Functional Organization of the Human Superior Temporal Sulcus

by
Ben Deen

B.S. Physics and Cognitive Science
Yale University, 2009

Submitted to the Department of Brain and Cognitive Sciences in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Neuroscience
at the Massachusetts Institute of Technology

February, 2016

Signature of Author: Signature redacted
Department of Brain and Cognitive Sciences
December 8, 2015

Certified by: Signature redacted
Rebecca Saxe
Professor of Brain and Cognitive Sciences
Thesis Supervisor

Certified by: Signature redacted
Nancy Kanwisher
Walter A. Rosenblith Professor of Cognitive Neuroscience
Thesis Supervisor

Certified by: Signature redacted
Matthew A. Wilson
Sherman Fairchild Professor of Neuroscience and Picower Scholar
Director of Graduate Education for Brain and Cognitive Sciences

Functional Organization of the Human Superior Temporal Sulcus

By Ben Deen

Submitted to the Department of Brain and Cognitive Sciences on December 1 in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Neuroscience

Abstract

As we observe and listen to other people, we interpret their actions in terms of a rich causal structure, driven by underlying mental states and dispositional traits. The ability to rapidly extract abstract social information from perceptual input, termed social perception, is critical to human social behavior. This thesis investigates the cognitive architecture of social perception by studying the functional organization and development of brain regions implicated in this process. In the first Chapters 2-4, I focus on a region that has been strongly implicated in social perception: the superior temporal sulcus (STS). In Chapter 2, I assess the overall functional organization of STS responses to different types of social stimuli, using fMRI. I find that the STS comprises a number of functionally specific subregions that process certain types of social information, such as body movement, vocal sounds, linguistic input, and abstract mental states, suggesting a functional division of labor in social perception. I also identify a multimodal region that responds both to face movement and to vocal sounds. Chapters 3-4 further explore the properties of this region, termed the fSTS. Analyzing spatial patterns of response in this region to different types of face movement, I find evidence that fSTS contains a parts-based representation of perceived face movement type, suggesting a representation more tied to face movement kinematics than implied mental state, but which generalizes across low-level visual properties (Chapter 3). Next, assessing responses to a range of naturalistic face movements and vocal sounds, I find that the fSTS responds strongly to virtually any face movement or vocal sound, irrespective of social relevance or speech content (Chapter 4). However, patterns of response in this region distinguish more and less socially relevant inputs, in a manner that generalizes across facial and vocal stimuli. Taken together, these results point to the fSTS as a mid-level region in social perceptual inference, with representations that are still tied to perceptual features, but begin to integrate visual and auditory inputs and make explicit high-level social distinctions. Lastly, in Chapter 5 I address a broader, related question: how do functionally specific brain regions such as the fSTS develop? To address this question, I develop methods for fMRI in awake infants viewing visual stimuli. I find that regions preferring specific visual categories (faces and scenes) exist by 4-6 months of age, but that responses in these regions are less selective than in adults. This suggests that functionally specific brain regions exist in some form at an early age, but that they become increasingly specialized throughout development.

Thesis Supervisors: Rebecca Saxe and Nancy Kanwisher
Titles: Professor of Brain and Cognitive Sciences & Walter A. Rosenblith Professor of Cognitive Neuroscience

Acknowledgements

Throughout my graduate career, I've had the pleasure of working and interacting with many of the most thoughtful, intelligent, and supportive people I have ever met. This page will express a tiny fraction of the gratitude I owe them.

First, I'm deeply grateful for my incredible advisors and scientific parents (and grandparent), Rebecca Saxe and Nancy Kanwisher. They are ideal role models in both science and life: incredibly passionate about and talented at what they do, tireless educators, mentors and supporters of their students, and also just people that I like to spend time with. I cannot imagine a better pair of mentors.

I'm also thankful to the other members of my thesis committee, Winrich Freiwald and Josh Tenenbaum, for teaching me about monkeys and priors, for providing useful feedback and novel perspectives along the way.

I'm grateful for all of my colleagues and friends in the Saxe Lab (Nick, Hilary, Jorie, Hyo, Liane, Zeynep, Emile, Todd, Marina, Elizabeth, Mina, Alex P, Amy, Julianne, Nir, Grace, Heather, Laura, A.J., Dorit, Lindsey, Stefano, Tyler, Dima, and Livia), and Kanwisher Lab (Danny, Brown, Julie, David P, Ev, Kami, Sarah, Josh, Sam, Alex K, Jason, Idan, Rosa, Matt, Sam G, Terri, Jenelle, Walid, Harris, Zeynep again, Caroline, Michael, Leyla, and of course Charlie), for their support, intellectual stimulation, coffee/beer breaks, and generally being some of the best people, with whom I look forward to maintaining lifelong friendships. Specific shout-outs go to Nick, for his dance moves; Hyo, for beer outings and 3AM scientific discussions; Alex P, for being the nicest person ever; Rosa, for hugs, jokes, and real talk; Matt, for bro dates; Alex K, for always being down to talk about consciousness; and Jorie, for all the things. I'm also grateful to friends from the McDermott, Tenenbaum, Schulz, DiCarlo, and Conway labs, as well as Harvard Psychology, for great discussions over the years, scientific and otherwise.

Thanks to the MRI team – Christina, Atsushi, Steve, and Sheeba – for making everything work, and quickly saving the day whenever things didn't work. Thanks also to participants in my studies – adults, infants, and parents of infants – for spending long hours lying still in a tube for the sake of science.

Thanks to the many lifelong friends I made before coming to grad school, including the cove crew and D12 gang.

And lastly, thanks so, so, so much to my parents, Robin Herbert and Darwin Deen, for making all of this possible in the first place, and for their love and encouragement throughout my life.

Table of Contents

Chapter 1: Introduction

1.1 Background and Motivation for Thesis Work

Humans are richly social creatures, spending much of our time interacting with or thinking about other people. As we observe and interact with others, we constantly make inferences about the underlying causes of their behavior, including their current mental states (thoughts, intentions, emotional states, and percepts) and more persistent dispositional traits (such as personality and competence). Remarkably, many of these inferences are made from perceptual cues – aspects of facial expressions, body posture and movements, and vocal sounds – rather than requiring explicit linguistic information. This ability to infer abstract social properties from perceptual input, termed social perception, is essential to human social interaction, and constitutes a case study of the rapid inductive inference at which humans excel.

Research in cognitive and social psychology over the past several decades has documented many impressive cases of social perceptual inference, drawing far-reaching conclusions from minimal cues. From simple animations of geometric shapes moving through space, such as the animations developed by Heider and Simmel, we can extract information about agents' intentions, knowledge states, emotional states (scared, angry), interactions (chase, escape), personality traits (bully), and relationships (friend, enemy) (Heider and Simmel 1944; Abell et al. 2000; Klin 2000). From animations of human body movements reduced to collections of point-lights placed on actors' limbs, we can extract information about others' emotional states (Dittrich et al. 1996; Pollick et al. 2001; Atkinson et al. 2004; Clarke et al. 2005). From brief, silent videos of other humans

("thin slices"), we can extract information about dispositional traits such as extraversion, trait anxiety, intelligence, professional competence, and sexual orientation (Gangestad et al. 1992; Ambady and Rosenthal 1993; Harrigan et al. 1996; Ambady et al. 1999; Murphy et al. 2003).

Despite this growing catalog of social perceptual inferences made by humans, we still understand little about how these inferences are performed. This inferential process begins with low-level perceptual inputs and ends with abstract social properties, but the path between these two remains largely a black box. In particular, we have little sense of the mid-level representations that lie between inputs and final outputs, or of the computations performed over these representations. To make progress on this problem, this thesis employs a strategy that has been successful in the domain of perception: probing the neural basis of a process as a way to provide insight into its cognitive and computational architecture. Specifically, I investigate the cognitive architecture of social perception by probing the functional organization and properties of a large brain region broadly implicated social perception: the human superior temporal sulcus (STS).

Regions of the STS have been implicated in numerous aspects of social perception and cognition, such as perceiving faces (Puce et al. 1996; Haxby et al. 2000; Pitcher et al. 2011), hand and body movements (Bonda et al. 1996; Allison et al. 2000; Grossman et al. 2000; Grossman and Blake 2002; Pelphrey, Mitchell, et al. 2003; Pelphrey et al. 2005), and vocal and speech sounds (Belin et al. 2000; Binder et al. 2000; Wright et al. 2003; Fecteau et al. 2004; Kreifelts et al. 2009; Overath et al. 2015), as well as understanding others' actions (Pelphrey, Singerman, et al. 2003; Pelphrey et al. 2004; Brass et al. 2007; Vander Wyk et al. 2009) and mental states (Fletcher et al. 1995;

Gallagher et al. 2000; Saxe and Kanwisher 2003; Saxe and Powell 2006; Ciaramidaro et al. 2007). This body of literature indicates an extensive and varied role of the STS in processing social information, and makes this region a prime candidate for a neural substrate of social perceptual inference. However, many questions about the nature of STS involvement in social perception remain open. Because most prior studies have only investigate a single or small number of domains of social stimuli (faces, bodies, voices, abstract mental states, etc), little is known about the relationship between STS responses to these different types of input. This leads to a basic but critical question: what are the functional subunits of the STS? Does this broad region contain subregions that selectively process certain types of social information, suggesting separate streams of processing? Does it contain regions that respond to multiple types of information and might play an integrative role? I address this question by analyzing STS responses to multiple types of social information within individual humans, and characterizing the overall functional organization of responses to these stimuli.

Having characterized this broad scale organization, and equally important question is what each functional subunit is doing. By targeting regions and characterizing their responses to more specific inputs, we can make inferences about their functional role and contribution to social perception. Here I focus on the face-responsive STS subregion (termed fSTS). This region has previously been interpreted as part of a network of face-selective visual regions, with the unique role of processing changeable aspects of faces, such as emotional expression and gaze direction, rather than identity (Haxby *et al.* 2000). Consistent with this view, studies using static face stimuli have indicated that this region encodes aspects of facial expression, more so than identity

(Andrews and Ewbank 2004; Winston et al. 2004; Harris et al. 2012). However, several recent findings suggest the need for a reconceptualization of this region. Unlike other face-preferring regions, the fSTS responds substantially more strongly to dynamic faces than static images of faces, suggesting a specialization for processing facial motion (Schultz and Pilz 2009; Pitcher *et al.* 2011; Schultz et al. 2013; Bernstein and Yovel 2015). Furthermore, I demonstrate (Chapter 2) that this region responds as strongly to vocal sounds as to dynamic faces. This intriguing response profile suggests that the fSTS would be well placed to play a role in extracting social information from dynamic facial and vocal cues, but also suggests that new stimuli and approaches will be needed to understand the functional role of this region. I therefore use a combination of dynamic facial and vocal stimuli to probe the properties of the fSTS.

