., 2012; Reisberg et al., 1987). Specifically, we have shown here that rehearsal immediately following speech input leads to greater activation in motor speech regions when the input contains visual speech, with large effects in primary motor regions for audiovisual speech in particular. One potential flaw in this conclusion, viz. that differences in Sensory activation between the input modalities produced the observed differences in motor system activation, deserves to be addressed explicitly here. Several recent studies suggest that perception of V or AV speech leads to increased motor system activation relative to perception of A speech (Callan et al., 2003; Matchin, Groulx, & Hickok, 2014; Skipper, Nusbaum, & Small, 2005; Skipper et al., 2007). However, this cannot be the source of activation differences observed in the current study. Firstly, although our perceive+rehearse task is inherently sensorimotor, activation to the Sensory phase of the task was factored out when computing activation to the Motor phase. For example, when comparing AV-Motor versus A-Motor the full contrast was (AV P+Reh – AV P+Rest) – (A P+Reh – A P+Rest), such that the

“P”, or perceptual component, was subtracted out separately for AV and A. Moreover, examination of the activation time-courses for Motor clusters in the VvsA-Motor and AvvsA-Motor interaction contrast maps (Figs. 5.4 & 5.5) reveals very little Sensory activation across input modalities. In other words, modality differences in Sensory activation cannot explain modality differences in Motor activation.

A distinct sensorimotor pathway for visual speech

Several of the sensorimotor brain regions showing an increased Motor response when the input stimulus contained visual speech also activated in the A-Motor contrast. These include the bilateral pSTS and left insula, ventral premotor cortex, and inferior parietal lobule. This set of regions responded *preferentially* to V, AV or both (relative to A). Other brain regions showing increased Motor activation following visual speech input responded *exclusively* to V, AV or both. Among these regions were the bilateral pre-central sulci and left central sulcus, striatum, IFG, and MTG. This set of regions may constitute a distinct sensorimotor pathway for visual speech that, when engaged in conjunction with auditory-motor networks by an audiovisual stimulus, produces increased activation of the speech motor system. If so, the influence of visual speech on production cannot be reduced to secondary activation of canonical auditory-motor pathways (i.e., via activation of auditory-phonological targets that interface with the speech motor system (Calvert et al., 1999; Calvert et al., 1997; Calvert, Campbell, & Brammer, 2000; K. Okada & Hickok, 2009; Kayoko Okada, Venezia, Matchin, Saberi, & Hickok, 2013)). In support of this conclusion, a recent TMS study using congruent (e.g., AV-ba and AV-ga) and incongruent (e.g., A-ba paired with V-ga and vice versa) audiovisual syllables suggests that, during perception, both the auditory and visual channels influence activity in speech motor cortex, but the two channels do not interact (Sato, Buccino, Gentilucci, & Cattaneo, 2010). Another study using an audiovisual perceive+rehearse paradigm with incongruent VCV syllables (A-aba paired with V-aga and vice versa) demonstrated that, when participants repeated the syllable from the auditory channel (attention was not directed to a particular channel), there was a shift in their production toward the syllable from the visual channel (evidenced by an f2 shift), even though participants were perceptually unaware of the incongruence (Gentilucci & Cattaneo, 2005).

Our results suggest that the left posterior MTG is a crucial node in the visual-to-motor speech pathway. This region was identified in the V-Sensorimotor and AV-Sensorimotor networks but not the A-Sensorimotor network, and responded significantly more (in fact only responded) to the Motor phase of the task when the input stimulus was V or AV. Sensory activation in the left MTG was much greater for V and AV inputs as well (Figs. 5.4 & 5.5). The left MTG figured prominently in a previous imaging study examining the effects of Speech Entrainment (SE) in nonfluent aphasics and normal subjects (Fridriksson et al., 2012). In both subject groups, there was significantly greater activation in the left MTG for SE (audiovisual shadowing) compared to spontaneous speech production, and probabilistic fiber tracking based on DTI data in the normal subjects indicated anatomical connections between the left MTG and left inferior frontal speech regions via the arcuate fasciculus. A recent voxel-based lesion-symptom mapping study examining conversational speech deficits (Borovsky, Saygin, Bates, & Dronkers, 2007) showed that damage to the left posterior MTG correlated with the token type ratio (a measure of the proportion of unique words generated). Crucially, conversational speech production was assessed in the context of one-on-one (presumably *face-to-face*) biographical interviews in a quiet room.

Visual and audiovisual speech inputs engage a similar rehearsal network

Behavioral increases in speech output are observed when subjects repeat audiovisual speech but not visual-only speech (Fridriksson et al., 2012; Reisberg et al., 1987). This is likely due to the fact that speechreading (i.e., of V alone) is perceptually demanding to the point that even the best speechreaders (with normal hearing) discern only 50% of the content from connected speech (MacLeod & Summerfield, 1987; Summerfield, 1992). As such, we may have expected to see differences in Motor activation depending on whether the input stimulus was AV or V. We did observe some such differences. Several small clusters were unique to the AV-Sensorimotor map and a cluster in the left cerebellum was unique to the V-Sensorimotor map. Moreover, the AvvsA-Motor interaction contrast emphasized pre- and post-central regions, while the VvsA-Motor interaction contrast emphasized medial regions including cingulate cortex and the caudate nucleus. However, when we specifically examined activation time-courses in these areas, it was generally the case that AvvsA-Motor regions also showed strong activation to the V-Motor contrast and vice versa for VvsA-Motor regions (in other words, motor areas that activated in AV also tended to activate in V; Figs. 5.4 & 5.5). There are two possible reasons for this phenomenon. First, we used a closed stimulus set with CV syllables that were easily distinguishable in the V modality, such that speechreading performance would be much higher than that observed for connected speech (i.e., effects of perceptual difficulty on rehearsal, and thus modality differences on this basis, would be diminished). Second, as observed above, there may be a distinct visual-to-motor pathway that interfaces visual speech with the motor system, and this pathway would be similarly engaged by visual and audiovisual speech.

It is worth noting that, although the visual syllables in this study were distinguishable, repetition in the V modality was certainly more taxing than in the A or AV modalities. This is particularly relevant with respect to certain regions that responded more in the Motor phase for V than AV or A. These include the left caudate nucleus, cingulate cortex, and IFG. The caudate is part of the basal ganglia, a group of subcortical regions theorized to be crucial for sequencing and timing in speech production (Bohland, Bullock, & Guenther, 2010; Fridriksson et al., 2005; Guenther, 2006; Lu, Chen, et al., 2010; Lu, Peng, et al., 2010; Pickett, Kuniholm, Protopapas, Friedman, & Lieberman, 1998; Stahl, Kotz, Henseler, Turner, & Geyer, 2011), while the left IFG and cingulate cortex are crucial for conflict monitoring and resolving among competing alternatives (Botvinick, Cohen, & Carter, 2004; Carter et al., 1998; January, Trueswell, & Thompson-Schill, 2009; Kerns et al., 2004; Novick, Trueswell, & Thompson-Schill, 2005; Novick, Trueswell, & Thompson-Schill, 2010). Each of these computations was likely taxed preferentially in the V perceive+rehearse task. The same cannot be argued for AV, which is the easiest version of the task. Still, there was one motor brain region that activated exclusively to rehearsal in the AV modality, the right pre-central sulcus. We can thus speculate that right hemisphere motor regions partially mediate behavioral improvements observed for SE in nonfluent aphasics (the right hemisphere is not damaged in these patients).

Why is visual speech linked to the motor system?

As alluded to in the Introduction, there is a well-accepted answer to this question in the auditory domain: speech sound representations are used to guide speech production. Development of the ability to speak constitutes the most intuitive evidence for this claim. In short, development of speech is a motor learning task that must take sensory speech as the input – at first from other users of the language and subsequently from self-generated babbling (Oller & Eilers, 1988). It has been suggested that a dorsal, auditory-motor processing stream functions to support language development, and that this stream continues to function into adulthood (G. Hickok & Poeppel, 2004; Gregory Hickok & Poeppel, 2000, 2007). More recent models suggest that for adult speakers auditory input functions primarily to tune internal feedback circuits that engage stored speech-sound representations to guide online speech production in real time (Gregory Hickok, 2012).

We have already cited evidence that perception of visual speech affects production (Fridriksson et al., 2009; Fridriksson et al., 2012; Gentilucci & Cattaneo, 2005; Reisberg et al., 1987; Sato et al., 2010). We have also asserted that the current imaging study supports the existence of a distinct visual-to-motor pathway for visual speech. Here we suggest that, like the dorsal auditory-motor stream, this visual-motor pathway begins to solidify during (and subserves) development of speech production. As described in the Introduction, the idea is that a distinct set of visual speech "targets" combine with auditory speech "targets" to facilitate motor control processes, in this case during development. Evidence suggests that sensory visual-speech "targets" are formed prior to the emergence of productive speech capacities. This is evidenced by existence of the McGurk effect in pre-linguistic infants (Burnham & Dodd, 2004; Rosenblum, Schmuckler, & Johnson, 1997). Pre-linguistic infants also match vowel sounds to facial displays of vowel articulation (Kuhl & Meltzoff, 1982), and show articulatory imitation of matching face/voice stimuli (Patterson & Werker, 1999). Moreover, visual speech improves phoneme discrimination and may lead to learning of category boundaries (Teinonen, Aslin, Alku, & Csibra, 2008). Finally, evidence suggests that visual speech representations (at least in the form of high-level motion and configuration information) are integrated with the motor system during development. Infants mimic facial gestures extensively, and they carry this out by correcting (tuning) their own motor behavior through a series of successive approximations to visual targets (i.e., representations of the face) (Meltzoff & Kuhl, 1994). This type of motor learning is precisely what would be required to "wire-up" motor control circuits for visual speech.

Conclusion

In summary, we have demonstrated that covert rehearsal following perception of syllable sequences results in increased speech motor activation when the input sequence contains visual speech. This increased activation is likely produced via recruitment of a visual-speech-specific network of sensorimotor brain regions. We presume this network functions to support speech motor control by providing a complementary set of visual speech "targets" that can be used in combination with auditory "targets" to guide production. This predicts that improvements in speech output will be observed when this visual-motor pathway is activated in conjunction with canonical auditory-motor speech pathways, a hypothesis in need of further testing. We have argued the visual-motor speech stream is formed during development for the purpose of motor control. Whether the visual-motor stream functions to support online production of spontaneous speech, as is true for the auditory-motor dorsal stream, remains to be determined.

References

Andersen, Richard A. (1997). Multimodal integration for the representation of space in the posterior parietal cortex. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 352*(1360), 1421-1428.

Arnold, Paul, & Hill, Fiona. (2001). Bisensory augmentation: A speechreading advantage when speech is clearly audible and intact. *British Journal of Psychology, 92*(2), 339-355.

Avants, B., Duda, J. T., Kim, J., Zhang, H., Pluta, J., Gee, J. C., & Whyte, J. (2008). Multivariate Analysis of Structural and Diffusion Imaging in Traumatic Brain Injury. *Academic Radiology, 15*(11), 1360-1375.

Avants, B., & Gee, J. C. (2004). Geodesic estimation for large deformation anatomical shape averaging and interpolation. *Neuroimage, 23*, 139-150.

Avants, Brian B, Tustison, Nicholas J, Song, Gang, Cook, Philip A, Klein, Arno, & Gee, James C. (2011). A reproducible evaluation of ANTs similarity metric performance in brain image registration. *Neuroimage, 54*(3), 2033-2044.

Bell, Andrew H, Meredith, M Alex, Van Opstal, A John, & Munoz, Douglas P. (2005). Crossmodal integration in the primate superior colliculus underlying the preparation and initiation of saccadic eye movements. *Journal of Neurophysiology, 93*(6), 3659-3673.

Bohland, Jason W, Bullock, Daniel, & Guenther, Frank H. (2010). Neural representations and mechanisms for the performance of simple speech sequences. *Journal of cognitive neuroscience, 22*(7), 1504-1529.

Borovsky, Arielle, Saygin, Ayse Pinar, Bates, Elizabeth, & Dronkers, Nina. (2007). Lesion correlates of conversational speech production deficits. *Neuropsychologia, 45*(11), 2525-2533.

Botvinick, Matthew M, Cohen, Jonathan D, & Carter, Cameron S. (2004). Conflict monitoring and anterior cingulate cortex: an update. *Trends in cognitive sciences, 8*(12), 539-546.

Buchsbaum, Bradley R, Hickok, Gregory, & Humphries, Colin. (2001). Role of left posterior superior temporal gyrus in phonological processing for speech perception and production. *Cognitive Science, 25*(5), 663-678.

Burnett, Theresa A, Freedland, Marcia B, Larson, Charles R, & Hain, Timothy C. (1998). Voice F0 responses to manipulations in pitch feedback. *The Journal of the Acoustical Society of America, 103*(6), 3153-3161.

Burnham, Denis, & Dodd, Barbara. (2004). Auditory–visual speech integration by prelinguistic infants: Perception of an emergent consonant in the McGurk effect. *Developmental psychobiology, 45*(4), 204-220.

Callan, D. E., Jones, J. A., Munhall, K., Callan, A. M., Kroos, C., & Vatikiotis-Bateson, E. (2003). Neural processes underlying perceptual enhancement by visual speech gestures. *Neuroreport, 14*(17), 2213-2218. doi: 10.1097/01.wnr.0000095492.38740.8f

Calvert, Gemma A, Brammer, Michael J, Bullmore, Edward T, Campbell, Ruth, Iversen, Susan D, & David, Anthony S. (1999). Response amplification in sensory-specific cortices during crossmodal binding. *Neuroreport, 10*(12), 2619.

Calvert, Gemma A, Bullmore, Edward T, Brammer, Michael J, Campbell, Ruth, Williams, Steven CR, McGuire, Philip K, . . . David, Anthony S. (1997). Activation of auditory cortex during silent lipreading. *Science, 276*(5312), 593-596.

Calvert, Gemma A, Campbell, Ruth, & Brammer, Michael J. (2000). Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. *Current Biology, 10*(11), 649-658.

Carter, Cameron S, Braver, Todd S, Barch, Deanna M, Botvinick, Matthew M, Noll, Douglas, & Cohen, Jonathan D. (1998). Anterior cingulate cortex, error detection, and the online monitoring of performance. *Science, 280*(5364), 747-749.

Chen, Gang, Saad, Ziad S, Nath, Audrey R, Beauchamp, Michael S, & Cox, Robert W. (2012). FMRI group analysis combining effect estimates and their variances. *Neuroimage, 60*(1), 747-765.

Cohen, Yale E, & Andersen, Richard A. (2002). A common reference frame for movement plans in the posterior parietal cortex. *Nature Reviews Neuroscience, 3*(7), 553-562.

Colonius, Hans, & Arndt, Petra. (2001). A two-stage model for visual-auditory interaction in saccadic latencies. *Perception & psychophysics, 63*(1), 126-147.

Corneil, BD, Van Wanrooij, M, Munoz, DP, & Van Opstal, AJ. (2002). Auditory-visual interactions subserving goal-directed saccades in a complex scene. *Journal of Neurophysiology, 88*(1), 438-454.

Diederich, Adele, & Colonius, Hans. (2004). Bimodal and trimodal multisensory enhancement: effects of stimulus onset and intensity on reaction time. *Perception & psychophysics, 66*(8), 1388-1404.

Erber, Norman P. (1969). Interaction of audition and vision in the recognition of oral speech stimuli. *Journal of Speech, Language, and Hearing Research, 12*(2), 423-425.

Fonov, Vladimir, Evans, Alan C, Botteron, Kelly, Almli, C Robert, McKinstry, Robert C, & Collins, D Louis. (2011). Unbiased average age-appropriate atlases for pediatric studies. *NeuroImage, 54*(1), 313-327.

Fonov, VS, Evans, AC, McKinstry, RC, Almli, CR, & Collins, DL. (2009). Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. *NeuroImage, 47*, S102.

Frens, Maarten A, Van Opstal, A John, & Van der Willigen, Robert F. (1995). Spatial and temporal factors determine auditory-visual interactions in human saccadic eye movements. *Perception & Psychophysics, 57*(6), 802-816.

Fridriksson, Julius, Baker, Julie M, Whiteside, Janet, Eoute, David, Moser, Dana, Vesselinov, Roumen, & Rorden, Chris. (2009). Treating visual speech perception to improve speech production in nonfluent aphasia. *Stroke, 40*(3), 853-858.

Fridriksson, Julius, Fillmore, Paul, Guo, Dazhou, & Rorden, Chris. (2014). Chronic Broca's Aphasia Is Caused by Damage to Broca's and Wernicke's Areas. *Cerebral Cortex*, bhu152.

Fridriksson, Julius, Hubbard, H Isabel, Hudspeth, Sarah Grace, Holland, Audrey L, Bonilha, Leonardo, Fromm, Davida, & Rorden, Chris. (2012). Speech entrainment enables patients with Broca's aphasia to produce fluent speech. *Brain, 135*(12), 3815-3829.

Fridriksson, Julius, Moss, Joel, Davis, Ben, Baylis, Gordon C, Bonilha, Leonardo, & Rorden, Chris. (2008). Motor speech perception modulates the cortical language areas. *Neuroimage, 41*(2), 605-613.

Fridriksson, Julius, Ryalls, Jack, Rorden, Chris, Morgan, Paul S, George, Mark S, & Baylis, Gordon C. (2005). Brain damage and cortical compensation in foreign accent syndrome. *Neurocase, 11*(5), 319-324.

Gentilucci, Maurizio, & Cattaneo, Luigi. (2005). Automatic audiovisual integration in speech perception. *Experimental Brain Research, 167*(1), 66-75.

Guenther, F. H. (2006). Cortical interactions underlying the production of speech sounds. *J Commun Disord, 39*(5), 350-365. doi: 10.1016/j.jcomdis.2006.06.013

Hardison, Debra M. (2003). Acquisition of second-language speech: Effects of visual cues, context, and talker variability. *Applied Psycholinguistics, 24*(04), 495-522.

Hasson, U., Skipper, J. I., Nusbaum, H. C., & Small, S. L. (2007). Abstract coding of audiovisual speech: beyond sensory representation. *Neuron, 56*(6), 1116-1126. doi: 10.1016/j.neuron.2007.09.037

Hickok, G., & Poeppel, D. (2004). Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language. *Cognition, 92*(1-2), 67-99. doi: 10.1016/j.cognition.2003.10.011

Hickok, Gregory. (2012). Computational neuroanatomy of speech production. *Nature Reviews Neuroscience, 13*(2), 135-145.

Hickok, Gregory. (2014). Towards an integrated psycholinguistic, neurolinguistic, sensorimotor framework for speech production. *Language, Cognition and Neuroscience, 29*(1), 52-59.

Hickok, Gregory, Buchsbaum, Bradley, Humphries, Colin, & Muftuler, Tugan. (2003). Auditory–motor interaction revealed by fMRI: speech, music, and working memory in area Spt. *Cognitive Neuroscience, Journal of, 15*(5), 673-682.

Hickok, Gregory, Houde, John, & Rong, Feng. (2011). Sensorimotor integration in speech processing: computational basis and neural organization. *Neuron, 69*(3), 407-422.

Hickok, Gregory, & Poeppel, David. (2000). Towards a functional neuroanatomy of speech perception. *Trends in cognitive sciences, 4*(4), 131-138.

Hickok, Gregory, & Poeppel, David. (2007). The cortical organization of speech processing. *Nat Rev Neurosci, 8*(5), 393-402.

Hughes, Howard C, Reuter-Lorenz, Patricia A, Nozawa, George, & Fendrich, Robert. (1994). Visual-auditory interactions in sensorimotor processing: saccades versus manual responses. *Journal of Experimental Psychology: Human Perception and Performance, 20*(1), 131.

Indefrey, Peter, & Levelt, Willem JM. (2004). The spatial and temporal signatures of word production components. *Cognition, 92*(1), 101-144.

Isenberg, A Lisette, Vaden, Kenneth I, Saberi, Kourosh, Muftuler, L Tugan, & Hickok, Gregory. (2012). Functionally distinct regions for spatial processing and sensory motor integration in the planum temporale. *Human brain mapping, 33*(10), 2453-2463.

January, David, Trueswell, John C, & Thompson-Schill, Sharon L. (2009). Co-localization of stroop and syntactic ambiguity resolution in Broca's area: implications for the neural basis of sentence processing. *Journal of Cognitive Neuroscience, 21*(12), 2434-2444.

Kerns, John G, Cohen, Jonathan D, MacDonald, Angus W, Cho, Raymond Y, Stenger, V Andrew, & Carter, Cameron S. (2004). Anterior cingulate conflict monitoring and adjustments in control. *Science, 303*(5660), 1023-1026.

Kleiner, Mario, Brainard, David, Pelli, Denis, Ingling, Allen, Murray, Richard, & Broussard, Christopher. (2007). What's new in Psychtoolbox-3. *Perception, 36*(14), 1.1-16.

Kuhl, Patricia K, & Meltzoff, Andrew N. (1982). *The bimodal perception of speech in infancy*.

Lu, Chunming, Chen, Chuansheng, Ning, Ning, Ding, Guosheng, Guo, Taomei, Peng, Danling, . . . Lin, Chunlan. (2010). The neural substrates for atypical planning and execution of word production in stuttering. *Experimental neurology, 221*(1), 146-156.

Lu, Chunming, Peng, Danling, Chen, Chuansheng, Ning, Ning, Ding, Guosheng, Li, Kuncheng, . . . Lin, Chunlan. (2010). Altered effective connectivity and anomalous anatomy in the basal ganglia-thalamocortical circuit of stuttering speakers. *Cortex, 46*(1), 49-67.

MacLeod, Alison, & Summerfield, Quentin. (1987). Quantifying the contribution of vision to speech perception in noise. *British journal of audiology, 21*(2), 131-141.

Massaro, Dominic W. (1998). *Perceiving talking faces: From speech perception to a behavioral principle* (Vol. 1): Mit Press.

Matchin, William, Groulx, Kier, & Hickok, Gregory. (2014). Audiovisual speech integration does not rely on the motor system: Evidence from articulatory suppression, the mcgurk effect, and fmri. *Journal of cognitive neuroscience, 26*(3), 606-620.

McCORMICK, BARRY. (1979). Audio-visual discrimination of speech*. *Clinical Otolaryngology & Allied Sciences, 4*(5), 355-361.

McGurk, Harry, & MacDonald, John. (1976). Hearing lips and seeing voices.

Meister, I. G., Wilson, S. M., Deblieck, C., Wu, A. D., & Iacoboni, M. (2007). The essential role of premotor cortex in speech perception. *Curr Biol, 17*(19), 1692-1696. doi: 10.1016/j.cub.2007.08.064

Meltzoff, Andrew N, & Kuhl, Patricia K. (1994). Faces and speech: Intermodal processing of biologically relevant signals in infants and adults. *The development of intersensory perception: Comparative perspectives*, 335-369.

Meredith, M Alex, Nemitz, James W, & Stein, Barry E. (1987). Determinants of multisensory integration in superior colliculus neurons. I. Temporal factors. *The Journal of Neuroscience, 7*(10), 3215-3229.

Meredith, M Alex, & Stein, Barry E. (1983). Interactions among converging sensory inputs in the superior colliculus. *Science*.

Möttönen, Riikka, & Watkins, Kate E. (2012). Using TMS to study the role of the articulatory motor system in speech perception. *Aphasiology, 26*(9), 1103-1118.

Neely, Keith K. (1956). Effect of visual factors on the intelligibility of speech. *The Journal of the Acoustical Society of America, 28*(6), 1275-1277.

Novick, Jared M, Trueswell, John C, & Thompson-Schill, Sharon L. (2005). Cognitive control and parsing: Reexamining the role of Broca's area in sentence comprehension. *Cognitive, Affective, & Behavioral Neuroscience, 5*(3), 263-281.

Novick, Jared M, Trueswell, John C, & Thompson-Schill, Sharon L. (2010). Broca's area and language processing: Evidence for the cognitive control connection. *Language and Linguistics Compass, 4*(10), 906-924.

Ojanen, V., Mottonen, R., Pekkola, J., Jaaskelainen, I. P., Joensuu, R., Autti, T., & Sams, M. (2005). Processing of audiovisual speech in Broca's area. *Neuroimage, 25*(2), 333-338. doi: 10.1016/j.neuroimage.2004.12.001

Okada, K., & Hickok, G. (2009). Two cortical mechanisms support the integration of visual and auditory speech: a hypothesis and preliminary data. *Neurosci Lett, 452*(3), 219-223. doi: 10.1016/j.neulet.2009.01.060

Okada, Kayoko, & Hickok, Gregory. (2006). Left posterior auditory-related cortices participate both in speech perception and speech production: Neural overlap revealed by fMRI. *Brain and Language, 98*(1), 112-117.

Okada, Kayoko, Venezia, Jonathan H, Matchin, William, Saberi, Kourosh, & Hickok, Gregory. (2013). An fMRI Study of Audiovisual Speech Perception Reveals Multisensory Interactions in Auditory Cortex. *PloS one, 8*(6), e68959.

Oller, D Kimbrough, & Eilers, Rebecca E. (1988). The role of audition in infant babbling. *Child development*, 441-449.

Patterson, Michelle L, & Werker, Janet F. (1999). Matching phonetic information in lips and voice is robust in 4.5-month-old infants. *Infant Behavior and Development, 22*(2), 237-247.

Pickett, Emily R, Kuniholm, Erin, Protopapas, Athanassios, Friedman, Joseph, & Lieberman, Philip. (1998). Selective speech motor, syntax and cognitive deficits associated with bilateral damage to the putamen and the head of the caudate nucleus: a case study. *Neuropsychologia, 36*(2), 173-188.

Purcell, David W, & Munhall, Kevin G. (2006). Compensation following real-time manipulation of formants in isolated vowels. *The Journal of the Acoustical Society of America, 119*(4), 2288-2297.

Rauschecker, Andreas M, Pringle, Abbie, & Watkins, Kate E. (2008). Changes in neural activity associated with learning to articulate novel auditory pseudowords by covert repetition. *Human brain mapping, 29*(11), 1231-1242.

Reisberg, Daniel, Mclean, John, & Goldfield, Anne. (1987). Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli.

Rorden, Christopher, Davis, Ben, George, Mark S, Borckardt, Jeffrey, & Fridriksson, Julius. (2008). Broca's area is crucial for visual discrimination of speech but not non-speech oral movements. *Brain stimulation, 1*(4), 383.

Rosenblum, Lawrence D, Schmuckler, Mark A, & Johnson, Jennifer A. (1997). The McGurk effect in infants. *Perception & Psychophysics, 59*(3), 347-357.

Ross, Lars A, Saint-Amour, Dave, Leavitt, Victoria M, Javitt, Daniel C, & Foxe, John J. (2007). Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments. *Cerebral Cortex, 17*(5), 1147-1153.

Sato, Marc, Buccino, Giovanni, Gentilucci, Maurizio, & Cattaneo, Luigi. (2010). On the tip of the tongue: modulation of the primary motor cortex during audiovisual speech perception. *Speech Communication, 52*(6), 533-541.

Schwartz, Jean-Luc, Basirat, Anahita, Ménard, Lucie, & Sato, Marc. (2012). The Perception-for-Action-Control Theory (PACT): A perceptuo-motor theory of speech perception. *Journal of Neurolinguistics, 25*(5), 336-354.

Skipper, J. I., Nusbaum, H. C., & Small, S. L. (2005). Listening to talking faces: motor cortical activation during speech perception. *Neuroimage, 25*(1), 76-89. doi: 10.1016/j.neuroimage.2004.11.006

Skipper, J. I., van Wassenhove, V., Nusbaum, H. C., & Small, S. L. (2007). Hearing lips and seeing voices: how cortical areas supporting speech production mediate audiovisual speech perception. *Cereb Cortex, 17*(10), 2387-2399. doi: 10.1093/cercor/bhl147

Stahl, Benjamin, Kotz, Sonja A, Henseler, Ilona, Turner, Robert, & Geyer, Stefan. (2011). Rhythm in disguise: why singing may not hold the key to recovery from aphasia. *Brain*, awr240.

Stein, Barry E, & Stanford, Terrence R. (2008). Multisensory integration: current issues from the perspective of the single neuron. *Nature Reviews Neuroscience, 9*(4), 255-266.

Stuart, Andrew, Kalinowski, Joseph, Rastatter, Michael P, & Lynch, Kerry. (2002). Effect of delayed auditory feedback on normal speakers at two speech rates. *The Journal of the Acoustical Society of America, 111*(5), 2237-2241.

Sumby, William H, & Pollack, Irwin. (1954). Visual contribution to speech intelligibility in noise. *The Journal of the Acoustical Society of America, 26*(2), 212-215.

Summerfield, Quentin. (1992). Lipreading and audio-visual speech perception. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 335*(1273), 71-78.

Teinonen, Tuomas, Aslin, Richard N, Alku, Paavo, & Csibra, Gergely. (2008). Visual speech contributes to phonetic learning in 6-month-old infants. *Cognition, 108*(3), 850-855.

Tourville, Jason A, Reilly, Kevin J, & Guenther, Frank H. (2008). Neural mechanisms underlying auditory feedback control of speech. *Neuroimage, 39*(3), 1429-1443.

Waldstein, Robin S. (1990). Effects of postlingual deafness on speech production: implications for the role of auditory feedback. *The Journal of the Acoustical Society of America, 88*(5), 2099-2114.