Lastly, a fundamental question about the functionally specialized brain regions studied here is how they develop. Do functionally specialized parts of the STS exist early in development, supporting early social behaviors, or do they develop gradually over time, perhaps in virtue of extensive experience with other humans? Studies using near-infrared spectroscopy (NIRS) have indicated the presence of responses to social over nonsocial stimuli in temporal cortex, within the first year of life (Grossmann et al. 2008; Lloyd-Fox et al. 2009). However, NIRS has very low spatial resolution and is limited to superficial regions, making it difficult to compare results from this method with the extensive neuroimaging literature in adults. The last part of this thesis therefore develops methods for fMRI in awake infants to study development, using the broader phenomenon of category-sensitive regions of visual cortex (including the fSTS but also regions outside of the STS) as a case study.

2.2 Organization and Overview of Thesis Work

As a first step toward understanding the role of the STS in social perception, Chapter 2 assesses the large-scale functional organization of STS responses to a range of social stimuli, including moving faces, moving bodies, vocal sounds, abstract descriptions of mental states, and linguistic inputs. I focus on whether STS subregions selective for specific types of social information can be identified. This analysis reveals regions selective for many key domains of social input, including body movements, vocal sounds, and mental states. These regions were spatially organized on a posterior to anterior axis along the length of the STS. Intriguingly, the STS subregion that responded most strongly to dynamic faces (fSTS) also responded strongly to vocal sounds, indicating that this region is not face-selective per se, but fundamentally multimodal. These results suggest that the STS is comprised of a number of functionally specific subunits that largely process distinct types of social information, and also point to a novel and intriguing multimodal response profile in the fSTS.

Chapters 3 and 4 further probe the functional properties of the fSTS. In Chapter 3 I investigate the nature of face motion representations in the fSTS. This study uses multivoxel pattern analysis (MVPA), which asks which stimulus distinctions evoke distinct spatial patterns of response in a region, as a way of probing which distinctions are represented in that region. Specifically, I ask whether the fSTS contains representations of face movement type that are sufficiently abstract to generalize across differences in low-level visual properties (visual position and actor), and whether these correspond to representations of movements of individual face parts (eyes, mouth) or holistic representations of full-face motion patterns. I find that this region contains

representations of face movement type, including subtle distinctions such as the difference between mouth opening, smiling, and frowning, and that these representations generalize across low-level visual properties. Furthermore, I find evidence that these representations are parts-based, rather than holistic: patterns of response evoked by specific combinations of eye and mouth movements can be accurately modeled as a linear combination of patterns of response to each individual part. This indicates that representations are sufficiently abstract to generalize across visual details, but still more tied to the kinematics of individual part movements, rather than full-face patterns or their associated social interpretation.

Chapter 4 assesses the fSTS response to a set range of more naturalistic stimuli, including face and hand movements and vocal sounds. To assess the functional significance of face/voice responses in this region, we asked whether responses would differ for different types of facial/vocal input, including speech and nonspeech inputs, and among nonspeech inputs, richly communicative, socially relevant actions, compared to noncommunicative, less socially relevant actions. We find that the fSTS responds similarly strongly to all face movements and vocal sounds presented, but doesn't respond strongly to hand movements or nonsocial inputs. This suggests that the fSTS selectively processes audiovisual inputs from the face region, but is broadly involved in processing any such input, whether speech or nonspeech, communicative or noncommunicative. However, using MVPA, we find that spatial patterns of response differ reliably across communicative and noncommunicative stimuli, in a manner that generalizes across facial and vocal stimuli, but does not generalize to hand movements. Thus, this region is sensitive to a high-level, socially meaningful distinction.

Lastly, in Chapter 5 I shift to a related but more general question: how do functionally specialized regions such as the fSTS develop? Are they largely innately specified and present at very early ages, consistent with the highly reliable properties of these regions in adults? Or do they develop gradually during childhood, perhaps as a result of extensive experience with particular types of stimuli? Here I use category-preferring visual regions (including the fSTS, as well as other regions outside of the STS) as a case study of this general question. I develop methods for awake infant fMRI and collect data from 2-8-month-old infants viewing videos of faces, objects, bodies and scenes. I find that category-sensitivity exists in these infants, observing face- and scene-preferring regions in similar anatomical locations as in adults. However, I do not find evidence for category-selective regions, which strongly prefer a single category over others, in infants. This result suggests that the large-scale spatial organization of category preferences exists early in development and may be in part innately specified, while development subsequently tunes regions within category-sensitive zones of cortex, ultimately leading to the strongly specialized regions observed in adults. Correspondingly, the computations underlying high-level vision may become increasingly specialized for specific categories of input.

1.3 References

Abell F, Happé F, Frith U. 2000. Do triangles play tricks? Attribution of mental states to animated shapes in normal and abnormal development. Cognitive Development 15:1-16.

Allison T, Puce A, McCarthy G. 2000. Social perception from visual cues: role of the STS region. Trends in cognitive sciences 4:267-278.

Ambady N, Hallahan M, Conner B. 1999. Accuracy of judgments of sexual orientation from thin slices of behavior. J Pers Soc Psychol 77:538.

Ambady N, Rosenthal R. 1993. Half a minute: Predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness. J Pers Soc Psychol 64:431.

Andrews TJ, Ewbank MP. 2004. Distinct representations for facial identity and changeable aspects of faces in the human temporal lobe. Neuroimage 23:905-913.

Atkinson AP, Dittrich WH, Gemmell AJ, Young AW. 2004. Emotion perception from dynamic and static body expressions in point-light and full-light displays. Perception-London 33:717-746.

Belin P, Zatorre RJ, Lafaille P, Ahad P, Pike B. 2000. Voice-selective areas in human auditory cortex. Nature 403:309-312.

Bernstein M, Yovel G. 2015. Two neural pathways of face processing: A critical evaluation of current models. Neurosci Biobehav Rev 55:536-546.

Binder JR, Frost JA, Hammeke TA, Bellgowan PS, Springer JA, Kaufman JN, Possing ET. 2000. Human temporal lobe activation by speech and nonspeech sounds. Cereb Cortex 10:512-528.

Bonda E, Petrides M, Ostry D, Evans A. 1996. Specific involvement of human parietal systems and the amygdala in the perception of biological motion. J Neurosci 16:3737-3744.

Brass M, Schmitt RM, Spengler S, Gergely G. 2007. Investigating action understanding: inferential processes versus action simulation. Curr Biol 17:2117-2121.

Ciaramidaro A, Adenzato M, Enrici I, Erk S, Pia L, Bara BG, Walter H. 2007. The intentional network: How the brain reads varieties of intentions. Neuropsychologia 45:3105-3113.

Clarke TJ, Bradshaw MF, Field DT, Hampson SE, Rose D. 2005. The perception of emotion from body movement in point-light displays of interpersonal dialogue. Perception-London 34:1171-1180.

Dittrich WH, Troscianko T, Lea SE, Morgan D. 1996. Perception of emotion from dynamic point-light displays represented in dance. Perception 25:727-738.

Fecteau S, Armony JL, Joanette Y, Belin P. 2004. Is voice processing species-specific in human auditory cortex? An fMRI study. Neuroimage 23:840-848.

Fletcher PC, Happe F, Frith U, Baker SC, Dolan RJ, Frackowiak RS, Frith CD. 1995. Other minds in the brain: a functional imaging study of "theory of mind" in story comprehension. Cognition 57:109-128.

Gallagher H, Happe F, Brunswick N, Fletcher P, Frith U, Frith C. 2000. Reading the mind in cartoons and stories: an fMRI study of "theory of mind" in verbal and nonverbal tasks. Neuropsychologia 38:11-21.

Gangestad SW, Simpson JA, DiGeronimo K, Biek M. 1992. Differential accuracy in person perception across traits: examination of a functional hypothesis. J Pers Soc Psychol 62:688.

Grossman E, Donnelly M, Price R, Pickens D, Morgan V, Neighbor G, Blake R. 2000. Brain areas involved in perception of biological motion. J Cogn Neurosci 12:711-720.

Grossman ED, Blake R. 2002. Brain areas active during visual perception of biological motion. Neuron 35:1167-1175.

Grossmann T, Johnson MH, Lloyd-Fox S, Blasi A, Deligianni F, Elwell C, Csibra G. 2008. Early cortical specialization for face-to-face communication in human infants. Proceedings of the Royal Society B: Biological Sciences 275:2803-2811.

Harrigan JA, Harrigan KM, Sale BA, Rosenthal R. 1996. Detecting anxiety and defensiveness from visual and auditory cues. J Pers 64:675-709.

Harris RJ, Young AW, Andrews TJ. 2012. Morphing between expressions dissociates continuous from categorical representations of facial expression in the human brain. Proceedings of the National Academy of Sciences 109:21164-21169.

Haxby JV, Hoffman EA, Gobbini MI. 2000. The distributed human neural system for face perception. Trends in cognitive sciences 4:223-233.

Heider F, Simmel M. 1944. An experimental study of apparent behavior. The American Journal of Psychology:243-259.

Klin A. 2000. Attributing social meaning to ambiguous visual stimuli in higher-functioning autism and Asperger syndrome: The social attribution task. Journal of Child Psychology and Psychiatry 41:831-846.

Kreifelts B, Ethofer T, Shiozawa T, Grodd W, Wildgruber D. 2009. Cerebral representation of non-verbal emotional perception: fMRI reveals audiovisual integration area between voice-and face-sensitive regions in the superior temporal sulcus. Neuropsychologia 47:3059-3066.

Lloyd-Fox S, Blasi A, Volein A, Everdell N, Elwell CE, Johnson MH. 2009. Social perception in infancy -- a near infrared spectroscopy study. Child Dev 80:986-999.

Murphy NA, Hall JA, Colvin CR. 2003. Accurate intelligence assessments in social interactions: Mediators and gender effects. J Pers 71:465-493.

Overath T, McDermott JH, Zarate JM, Poeppel D. 2015. The cortical analysis of speech-specific temporal structure revealed by responses to sound quilts. Nat Neurosci 18:903-911.

Pelphrey KA, Mitchell TV, McKeown MJ, Goldstein J, Allison T, McCarthy G. 2003. Brain activity evoked by the perception of human walking: Controlling for meaningful coherent motion. J Neurosci 23:6819-6825.

Pelphrey KA, Morris JP, Mccarthy G. 2004. Grasping the intentions of others: the perceived intentionality of an action influences activity in the superior temporal sulcus during social perception. J Cogn Neurosci 16:1706-1716.

Pelphrey KA, Morris JP, Michelich CR, Allison T, McCarthy G. 2005. Functional anatomy of biological motion perception in posterior temporal cortex: an fMRI study of eye, mouth and hand movements. Cereb Cortex 15:1866-1876.

Pelphrey KA, Singerman JD, Allison T, McCarthy G. 2003. Brain activation evoked by perception of gaze shifts: the influence of context. Neuropsychologia 41:156-170.

Pitcher D, Dilks DD, Saxe RR, Triantafyllou C, Kanwisher N. 2011. Differential selectivity for dynamic versus static information in face-selective cortical regions. Neuroimage 56:2356-2363.

Pollick FE, Paterson HM, Bruderlin A, Sanford AJ. 2001. Perceiving affect from arm movement. Cognition 82:B51-B61.