Watkins, K. E., Strafella, A. P., & Paus, T. (2003). Seeing and hearing speech excites the motor system involved in speech production. *Neuropsychologia, 41*(8), 989-994. doi: 10.1016/s0028-3932(02)00316-0

Wildgruber, D, Ackermann, H, & Grodd, W. (2001). Differential contributions of motor cortex, basal ganglia, and cerebellum to speech motor control: effects of syllable repetition rate evaluated by fMRI. *Neuroimage, 13*(1), 101-109.

Yates, Aubrey J. (1963). Delayed auditory feedback. *Psychological Bulletin, 60*(3), 213.

CHAPTER 6

Primer

Throughout the preceding chapters I have worked under the assumption that speech perception is fundamentally an auditory process (and, incidentally, that speech production is fundamentally an auditory-motor process). Yet, I have not dedicated any space to a direct investigation of the mechanisms underlying perception of auditory speech. Perhaps this is no surprise. Auditory speech perception is a very difficult problem that can be approached from many levels. In the current chapter, I begin to tackle that problem. I have chosen (along with collaborators) to be start with a bottom-up approach to speech perception. To be specific, the initial goal of this program is to understand the nature of cortical representations of low-level auditory features. I assume that speech sound representations are built hierarchically out of these lower-level auditory representations. As such, it will be of considerable use to describe the organization of low-level auditory cortical brain regions. In other work, we have begun to develop a technique that allows for mapping of individual cortical auditory fields with an unprecedented level of detail (Barton, Venezia, Saberi, Hickok, & Brewer, 2012). Here, I investigate the fine-grained structure of these auditory field maps with respect to a particular low-level auditory feature: low frequency temporal modulations. This feature appears to be particularly relevant for processing of speech sounds, as will be motivated below.

Periodicity coding in human auditory cortex

Jonathan H. Venezia, Brian Barton, Kourosh Saberi, Alyssa Brewer and Gregory Hickok

Introduction

Human speech, like many natural sounds, is a highly periodic signal (Nelken, Rotman, & Yosef, 1999; Singh & Theunissen, 2003). The periodicity of a signal is characterized by rhythmic fluctuations in the amplitude envelope. Behavioral evidence suggests that the envelope, or contour, of a sound is of particular importance for decoding human speech (Ahissar et al., 2001; Luo & Poeppel, 2007; Nourski et al., 2009; Saberi & Perrott, 1999; Shannon, Zeng, Kamath, Wygonski, & Ekelid, 1995), and it has been proposed that the speech envelope is extracted via one or another form of phase locking with periodicities in the speech stream, reflected in or driven by neural oscillations (A. L. Giraud & Poeppel, 2012). Indeed, some assert that extraction and representation (i.e., coding) of envelope periodicities is fundamental to the auditory nervous system, on par with spectral decomposition and maintenance of orderly tonotopic representations (Ahissar et al., 2001; Attias & Schreiner, 1997; T. Dau, Verhey, & Kohlrausch, 1999; Z. M. Smith, Delgutte, & Oxenham, 2002).

There is evidence of such periodicity coding in animal models. Beginning at the auditory nerve and continuing from subcortical auditory centers into primary auditory cortex, periodicity is coded in either the temporally modulated firing pattern or mean firing rate (or both) of individual cells (Joris, Schreiner, & Rees, 2004). In short, neurons show a preference for particular amplitude modulation (AM) frequencies in terms of the reliability of their phase-locked responses or in the overall strength of their responses. Distributions of best modulation frequency (BMF) across large samples of cells describe important properties of the periodicity code (Bieser & Müller-Preuss, 1996; Joris et al., 2004; Liang, Lu, & Wang, 2002). Across a variety of species, mean BMF generally decreases moving from the auditory brainstem (100-500 Hz) to the auditory midbrain (40-250 Hz) and into auditory cortex (8-50 Hz) (Joris et al., 2004). Recent evidence in marmosets suggests this trend may continue into the cortical hierarchy (Bendor & Wang, 2008). However, these data stem largely from unit physiology studies, which are subject to ascertainment biases and may not faithfully represent the neural populations in these regions. To be sure, it is common practice to pool unit data across hemispheres and individuals when describing population-level properties, which may obscure the large-scale organization.

As such, several studies have aimed specifically at describing the large-scale structure of periodicity codes. Recent data involving harmonic and AM stimuli, and utilizing optical imaging, MEG, and fMRI across several species, indicate the existence of an orderly spatial representation of periodicity in the auditory midbrain (Baumann et al., 2011) and primary auditory cortex (Langner, Dinse, & Godde, 2009; Langner, Sams, Heil, & Schulze, 1997; Schulze, Hess, Ohl, & Scheich, 2002). This periodotopic place map for modulation frequency runs from low to high in a gradient tilted orthogonally to the well-known tonotopic gradient. A recent fMRI study extended the characterization of periodotopic maps to multiple regions of human auditory cortex, showing that several reversals of the periodotopic gradient occur within a single, shared representation of tonotopic frequency space, delineating the borders of individual, orthogonally-organized auditory field maps (AFMs) (Barton et al., 2012). This study defined 11 AFMs in core and belt auditory regions, each of which demonstrated spatial periodicity coding across a range of AM rates (2-256 Hz).

Some authors have suggested that population-level properties of the cortical periodicity code may reflect so-called "temporal integration windows"(Wang, Lu, & Liang, 2003). In short, the population BMF of a cortical field (or AFM) may reflect its temporal integration window – the length of time over which separate events are not distinguished in the neural output. Also, it may be that two (or more) temporal integration windows are reflected as peaks in the population-level distribution of BMF over different ranges of the periodicity code. This idea relates neatly to prominent theories of speech perception, wherein Poeppel has emphasized two critical time scales – a slower syllable scale and a faster phoneme scale – that produce distinct periodicities in the speech stream (~5 Hz versus ~35 Hz, respectively) (D. Poeppel, 2003). Several functional-anatomic models have suggested that the left and right auditory cortices differ in their sensitivity to faster versus slower features of the acoustic signal. Zatorre's (R.J. Zatorre, Belin, & Penhune, 2002; Robert J Zatorre, 1997) temporal vs. spectral model and Poeppel's asymmetric sampling in time model (A. L. Giraud & Poeppel, 2012; D. Poeppel, 2003) (as well as earlier claims by Tallal and others (Tallal, Miller, & Fitch, 1993)) have argued that the left hemisphere is tuned to fast temporal processing (25-50 ms time scale = 20-40 Hz), whereas the right hemisphere is tuned to slower temporal processing, which enhances spectral information (150-300 ms time scale = 3.33-6.67 Hz).

Here we characterize the organization of the periodicity code in human auditory cortex. We use an fMRI technique that produces voxel-wise maps of BMF in core and belt auditory regions. The use of fMRI offers a unique opportunity to observe macro-organizational properties of the periodicity code, including the extent of cortical surface area (SA) dedicated to representation of particular BMFs, free from ascertainment bias present in single- and multi-unit recoding studies. Moreover, it allows us to compare periodicity coding across an unprecedented number of auditory cortical fields and between hemispheres. Thus, focusing on the distribution of cortical SA (relative to functionally-defined BMF), we aim to describe the periodicity code in human auditory cortex over 11 core and belt auditory fields and between hemispheres. We combine a descriptive approach with exploratory statistical analyses to investigate the following questions: (i) are periodicity codes uniform across AM rates or is there cortical magnification of a particular range (or ranges) of BMFs? If they are not uniform, (ii) do the two hemispheres differ in their preferred rates as the aforementioned theories suggest?, and (iii) does the decreasing BMF (brainstem to cortex) observed in animal models continue into the human cortical processing hierarchy?

Methods

This section will include information relevant to the construction of both tonotopic and periodotopic maps in human auditory cortex and, in addition, some brief details on how these maps are combined to delineate individual AFMs. However, the focus of the current study is to characterize the periodotopic component of AFMs (i.e., the spatial periodicity code). Methodological details that are directly relevant to this focus are given in the two final subsections of the Methods.

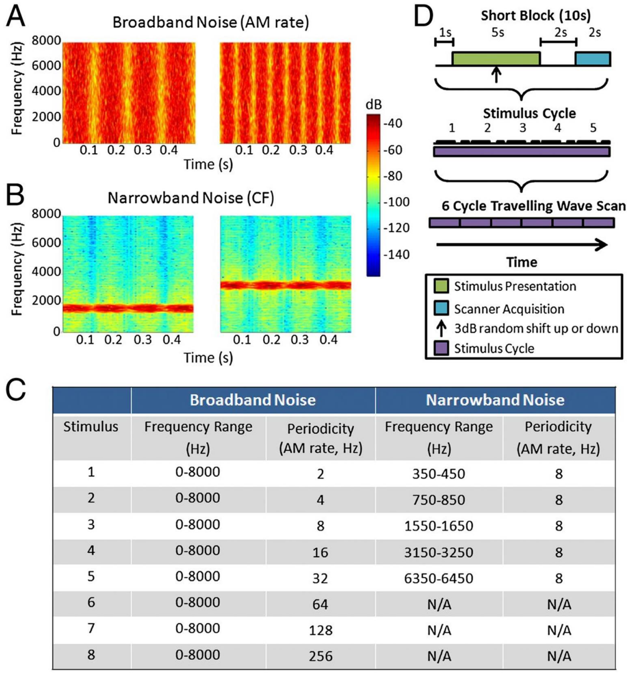
Subjects

Four human subjects (one female and three males, aged 26–38) from the University of California, Irvine, served as participants for the present study. The experimental protocol was approved by the Institutional Review Board at University of California, Irvine, and informed consent was obtained from all subjects.

Experimental Design

Auditory stimuli were presented in a block fMRI design following the traveling wave method, which is the standard fMRI paradigm employed in visual field mapping (A.A. Brewer & Barton, 2012; Alyssa A Brewer, Liu, Wade, & Wandell, 2005; DeYoe et al., 1996; Engel, 1994; Sereno et al., 1995; B. A. Wandell, Brewer, & Dougherty, 2005; Brian A Wandell & Winawer, 2011). In the traveling wave method, stimuli that vary on a single dimension are presented in consecutive blocks at discrete values that span the entire 'stimulus space' (e.g., visual eccentricity from 0-20°). Blocks are presented in order from low to high stimulus values (e.g., 5°→10°→20°). This presentation scheme produces a "traveling wave" of activation that allows for specification of the preferred stimulus value at each voxel in topographically organized brain regions. Here, we modified the traveling wave procedure to a sparse acquisition sequence designed for auditory presentation – stimulus blocks were presented in the silent periods between single volume acquisitions.

We employed two classes of amplitude-modulated (AM) stimuli, narrowband and broadband noise, which varied along the stimulus dimensions of center frequency (CF; tonotopy) and AM rate (periodotopy), respectively. These stimuli were presented in short blocks with a 1s silent period, then 5s of AM noise at a single stimulus value (50 ms cosine rise/fall envelope), followed by a 2s silent period, and a 2s slice acquisition period. Participants maintained fixation on a small white central cross (~0.5° of visual angle) on a black background. Stimuli were presented at a mean level of 65 dB SPL (A), with a 3 dB level increase or decrease (selected randomly) at the midpoint of each presentation. Participants were asked to indicate the direction of the shift with a button press. Tonotopic stimuli consisted of 100-Hz-wide bands of noise with varying CF – 400, 800, 1,600, 3,200, and 6,400 Hz – each of which were amplitude modulated at 8 Hz (80% depth). Periodotopic stimuli were broadband noise segments (0–8,000 Hz) amplitude modulated (80% depth) at 2, 4, 8, 16, 32, 64, 128, and 256 Hz (Figure 6.1). Separate functional scans were dedicated to tonotopic and periodotopic stimuli. In these scans, the entire range of stimuli was covered by consecutive blocks presented in order from low to high (CF or AM rate) in what is referred to as one stimulus cycle. Six stimulus cycles were presented sequentially in each functional scan, and each subject underwent 6-8 functional scans in each stimulus class (see Fig. 6.1D for a schematic of the experimental design).



	Broadband Noise		Narrowband Noise	
Stimulus	Frequency Range (Hz)	Periodicity (AM rate, Hz)	Frequency Range (Hz)	Periodicity (AM rate, Hz)
1	0-8000	2	350-450	8
2	0-8000	4	750-850	8
3	0-8000	8	1550-1650	8
4	0-8000	16	3150-3250	8
5	0-8000	32	6350-6450	8
6	0-8000	64	N/A	N/A
7	0-8000	128	N/A	N/A
8	0-8000	256	N/A	N/A

Figure 6.1. Experimental stimuli and design. The sound spectrogram across frequencies (vertical axes) and time (horizontal axes). Increasing sound energy is represented as increasingly "warmer" colors. (A) Example broadband noise stimuli with amplitude modulation (AM) rates of 8 (Left) and 16 Hz (Right). (B) Example narrowband noise stimuli with center frequencies (CF) of 1,600 (Left) and 3,200 Hz (Right). (C) All experimental stimuli. Broadband noise stimuli maintain constant frequency information and vary periodicity, whereas narrowband noise stimuli hold periodicity constant and vary frequency. (D) Sparse sampling traveling wave experimental design.

Stimulus Presentation

Sounds were presented over MR-compatible, insert-style headphones (Sensimetrics model S14) powered by a 15 watt-per-channel stereo amplifier (Dayton model DTA-1). This style of headphone utilizes a disposable "earbud" insert that serves as both a sound attenuation device (earplug) and sound delivery apparatus, allowing sounds to be presented directly to participants' ear canals, so no transfer function need be applied to the stimulus. The headset can provide output levels in the ear canal up to 110 dB SPL and has a flat frequency response over the stimulus frequency range 0Hz – 8kHz , covering the range used in these experiments. During scanning, a secondary protective ear cover (Pro Ears Ultra 26) was placed over the earbuds for additional attenuation of scanner noise. Stimulus delivery and timing were controlled using Cogent 2000 software (http://www.vislab.ucl.ac.uk/cogent 2000.php) implemented in Matlab R12 (Mathworks, Inc, USA).

Visual displays of the task instructions and the fixation cross were generated using the Cogent 2000 software and back-projected via a Christie DLV1400-DX DLP projector onto a screen at the head end of the bore of the magnet (spatial resolution: 1024x768 pixels; refresh rate: 60 Hz). Subjects viewed the display on an angled front surface mirror mounted on the head coil close to the eyes with a viewing distance of approximately 70 cm. Head movements were minimized with padding and tape.

Anatomical Data Acquisition

Scanning was conducted on the 3T Philips Achieva MR scanner at the University of California, Irvine, with an 8 channel SENSE imaging head coil. One high-resolution whole-brain anatomical dataset was acquired for each subject (T1-weighted 3D MPRAGE, 1 mm^3 voxels, TR = 8.4 ms, TE = 3.7 ms, flip = 8˚, SENSE factor = 2.4), which uses a fast gradient echo T1-weighted inversion pulse sequence (MPRAGE) in conjunction with parallel imaging to maximize the image contrast between white and gray cortical matter. In addition, one anatomical in-plane image was acquired before each set of functional scans, with the same slice prescription as the functional scans but with a higher spatial resolution (1 mm x 1 mm x 3 mm voxels).