Puce A, Allison T, Asgari M, Gore JC, McCarthy G. 1996. Differential sensitivity of human visual cortex to faces, letterstrings, and textures: a functional magnetic resonance imaging study. J Neurosci 16:5205-5215.

Saxe R, Kanwisher N. 2003. People thinking about thinking people: The role of the temporo-parietal junction in "theory of mind". Neuroimage 19:1835-1842.

Saxe R, Powell LJ. 2006. It's the Thought That Counts: Specific Brain Regions for One Component of Theory of Mind. Psychol Sci 17:692-699.

Schultz J, Brockhaus M, Bülthoff HH, Pilz KS. 2013. What the human brain likes about facial motion. Cereb Cortex 23:1167-1178.

Schultz J, Pilz KS. 2009. Natural facial motion enhances cortical responses to faces. Exp Brain Res 194:465-475.

Vander Wyk BC, Hudac CM, Carter EJ, Sobel DM, Pelphrey KA. 2009. Action understanding in the superior temporal sulcus region. Psychol Sci 20:771-777.

Winston JS, Henson R, Fine-Goulden MR, Dolan RJ. 2004. fMRI-adaptation reveals dissociable neural representations of identity and expression in face perception. J Neurophysiol 92:1830-1839.

Wright TM, Pelphrey KA, Allison T, McKeown MJ, McCarthy G. 2003. Polysensory interactions along lateral temporal regions evoked by audiovisual speech. Cereb Cortex 13:1034-1043.

Chapter 2: Large-scale functional organization of the human superior temporal sulcus[1]

The superior temporal sulcus (STS) is considered a hub for social perception and cognition, including the perception of faces and human motion, as well as understanding others' actions, mental states, and language. However, the functional organization of the STS remains debated: is this broad region composed of multiple functionally distinct modules, each specialized for a different process, or are STS subregions multifunctional, contributing to multiple processes? Is the STS spatially organized, and if so, what are the dominant features of this organization? We address these questions by measuring STS responses to a range of social and linguistic stimuli in the same set of human participants, using fMRI. We find a number of STS subregions that respond selectively to certain types of social input, organized along a posterior-to-anterior axis. We also identify regions of overlapping response to multiple contrasts, including regions responsive to both language and theory of mind, faces and voices, and faces and biological motion. Thus, the human STS contains both relatively domain-specific areas, and regions that respond to multiple types of social information.

[1] The content of this chapter has been published as Deen B, Koldewyn K, Kanwisher N, Saxe R. 2015. Functional organization of social perception and cognition in the superior temporal sulcus. Cereb Cortex 25(11): 4596-4609.

2.1 Introduction

Humans are profoundly social beings, and accordingly devote considerable cortical territory to social cognition, including lower-tier regions specialized for perceiving the shapes of faces and bodies (Kanwisher 2010), and high-level regions specialized for understanding the meaning of sentences (Binder et al. 1997; Fedorenko et al. 2011) and the contents of other people's thoughts (Saxe and Kanwisher 2003; Saxe and Powell 2006). Yet between these two extremes lies a rich space of intermediate social processes, including the ability to discern the goal of an action, the significance of a fleeting facial expression, the meaning of a tone of voice, and the nature of the relationships and interactions in a social group. How do we so quickly and effortlessly extract this multifaceted social information from "thin slices" of social stimuli (Ambady and Rosenthal 1992)? Here we investigate the functional organization of our computational machinery for social cognition by using fMRI to target a brain region that has long been implicated as a nexus of these processes: the superior temporal sulcus (STS).

The STS is one of the longest sulci in the brain, extending from the inferior parietal lobe anteriorly along the full length of the temporal lobe. Subregions of the STS have been implicated in diverse aspects of social perception and cognition, including the perception of faces (Puce et al. 1996; Haxby et al. 2000; Pitcher et al. 2011), voices (Belin et al. 2000), and biological motion (Bonda et al. 1996; Allison et al. 2000; Grossman et al. 2000; Grossman and Blake 2002; Pelphrey, Mitchell, et al. 2003; Pelphrey et al. 2005), and understanding of others' actions (Pelphrey, Singerman, et al. 2003; Pelphrey, Morris, et al. 2004; Brass et al. 2007; Vander Wyk et al. 2009), and

mental states (Fletcher et al. 1995; Gallagher et al. 2000; Saxe and Kanwisher 2003; Saxe and Powell 2006; Ciaramidaro et al. 2007). Regions of the STS have also been implicated in linguistic processing (Binder *et al.* 1997; Vigneau et al. 2006; Fedorenko et al. 2012), as well as basic perceptual and attentional functions, such as audiovisual integration (Calvert et al. 2001; Beauchamp et al. 2004; Taylor et al. 2006), and the control of visual attention (Corbetta and Shulman 2002). But because most prior studies have investigated only a small subset of these mental processes, the relationship between the regions involved in each process remains unknown.

One possibility is that the STS is composed of a number of distinct, functionally specialized subregions, each playing a role in one of these domains of processing and not others. This would point to a modular organization, and would further point to separate streams of processing for the domains listed above – processing faces, voices, mental states, etc. Another possibility is that responses to these broad contrasts overlap. Overlap could either point to 1) STS subregions involved in multiple processes, indicating a nonmodular organization; or 2) a response driven by an underlying process shared across multiple tasks, such as integration of information from multiple modalities or domains.

Hein and Knight (Hein and Knight 2008) performed a meta-analysis assessing locations of peak coordinates of STS responses from biological motion perception, face perception, voice perception, theory of mind, and audiovisual integration tasks. They found that peak coordinates from different tasks didn't fall into discrete spatial clusters, and thus argued 1) that the STS consists of multifunctional cortex, whose functional role

at a given moment depends on coactivation patterns with regions outside of the STS; and 2) that there is little spatial organization to the STS response to different tasks.

However, meta-analyses cannot provide strong evidence for overlap between functional regions. Because the anatomical location of each functional region varies across individual subjects, combining data across subjects in a standard stereotactic space can lead to findings of spurious overlap (Brett et al. 2002; Saxe et al. 2006; Fedorenko and Kanwisher 2009). These issues are compounded in meta-analyses, which combine data across studies using different normalization algorithms and stereotactic coordinate systems. Furthermore, to investigate overlap between regions responding to distinct tasks, we would ideally want to study the full spatial extent of these regions, rather than simply to peak coordinates.

The present study addresses these limitations by scanning the same set of subjects while they engage in face perception, biological motion perception, mental state understanding (termed theory of mind, ToM), linguistic processing, and voice perception. Within individual subjects, we compare STS responses across different tasks. We show that distinct input domains evoke distinct patterns of activation along the STS, pointing to different processes engaged by each type of input. In particular, we find that a dominant feature of this spatial organization consists of differences in response profile along the anterior-posterior axis of the STS. Investigating focal regions that respond maximally to each contrast within individual subjects, we are able to find strongly selective regions for most processes assessed, including biological motion perception, voice perception, theory of mind, and language. These selective regions are also characterized by distinct patterns of whole-brain functional connectivity, and similarity in functional connectivity profiles

across regions is predictive of similarity in task responses. In addition to these selective regions, we identify a number of regions that respond reliably to multiple contrasts, including language and theory of mind, faces and voices, and faces and biological motion. Thus, the STS appears to contain both subregions specialized for particular domains of social processing, as well as areas responsive to information from multiple domains, potentially playing integrative roles.

2.2 Methods

Participants

Twenty adult subjects (age 19-31, 11 female, all right-handed) participated in the study. Participants had no history of neurological or psychiatric impairment, had normal or corrected vision, and were native English speakers. All participants provided written, informed consent.

Paradigm

Each participant performed five tasks over the course of 1-3 scan sessions. These included a theory of mind (ToM) task, biological motion perception task, face perception task, voice perception task, and an auditory story task that yielded multiple contrasts of interest. The paradigms were designed such that roughly five minutes of data was collected for each condition within each experiment.

In the ToM task, participants read brief stories describing either false beliefs (ToM condition) or false physical representations (control condition), and then answered true/false questions about these stories. Stories were chosen based on a prior study that identified false belief stories that elicited the largest response in the right temporo-parietal

junction (Dodell-Feder et al. 2011). Stories were presented for 10s, followed by a 4s question phase, and 12s fixation period, with an additional 12s fixation at the start of the run. 20 stories (10 per condition) were presented over two runs, each lasting 4:32 minutes. Conditions were presented in a palindromic order (e.g. 1, 2, 2, 1), counterbalanced across runs and subjects. The stimuli and experimental scripts are available on our lab's website (http://saxelab.mit.edu/superloc.php).

In the biological motion task, participants watched brief point-light-display (PLD) animations that either depicted various human movements (walking, jumping, waving, etc) or rotating rigid three-dimensional objects with point-lights at vertices (Vanrie and Verfaillie 2004). Animations consisted of white dots moving on a black background. Individual animations lasted 2s, and were presented in blocks of nine, with a .25s gap between animations. Participants performed a one-back task on individual animations, pressing a button for repeated stimuli, which occurred once per block. Additionally, four other conditions were included which are not reported here; these included spatially scrambled versions of human and object PLDs, linearly moving dots, and static dots. In each of six runs, two blocks per condition were presented in palindromic order, with condition order counterbalanced across runs and subjects. Runs consisted of twelve 20.25s blocks as well as 18s fixation blocks at the start, middle, and end, for a total run time of 4:57 minutes. Due to timing constraints, one subject did not complete the biological motion task.

In the face perception task, participants passively viewed 3s movie clips of faces or of moving objects, using stimuli that have been previously described (Pitcher *et al.* 2011). We chose to use dynamic as opposed to static face stimuli as dynamic stimuli

have been shown to yield a substantially stronger response in the face area of posterior STS (Pitcher *et al.* 2011), facilitating the ability to identify face-responsive regions of the STS in individual subjects. Stimuli were presented in blocks of 6 clips with no interval between clips. Additionally, a third condition presenting movies of bodies was included, but not assessed in this report. In each of three runs, four blocks per condition were presented in palindromic order, with condition order counterbalanced across runs and subjects. Runs consisted of twelve 18s stimulus blocks as well as 18s rest blocks at the start, middle, and end, for a total run time of 4:30 minutes.

In the voice perception task, participants passively listened to audio clips consisting either of human vocal sounds (e.g. coughing, laughing, humming, sighing, speech sounds), or nonvocal environmental sounds (e.g. sirens, doorbells, ocean sounds, instrumental music). Stimuli were taken from a previous experiment that identified a voice-responsive region of STS (Belin et al. 2000) (http://vnl.psy.gla.ac.uk/resources.php). Clips were presented in 16s blocks that alternated between the two conditions, with a 12s fixation period between blocks and 8s of fixation at the start of the experiment. Condition order was counterbalanced across subjects. A single run was given, with 10 blocks per condition, lasting 9:28. Due to timing constraints, four subjects did not receive the voice perception task.