Functional Data Acquisition

Functional MR data were acquired on the same scanner as the anatomical data, with 35 axial slices oriented close to parallel to the STG (T2-weighted, gradient echo imaging, TR = 10s, TA = 2s, TE = 30 ms, flip = 90˚, SENSE factor = 1.7, reconstructed voxel size of 1.875 x 1.875 x 3 mm, no gap).

Anatomical Data Analysis

The T1-weighted slices were physically in register with the functional slices and were used to align the functional data with the high-resolution anatomical data, first by a manual co-registration and then by a semi-automated three-dimensional (3D) co-registration algorithm, a mutual information method (Maes, Collignon, Vandermeulen, Marchal, & Suetens, 1997; Nestares & Heeger, 2000). In addition, the high-resolution anatomical volume was corrected for inhomogeneity and linearly transformed with no rescaling and no distortion to align with the Talairach reference brain, using tools from the FMRIB software library (http://www.fmrib.ok.ac.uk/fsl/).

For analysis of neuroimaging data for individual subjects, we used a Matlab-based signal processing software package developed by the Wandell lab at Stanford University (*mrVista* is open-source software and is publicly available online at http://white.stanford.edu/software/). With this software, the location of the cortical gray matter for each subject was identified ("segmented") in the high-resolution anatomical scan using the *mrVista* automated algorithm followed by hand-editing to minimize errors for individual subject analyses (Teo, Sapiro, &

Wandell, 1997). Gray matter was then grown from the segmented white matter to form a 3-4 mm layer covering the white matter surface. To improve sensitivity, only data from this identified gray matter were analyzed. The gray matter was then rendered in 3D close to the gray-white matter boundary or unfolded into a continuous, flat sheet to allow visualization of functional activity within the sulci, with light gray regions indicating gyri and dark gray regions representing sulci. After registration to the high-resolution anatomy, the functional activity can be visualized either in its original coordinate frame ('inplanes'), on the segmented gray matter in anatomical volume slices, or on inflated or flattened representations of the cortical surface to allow for optimal definition of auditory field map boundaries (B. A. Wandell, S. Chial, & B. T. Backus, 2000).

Functional Data Analysis

We analyzed the functional MRI data using the same *mrVista* custom Matlab software described above for anatomical data analysis (http://white.stanford.edu/software). For each subject, data in each fMRI session were analyzed voxel-by-voxel with no spatial smoothing. The mean value maps of the BOLD signal were compared across scans to check for potential head movements. Because all scans had less than one voxel of head motion, no motion correction algorithm was applied. The time series from each scan was high-pass filtered to remove low-frequency sources of physiological noise and averaged together to form one mean time series for each subject. A Fourier coherence analysis was applied to every voxel to separate activity due to the stimuli (at the frequency of six cycles per scan) from activity due to random and physiological noise (at the other frequencies in the cycles-per-scan domain). A phase value was assigned to each voxel with activity above a standard coherence threshold of 0.20 (Alyssa A Brewer et al., 2005; B. A. Wandell, Dumoulin, & Brewer, 2007). This phase value corresponds to that voxel's preferred point in the relevant stimulus space. The resultant phase maps correspond to voxel-wise descriptions of *best frequency* (BF; narrowband noise; tonotopy) and *best modulation frequency* (BMF; broadband noise; periodotopy). It should be noted that phase maps allow smooth interpolation of BF and BMF across the entire range of stimuli presented in the study.

Definition of Auditory Field Maps

Details are given elsewhere (Barton et al., 2012). Briefly, AFMs were defined individually for each hemisphere of each subject on a flattened representation of the cortical surface centered on Heschl's gyrus (HG). By combining tonotopic and periodotopic maps – i.e., using reversals in tonotopic maps to define one set of functional boundaries and using reversals in periodotopic maps to define another set of functional boundaries – we defined 11 AFMs in each hemisphere of each subject. Tonotopic responses included a large low-frequency region oriented parallel to HG, which was encircled by and transitioned to a high frequency region, while periodotopic responses formed "isoperiodotopic" bands organized in radial spokes within HG, orthogonal to the tonotopic gradient (Figure 6.2 C, D).

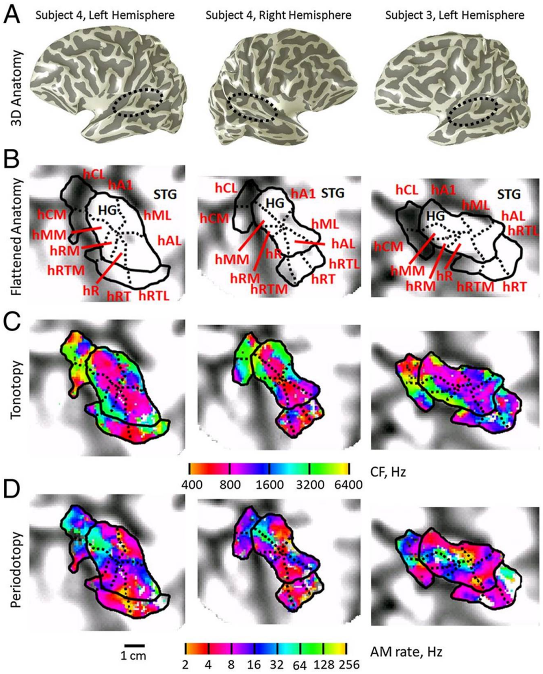


Figure 6.2. Anatomical and functional data in auditory core and belt. (Left) Data in subject 4's (S4's) left hemisphere and (Middle) in S4's right hemisphere. (Right) Data in S3′s left hemisphere. Light gray indicates gyri; dark gray indicates sulci. (A) A 3D rendering of individual cortical surfaces. Circles indicate HG and surrounding regions presented in (B). (B) Flattened cortical surface of HG and surrounding regions for each hemisphere, orientated to align STG. Solid black lines indicate AFM boundaries between maps along mirror-symmetric tonotopic reversals. Dotted black lines indicate AFM boundaries between maps along mirror-symmetric periodotopic reversals. Red text indicates AFM names; black text indicates gyri names. (C) Tonotopy mapped using narrowband noise stimuli. Colors indicate the preferred center frequency for each voxel (CF, in hertz). (D) Periodotopy mapped using broadband noise stimuli. Colors indicate the preferred period for each voxel (AM rate, in hertz). Each voxel is measured independently with no spatial or temporal smoothing and no motion correction. Voxels presented have coherence above the statistical threshold of 0.20 and are within one of the 11 AFMs presently studied. Scale bar denotes 1 cm along the flattened cortical surface.

The 11 observed AFMs correspond nicely to proposed auditory fields in the macaque auditory core and belt, and so we refer to our human AFMs as hA1, hR, hRT, hCM, hMM, hRM, hRTM, hCL, hAL, hRL, and hRTL, where "h" stands for "human" and the other letters indicate proposed homology to the monkey auditory subfield names. It should be noted that moving rostrally in the monkey anatomy (e.g., from A1, to R and RT) corresponds to moving laterally in the homologous human AFMs (hA1, hR, hRT; Fig. 6.2). However, we organize human AFMs similarly into core (hA1, hR, hRT), medial belt (hCM, hMM, hRM, hRTM) and lateral belt (hCL, hML, hAL, hRTL) subregions of auditory cortex. By definition, voxels within each AFM cover the entire narrowband and broadband stimulus ranges in terms of BF and BMF, respectively, in orthogonal topographic gradients from low to high (Barton et al., 2012).

Surface Area Measurements

Voxel-wise maps of BMF (i.e., periodotopic maps) were first identified on a 2-d flattened region ('flat map') representing the cortical surface near HG. Measurements of BMF in periodotopic maps were interpolated smoothly across the range of AM values expressed in our broadband noise stimuli (2-256 Hz). To construct surface area distributions, we discretized the range of stimulus values by constructing periodicity bins around the AM rates of the actual stimuli presented in the experiment (2, 4, 8, 16, 32, 64, 128, 256 Hz). The stimulus values 4, 8, 16, 32, 64, and 128 Hz served as bin centers on a linear scale, such that the bins were equal width on either side of the center (the 4 Hz bin ranged from 3-5 Hz, the 8 Hz bin ranged from 5-11 Hz, etc). The stimulus values 2 and 256 Hz served as lower and upper bin boundaries for bins at the lowest and highest AM rates, respectively, and these bins were half-width relative to the others. Once bins were formed, individual regions of interest (ROIs) were identified within each AFM to represent all 8 of the bins separately – that is, for a particular bin, only those voxels with BMFs that fell within the bin range were assigned to the corresponding ROI. Surface area was calculated for each of the 8 periodicity-bin ROIs in all 11 AFMs in each subject. The 2-d coordinates of the ROIs were identified on the flattened representation of the cortical surface.

Because the flattening process inevitably distorts distance and area measurements, all surface area measurements were made along the 3-d cortical manifold by mapping the 2-d coordinates back to the 3-d manifold. The surface area is then measured along the boundary between gray and white matter, which allows more reliable boundary definition than the outer surface of the gray matter or any particular cortical layer (Teo et al., 1997; B. A. Wandell, S. Chial, & B. Backus, 2000). For extended details about the algorithm used in these measurements, see (R. F. Dougherty et al., 2003). Raw surface area measurements in mm^2 were then converted to proportional measurements by dividing the surface area of each periodicity-bin ROI by the total surface area of its parent AFM. This resulted in 88 discrete distributions of proportional cortical surface area relative to BMF (11 AFMs x 2 hemispheres x 4 participants).

Exploratory statistical analyses of differences in SA distributions

We aimed to conduct exploratory statistical analyses in order to examine two research questions of interest: (i) are there hemispheric differences in periodicity coding (i.e., do SA distributions differ between the left and right hemispheres)?, and (ii) are differences in periodicity coding observed moving through the cortical hierarchy (i.e., do SA distributions differ between individual AFMs)? Of note, our data were doubly multivariate with 8 measurements of proportional surface area (for each of the 8 periodicity bins) repeated in 22 distinct auditory fields per participant (11 AFMs x 2 hemispheres). Since we had data from only four participants, we chose to reduce the data by focusing on the periodicity bins containing the large majority of total surface area (4-32 Hz; see Results below). Further, we chose to restrict our exploratory analyses to only the three AFMs that make up the auditory core (hA1, hR, hRT).

Finally, we avoided doubly multivariate analysis by testing for differences only within individual periodicity bins.

All statistical analyses were conducted in IBM SPSS Statistics (release 20.0.0). We tested for hemispheric differences by conducting four paired t-tests (4 Hz left vs. right; 8 Hz left vs. right; 16 Hz left vs. right; 32 Hz left vs. right) in each of the three core AFMs (hA1, hR, hRT). Individual t-tests were thresholded at $p < 0.05$ and family-wise error was controlled via Bonferroni correction (12 total tests, corrected threshold $p < 0.004167$). We tested for 'cortical hierarchy' differences by first collapsing SA distributions across hemispheres and then constructing separate linear mixed models for each of the four periodicity-bins (4 Hz, 8 Hz, 16 Hz, 32 Hz). Each mixed model had a single fixed factor, *field*, with three levels (hA1, hR, hRT). Surface area data from each AFM were treated as repeated measures and so *field* was entered as a repeated factor with an unstructured covariance matrix (analogous to the MANOVA approach to repeated measures (Bagiella, Sloan, & Heitjan, 2000)). All pairwise comparisons (paired t-tests) were also performed for each periodicity bin. Individual tests were thresholded at $p < 0.05$ and family-wise error was controlled via Bonferoni correction (3 fixed effects tests on *field*, corrected $p < 0.0167$; 12 pairwise comparisons, corrected $p < 0.004167$).

A follow-up analysis was performed in order to further explore 'cortical hierarchy' differences in SA distributions. Rather than reducing the data by focusing on the 4-32 Hz periodicity bins, we parameterized SA distributions by fitting a rounded exponential (RoEx) function (Moore & Glasberg, 1983) to each individual distribution (again, collapsed across hemispheres) and measuring the width of its passband (henceforth called *bandwidth*) in Hz (see Appendix). The RoEx function is typically used to model peripheral auditory filters. As such, it provides a good fit to distributions with a passband (a single peak) that gives way to shallower tails, as our SA distributions do. We chose to measure bandwidth (BW) in light of the observed shape of the SA distributions – i.e., considering the unimodal character of the SA distributions, including common peaks at or near 8 Hz, a smaller bandwidth indicates greater accumulation of SA around the peak (or *center modulation frequency*, CMF) with a smaller proportion of SA dedicated to fast modulations (and very slow modulations). The results of our single-bin analyses suggested that BW might decrease moving laterally through the core AFMs (hA1→hR→hRT; see Results below). To test this, we repeated the linear mixed model analysis described above with BW as the dependent measure. The fixed effect test for *field* was thresholded at $p < 0.05$. In addition to BW and CMF, we report a centroid measure (CEN, Hz) for each of the 11 AFMs corresponding to the point at which 50% of the area under the RoEx curve was accumulated.

Results

Descriptive characterization of periodicity distributions

We obtained 88 individual distributions of cortical SA relative to BMF (4 participants x 2 hemispheres x 11 AFMs). These distributions are pictured averaged across the four participants (Figure 6.3). Cursory examination of SA distributions reveals two important features. First, the distributions in all 11 AFMs are not uniform. Rather, they are unimodal with peaks at or near the 8 Hz periodicity bin (in other words, cortical magnification is observed over a certain range of periodicities; Fig. 6.3, line graphs). Second, differences between hemispheres are small compared to differences between AFMs within each hemisphere. Also, the within-hemisphere differences between AFMs appear to be similar in the left and right hemispheres, and tend to occur primarily within a particular range of the SA distributions (4-32Hz; Fig. 6.3, pie charts show SA data collapsed across hemispheres). This range contains the overwhelming majority of the total mass for each distribution, which is significant because it contains AM rates present in the most common envelope modulations observed in natural human speech (Chandrasekaran, Trubanova, Stillittano, Caplier, & Ghazanfar, 2009). Below we report exploratory statistical tests for hemispheric differences in SA distributions as well as differences between individual AFMs.

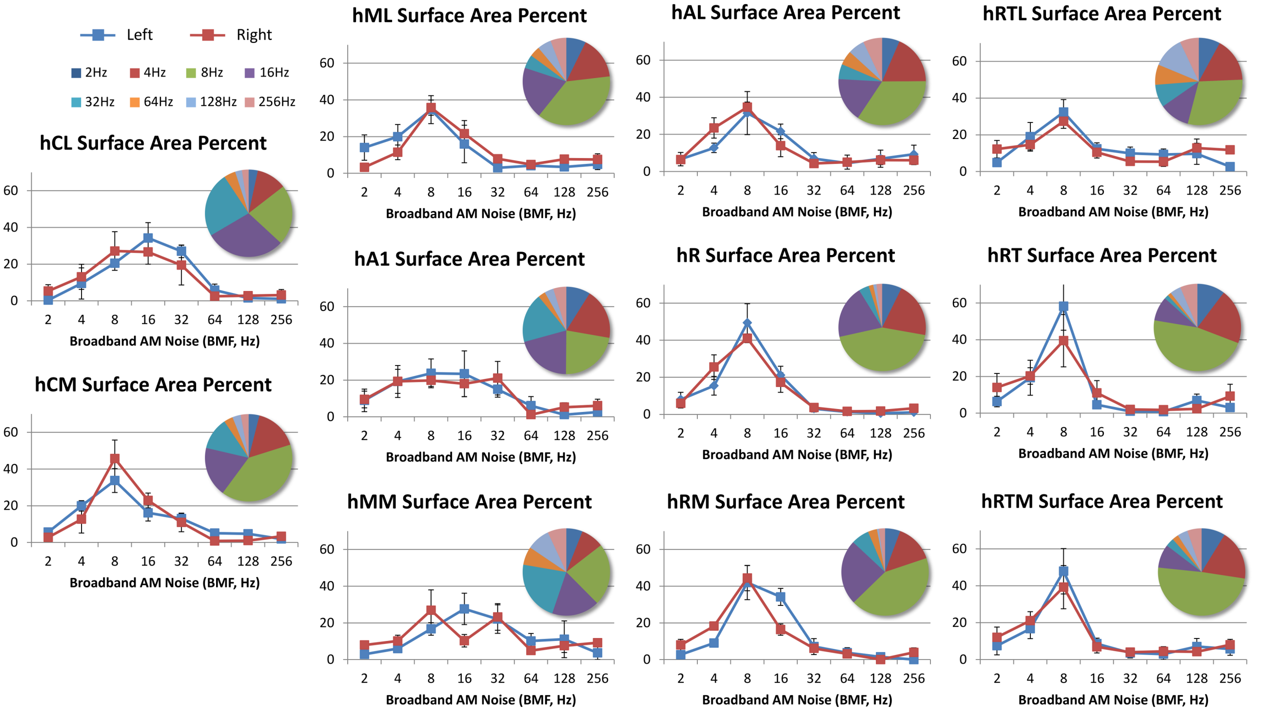


Figure 6.3. Distributions of cortical surface area relative to best modulation frequency (periodicity distributions). Periodicity distributions are shown for all 11 core and belt AFMs. (Top Row + hCL) Lateral belt AFMs. (Bottom Row + hCM) Medial belt AFMs. (Middle Row) Core AFMs. Columns moving left to right reflect anatomical position moving medial to lateral. Periodicity distributions were measured by tabulating the surface area in each of 8 periodicity bins. The 8 stimulus values (AM rate; 2, 4, 8, 16, 32, 64, 128, 256 Hz) served as bin centers with equal width to the left and right (the 2 and 256 Hz bins were half-width relative to the others). Surface area is expressed as a percent of the total surface area in each AFM. (Line Graphs) Blue and red lines reflect periodicity distributions in the left and right hemispheres, respectively, in each AFM. (Pie Charts) Periodicity distribution collapsed across hemispheres. Notice two features: (1) the periodicity distributions in the left and right hemispheres are nearly overlapping in all 11 AFMs, and (2) periodicity distributions vary from one AFM to another within the core, medial belt and lateral belt, especially in the range of 8-32Hz.

Exploratory Tests – Hemispheric Differences in Periodicity Coding in Core AFMs

As mentioned above, a widely held view is that the left and right auditory cortices differ in their sensitivity to faster versus slower periodicities. These periodicities may be related to distinct processing time scales in speech perception. The AM rates that fall within these time scales are covered predominantly by our SA bands centered on 4 and 32 Hz (Figure 6.4). Here we report tests for differences in mean proportional SA between the left and right hemispheres in the 4, 8, 16 and 32 Hz periodicity bins. Tests were restricted to AFMs that make up the auditory core (hA1, hR, hRT). The results of each test are given in Table 6.1. In short, no test was significant at the uncorrected (p<0.05) or Bonferroni-corrected (p<0.004167) level of significance. The only test to approach the uncorrected significance level was the 32 Hz bin in field hRT (p=0.063). However, the mean proportional SA in this bin was negligible at 0.0149 (1.5% of the total SA in hRT), and the difference in means (left vs. right) was -0.0098. This difference was also negligible and in the opposite direction of that predicted by hemispheric-differences hypotheses. Thus, these results support the conclusion drawn from qualitative inspection of the data – specifically, there is no evidence of hemispheric differences in periodicity coding.

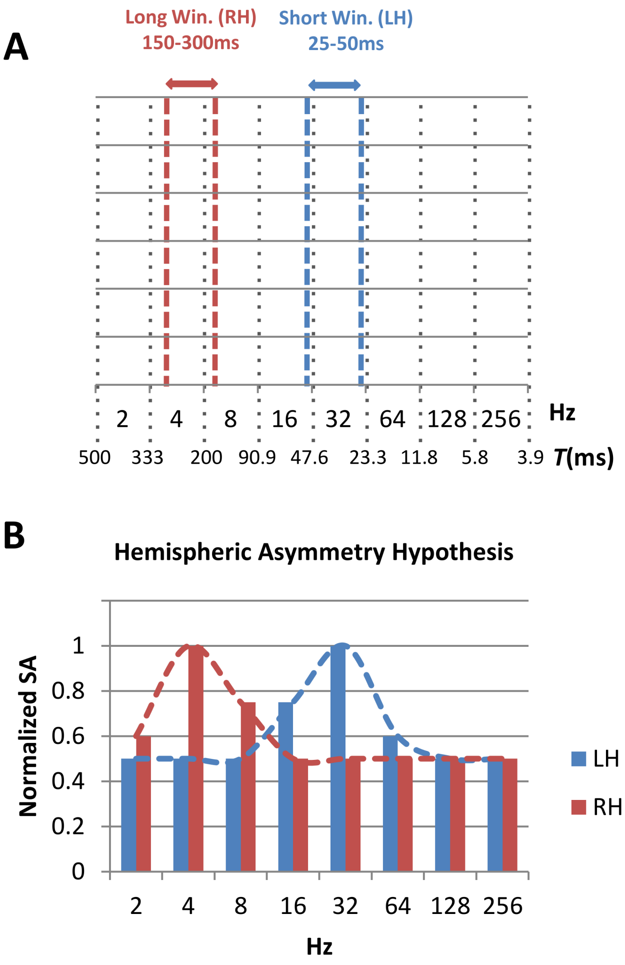


Figure 6.4. Schematic of the hemispheric asymmetry hypothesis. Poeppel's asymmetric sampling in time model argues that the left hemisphere is tuned to fast temporal processing (short window, 25-50 ms time scale = 20-40 Hz), whereas the right hemisphere is tuned to slower temporal processing, which enhances spectral information (long window, 150-300 ms time scale = 3.33-6.67 Hz). (A) Diagram of our 8 periodicity bins in which dotted lines are placed at the bin boundaries. Each bin is labeled with the bin center (Hz) and bin boundaries (period, T, ms; period = 1000/Hz). Poeppel's short (blue) and long (red) windows are overlaid on the diagram with boundaries marked by dashed lines. It is clear from the diagram that these windows strongly overlap the 32 and 4 Hz bins, respectively. (B) Predicted effect of hemispheric asymmetry on periodicity distributions, depicted as a bar graph with dashed fit-lines. The left hemisphere (blue) has a peak in surface area at the 32 Hz bin. The right hemisphere (red) has a peak in surface area at the 4 Hz bin. The predicted distributions account for possible carry-over effects in neighboring bins.

Table 6.1

Surface Area Differences: Between-Hemisphere Paired T-tests

Periodicity Bin	AFM*	Diff	SEM	95% Confidence Interval of the Difference		t	df	p**
				Lower	Upper			
4 Hz	hA1	-0.070	7.338	-23.424	23.284	-0.010	3	0.993
	hR	-10.058	5.344	-27.065	6.950	-1.882	3	0.156
	hRT	-0.948	9.234	-30.333	28.438	-0.103	3	0.925
8 Hz	hA1	3.873	7.996	-21.574	29.319	0.484	3	0.661
	hR	8.340	8.419	-18.452	35.132	0.991	3	0.395
	hRT	18.750	25.411	-62.120	99.620	0.738	3	0.514
16 Hz	hA1	5.433	5.783	-12.971	23.836	0.939	3	0.417
	hR	3.935	5.912	-14.879	22.749	0.666	3	0.553
	hRT	-6.470	7.699	-30.971	18.031	-0.840	3	0.462
32 Hz	hA1	-6.103	9.671	-36.879	24.674	-0.631	3	0.573
	hR	-0.565	2.552	-8.688	7.558	-0.221	3	0.839
	hRT	-0.980	0.339	-2.060	0.100	-2.888	3	0.063

*All contrasts are Left vs Right

**Bonferroni-corrected significance level p < 0.004167

Exploratory Tests – Differences in Periodicity Coding between Individual Core AFMs

Previously we tested for hemispheric differences in periodicity coding. Here, we test for differences in periodicity coding as we move through the cortical processing hierarchy in the auditory core – that is, from hA1 to hR to hRT. Previous research has revealed differences in temporal processing along this gradient ((Bendor & Wang, 2008; Camalier, D'Angelo, Sterbing-D'Angelo, Lisa, & Hackett, 2012). To test for such differences, we collapsed our SA distributions across hemispheres and constructed linear mixed models with fixed factor *field* (hA1, hR, hRT) separately for the 4, 8, 16 and 32 Hz periodicity bins. The effect of *field* was not significant for the 4 Hz bin [F(2,3) = 0.235, p = 0.804]. However, the effect of *field* was significant at both the uncorrected (p < 0.05) and Bonferroni-corrected (P < 0.0167) levels for the 8 Hz bin [F(2,3) = 22.561, p = 0.160], and was significant at the uncorrected level (but did not survive Bonferroni correction) for the 16 Hz bin [F(2,3) = 12.172, p = 0.360]. Once again, the effect of *field* was not significant for the 32 Hz bin but did approach significance at the uncorrected level [F(2,3) = 8.412, p = 0.059].

None of the pairwise comparisons survived Bonferroni correction (p < 0.004167). However, the pattern of the results remains informative (Table 6.2). In the 8 Hz bin, SA was significantly smaller (uncorrected) in hA1 versus hR [t(3) = -6.692, p = 0.007] and hRT [t(3) = -3.517, p = 0.039]. In the 16 Hz bin, SA was significantly larger (uncorrected) in hR versus hRT [t(3) = 3.972, p = 0.029]. In the 32 Hz bin, SA was significantly larger (uncorrected) in hA1 versus hR [t(3) = 3.331, p = 0.045] and there was a trend in this direction for hA1 versus hRT [t(3) = 2.638, p = 0.078]. Thus, SA appears to be spread more evenly across periodicities in the 8-32Hz range in hA1, while SA in hR in hRT is dedicated in increasing proportion to periodicities near 8 Hz (Figure 6.5).

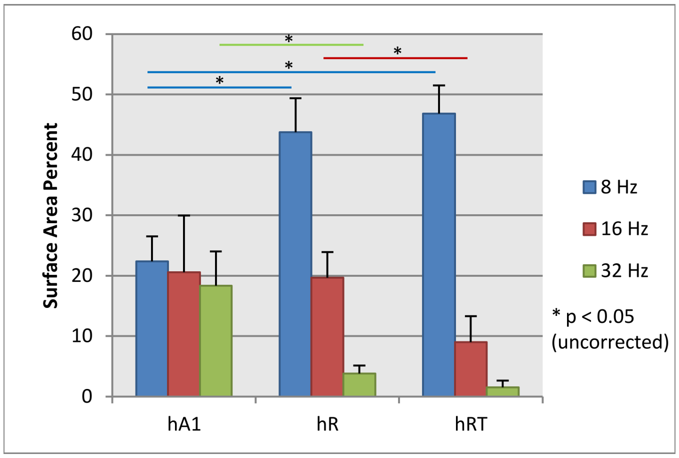


Figure 6.5. Differences in a subsection (8-32 Hz) of the periodicity distributions in the core auditory fields, depicted as a bar graph. The three clusters of bars running left to right reflect surface area in fields hA1, hR and hRT. Surface area is expressed as a percent of the total surface area in each AFM, collapsed across hemispheres and averaged over all 4 participants (error bars are ± 1 SEM). It is clear from the graph that a more restricted representation of periodicity emerges moving from hA1 to hR and hRT – an increasing proportion of surface is dedicated to the 8 Hz bin at the expense of the 16 and 32 Hz bins. Pairwise comparisons (paired t-tests within periodicity bins and between AFMs) that were significant at the uncorrected level of p < 0.05 are shown (colored lines marked with stars; the color indicates the periodicity bin for which the comparison was made). No tests survived Bonferroni correction (p < 0.004167).

Table 6.2

Surface Area Differences: Between-AFM Paired T-tests

Periodicity Bin	Contrast	Diff	SEM	95% Confidence Interval of the Difference		t	df	p*
				Lower	Upper			
4 Hz	hA1 vs hR	-1.705	3.878	-14.045	10.635	-0.440	3	0.690
	hA1 vs hRT	-1.708	2.540	-9.791	6.376	-0.672	3	0.550
	hR vs hRT	-0.003	2.470	-7.865	7.860	-0.001	3	0.999
8 Hz	hA1 vs hR	-21.348	3.190	-31.500	-11.195	-6.692	3	0.007
	hA1 vs hRT	-24.423	6.945	-46.523	-2.322	-3.517	3	0.039
	hR vs hRT	-3.075	6.212	-22.845	16.695	-0.495	3	0.655
16 Hz	hA1 vs hR	0.925	7.028	-21.440	23.290	0.132	3	0.904
	hA1 vs hRT	11.573	5.919	-7.266	30.411	1.955	3	0.146
	hR vs hRT	10.648	2.681	2.117	19.178	3.972	3	0.029

32 Hz	hA1 vs hR	14.550	5.516	-3.004	32.104	2.638	3	0.078
	hA1 vs hRT	16.823	5.050	0.751	32.894	3.331	3	0.045
	hR vs hRT	2.273	1.392	-2.158	6.703	1.632	3	0.201

*Bonferroni-corrected significance level $p < 0.004167$

Follow-up Analysis: Differences in Bandwidth of Periodicity Distributions

In order to characterize the bandwidth (BW), or spread, of SA distributions, we fit RoEx functions to the distribution for each of the 11 AFMs (collapsed across hemispheres) for all four participants. The RoEx function expresses each distribution as a filter in terms of "gain" relative to the maximum proportional SA, whereas this maximum is set to occur at the bin center of the periodicity bin with the largest proportional SA (2, 4, 8, 16, 32, 64, 128 or 256 Hz). Example fits are shown for the core AFMs in (Figure 6.6). For each fit, we report in Table 6.3 the center modulation frequency (CMF), BW, and a centroid (CEN) value (modulation frequency at which half of the area under the RoEx function is accumulated). We chose to analyze BW due to the observation that the spread of periodicity distributions appears to change in an orderly fashion (e.g., from hA1 to hR to hRT as seen in the immediately preceding section). The advantage of the BW measure is that SA data from all eight periodicity bins affect the measure – i.e., BW is a summary measure that parameterizes the entire SA distribution. Here we report the results of a linear mixed model testing for the effect of *field* (hA1, hR, hRT) on BW in core AFMs. We also present summary information relevant to differences in BW across all 11 AFMs.