During the auditory story task, participants listened to either stories or music. Four conditions were included: ToM stories, physical stories, jabberwocky, and music. This task provides a language contrast (physical stories versus jabberwocky), a second ToM contrast (ToM versus physical stories), and a second voice contrast (jabberwocky versus music). Two additional conditions, stories depicting physical and biological

movements, were included for separate purposes, and are not analyzed in the present report. ToM stories consisted of stories describing the false belief of a human character, with no explicit descriptions of human motion. Physical stories described physical events involving no object motion (e.g., streetlights turning on at night), and no human characters. All stories consisted of three sentences, and stories were matched across conditions on number of words, mean syllables per word, Flesch reading ease, number of noun phrases, number of modifiers, number of higher level constituents, number of words before the first verb, number of negations, and mean semantic frequency (log Celex frequency). Jabberwocky stimuli consisted of English sentences with content words replaced by pronounceable nonsense words, and with words temporally reordered. This condition has minimal semantic and syntactic content, but preserves prosody, phonology, and vocal content. Music stimuli consisted of clips from instrumental classical and jazz pieces, with no linguistic content. Each auditory stimulus lasted 9s. After a 1s delay, participants performed a delayed-match-to-sample task, judging whether a word (or music clip) came from the prior stimulus. Each run consisted of four trials per condition, for a total of 24 trials. Four runs of 8:08 minutes were given. Stimuli were presented in a jittered, slow event-related design, with stimulus timing determined using Freesurfer's optseq2 to optimize power in comparing conditions.

Additionally, resting-state data were acquired to investigate functional connectivity of STS subregions. For these scans, participants were asked to keep their eyes open, avoid falling asleep, and stay as still as possible. These scans lasted 10 minutes.

Data acquisition

Data were acquired using a Siemens 3T MAGNETOM Tim Trio scanner (Siemens AG, Healthcare, Erlangen, Germany). High-resolution T1-weighted anatomical images were collected using a multi-echo MPRAGE pulse sequence (repetition time [TR] = 2.53s; echo time [TE] = 1.64ms, 3.5ms, 5.36ms, 7.22ms, flip angle $\alpha = 7°$, field of view [FOV] = 256mm, matrix = 256x256, slice thickness = 1mm, 176 near-axial slices, acceleration factor = 3, 32 reference lines). Task-based functional data were collected using a T2*-weighted echo planar imaging (EPI) pulse sequence sensitive to blood-oxygen-level-dependent (BOLD) contrast (TR = 2s, TE=30ms, $\alpha = 90°$, FOV = 192mm, matrix = 64x64, slice thickness = 3mm, slice gap = .6mm, 32 near-axial slices). Resting-state functional data were also collected using a T2*-weighted EPI sequence (TR = 6s, TE = 30ms, $\alpha = 90°$, FOV = 256mm, matrix = 128x128, slice thickness = 2mm, 67 near-axial slices). Resting data were acquired at higher resolution (2mm isotropic) to reduce the relative influence of physiological noise (Triantafyllou et al. 2005; Triantafyllou et al. 2006).

Data preprocessing and modeling

Data were processed using the FMRIB Software Library (FSL), version 4.1.8, supplemented by custom MATLAB scripts. Anatomical and functional images were skull-stripped using FSL's brain extraction tool. Functional data were motion corrected using rigid-body transformations to the middle image of each run, corrected for interleaved slice acquisition using sinc interpolation, spatially smoothed using an isotropic Gaussian kernel (5mm FWHM unless otherwise specified), and high-pass

filtered (Gaussian-weighted least squares fit straight line subtraction, with σ = 50s (Marchini and Ripley 2000)). Functional images were registered to anatomical images using a rigid-body transformation determined by Freesurfer's bbregister (Greve and Fischl 2009). Anatomical images were in turn normalized to the Montreal Neurological Instititute-152 template brain (MNI space), using FMRIB's nonlinear registration tool (FNIRT). Further details on the preprocessing and modeling of resting-state data are provided below (see Resting state functional connectivity analysis).

Whole-brain general linear model (GLM)-based analyses were performed for each subject, run, and task. Regressors were defined as boxcar functions convolved with a canonical double-gamma hemodynamic response function. For the theory of mind task, the story and response periods for each trial were modeled as a single event, lasting 14s. For the auditory story task, 9s-long stories were modeled as single events; these did not include the response period, as the response was unrelated to the processes of interest for this task. For the face, biological motion, and voice perception tasks, the regressor for a given condition included each block from that condition. Temporal derivatives of each regressor were included in all models, and all regressors were temporally high-pass filtered. FMRIB's improved linear model (FILM) was used to correct for residual autocorrelation, to provide valid statistics at the individual subject level (Woolrich et al. 2001).

Subsequently, data were combined across runs for each subject using 2^{nd}-level fixed effects analyses, after transforming beta maps to MNI space. For split-half analyses (further described below), data were combined across even and odd runs separately. For the voice localizer, which only had a single run, the data were temporally split into first

and second halves, each with five blocks per condition, and these were analyzed as if they were separate runs.

The contrasts analyzed were as follows: from the theory of mind task, false belief versus false physical representation stories (termed ToM 1); from the face perception task, faces versus objects (Faces); from the biological motion task, biological motion versus rigid object motion (Biological Motion); from the voice perception task, vocal versus nonvocal sounds (Voice 1); and from the auditory story task, false belief versus physical stories (ToM 2), physical stories versus jabberwocky (Language), and jabberwocky versus music (Voice 2). Note that the ToM 2 and Language contrasts are nonorthogonal and thus statistically dependent, as are the Language and Voice 2 contrasts. As a result, these pairs of contrasts are biased toward finding non-overlapping sets of voxels: for instance, voxels with high responses in the physical condition are less likely have significant effects of ToM2, and more likely to have significant effects of Language. However, in both of these cases we have contrasts from separate datasets (ToM 1 and Voice 1 contrasts) to validate the results.

Because we were specifically interested in responses within the STS, 2^{nd}-level analyses were restricted to voxels within a bilateral STS mask, defined by drawing STS gray matter on the MNI template brain. Posteriorly, the STS splits into two sulci surrounding the angular gyrus. Our mask included both of these sulci as well as gray matter in the angular gyrus, because responses to ToM contrasts have previously been observed on the angular gyrus. Statistical maps were thresholded using a false discovery rate (FDR) of $q < .01$, which controls the proportion of positive results that are expected

to be false positives, to correct for multiple comparisons; supplementary analyses also used different thresholds to determine the effect on overlap estimates.

Anterior-posterior organization

We first investigated the large-scale spatial organization of STS responses to different contrasts, by assessing how responses to each contrast vary as a function of position along the length of the sulcus. We sought to define a series of regions-of-interest (ROIs) that carved the STS into slices along its length. Prior studies have analyzed responses in coronal slices of the STS, assessing how responses vary as a function of the y-coordinate in MNI space (Pelphrey, Mitchell, *et al.* 2003; Pelphrey, Singerman, *et al.* 2003; Pelphrey, Viola, et al. 2004; Morris et al. 2005). However, the STS has an oblique orientation in the y-z plane of MNI space, and we wished to define ROIs that extended perpendicularly to the local direction of the STS in the y-z plane.

To this end, we used our STS mask to estimate the local orientation of the sulcus at different y-coordinates. Mask coordinates were averaged across the x- and z-dimensions, to obtain a function specifying the mean z-coordinate of the STS for a given y-coordinate. Next, for each y-coordinate, the local slope of the STS was determined by fitting a linear regression to z-coordinates in a 1cm window along the y-dimension. This slope was used to define "slice" ROIs along the length of the STS, by constructing an anisotropic Gaussian ROI and intersecting this with the STS mask (sample ROIs are shown in Figure 2). Note that for the posterior segment of the STS, where it splits into two sulci, our approach does not treat these sulci separately, instead computing the local slope of a mask including both sulci and the angular gyrus. For each ROI, hemisphere,

subject, and contrast, percent signal change values were extracted, and plotted as a function of y-coordinate.

Additionally, we asked whether these patterns differed across the upper and lower banks of the STS. The upper and lower banks of the STS were drawn on individual subjects' cortical surface representations, and intersected with the slice ROIs defined above. Percent signal change values were extracted from the resulting "upper and lower slice" ROIs. For this analysis, we only considered portions of the STS that were anterior to the point at which the STS splits into two sulci posteriorly, and data were only smoothed at 3mm-FWHM to minimize bleeding across upper and lower banks.

These analyses revealed that the positions of regions with the strongest response to each contrast were ordered as follows, from posterior to anterior: ToM, biological motion, faces, voices, language. We next aimed to statistically assess these differences in spatial position. Although some contrasts elicited responses in multiple regions along the STS, we aimed to compare responses specifically within the region of maximal response to each task, and thus assessed active regions in individual subjects within spatially constrained group-level search spaces. Search spaces for each contrast were defined from group-level activation maps, computed using a mixed-effects analysis (Woolrich et al. 2004), as the set of active voxels within a 15mm-sphere around a peak coordinate (shown in Supplementary Figure 1). For each spatially adjacent pair of search spaces (e.g., ToM and biological motion, biological motion and faces, etc) we combined the two search spaces, and identified regions of activation in individual subjects to each of the two contrasts within this combined search space. We then computed the center of mass

of these regions, and compared their y-coordinates across the two contrasts using a paired, two-sample, two-tailed *t*-test.

Responses in maximally sensitive regions

We next asked whether the STS contains selective regions, responding to one contrast and not others. We focused on small regions that were maximally responsive to each contrast in a given subject, to increase the likelihood of finding selective responses, and extracted responses in these regions across all conditions. ROIs were defined using data from odd runs of each task. For the face localizer, which had three runs, runs one and three were used to define ROIs. The group-level search spaces defined above (Anterior-posterior organization section) were used to spatially constrain ROI definition. For each contrast, hemisphere, and subject, we identified the coordinate with the global maximum response across a given search space, placed a 5mm-radius sphere around this coordinate, and intersected this sphere with the individual subject's activation map for that contrast.

Responses to each condition in each task were then extracted from these ROIs. For the task used to define the ROI, data were extracted from even runs, while for other tasks the full dataset was used, such that the extracted responses were always independent form data used to define the ROI. Percent signal change was extracted by averaging beta values across each ROI and dividing by mean BOLD signal in the ROI. *T*-tests were used to test for an effect of each of the seven contrasts of interest in each ROI, with a threshold of $P < .01$ (one-tailed). Participants who lacked a certain ROI were not included in the statistical analysis for that ROI.

Resting-state functional connectivity analysis

We next probed another aspect of spatial organization in the STS: do subregions of the STS have different patterns of functional connectivity, and do these patterns relate to the task response profile of that region? Specifically, we assessed resting-state functional connectivity of functionally defined STS subregions.

For resting-state data, several additional preprocessing steps were performed to diminish the influence of physiological and motion-related noise. Time series of six motion parameter estimates, computed during motion correction, were removed from the data via linear regression. Additionally, time series from white matter and cerebrospinal fluid (CSF) were removed using the CompCorr algorithm (Behzadi et al. 2007; Chai et al. 2012). White matter and CSF ROIs were defined using FSL's Automated Segmentation Tool (FAST) and eroded by one voxel. The mean and first four principal components of time series from these masks were computed and removed from the data via linear regression.

We again focused on regions with maximum responses to a given contrast, to isolate spatially distinct subregions of the STS. ROIs were defined in the same way as above (see Responses in maximally sensitive regions): as the set of active voxels within a 5mm-radius sphere around the peak coordinate of response from a given task and participant. Although these ROIs were defined using the same procedure as described above, they were defined using the full dataset from each task, rather than half of the data, and thus differed slightly from the ROIs used above. Time series were extracted from each ROI, and correlations were computed with time series from every voxel in a

Freesurfer-derived gray matter mask (excluding within-hemisphere STS voxels), to derive a whole-brain functional connectivity map for each region. We then computed correlations between whole-brain functional connectivity maps from different regions, to determine the degree of similarity of functional connectivity maps from different STS subregions (functional connectivity similarity). We also computed task responses of these ROIs across the 12 conditions assessed in this study, and computed correlations between these task response vectors across ROIs, to assess similarity of response profiles (response similarity).

Lastly, we assessed the relationship between functional connectivity similarity and response similarity, after accounting for effects of spatial proximity of ROIs. Within each hemisphere, a linear mixed model was performed with functional connectivity similarity values (Fisher-transformed) across ROI pairs and subjects as the dependent variable. To avoid pairs of ROIs with trivially similar response profiles due to similarity in physical location, we excluded ROIs from the ToM 2 and Voice 2 contrasts, as well as any pair of ROIs whose centers were closer than 1cm, which prevents any overlap between pairs of ROIs. Response similarity (Fisher-transformed) was used as the explanatory variable of interest. Additionally, physical distance between ROIs, as well as the square and cube roots thereof, were included as nuisance regressors to account for effects of spatial proximity on similarity of functional connectivity maps. These specific nonlinear functions of physical distance were found to accurately model the relation between physical distance and functional connectivity similarity values. A mixed model with random effect terms for the intercept and the effect of response similarity was used. Parameters were estimated using an approximate maximum likelihood method

(Lindstrom and Bates 1990), implemented using MATLAB's nlmefit function. A Wald test was used to assess the relationship between response similarity and functional connectivity similarity in each hemisphere; the use of a normal approximation is justified by the large number of data points in each model (N=291, right hemisphere; N=233, left hemisphere).

Overlap analysis

Having probed the response profile of focal, maximally responsive ROIs for each contrast, we next investigated the full spatial extent of responses to each contrast. Specifically, we assessed the amount of overlap between significantly active STS voxels in each hemisphere across contrasts. To illustrate our method for quantifying overlap, suppose we have two regions, called A and B, defined by two different contrasts, and let AB denote the region of overlap between A and B. We compute two quantities to assess the overlap between A and B: the size of AB divided by the size of A, and the size of AB divided by the size of B. In addition to the amount of overlap, these measures provide some insight into the type of overlap occurring. For instance, if region A encompasses and extends beyond region B, then size(AB)/size(B) = 1, while size(AB)/size(A) < 1. In contrast, if the regions are of equal size, then size(AB)/size(B) = size(AB)/size(A). Furthermore, these quantities have an intuitive and straightforward interpretation, as the proportion of one region (A or B) that overlaps with the other. Overlap values were averaged across subjects. For each pair of contrasts and each hemisphere, subjects who lacked any STS response to one or both contrast were excluded from this average.

fMRI overlap values depend on both the extent of spatial smoothing applied, as well as the statistical threshold used to define the extent of active regions. For this reason, we additionally computed overlap values at a range of different thresholds ($q < .05$, $q < .01$, and $q < .005$) as well as smoothing kernels (5mm, 3mm, and 0mm FWHM) to ask whether overlap could be consistently observed across these different parameters.

Lastly, we investigated the response profiles of overlapping regions, focusing on pairs of contrasts for which substantial overlap was observed. Specifically, we focused on regions responsive to language and theory of mind, faces and voices, and faces and biological motion. We used a split-half analysis approach. Regions were defined in the first half of the dataset as the set of all voxels that responded to two given contrasts. Responses were then extracted in the second half of the data for the two tasks used to define the region, or the full dataset for other tasks, such that responses were always extracted from data that was independent of those used to define the ROI. Unsmoothed data were used to extract responses, such that overlapping responses could not be introduced by spatial smoothing. *T*-tests were used to test for an effect of each of the seven contrasts of interest in each ROI, with a threshold of $P < .01$ (one-tailed).

2.3 Results

Individual subject activations

Individual subject activations are shown in Figure 1, and mean peak coordinates of response (within search spaces for each contrast) are given in Table 1. For the theory of mind contrasts, the most commonly observed response bilaterally was in the angular gyrus or one of the two branching sulci of the STS, a region previously termed the temporo-parietal junction (TPJ; (Saxe and Kanwisher 2003)). Additionally, responses in

middle and anterior STS region were often observed. In some subjects, these responses were relatively focal (e.g. Subject 2 in Figure 1), while in others this response encompassed a large portion of middle to anterior STS (e.g. Subject 1 in Figure 1).

Responses to the language contrast were generally stronger in the left than right hemisphere. In the left hemisphere, most subjects had several distinct regions of activation along the STS, with variable positions across subjects, ranging from angular gyrus to middle and anterior STS. In the right hemisphere, the most commonly observed response was a single region of far-anterior STS, as seen in both subjects in Figure 1.

For the voice contrast, activations were generally centered in the middle STS, and were typically stronger in the upper than lower bank of the STS. These responses varied substantially in extent across subjects, with some being relatively focal (e.g. Subject 2 in Figure 1), and some extending along nearly the full length of the STS anterior to the angular gyrus (e.g. Subject 1 in Figure 1).

Face responses were most commonly observed in a region at and/or just anterior to the point at which the STS breaks into two sulci, previously termed the posterior STS (pSTS; (Pelphrey, Mitchell, *et al.* 2003; Shultz et al. 2011)). Additionally, many subjects had several other discrete face-sensitive regions both anterior and posterior to this pSTS region, with locations varying across individuals.

Biological motion responses were typically observed in a similar region of pSTS. This region was typically overlapping with the face-responsive pSTS region, but was centered slightly posteriorly in most subjects (e.g. Subject 2 in Figure 1; also see Anterior-posterior organization section).

Anterior-posterior organization

To summarize the large-scale organization of responses to each contrast, we next analyzed the strength of BOLD responses to each contrast as a function of position along the length of the STS.

Results from this analysis are shown in Figure 2 (ToM 2 and Voice 2 contrasts are omitted for visualization purposes); results separated across the upper and lower banks are shown in Supplementary Figure 2. These results are consistent with the qualitative descriptions of individual-subject activation patterns described above, and provide a visualization of these effects at the group level. Theory of mind responses are strongest in the posterior-most part of the STS (in the angular gyrus and surrounding sulci), and were also observed in middle-to-anterior STS. Biological motion responses peaked in a pSTS region just anterior to the angular gyrus ToM response. Face responses peaked in a further-anterior pSTS region, with weaker responses also observed in middle-to-anterior STS. The voice response was very broad, encompassing much of the STS, and centered on middle STS. Lastly, language responses in the left hemisphere were observed along the extent of the STS, with peaks in posterior and anterior STS regions. In the right STS, only an anterior region of language activation was observed. Voice responses were substantially stronger in the upper bank than the lower bank of the STS, while responses to ToM, biological motion, faces, and language were largely symmetric across the two banks (with a slightly stronger right anterior ToM response in the lower bank).

To determine the reliability of these anterior-posterior spatial relations across individual subjects, we statistically assessed differences in center of mass along the y-axis of spatially adjacent regions. This difference was significant for all pairs of regions,

including ToM 1 and biological motion regions (LH: $t = 2.34$, $P < .05$, RH: $t = 5.65$, $P < 10^{-4}$), biological motion and face regions (LH: $t = 7.30$, $P < 10^{-5}$, RH: $t = 3.17$, $P < .01$), face and voice 1 regions (LH: $t = 2.54$, $P < .05$, RH: $t = 6.10$, $P < 10^{-4}$), and voice 1 and language regions (LH: $t = 7.46$, $P < 10^{-5}$, RH: $t = 6.91$, $P < 10^{-4}$). This result demonstrates a reliable, bilateral anterior-to-posterior ordering of responses: the TPJ response to ToM, pSTS response to biological motion, pSTS response to faces, middle STS response to voices, and anterior STS response to language.

Responses in maximally sensitive regions

Do STS subregions of maximal sensitivity to a given contrast also exhibit selectivity – a response to one contrast but not others? We tested this by extracting responses across all conditions from small ROIs surrounding peak coordinates for the response to each contrast (Figure 3), using data independent of those used to define the ROI. Responses in regions defined by ToM 2 and Voice 2 contrasts are omitted for brevity, as the locations of these regions and their response profiles were highly similar to the ToM 1 and Voice 1 ROIs.

Among regions defined by the theory of mind (ToM 1) contrast, the left-hemisphere region responded strongly to the ToM 1 ($t(18) = 7.13$, $P < 10^{-6}$) and ToM 2 ($t(18) = 7.45$, $P < 10^{-6}$) contrasts, but also responded significantly to the language contrast ($t(18) = 4.88$, $P < 10^{-4}$). The right-hemisphere region, by contrast, exhibited a selective response, with significant effects only for the ToM 1 ($t(18) = 6.84$, $P < 10^{-6}$) and ToM 2 ($t(18) = 4.00$, $P < 10^{-3}$) contrasts.

Language regions were defined in a relatively small number of subjects (4 for the right hemisphere and 8 for the left, in split-half data). This likely reflects the fact that this contrast was generally somewhat weaker than the others, as well as more spatially variable, leading to a substantial portion of subjects with no significant response in the anterior STS language search space when only odd runs were analyzed. Nevertheless, significant effects of the language contrast were observed in both the left ($t(7) = 6.03$, $P < .001$) and right ($t(3) = 5.74$, $P < .01$) hemispheres. An effect of the voice 2 contrast was observed in the left hemisphere ($t(7) = 3.65$, $P < .01$). Note that the jabberwocky condition used to define this contrast involves phonemic and prosodic information; this difference could thus still reflect linguistic processing. However, an effect of the ToM 2 contrast was also observed in the left hemisphere ($t(7) = 3.73$, $P < .01$), in addition to a marginal effect of ToM 2 in the right hemisphere ($t(3) = 2.16$, $P < .05$) and marginal effects of ToM 1 in the left ($t(7) = 2.63$, $P < .05$) and right ($t(3) = 3.73$, $P < .05$) hemispheres. These effects reflected moderately sized differences in percent signal change (.2-.4%), but were nevertheless statistically marginal due to the small number of subjects with a defined region. Thus, although this region appears largely language selective—with a stronger response to language conditions relative to a range of nonlinguistic visual and auditory conditions—it appears to also be modulated by mental state content.