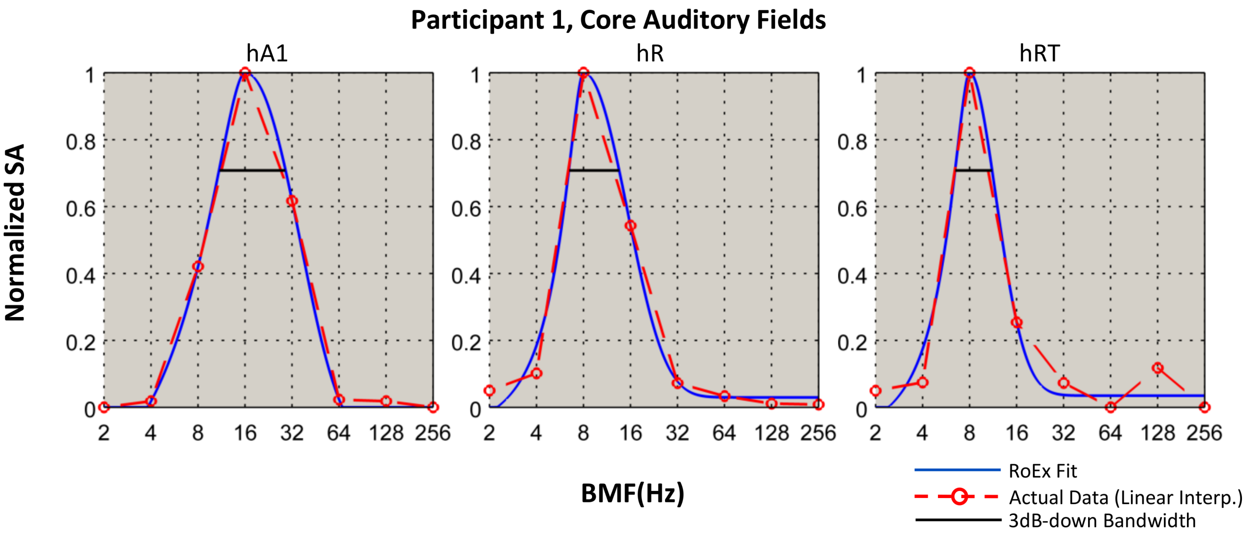


Figure 6.6. Example RoEx fits in a representative subject (Participant 1). Fits are shown for the core auditory fields (hA1, hR, hRT). Red circles mark the actual data composing periodicity distributions, depicted here as normalized surface area. The blue lines depict the best-fitting RoEx function for each distribution (nonlinear least squares). The black line shows the bandwidth of each distribution, measured as the full width of the RoEx function at 0.707 (3dB down from the peak). The center modulation frequencies (i.e., peaks) for hA1, hR, and hRT are 16, 8, and 8 Hz, respectively.

Table 6.3 RoEx fits to surface area data averaged across the two hemispheres. Within each subregion (core, medial belt, lateral belt), the table is organized such that AFMs that run anatomically medial to lateral are positioned left to right (e.g., hA1→hR→hRT). CMF = center modulation frequency, BW = bandwidth at three decibels down from the peak, CEN = centroid (point at which half of the total area under the RoEx curve is accumulated). CMF, BW, and CEN are reported in Hz.

Summary Measures: RoEx Fits in Individual Subjects												
	Core											
				hA1			hR			hRT		
Participant				CMF	BW	CEN	CMF	BW	CEN	CMF	BW	CEN
1				16	18.0	17.3	8	7.0	9.8	8	4.6	8.8
2				4	3.9	5.5	8	10.4	7.4	8	5.6	6.6
3				4	18.0	10.6	8	8.1	8.4	8	4.2	6.4
4				8	15.0	10.0	8	5.6	6.9	8	5.0	6.4
	Medial Belt											
	hCM			hMM			hRM			hRTM		
	CMF	BW	CEN	CMF	BW	CEN	CMF	BW	CEN	CMF	BW	CEN
1	16	25.5	13.8	32	48.9	27.3	8	11.0	9.3	8	5.5	10.8
2	8	6.7	8.0	8	5.7	8.2	8	13.8	11.4	8	3.9	6.7
3	8	5.5	10.2	32	25.2	22.5	8	8.5	10.2	8	4.3	7.7
4	8	6.3	9.5	16	29.2	16.6	8	4.7	8.2	8	6.4	6.7
	Lateral Belt											
	hCL			hML			hAL			hRTL		
	CMF	BW	CEN	CMF	BW	CEN	CMF	BW	CEN	CMF	BW	CEN
1	32	22.3	21.5	16	11.1	14.5	8	8.1	15.8	8	4.2	14.2
2	4	3.7	7.5	8	7.1	7.6	8	8.4	9.8	8	7.8	10.5
3	16	19.8	15.0	8	5.0	8.4	8	6.1	7.4	8	4.3	8.4
4	8	16.6	11.8	8	6.9	7.9	8	11.4	9.0	4	5.5	9.8

In our linear mixed model investigating differences in BW among the core AFMs, the effect of *field* was significant [F(2,3) = 16.206, p = 0.025] with BW decreasing from hA1 to hR and then hRT (13.740→7.780→4.840). In post-hoc testing, only the difference between hR and hRT was significant [t(3) = 3.220, p = 0.049], and only at the uncorrected level (p < 0.05; Bonferroni corrected level is p < 0.01667). Comparisons involving hA1 did not reach significance due to large variance in BW for hA1, stemming from a single participant (Participant 2) that did not follow group pattern (see Table 6.3). Overall, there appears to be a trend, based on BW and SA measurements, in which BMFs decrease moving from hA1 to hR and hRT. In anatomical terms, a vector connecting hA1, hR and hRT moves medial (hA1) to lateral (hRT), from Heschl's gyrus toward the superior temporal gyrus. In other words, based on the trend in the auditory core, medial AFMs tend to represent a broader range of periodicities including a greater proportion of SA representing large BMFs.

To explore this further, we averaged cortical SA distributions across the four participants and fit RoEx functions to these averaged distributions (11 total, one for each AFM). Once again, we calculated CMF, BW, and CEN. AFMs were split into *medial* and *lateral* categories according to the schematic in Figure 6.7A. In Figure 6.7B, we plot CEN against BW for each of the 11 AFMs (see also Table 6.3). Medial AFMs (hCM, hCL, hMM, hA1, hML) are plotted in red and laterl AFMs (hRM, hR, hAL, hRTM, hRT, hRTL) are plotted in blue. In Figure 6.7C, we reproduce Figure 6.7B based on RoEx fits in individual participants. Two features are prominent in the graphs: first, centroid estimates are highly correlated with bandwidth estimates, and second, *lateral* AFMs cluster toward the bottom left while *medial* AFMs cluster to the top right – that is, the cortical SA distributions of *lateral* AFMs are narrower and have less area dedicated to large BMFs than those of *medial* AFMs. (Recall that CEN measures the point at which half of the area under the RoEx function is accumulated, such that smaller values of CEN indicate proportionally less SA dedicated to large BMFs). Indeed, the three points located closest to the bottom left of the graph are the lateral-most AFMs (hRTM, hRT, hRTL). The implication of this result is discussed below.

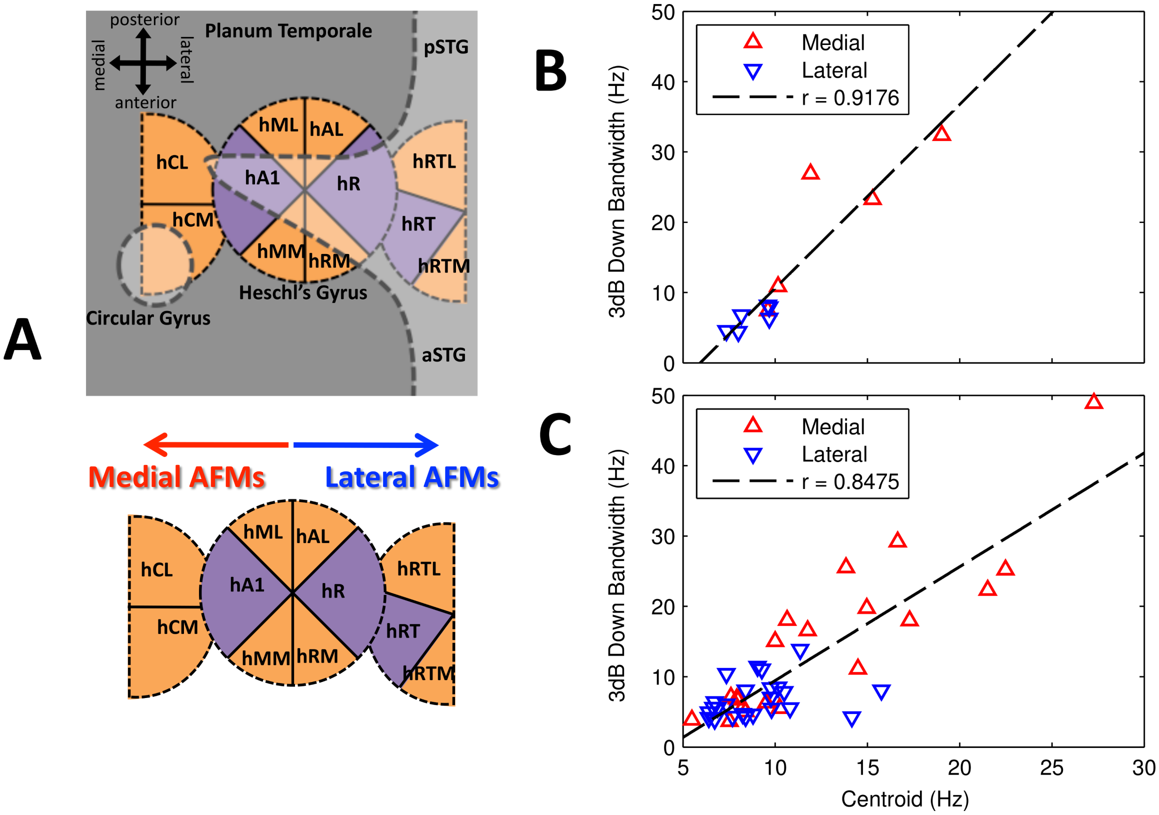


Figure 6.7. Schematic of the organization of human core and belt auditory fields with respect to the relevant anatomy, depicted on a flattened representation of the cortical surface. Dark gray indicates sulci or the plane of the Sylvian fissure, while light gray indicates gyri. Purple regions represent auditory core. Orange regions indicate auditory belt. (A, bottom) Core and belt AFMs are pictured moving medial to lateral from left to right. For descriptive purposes, AFMs were split down the middle into medial and lateral groups (red and blue arrows, respectively). (B) Graphical plot of the characteristics of RoEx fits to the data. Centroid (Hz) is plotted against bandwidth (Hz). In this plot, data points reflect paired observations drawn from RoEx fits to surface area data collapsed across hemispheres and participants in all 11 AFMs. Medial AFMs are plotted in red while lateral AFMs are plotted in blue. The three data points (blue triangles) at the bottom left of the graph represent hRT, hRTL and hRTM. The three data points at the top right of the graph represent hA1, hMM, and hCL. The dotted black line represents the linear relationship between centroid and bandwidth, which are highly correlated (see legend). (C) Structured identically to (B). In this plot, data points reflect paired observations drawn from RoEx fits to surface area data in each individual participant collapsed across hemispheres (4 participants x 11 AFMs = 44 data points).

Discussion

In previous work, we demonstrated the existence of an orderly, topographic gradient in the representation of periodicity in human auditory cortex (Barton et al., 2012). Here, we have argued that the representational properties of this periodotopic place code can be elucidated in terms of the amount cortical surface area dedicated to different modulation frequencies. This notion – i.e., that important functional properties can be captured by cortical magnification of particular regions of sensory space – is common in sensory neurophysiology. Indeed, the work of Penfield and Rasmussen reveals cortical magnification as a key organizational feature in the neural representation of the human body (Penfield & Rasmussen, 1950). Modern visual field mapping techniques make regular use of cortical SA measurements to describe absolute and relative magnification of visual areas (A.A. Brewer & Barton, 2012; Alyssa A Brewer et al., 2005; Robert F Dougherty et al., 2003; Ejima et al., 2003; Horton & Hoyt, 1991), and this research shows that cortical magnification is inversely related to receptive field size (Dumoulin & Wandell, 2008; Kastner et al., 2001; A. T. Smith, Singh, Williams, & Greenlee, 2001; Tootell et al., 1997). Additionally, cortical magnification in V1 correlates with acuity thresholds (Duncan & Boynton, 2003). We believe that the current characterization of SA distributions with respect to temporal auditory information holds similar functional significance.

Periodicity Coding – Cortical Magnification

We initially set out to determine whether periodicity codes in human auditory cortex were uniform across the range of AM rates tested (i.e., whether or not there was cortical magnification of particular AMs). We observed several key features in the distribution of surface area relative to BMF (henceforth “periodicity distribution”) that inform this question. First, the periodicity distribution was not uniform across all modulation frequencies – rather, there was a disproportionate representation of a subrange of frequencies distributed in unimodal form in all 11 core and belt AFMs currently measured. Second, there was relatively little cortical

SA dedicated to very high modulation rates (>32 Hz). Last, the bulk of cortical SA was dedicated to modulation rates between 4 and 32 Hz (i.e., we observed cortical magnification in this region), with peaks in the periodicity distribution typically falling at or near 8 Hz (see Table 6.3, Center Frequency).

These observations are consistent with the animal literature. In cats, the composite temporal modulation transfer function (tMTF) of A1 neurons, computed as the weighted sum of individual tMTFs, is bandpass (unimodal) with a peak at 12.8 Hz (Miller, Escabí, Read, & Schreiner, 2002), and the distributions of BMF in several auditory fields, which are based on counts of single-neuron BMFs, are unimodal for temporal (Schreiner & Urbas, 1988) and rate-based (Miller et al., 2002; Schreiner & Urbas, 1988) coding of modulation rate (although the reliability of this measure depends largely on the number of cells recorded from). The distribution of BMFs extracted from multi-electrode patterns of neural firing in cat A1 also contains a large peak at 8 Hz (Gourevitch & Eggermont, 2010). In squirrel monkeys, the most commonly observed form of MTF in individual cortical neurons is bandpass, with a peak in the distribution of BMFs at 8 Hz in A1 (Bieser & Müller-Preuss, 1996). Single-cell recordings in marmoset A1 reveal unimodal distributions of BMF centered at 8 Hz (temporal coding) or 16 Hz (rate coding) (Liang et al., 2002). In macaques, population MTFs based on firing rate show weak tuning that nonetheless appears to be bandpass with a peak near 8 Hz (Scott, Malone, & Semple, 2011) (although distributions of BMF are not clearly unimodal (Yin, Johnson, O'Connor, & Sutter, 2011)).