Voice-sensitive regions in middle STS showed a clearly selective profile of responses bilaterally. These regions responded to the voice 1 (left: $t(13) = 6.78$, $P < 10^{-5}$; right: $t(14) = 11.12$, $P < 10^{-8}$) and voice 2 (left: $t(13) = 7.02$, $P < 10^{-5}$; right: $t(14) = 6.95$, $P < 10^{-5}$) contrasts, and did not respond significantly to any other contrast.

Strikingly, the face-sensitive region of posterior STS responded strongly to both face and voice contrasts, bilaterally. The expected response to faces was found in the left ($t(6) = 6.51$, $P < .001$) and right ($t(13) = 7.50$, $P < 10^{-5}$) hemispheres. Additionally, there was a bilateral effect of the voice 1 (left: $t(5) = 7.42$, $P < .001$; right: $t(11) = 5.13$, $P < 10^{-6}$) and voice 2 (left: $t(6) = 4.13$, $P < .01$; right: $t(13) = 3.86$, $P < .01$) contrasts. Weaker effects of the ToM 2 contrast ($t(14) = 2.80$, $P < .01$) and the biological motion contrast ($t(12) = 2.90$, $P < .01$) were also observed in the right hemisphere.

Lastly, the posterior STS regions defined by the biological motion contrast responded to biological motion bilaterally (left: $t(16) = 15.20$, $P < 10^{-10}$; right: $t(16) = 13.76$, $P < 10^{-10}$). There was also an effect of the voice 1 contrast in the right hemisphere ($t(13) = 2.78$, $P < .01$) and marginally in the left hemisphere ($t(13) = 2.43$, $P < .02$), although neither region showed an effect of the voice 2 contrast. This indicates the presence of an STS subregion that is quite selective for processing biological motion.

Could differences in effect sizes across regions in this analysis reflect differences in signal quality from different subregions of the STS? To address this, we computed temporal signal-to-noise ratios (tSNR) for the ROIs assessed here (Supplementary Figure 3). These results indicate that tSNR values are largely similar across ROIs, with a slight (~25%) decrease in values for voice ROIs. This suggests against the possibility that these functional dissociations result from signal quality differences.

Resting-state functional connectivity analysis

We next asked about another dimension of spatial organization: do functionally and anatomically distinct subregions of the STS have differing patterns of functional

connectivity with the rest of the brain? We identified regions of maximum response to each task within individual subjects, and assessed similarity between their functional connectivity maps, as well as their task response profiles.

Matrices of functional connectivity and response similarity are shown in Figure 4 (additionally, whole-brain functional connectivity maps for each seed region are shown in Supplementary Figure 4). Generally, positive correlations were observed between functional connectivity patterns. Excluding correlations between regions defined by similar contrasts (ToM1/ToM2, Voice1/Voice2), these correlations ranged from .05 to .58 (LH), and .11 to .60 (RH). The broad range in these correlation values indicates that some pairs of STS subregions share similar functional connectivity patterns, while others diverge.

Does this variability in functional connectivity similarity relate to variability in the response similarity of pairs of regions? A linear mixed model showed a significant relationship both in the left hemisphere ($z = 2.52$, $P < .05$) and the right hemisphere ($z = 3.29$, $P < .001$), after accounting for effects of spatial proximity. These findings indicate that there are multiple functional connectivity patterns and response profiles associated with STS subregions, and that pairwise similarity along these two measures is related: regions with more similar response profiles also have more similar patterns of functional connectivity.

Overlap analysis

Having investigated the response profiles of maximally responsive focal regions in individual subjects, we next asked whether the full STS response to each contrast is

spatially distinct or overlapping across contrasts. To answer this question, we computed the proportion of the STS activation to one contrast that overlapped with activations to each other contrast.

Results from the overlap analysis are shown in Figure 5. As expected, the strongest overlap values (47-75%) were found for the ToM 1 and ToM 2 contrasts, as well as the Voice 1 and Voice 2 contrasts, intended to elicit activity in similar regions. Overlap values for other pairs of contrasts ranged from 4-59%. Particularly strong overlap was observed between face and voice responses (19-59%, mean = 36%), face and biological motion responses (15-39%, mean = 30%), and ToM and language responses (11-50%, mean = 29%). Relatively high values were also observed for overlap between ToM and face responses (16-39%, mean = 24%), and ToM and voice responses (9-35%, mean = 20%). This indicates in addition to focal regions with selective response profiles, the STS contains parts of cortex that respond significantly to social information from multiple domains.

We next assessed how overlap values vary as a function of amount of smoothing and the statistical threshold used to define regions. Supplementary figures 5 and 6 show overlap matrices at smoothing kernels of 5mm, 3mm, and 0mm FWHM, and thresholds of $q < .05$, $q < .01$, and $q < .005$. As expected, using stricter thresholds and less spatial smoothing lead to numerically smaller overlap values, with spatial smoothing appearing to have a greater influence in the range of parameters tested. For example, face/voice overlap had a mean value of 42% at $q < .05$ and 5mm-FWHM, and a mean value of 28% at $q < .005$ and no smoothing. Nevertheless, a similar pattern of overlap values across pairs of contrasts was observed across thresholds and smoothing kernels. In particular,

relatively strong overlap between responses language and ToM, faces and voices, and faces and biological motion were consistently observed. In contrast, overlap between responses to ToM and faces, as well as ToM and voices, decreased substantially as less smoothing was used. This may indicate that this overlap was introduced by spatial blurring of distinct regions; alternatively, it is possible that the increased signal-to-noise ratio afforded by smoothing is necessary to detect this overlap.

Lastly, we investigated response profiles of overlapping regions, by defining overlapping ROIs in one half of the dataset, and extracting responses from left-out, unsmoothed data (Figure 6). We focused on pairs of contrasts for which substantial overlap was consistently observed.

The language and ToM region responded substantially above baseline for all language conditions, with a near-zero response to all other conditions, and was additionally modulated by the presence of mental state content. Significant effects of the language contrast were observed in both hemispheres (RH: $t(10) = 6.31$, $P < 10^{-4}$; LH: $t(11) = 4.56$, $P < 10^{-3}$). Effects of the ToM 2 contrast were also significant bilaterally (RH: $t(10) = 4.59$, $P < 10^{-3}$; LH: $t(11) = 2.74$, $P < .01$), while effects of ToM 1 were significant in the right hemisphere ($t(10) = 3.08$, $P < .01$) and marginal in the left ($t(11) = 1.86$, $P < .05$).

The face and voice region had a roughly similar response profile to the pSTS region defined by a face contrast, with a moderately sized response to both dynamic faces and vocal sounds. The effect of the voice 1 contrast was significant bilaterally (RH: $t(12) = 9.00$, $P < 10^{-6}$; LH: $t(11) = 5.38$, $P < 10^{-3}$), as was the effect of the voice 2 contrast (RH: $t(12) = 4.50$, $P < 10^{-3}$; LH: $t(11) = 2.93$, $P < .01$). There was also a significant

effect of faces over objects bilaterally (RH: $t(12) = 5.12$, $P < 10^{-3}$; LH: $t(11) = 4.68$, $P < 10^{-3}$). This region also had a weak but reliable response to the theory of mind contrasts: ToM 2 bilaterally (RH: $t(12) = 3.50$, $P < .01$; LH: $t(11) = 4.45$, $P < 10^{-3}$) and ToM 1 in the right hemisphere ($t(12) = 3.75$, $P < .01$), and marginally in the left hemisphere ($t(11) = 2.05$, $P < .05$). This response appeared to be driven by overlap with the mid-STS theory of mind response.

The face and biological motion region had a moderate response to faces, a relatively weak response to biological motion, and a response to vocal sounds of variable effect size. This region responded significantly to faces over objects bilaterally (RH: $t(15) = 10.15$, $P < 10^{-7}$; LH: $t(14) = 6.64$, $P < 10^{-5}$). There was also a significant effect of biological motion in the left hemisphere ($t(15) = 3.33$, $P < .01$), but not in the right. The lack of an effect in the right hemisphere, however, appeared to be driven by a single outlier with a strongly negative effect of biological motion; there was a significant effect after removing this participant ($t(13) = 3.82$, $P < .01$). Additionally, there was a significant effect of the voice 1 contrast bilaterally (RH: $t(13) = 5.01$, $P < 10^{-3}$; LH: $t(12) = 4.61$, $P < 10^{-3}$), and the voice 2 contrast in the left hemisphere ($t(15) = 3.15$, $P < .01$) and marginally in the right hemisphere ($t(14) = 1.83$, $P < .01$). Lastly, there was a weak but significant effect of the ToM 1 contrast in the left hemisphere ($t(15) = 2.84$, $P < .01$). These results indicate that while there is a pSTS region responsive to both dynamic faces and biological motion, it only responds weakly to biological motion, with a much stronger response observed in the more selective biological motion area. Furthermore, this region appears to also have a substantial response to vocal sounds, of magnitude similar to or stronger than the response to biological motion.

2.4 Discussion

We investigated STS responses to a number of social cognitive and linguistic contrasts, and found that patterns of response along the length of the STS differed substantially across each contrast. Furthermore, we found largely selective subregions of the STS for theory of mind, biological motion, voice perception, and linguistic processing, as well as a region that specifically responds to dynamic faces and voices. Contrary to claims that there is little systematic spatial organization to the STS response to different tasks and inputs (Hein and Knight 2008), these findings argue for a rich spatial structure within the STS. In addition to these selective areas, regions responsive to multiple contrasts were observed, most clearly for responses to language and ToM, faces and voices, and faces and biological motion. These results indicate that the STS contains both domain-specific regions that selectively process a specific type of social information, as well as multifunctional regions involved in processing information from multiple domains.

Our analysis of resting-state functional connectivity of STS subregions supports the argument for systematic spatial organization in the STS. Our results point to distinct patterns of functional connectivity within the STS, suggesting the presence of fine-grained distinctions within the 2-3 patterns observed in prior studies (Power et al. 2011; Shih et al. 2011; Yeo et al. 2011). Furthermore, we show that these connectivity differences are linked to differences in response profiles, consistent with the broad claim that areas of common functional connectivity also share common function (Smith et al. 2009). Contrary to the claim that STS subregions are recruited for different functions based on their spontaneous coactivation with other brain regions (Hein and Knight 2008),

these results paint a picture in which STS subregions have stable, distinct response profiles and correspondingly distinct patterns of coactivation with the rest of the brain.

The STS regions found to respond to each contrast in the current study are broadly consistent with regions reported in prior studies. Prior studies using ToM tasks have most consistently reported the posterior-most TPJ region (Fletcher *et al.* 1995; Gallagher *et al.* 2000; Saxe and Kanwisher 2003; Saxe and Powell 2006; Ciaramidaro *et al.* 2007; Saxe et al. 2009; Dodell-Feder *et al.* 2011; Bruneau et al. 2012; Gweon et al. 2012), but some have also reported middle and anterior STS regions like those observed in the present study (Dodell-Feder *et al.* 2011; Bruneau *et al.* 2012; Gweon *et al.* 2012). At least two prior studies have investigated responses to both ToM and biological motion, and both found that the response to biological motion was anterior to the TPJ region elicited by ToM tasks, as observed in the present results (Gobbini et al. 2007; Saxe *et al.* 2009). Consistent with prior arguments, the right TPJ region observed in the current study was strongly selective for mental state reasoning, among the tasks used here.