Psychophysical measurements provide additional support for the current data. Human behavioral MTFs, based on threshold performance in a multiple-interval modulation detection task, display a lowpass characteristic for broadband noise carriers wherein performance begins to decline at ~10 Hz (Viemeister, 1979). Using narrowband noise carriers, human behavioral MTFs take on more of a bandpass (unimodal) character with peaks often occurring at 8 Hz (Torsten Dau, Kollmeier, & Kohlrausch, 1997). Macaques have a bandpass behavioral MTF with a peak in sensitivity near 20 Hz (Moody, 1994). Human modulation difference limens (detecting increments in modulation frequency) begin to increase (performance declines) at 10 Hz and increase dramatically from 60 Hz (difference limen = 2.61 Hz) to 400 Hz (difference limen = 122 Hz) (Formby, 1985). Macaque difference limens follow a similar pattern although they do not show the same precipitous increase at very high modulation rates (Moody, 1994).

Periodicity Coding – Absence of Hemispheric Differences

Given that all of our sampled AM rates (periodicities) were not equally represented in auditory cortex, we next turn to the question of whether hemispheric differences in periodicity coding can be observed. We failed to identify such hemispheric differences in our measurements, which were drawn from 11 AFMs spanning the auditory core and belt. Of note, theories predicting hemispheric asymmetries in temporal auditory processing are either unclear with respect to the anatomical location and/or processing stage at which these asymmetries should occur (R.J. Zatorre et al., 2002), or clearly predict that asymmetries should occur in nonprimary auditory areas (D. Poeppel, 2003; David Poeppel, 2001). Indeed, recent functional imaging evidence bears out the prediction that hemispheric differences occur in downstream auditory cortical regions, with right hemisphere preferences for longer segment duration (slower temporal modulation) being observed only in posterior temporal regions (including pSTG/STS and planum temporale) in segmented noise (Boemio, Fromm, Braun, & Poeppel, 2005) and speech (Liem, Hurschler, Jäncke, & Meyer, 2013). However, a number of other recent studies support the emergence of hemispheric asymmetries as early as Heschl's gyrus (Belin et al., 1998; A.-L. Giraud et al., 2007; Liégeois-Chauvel, de Graaf, Laguitton, & Chauvel, 1999), including a specific relationship between the cortical volume of HG and observed functional asymmetries (Warrier et al., 2009). In any case, from the present results we can conclude that, should such hemispheric asymmetries exist, they do not result from differential cortical magnification in the two hemispheres at the level of the core and belt auditory regions.

To reconcile with existing data, we should consider where hemispheric differences in temporal processing might in fact lie. Indeed, hemispheric differences may not be detectable in the distribution of cortical SA. Rather, it may be the case that place representations of periodicity, which are symmetric across hemispheres in the current data, are engaged differentially by downstream areas (or interact differentially within core and belt areas) during active stimulus processing in the left versus right hemisphere. It has been hypothesized that these types of interactions are reflected in population level neural oscillations, which vary between hemispheres in auditory cortex (A. L. Giraud & Poeppel, 2012). In addition, as mentioned above, some theories that emphasize hemispheric asymmetries in temporal processing are concerned with temporal integration windows (D. Poeppel, 2003), which are often assessed at the single-neuron level by measuring a cell's upper limit of synchronization (Lu, Liang, & Wang, 2001; Wang et al., 2003). This measure hinges on temporal firing patterns rather than overall firing rate – whereas fine-grained temporal information is not available in the BOLD signal as measured presently – and is often drawn in relation to discrete stimuli (e.g., click trains). However, population temporal integration windows may also be reflected in the maximal firing rate response to a continuous modulation (Wang et al., 2003), analogous to the voxel-wise measure of BMF in the current study.

We should also consider the possibility that differences in cortical magnification exist on a very fine scale. For instance, the critical range of periodicities over which hemispheric differences are hypothesized to occur (~4-40 Hz (David Poeppel, 2001)) is sampled over three stimulus values in the current study. Although our analysis method allows for smooth interpolation over these values (i.e., when calculating BMF), it may be that subtle features of cortical SA distributions were obscured. Furthermore, our power to detect hemispheric differences was limited by sample size, and our exploratory statistical analyses were restricted to the core auditory fields (hA1, hR, hRT). It is feasible that hemispheric differences in cortical magnification exist at a scale detectable only with a very large sample, and that such differences may be present in the belt fields that were not formally assessed presently (however, visual inspection of the data suggests that periodicity codes are quite uniform across hemispheres in these fields, as well). Overall, the effect of hemisphere appears to be quite weak even given our level of power, and we were able to detect other effects, which predominate at the same level of power. As such, the current data lead us to hypothesize that hemispheric differences will not be observed in cortical magnification of particular periodicities even at a finer scale or in a larger sample, which is nontrivial given the current state of the literature.

Periodicity Coding – A Medial to Lateral Gradient in Humans

A second question we set out to address with our data concerned whether or not a decrease in BMF can be observed between AFMs, i.e., does the general trend toward slower periodicities as one moves from the peripheral to the central auditory system continue in the cortical hierarchy? Bendor and Wang recently formalized a model in which the "temporal processing pathway" in primate auditory cortex follows the caudal-to-rostral axis, with increasing temporal integration windows moving rostrally (Bendor & Wang, 2008). This model was based on single-unit data collected from the auditory core (A1, R, RT) of marmosets in response to AM tones. Differences were observed in the temporal coding of AM: population synchronization to AMs was weaker in R and RT than in A1, a larger proportion of A1 neurons synchronized to AMs than in R or RT (and RT had a larger proportion of nonsynchronized neurons than A1 or R), the distribution of temporal BMFs was significantly different in A1 and RT (greater proportion of tBMFs at high modulation frequencies in A1), and a larger proportion of neurons in A1 had maximum synchronization frequencies at high modulation rates compared to R and RT. Moreover, differences were observed in the firing rate response to AM: the average bandwidth of rate-based modulation transfer functions was significantly greater in A1 than RT, and among nonsynchronized neurons average bandwidth was significantly smaller in R and RT compared to A1.

These data match the observations in the current study, in which we have demonstrated that the "bandwidth" of periodicity distributions decreases moving laterally, including clear decreases moving from hA1 to hR and hRT. (Of note, the human AFMs that fall on the medial-to-lateral pathway are homologous with monkey AFMs that fall on the caudal-to-rostral pathway.) A decreasing bandwidth implies the emergence of a more restricted representation of periodicity in the lateral fields, with most of the cortical SA dedicated to populations of neurons with BMFs near 8 Hz. This proportional increase in number of cells that respond maximally over relatively long temporal windows may be related to auditory object information along a medial-to-lateral temporal processing gradient.

Further support for this temporal processing gradient is provided in measurements of the response latency of auditory cortical cells. Response latency, which correlates inversely with maximum synchronization frequency in the response to AM (Bendor & Wang, 2008; Scott et al., 2011), increases moving rostrally in the core, medial belt, and lateral belt of macaques (Camalier et al., 2012). Moreover, while response latencies in A1 appear to drop with increasing tonotopic frequency, neurons in R demonstrate a relatively constant response latency across the tonotopic gradient, perhaps "facilitating the integration of frequency components into a unified auditory object" (Scott et al., 2011) (p. 728). Another recent study confirms increases in response latency and demonstrates increasing stimulus selectivity to monkey calls and other complex sounds moving along the rostral pathway from A1 to RT and further to the more rostral supratemporal plane (Kikuchi, Horwitz, & Mishkin, 2010). These data provide loose support for a prominent model of human auditory processing in which auditory object formation follows a ventral stream along an anterolateral pathway from primary auditory cortex (Rauschecker & Scott, 2009). However, the aforementioned difference between primate and human anatomy must be respected: the current data support a temporal processing pathway that moves medial to lateral in human auditory cortex. This characterization is supported by localization of complex auditory processes such as speech perception to the superior temporal gyrus and sulcus lateral to core and belt auditory cortex (Binder et al., 2000; Hickok & Poeppel, 2007; Price, 2010).

Appendix

The rounded exponential or "RoEx" filter has been successfully used to model the response of peripheral auditory filters (Moore & Glasberg, 1983). Its shape is given by the equation

$$W(g) = (1 - r)(1 + pg)e^{-pg} + r$$

where g is the deviation from the filter center frequency (or, in our case, the *center modulation frequency*, CMF) divided by the CMF, p determines the passband of the filter, and r determines the point at which the passband gives way to the shallower tails of the filter. This shape closely approximates the shape of our proportional surface area distributions. In order to fit the RoEx function to a surface area distribution in a given AFM, proportional surface area measurements in each periodicity bin were normalized such that the bin with the largest proportional surface area was assigned a value of 1. The distributions were thus expressed in terms of "gain" relative to the peak value of 1. The periodicity (i.e., the AM rate of the stimulus) at the center of the peak bin was treated as the CMF, and values for g were thus determined by subtracting the CMF from each bin center (our AM stimulus values: 2, 4, 8, 16, 32, 64, 128, 256 Hz). In one case, the CMF was forced to the center of three bins with gain values close to 1. Values for p and r were determined by performing a nonlinear least squares fit of g on the gain values describing the surface area distribution. Separate fits were performed to the left and right of the CMF such that p and r were estimated separately for each side of the RoEx filter. Fits were carried out using the MATLAB curve-fitting toolbox with the 'Trust-Region' algorithm and lower bounds of [0, -5] on *p* and *r*, respectively, upper bounds of [5, 50] on *p* and *r*, respectively, and starting points [3, 0] on *p* and *r*, respectively. In cases for which a negative value for *r* provided the best fit, values of *W(g)* were forced to a minimum of zero. Additionally, fits were truncated at the first zero crossing (bins to the left or right of the CMF for which gain was less than 0.05). In some cases, there were not sufficient degrees of freedom to estimate the *r* parameter and only a value for *p* was estimated. If the CMF was located at an endpoint bin, parameters were estimated for only one side of the RoEx filter. Fits with a total adjusted R^2 less than 0.3 were discarded. Once the full set of parameters was estimated, the function *W(g)* was specified computationally over the full range of *g* with a granularity of 0.01. The function was then shifted by converting the vector of *g* values back to values of AM rate in Hz. Bandwidth was calculated computationally as the full width (Hz) of the estimated RoEx equation at 3 dB down from the peak (from 0.707 on the left of the CMF to 0.707 on the right of the CMF, where the peak at the CMF is set to 1). A value for the centroid (Hz) of each distribution was calculated computationally as the value at which the cumulative area under the curve *W(g)* reached 50% of its maximum.

References

Ahissar, E., Nagarajan, S., Ahissar, M., Protopapas, A., Mahncke, H., & Merzenich, M.M. (2001). Speech comprehension is correlated with temporal response patterns recorded from auditory cortex. *Proceedings of the National Academy of Sciences, 98*(23), 13367-13372.

Attias, H., & Schreiner, CE. (1997). Temporal low-order statistics of natural sounds. *Advances in neural information processing systems*, 27-33.

Bagiella, Emilia, Sloan, Richard P, & Heitjan, Daniel F. (2000). Mixed-effects models in psychophysiology. *Psychophysiology, 37*(1), 13-20.

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

Baumann, S., Griffiths, T. D., Sun, L., Petkov, C. I., Thiele, A., & Rees, A. (2011). Orthogonal representation of sound dimensions in the primate midbrain. *Nat Neurosci, 14*(4), 423-425. doi: 10.1038/nn.2771

Belin, Pascal, Zilbovicius, Monica, Crozier, Sophie, Thivard, Lionel, Fontaine, Anne, Masure, Marie-Cécile, & Samson, Yves. (1998). Lateralization of speech and auditory temporal processing. *Journal of Cognitive Neuroscience, 10*(4), 536-540.

Bendor, D., & Wang, X. (2008). Neural response properties of primary, rostral, and rostrotemporal core fields in the auditory cortex of marmoset monkeys. *J Neurophysiol, 100*(2), 888-906. doi: 10.1152/jn.00884.2007

Bieser, A., & Müller-Preuss, P. (1996). Auditory responsive cortex in the squirrel monkey: neural responses to amplitude-modulated sounds. *Experimental Brain Research, 108*(2), 273-284.

Binder, JR, Frost, JA, Hammeke, TA, Bellgowan, PSF, Springer, JA, Kaufman, JN, & Possing, ET. (2000). Human temporal lobe activation by speech and nonspeech sounds. *Cerebral Cortex, 10*(5), 512-528.

Boemio, A., Fromm, S., Braun, A., & Poeppel, D. (2005). Hierarchical and asymmetric temporal sensitivity in human auditory cortices. *Nat Neurosci, 8*(3), 389-395. doi: 10.1038/nn1409

Brewer, A.A., & Barton, B. (2012). Visual field map organization in human visual cortex. *Visual Cortex, InTech*.

Brewer, Alyssa A, Liu, Junjie, Wade, Alex R, & Wandell, Brian A. (2005). Visual field maps and stimulus selectivity in human ventral occipital cortex. *Nat Neurosci, 8*(8), 1102-1109.

Camalier, C.R., D'Angelo, W.R., Sterbing-D'Angelo, S.J., Lisa, A., & Hackett, T.A. (2012). Neural latencies across auditory cortex of macaque support a dorsal stream supramodal timing advantage in primates. *Proceedings of the National Academy of Sciences, 109*(44), 18168-18173.

Chandrasekaran, Chandramouli, Trubanova, Andrea, Stillittano, Sébastien, Caplier, Alice, & Ghazanfar, Asif A. (2009). The natural statistics of audiovisual speech. *PLoS computational biology, 5*(7), e1000436.

Dau, T., Verhey, J., & Kohlrausch, A. (1999). Intrinsic envelope fluctuations and modulation-detection thresholds for narrow-band noise carriers. *The Journal of the Acoustical Society of America, 106*, 2752.

Dau, Torsten, Kollmeier, Birger, & Kohlrausch, Armin. (1997). Modeling auditory processing of amplitude modulation. I. Detection and masking with narrow-band carriers. *The Journal of the Acoustical Society of America, 102*, 2892.

DeYoe, Edgar A, Carman, George J, Bandettini, Peter, Glickman, Seth, Wieser, Jon, Cox, Robert, . . . Neitz, Jay. (1996). Mapping striate and extrastriate visual areas in human cerebral cortex. *Proceedings of the National Academy of Sciences, 93*(6), 2382-2386.

Dougherty, R. F., Koch, V. M., Brewer, A. A., Fischer, B., Modersitzki, J., & Wandell, B. A. (2003). Visual field representations and locations of visual areas V1/2/3 in human visual cortex. *J Vis, 3*(10), 586-598. doi: 10:1167/3.10.1

/3/10/1/ [pii]

Dougherty, Robert F, Koch, Volker M, Brewer, Alyssa A, Fischer, Bernd, Modersitzki, Jan, & Wandell, Brian A. (2003). Visual field representations and locations of visual areas V1/2/3 in human visual cortex. *Journal of Vision, 3*(10).

Dumoulin, Serge O, & Wandell, Brian A. (2008). Population receptive field estimates in human visual cortex. *Neuroimage, 39*(2), 647-660.

Duncan, Robert O, & Boynton, Geoffrey M. (2003). Cortical magnification within human primary visual cortex correlates with acuity thresholds. *Neuron, 38*(4), 659-671.

Ejima, Yoshimichi, Takahashi, Shigeko, Yamamoto, Hiroki, Fukunaga, Masaki, Tanaka, Chuzo, Ebisu, Toshihiko, & Umeda, Masahiro. (2003). Interindividual and interspecies variations of the extrastriate visual cortex. *Neuroreport, 14*(12), 1579-1583.

Engel, Stephen A. (1994). fMRI measurements of human visual cortex.

Formby, C. (1985). Differential sensitivity to tonal frequency and to the rate of amplitude modulation of broadband noise by normally hearing listeners. *The Journal of the Acoustical Society of America, 78*, 70.

Giraud, A. L., & Poeppel, D. (2012). Cortical oscillations and speech processing: emerging computational principles and operations. *Nat Neurosci, 15*(4), 511-517. doi: 10.1038/nn.3063

Giraud, Anne-Lise, Kleinschmidt, Andreas, Poeppel, David, Lund, Torben E, Frackowiak, Richard SJ, & Laufs, Helmut. (2007). Endogenous cortical rhythms determine cerebral specialization for speech perception and production. *Neuron, 56*(6), 1127-1134.

Gourevitch, B., & Eggermont, J. J. (2010). Maximum decoding abilities of temporal patterns and synchronized firings: application to auditory neurons responding to click trains and amplitude modulated white noise. *J Comput Neurosci, 29*(1-2), 253-277. doi: 10.1007/s10827-009-0149-3

Hickok, Gregory, & Poeppel, David. (2007). The cortical organization of speech processing. *Nat Rev Neurosci, 8*(5), 393-402.

Horton, Jonathan C, & Hoyt, William F. (1991). The representation of the visual field in human striate cortex. A revision of the classic Holmes map. *Archives of ophthalmology, 109*(6), 816.

Joris, PX, Schreiner, CE, & Rees, A. (2004). Neural processing of amplitude-modulated sounds. *Physiological Reviews, 84*(2), 541-577.

Kastner, Sabine, De Weerd, Peter, Pinsk, Mark A, Elizondo, M Idette, Desimone, Robert, & Ungerleider, Leslie G. (2001). Modulation of sensory suppression: implications for receptive field sizes in the human visual cortex. *J Neurophysiol, 86*(3), 1398-1411.

Kikuchi, Yukiko, Horwitz, Barry, & Mishkin, Mortimer. (2010). Hierarchical auditory processing directed rostrally along the monkey's supratemporal plane. *The Journal of Neuroscience, 30*(39), 13021-13030.

Langner, G., Dinse, H. R., & Godde, B. (2009). A map of periodicity orthogonal to frequency representation in the cat auditory cortex. *Front Integr Neurosci, 3*, 27. doi: 10.3389/neuro.07.027.2009

Langner, G., Sams, M., Heil, P., & Schulze, H. (1997). Frequency and periodicity are represented in orthogonal maps in the human auditory cortex: evidence from magnetoencephalography. *Journal of Comparative Physiology A: Neuroethology, Sensory, Neural, and Behavioral Physiology, 181*(6), 665-676.

Liang, L., Lu, T., & Wang, X. (2002). Neural representations of sinusoidal amplitude and frequency modulations in the primary auditory cortex of awake primates. *J Neurophysiol, 87*(5), 2237-2261.

Liégeois-Chauvel, Catherine, de Graaf, Jozina B, Laguitton, Virginie, & Chauvel, Patrick. (1999). Specialization of left auditory cortex for speech perception in man depends on temporal coding. *Cerebral Cortex, 9*(5), 484-496.

Liem, Franziskus, Hurschler, Martina A, Jäncke, Lutz, & Meyer, Martin. (2013). On the planum temporale lateralization in suprasegmental speech perception: Evidence from a study investigating behavior, structure, and function. *Hum Brain Mapp*.

Lu, T., Liang, L., & Wang, X. (2001). Temporal and rate representations of time-varying signals in the auditory cortex of awake primates. *Nat Neurosci, 4*(11), 1131-1138. doi: 10.1038/nn737

Luo, H., & Poeppel, D. (2007). Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. *Neuron, 54*(6), 1001.

Maes, F., Collignon, A., Vandermeulen, D., Marchal, G., & Suetens, P. (1997). Multimodality image registration by maximization of mutual information. *IEEE Trans Med Imaging, 16*(2), 187-198. doi: 10.1109/42.563664

Miller, Lee M, Escabí, Monty A, Read, Heather L, & Schreiner, Christoph E. (2002). Spectrotemporal receptive fields in the lemniscal auditory thalamus and cortex. *J Neurophysiol, 87*(1), 516-527.

Moody, David B. (1994). Detection and discrimination of amplitude-modulated signals by macaque monkeys. *The Journal of the Acoustical Society of America, 95*, 3499.

Moore, Brian CJ, & Glasberg, Brian R. (1983). Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. *The Journal of the Acoustical Society of America, 74*, 750.

Nelken, Israel, Rotman, Yaron, & Yosef, Omer Bar. (1999). Responses of auditory-cortex neurons to structural features of natural sounds. *Nature, 397*(6715), 154-157.

Nestares, O., & Heeger, D. J. (2000). Robust multiresolution alignment of MRI brain volumes. *Magn Reson Med, 43*(5), 705-715. doi: 10.1002/(SICI)1522-2594(200005)43:5<705::AID-MRM13>3.0.CO;2-R [pii]

Nourski, K.V., Reale, R.A., Oya, H., Kawasaki, H., Kovach, C.K., Chen, H., . . . Brugge, J.F. (2009). Temporal envelope of time-compressed speech represented in the human auditory cortex. *The Journal of Neuroscience, 29*(49), 15564-15574.

Penfield, Wilder, & Rasmussen, Theodore. (1950). The cerebral cortex of man; a clinical study of localization of function.

Poeppel, D. (2003). The analysis of speech in different temporal integration windows: cerebral lateralization as 'asymmetric sampling in time'. *Speech Communication, 41*(1), 245-255.

Poeppel, David. (2001). Pure word deafness and the bilateral processing of the speech code. *Cognitive Science, 25*(5), 679-693.

Price, Cathy J. (2010). The anatomy of language: a review of 100 fMRI studies published in 2009. *Annals of the New York Academy of Sciences, 1191*(1), 62-88.

Rauschecker, Josef P, & Scott, Sophie K. (2009). Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. *Nature neuroscience, 12*(6), 718-724.

Saberi, K., & Perrott, D.R. (1999). Cognitive restoration of reversed speech. *Nature, 398*(6730), 760-760.

Schreiner, C.E., & Urbas, J.V. (1988). Representation of amplitude modulation in the auditory cortex of the cat. II. Comparison between cortical fields. *Hear Res, 32*(1), 49-63.

Schulze, H., Hess, A., Ohl, F.W., & Scheich, H. (2002). Superposition of horseshoe-like periodicity and linear tonotopic maps in auditory cortex of the Mongolian gerbil. *European Journal of Neuroscience, 15*(6), 1077-1084.

Scott, B. H., Malone, B. J., & Semple, M. N. (2011). Transformation of temporal processing across auditory cortex of awake macaques. *J Neurophysiol, 105*(2), 712-730. doi: 10.1152/jn.01120.2009

Sereno, Martin I, Dale, AM, Reppas, JB, Kwong, KK, Belliveau, JW, Brady, TJ, . . . Tootell, RB. (1995). Borders of multiple visual areas in humans revealed by functional magnetic resonance imaging. *Science, 268*(5212), 889-893.

Shannon, R.V., Zeng, F.G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *SCIENCE-NEW YORK THEN WASHINGTON-*, 303-303.

Singh, Nandini C, & Theunissen, Frédéric E. (2003). Modulation spectra of natural sounds and ethological theories of auditory processing. *The Journal of the Acoustical Society of America, 114*, 3394.

Smith, Andrew T, Singh, Krish Devi, Williams, AL, & Greenlee, MW. (2001). Estimating receptive field size from fMRI data in human striate and extrastriate visual cortex. *Cerebral Cortex, 11*(12), 1182-1190.

Smith, Z.M., Delgutte, B., & Oxenham, A.J. (2002). Chimaeric sounds reveal dichotomies in auditory perception. *Nature, 416*(6876), 87-90.

Tallal, Paula, Miller, Steve, & Fitch, Roslyn Holly. (1993). Neurobiological basis of speech: a case for the preeminence of temporal processing. *Ann N Y Acad Sci, 682*(1), 27-47.

Teo, P. C., Sapiro, G., & Wandell, B. A. (1997). Creating connected representations of cortical gray matter for functional MRI visualization. *IEEE Trans Med Imaging, 16*(6), 852-863.

Tootell, Roger BH, Mendola, Janine D, Hadjikhani, Nouchine K, Ledden, Patrick J, Liu, Arthur K, Reppas, John B, . . . Dale, Anders M. (1997). Functional analysis of V3A and related areas in human visual cortex. *The Journal of Neuroscience, 17*(18), 7060-7078.

Viemeister, Neal F. (1979). Temporal modulation transfer functions based upon modulation thresholds. *The Journal of the Acoustical Society of America, 66*, 1364.

Wandell, B. A., Brewer, A. A., & Dougherty, R. F. (2005). Visual field map clusters in human cortex. *Philos Trans R Soc Lond B Biol Sci, 360*(1456), 693-707. doi: 10.1098/rstb.2005.1628

Wandell, B. A., Chial, S., & Backus, B. (2000). Visualization and Measurement of the Cortical Surface. *J. of Cognitive Neuroscience, 12*(5), 739-752.

Wandell, B. A., Chial, S., & Backus, B. T. (2000). Visualization and measurement of the cortical surface. *J Cogn Neurosci, 12*(5), 739-752.

Wandell, B. A., Dumoulin, S. O., & Brewer, A. A. (2007). Visual field maps in human cortex. *Neuron, 56*(2), 366-383. doi: 10.1016/j.neuron.2007.10.012

Wandell, Brian A, & Winawer, Jonathan. (2011). Imaging retinotopic maps in the human brain. *Vision research, 51*(7), 718-737.

Wang, Xiaoqin, Lu, Thomas, & Liang, Li. (2003). Cortical processing of temporal modulations. *Speech Communication, 41*(1), 107-121. doi: 10.1016/s0167-6393(02)00097-3

Warrier, Catherine, Wong, Patrick, Penhune, Virginia, Zatorre, Robert, Parrish, Todd, Abrams, Daniel, & Kraus, Nina. (2009). Relating structure to function: Heschl's gyrus and acoustic processing. *The Journal of Neuroscience, 29*(1), 61-69.

Yin, P., Johnson, J. S., O'Connor, K. N., & Sutter, M. L. (2011). Coding of amplitude modulation in primary auditory cortex. *J Neurophysiol, 105*(2), 582-600. doi: 10.1152/jn.00621.2010

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

Zatorre, Robert J. (1997). Hemispheric specialization of human auditory processing: Perception of speech and musical sounds. *Advances in Psychology, 123*, 299-323.

CONCLUDING REMARKS

I will make no great efforts to provide a comprehensive link between the results from each of the preceding investigations. The motivation for the research program as a whole and for each of the individual investigation has been given in the Introduction and Primers for each chapter, respectively. In truth, the three approaches to understanding human speech processing adopted in the current work are largely unrelated at a detailed level. They unite under the broad goal of describing the neural and computational mechanisms underlying speech perception and production, but the contributions to this description under each approach are rather unique. As such, I will only return, briefly, to the objectives for each approach as outlined in the Introduction, so as to provide a report on overall progress.

(1) Clarify what speech perception *is* by establishing what speech perception *is not*.

Chapter 1, which was merely a synthesis of existing work, established (or reiterated, at least) that the motor system does not contribute meaningfully to speech perception. As such, we can conclude that the objects of speech perception *are not* articulatory gestures (barring extreme interpretations of how articulatory gestures are perceived; if they are perceived by the auditory system, why quibble). Chapter 2 provided a convincing explanation for motor system contributions to laboratory speech tasks – namely, activity in the speech motor system is modulated by top-down decision mechanisms in (unnatural) tasks in which an overt response must be generated. Overall, I think this line of work can largely be left behind. It is clear that speech perception is a sensory process and it will be most worthwhile to examine it as such.

(2) Establish how speech systems interface with signals from multiple sensory modalities.

Chapters 3-5 approached this objective by examining audiovisual speech processing. Chapter 3 demonstrated that dynamic visual features corresponding to individual articulatory events are integrated with auditory speech signal to support perception. This appeared to be done in a bottom-up fashion – the weight of each visual feature was determined by its salience and temporal separation from the auditory signal. Chapter 4 identified potential neural mechanisms for both visual feature extraction and audiovisual speech integration in the superior temporal sulcus. Chapter 5 suggested that visual speech, like auditory speech, functions to support motor control for speech production. New neuroimaging evidence supported the existence of a visual-to-motor integration pathway for speech. I speculated that this pathway forms during development of productive speech capacities and remains functional into adulthood (to service vocal tract motor control). Overall, there is much still to be done in the area of multisensory integration in speech. In particular, the processing stage at which visual and auditory speech signals are combined and precisely where in the brain these signal converge on high-level speech sound representations are facts that remain elusive. The precise organization and function of visual-to-motor speech integration networks remains to be specified. Everything we know at this point is based on tasks involving repetition of (audio)visual speech, which is fairly unnatural. The contributions of sensorimotor integration networks for visual speech will need to be assessed using a larger variety of tasks.

(3) Establish the organization of central representations of auditory signals.

This line of work is truly in its infancy. We have only just developed techniques for high resolution mapping of cortical auditory representations. The results of Chapter 6 provided some basic facts about the organization of central representations for one particular feature of sound – low frequency temporal modulations. While these facts were informative, descriptions of the large-scale organization of the auditory cortical processing hierarchy were mostly speculative. Thus far, data have only been collected from a handful of subjects using one set of stimuli. Moreover, our state-of-the-art imaging techniques will soon be improved upon (i.e., with experience and fine-tuning). The future for this approach to understanding human speech processing is bright and wide open.

And that's all.