Studies of face and biological motion perception have found responses in a similar region of posterior STS (pSTS) (Allison *et al.* 2000; Grossman *et al.* 2000; Pelphrey, Mitchell, *et al.* 2003; Pitcher *et al.* 2011; Engell and McCarthy 2013). In the present study, we observe overlapping responses to faces and biological motion in the pSTS, consistent with a recent study on responses to biological motion and faces in a large set of participants (Engell and McCarthy 2013). However, we find that the pSTS region responding to faces is slightly but reliably anterior to the region responding to biological motion, and that it is possible to define a maximally biological-motion-sensitive region of pSTS that has no response to faces over objects. Thus, pSTS

responses to faces and biological motion, while overlapping, also differ reliably. The finding of a consistent difference in position of face and biological motion responses diverges from the results of Engell and McCarthy (2013); this difference could result from the use of dynamic face stimuli in our study. While studies using static faces have typically observed a single face-responsive region of posterior STS, the current results and prior evidence (Pitcher *et al.* 2011) indicate that dynamic faces engage several regions along the length of the STS, and may engage a posterior STS region with slightly different spatial properties.

The most striking case of overlap observed in the current study was that between responses to dynamic faces and vocal sounds. This result manifested as a strong voice response in a region defined to be maximally responsive to faces, and substantial overlap between the set voxels responding significantly to each contrast. This finding is consistent with prior work finding a region of posterior STS that responds to faces and voices (Wright *et al.* 2003; Kreifelts *et al.* 2009; Watson, Latinus, Charest, et al. 2014), as well as individual neurons in macaque STS that respond to faces and voices (Barraclough et al. 2005; Ghazanfar et al. 2008; Perrodin et al. 2014). The strikingly high voice response of a region defined by a face contrast was nevertheless unexpected, and indicates that this region should not be characterized as a "face region," but rather a fundamentally audiovisual area. This case of overlap seems most plausibly interpreted as suggesting a common underlying process elicited by the two broad contrasts used in this study. One possibility is that this region is involved in human identification using both facial and vocal cues. However, this hypothesis cannot easily explain the strong preference for dynamic over static faces in this region (Pitcher *et al.* 2011). Another

possibility is that this region is involved in audiovisual processing of speech and/or other human vocalizations. Alternatively, this region might be more generally involved in processing communicative stimuli of different modalities. Further research will be necessary to tease apart these possibilities.

In addition to a region of overlapping response to faces and voices, a further anterior region responded highly selectively to vocal stimuli. This region was centered on the upper bank of the middle STS, consistent with prior reports (Belin *et al.* 2000).

Language responses have been observed along the entire length of the left STS (Binder *et al.* 1997; Fedorenko *et al.* 2012), consistent with our results. In our data, language responses were strongest bilaterally in a far-anterior region of STS, a pattern that prior studies have not noted. Unlike most prior studies, the sentences used in our language contrast involved no human characters whatsoever (nor any living things at all), to dissociate language from social reasoning. This difference might account for the slight divergence between our results and those of prior studies.

Another case of particularly strong overlap occurred between responses to language and ToM. The left TPJ region defined by a ToM contrast was modulated by linguistic content, and the bilateral anterior language regions were both modulated by mental state content. We also observed substantial overlap between language and ToM responses across the STS: roughly half of voxels with a language response also had an effect of ToM content. While not observed previously, this observation is consistent with prior reports that regions elicited by ToM and semantic contrasts both bear a rough, large-scale resemblance to default mode areas (Buckner and Carroll 2007; Binder 2012). While relationships between language and ToM in development have been extensively

documented (De Villiers 2007), lesion evidence indicates that these functions can be selectively impaired in adults (Apperly et al. 2004; Apperly et al. 2006). Nevertheless, the finding of strong overlap between language and ToM responses is intriguing and should be explored further. One potential account of the effect of ToM in language regions is that the presence of agents is an organizing principle of semantic representations in these regions.

Based on fMRI overlap results, we have argued that at the spatial resolution of the present study, STS responses to certain types of social information overlap, in some cases substantially. However, a number of caveats must be made regarding the interpretation of fMRI overlap data. For a number of reasons, it is impossible to directly infer the presence of overlap in underlying neural responses from overlap in fMRI responses, which measure a hemodynamic signal at relatively low resolution. For one, fMRI measures signal from blood vessels, and stronger signal is obtained from larger vessels that pool blood from larger regions of cortex (Polimeni et al. 2010). Next, fMRI is a relatively low-resolution measure (typically ~3mm), pooling responses over hundreds of thousands of individual neurons, and further structure likely exists within the resolution of a single voxel. Lastly, fMRI data is both intrinsically spatially smoothed by the use of *k*-space sampling for data acquisition, and typically smoothed further in preprocessing, to increase signal-to-noise ratio. In the present study, we mitigate this concern by measuring overlap at different smoothing kernels, and assessing response profiles of overlapping regions using spatially unsmoothed data.

The above points establish that fMRI can miss spatial structure at a fine scale. Thus, in principle, functionally specific STS subregions within the overlapping regions

observed here could exist at finer spatial scales. Nevertheless, the present results argue for substantial overlap at the typical resolution of fMRI studies, which isn't merely induced by spatial smoothing during preprocessing, and is robust to differences in the statistical threshold used to define regions.

We have argued that the STS contains subregions with distinct profiles of fMRI responses, resulting from distinct underlying neuronal selectivity profiles. Could the differences observed here instead relate to other influences on contrast-to-noise ratio (CNR), such as differences in signal reliability across regions, or differences in the efficacy of different stimulus sets in driving neural responses? It is unlikely that signal reliability contributes substantially to regional differences, given that these regions are roughly matched on tSNR (and despite slightly lower tSNR in voice regions, these areas had among the strongest category-selective responses observed). The paradigms we used were designed to be similar in basic ways (e.g., mostly blocked designs, with ~5 minutes of stimulation per condition), but could nevertheless differ in their ability to drive strong responses. For instance, they might differ in variability along stimulus dimensions encoded in a given region, which is impossible to judge without knowing these dimensions. We consider it unlikely that this accounts for the substantial differences across regions observed here, given that 1) each task drove strong responses in at least some STS subregion; and 2) for tasks that evoked the most spatially extensive responses (ToM and voices), we found similar responses across two tasks, suggesting that these were somewhat robust to stimulus details. Nevertheless, it is important to note that differences in the efficacy of different tasks could have influenced the response magnitudes reported here.

Another general limitation of this study is that while we would like to identify regions of the STS that are involved in specific cognitive and perceptual processes, we instead use the proxy of identifying regions of the STS that respond in a given task contrast, as in any fMRI study. There is presumably no simple one-to-one mapping between processes or representations and broad pairwise contrasts. Given that we have little knowledge of the specific processes underlying social perception and cognition, it is difficult to determine the appropriate mapping. For instance, our face stimuli presumably elicit processes related to the perceptual processing of faces (which itself is a high-level description that likely comprises multiple computations), but they also contain specific types of biological motion (eye and mouth motion), which might be processed via separate mechanisms. These stimuli could also trigger the analysis of intentions or emotional states of the characters in the clips. Likewise, our theory of mind contrasts are intended to target mental state reasoning, but the theory of mind condition in both contrasts is more focused on human characters, and thus some of the responses we observe may relate to general conceptual processing of humans or imagery of human characters (although note that prior studies have found TPJ responses using tighter contrasts (Saxe and Kanwisher 2003; Saxe and Powell 2006)). Generally, it is important to emphasize that the activation maps we describe presumably comprise responses evoked by a number of distinct processes, which future research will hopefully tease apart.

In sum, the present study converges with prior research to indicate that the STS is a key hub of social and linguistic processing. Our method of testing each subject on multiple cotnrasts enables us to rule out prior claims that the STS represents a largely

homogenous and multifunctional region (Hein and Knight 2008), instead revealing rich spatial structure and functional heterogeneity throughout the STS. Specifically, the STS appears to contain both subregions that are highly selective for processing specific types of social stimuli, as well as regions that respond to social information from multiple domains. These findings paint a picture in which the extraordinary human capacity for social cognition relies in part on a broad region that computes multiple dimensions of social information over a complex, structured functional landscape. This work opens up myriad questions for future research, including which specific computations are performed within each subregion of the STS, how the functional landscape of the STS differs in disorders that selectively disrupt or selectively preserve social cognition (e.g., autism and Williams' syndrome respectively), and whether the spatial overlap observed here is functionally relevant, or whether it reflects distinct but spatially interleaved neural populations.

2.5 References

Allison T, Puce A, McCarthy G. 2000. Social perception from visual cues: role of the STS region. Trends in cognitive sciences 4:267-278.

Ambady N, Rosenthal R. 1992. Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. Psychol Bull 111:256.

Apperly IA, Samson D, Carroll N, Hussain S, Humphreys G. 2006. Intact first-and second-order false belief reasoning in a patient with severely impaired grammar. Social neuroscience 1:334-348.

Apperly IA, Samson D, Chiavarino C, Humphreys GW. 2004. Frontal and temporo-parietal lobe contributions to theory of mind: neuropsychological evidence from a

false-belief task with reduced language and executive demands. J Cogn Neurosci 16:1773-1784.

Barraclough NE, Xiao D, Baker CI, Oram MW, Perrett DI. 2005. Integration of visual and auditory information by superior temporal sulcus neurons responsive to the sight of actions. J Cogn Neurosci 17:377-391.

Beauchamp MS, Lee KE, Argall BD, Martin A. 2004. Integration of auditory and visual information about objects in superior temporal sulcus. Neuron 41:809-824.

Behzadi Y, Restom K, Liau J, Liu TT. 2007. A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. Neuroimage 37:90.

Belin P, Zatorre RJ, Lafaille P, Ahad P, Pike B. 2000. Voice-selective areas in human auditory cortex. Nature 403:309-312.

Belin P, Zatorre RJ, Lafaille P, Ahad P, Pike B. 2000. Voice-selective areas in human auditory cortex. Nature 403:309-312.

Binder JR. 2012. Task-induced deactivation and the" resting" state. Neuroimage 62:1086-1091.

Binder JR, Frost JA, Hammeke TA, Cox RW, Rao SM, Prieto T. 1997. Human brain language areas identified by functional magnetic resonance imaging. J Neurosci 17:353-362.

Bonda E, Petrides M, Ostry D, Evans A. 1996. Specific involvement of human parietal systems and the amygdala in the perception of biological motion. J Neurosci 16:3737-3744.

Brass M, Schmitt RM, Spengler S, Gergely G. 2007. Investigating action understanding: inferential processes versus action simulation. Curr Biol 17:2117-2121.

Brett M, Johnsrude IS, Owen AM. 2002. The problem of functional localization in the human brain. Nature reviews neuroscience 3:243-249.

Bruneau EG, Dufour N, Saxe R. 2012. Social cognition in members of conflict groups: behavioural and neural responses in Arabs, Israelis and South Americans to each other's misfortunes. Philosophical Transactions of the Royal Society B: Biological Sciences 367:717-730.

Buckner RL, Carroll DC. 2007. Self-projection and the brain. Trends in Cognitive Sciences 11:49-57.

Calvert GA, Hansen PC, Iversen SD, Brammer MJ. 2001. Detection of audio-visual integration sites in humans by application of electrophysiological criteria to the BOLD effect. Neuroimage 14:427-438.

Chai XJ, Castañón AN, Ongür D, Whitfield-Gabrieli S. 2012. Anticorrelations in resting state networks without global signal regression. Neuroimage 59:1420-1428.

Ciaramidaro A, Adenzato M, Enrici I, Erk S, Pia L, Bara BG, Walter H. 2007. The intentional network: How the brain reads varieties of intentions. Neuropsychologia 45:3105-3113.

Corbetta M, Shulman GL. 2002. Control of goal-directed and stimulus-driven attention in the brain. Nature reviews neuroscience 3:201-215.

De Villiers J. 2007. The interface of language and theory of mind. Lingua 117:1858-1878.

Dodell-Feder D, Koster-Hale J, Bedny M, Saxe R. 2011. fMRI item analysis in a theory of mind task. Neuroimage 55:705-712.

Engell AD, McCarthy G. 2013. Probabilistic atlases for face and biological motion perception: An analysis of their reliability and overlap. Neuroimage 74:140-151.

Fedorenko E, Behr MK, Kanwisher N. 2011. Functional specificity for high-level linguistic processing in the human brain. Proceedings of the National Academy of Sciences 108:16428-16433.

Fedorenko E, Kanwisher N. 2009. Neuroimaging of Language: Why Hasn't a Clearer Picture Emerged? Language and Linguistics Compass 3:839-865.

Fedorenko E, Nieto-Castanon A, Kanwisher N. 2012. Lexical and syntactic representations in the brain: An fMRI investigation with multi-voxel pattern analyses. Neuropsychologia 50:499-513.

Fletcher PC, Happe F, Frith U, Baker SC, Dolan RJ, Frackowiak RS, Frith CD. 1995. Other minds in the brain: a functional imaging study of "theory of mind" in story comprehension. Cognition 57:109-128.

Gallagher H, Happe F, Brunswick N, Fletcher P, Frith U, Frith C. 2000. Reading the mind in cartoons and stories: an fMRI study of "theory of mind" in verbal and nonverbal tasks. Neuropsychologia 38:11-21.

Ghazanfar AA, Chandrasekaran C, Logothetis NK. 2008. Interactions between the superior temporal sulcus and auditory cortex mediate dynamic face/voice integration in rhesus monkeys. J Neurosci 28:4457-4469.

Gobbini MI, Koralek AC, Bryan RE, Montgomery KJ, Haxby JV. 2007. Two takes on the social brain: A comparison of theory of mind tasks. J Cogn Neurosci 19:1803-1814.

Greve DN, Fischl B. 2009. Accurate and robust brain image alignment using boundary-based registration. Neuroimage 48:63.

Grossman E, Donnelly M, Price R, Pickens D, Morgan V, Neighbor G, Blake R. 2000. Brain areas involved in perception of biological motion. J Cogn Neurosci 12:711-720.

Grossman ED, Blake R. 2002. Brain areas active during visual perception of biological motion. Neuron 35:1167-1175.

Gweon H, Dodell-Feder D, Bedny M, Saxe R. 2012. Theory of Mind Performance in Children Correlates With Functional Specialization of a Brain Region for Thinking About Thoughts. Child Dev 83:1853-1868.

Haxby JV, Hoffman EA, Gobbini MI. 2000. The distributed human neural system for face perception. Trends in cognitive sciences 4:223-233.

Hein G, Knight RT. 2008. Superior Temporal Sulcus-It's My Area: Or Is It? J Cogn Neurosci 20:2125-2136.

Kanwisher N. 2010. Functional specificity in the human brain: a window into the functional architecture of the mind. Proceedings of the National Academy of Sciences 107:11163-11170.

Kreifelts B, Ethofer T, Shiozawa T, Grodd W, Wildgruber D. 2009. Cerebral representation of non-verbal emotional perception: fMRI reveals audiovisual integration area between voice-and face-sensitive regions in the superior temporal sulcus. Neuropsychologia 47:3059-3066.

Lindstrom MJ, Bates DM. 1990. Nonlinear mixed effects models for repeated measures data. Biometrics 46:673-687.

Marchini JL, Ripley BD. 2000. A new statistical approach to detecting significant activation in functional MRI. Neuroimage 12:366-380.

Morris JP, Pelphrey KA, Mccarthy G. 2005. Regional brain activation evoked when approaching a virtual human on a virtual walk. J Cogn Neurosci 17:1744-1752.

Pelphrey KA, Mitchell TV, McKeown MJ, Goldstein J, Allison T, McCarthy G. 2003. Brain activity evoked by the perception of human walking: Controlling for meaningful coherent motion. J Neurosci 23:6819-6825.

Pelphrey KA, Morris JP, Mccarthy G. 2004. Grasping the intentions of others: the perceived intentionality of an action influences activity in the superior temporal sulcus during social perception. J Cogn Neurosci 16:1706-1716.

Pelphrey KA, Morris JP, Michelich CR, Allison T, McCarthy G. 2005. Functional anatomy of biological motion perception in posterior temporal cortex: an fMRI study of eye, mouth and hand movements. Cereb Cortex 15:1866-1876.

Pelphrey KA, Singerman JD, Allison T, McCarthy G. 2003. Brain activation evoked by perception of gaze shifts: the influence of context. Neuropsychologia 41:156-170.

Pelphrey KA, Viola RJ, McCarthy G. 2004. When Strangers Pass Processing of Mutual and Averted Social Gaze in the Superior Temporal Sulcus. Psychol Sci 15:598-603.

Perrodin C, Kayser C, Logothetis NK, Petkov CI. 2014. Auditory and visual modulation of temporal lobe neurons in voice-sensitive and association cortices. J Neurosci 34:2524-2537.

Pitcher D, Dilks DD, Saxe RR, Triantafyllou C, Kanwisher N. 2011. Differential selectivity for dynamic versus static information in face-selective cortical regions. Neuroimage 56:2356-2363.

Polimeni JR, Fischl B, Greve DN, Wald LL. 2010. Laminar analysis of 7T BOLD using an imposed spatial activation pattern in human V1. Neuroimage 52:1334-1346.

Power JD, Cohen AL, Nelson SM, Wig GS, Barnes KA, Church JA, Vogel AC, Laumann TO, Miezin FM, Schlaggar BL. 2011. Functional network organization of the human brain. Neuron 72:665-678.

Puce A, Allison T, Asgari M, Gore JC, McCarthy G. 1996. Differential sensitivity of human visual cortex to faces, letterstrings, and textures: a functional magnetic resonance imaging study. J Neurosci 16:5205-5215.

Saxe R, Brett M, Kanwisher N. 2006. Divide and conquer: a defense of functional localizers. Neuroimage 30:1088-1096.

Saxe R, Kanwisher N. 2003. People thinking about thinking people: The role of the temporo-parietal junction in "theory of mind". Neuroimage 19:1835-1842.

Saxe R, Powell LJ. 2006. It's the Thought That Counts: Specific Brain Regions for One Component of Theory of Mind. Psychol Sci 17:692-699.

Saxe RR, Whitfield-Gabrieli S, Scholz J, Pelphrey KA. 2009. Brain Regions for Perceiving and Reasoning About Other People in School-Aged Children. Child Dev 80:1197-1209.

Shih P, Keehn B, Oram JK, Leyden KM, Keown CL, Müller RA. 2011. Functional Differentiation of Posterior Superior Temporal Sulcus in Autism: A Functional Connectivity Magnetic Resonance Imaging Study. Biol Psychiatry 70:270-277.

Shultz S, Lee SM, Pelphrey K, McCarthy G. 2011. The posterior superior temporal sulcus is sensitive to the outcome of human and non-human goal-directed actions. Social cognitive and affective neuroscience 6:602-611.

Smith SM, Fox PT, Miller KL, Glahn DC, Fox PM, Mackay CE, Filippini N, Watkins KE, Toro R, Laird AR. 2009. Correspondence of the brain's functional architecture during activation and rest. Proceedings of the National Academy of Sciences 106:13040-13045.

Taylor KI, Moss HE, Stamatakis EA, Tyler LK. 2006. Binding crossmodal object features in perirhinal cortex. Proceedings of the National Academy of Sciences 103:8239-8244.

Triantafyllou C, Hoge RD, Krueger G, Wiggins CJ, Potthast A, Wiggins GC, Wald LL. 2005. Comparison of physiological noise at 1.5 T, 3 T and 7 T and optimization of fMRI acquisition parameters. Neuroimage 26:243-250.

Triantafyllou C, Hoge RD, Wald LL. 2006. Effect of spatial smoothing on physiological noise in high-resolution fMRI. Neuroimage 32:551-557.

Vander Wyk BC, Hudac CM, Carter EJ, Sobel DM, Pelphrey KA. 2009. Action understanding in the superior temporal sulcus region. Psychol Sci 20:771-777.

Vanrie J, Verfaillie K. 2004. Perception of biological motion: A stimulus set of human point-light actions. Behavior Research Methods 36:625-629.

Vigneau M, Beaucousin V, Herv P, Duffau H, Crivello F, Houd O, Mazoyer B, Tzourio-Mazoyer N. 2006. Meta-analyzing left hemisphere language areas: phonology, semantics, and sentence processing. Neuroimage 30:1414-1432.

Watson R, Latinus M, Charest I, Crabbe F, Belin P. 2014. People-selectivity, audiovisual integration and heteromodality in the superior temporal sulcus. Cortex 50:125-136.

Woolrich MW, Behrens TEJ, Beckmann CF, Jenkinson M, Smith SM. 2004. Multilevel linear modelling for FMRI group analysis using Bayesian inference. Neuroimage 21:1732-1747.

Woolrich MW, Ripley BD, Brady M, Smith SM. 2001. Temporal autocorrelation in univariate linear modeling of FMRI data. Neuroimage 14:1370-1386.

Wright TM, Pelphrey KA, Allison T, McKeown MJ, McCarthy G. 2003. Polysensory interactions along lateral temporal regions evoked by audiovisual speech. Cereb Cortex 13:1034-1043.

Yeo BT, Krienen FM, Sepulcre J, Sabuncu MR, Lashkari D, Hollinshead M, Roffman JL, Smoller JW, Zöllei L, Polimeni JR et al. 2011. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. J Neurophysiol 106:1125-1165.

2.6 Tables

Table 1: Peak coordinates (in MNI space) of response to each task. Coordinates were defined in individual participants (within search spaces for each task) and then averaged across participants.

ROI	x	y	z
LH			
ToM 1	47.4	-58.7	22.9
ToM 2	45.6	-59.6	24.2
Bio Motion	49.9	-59.7	11.4
Faces	55.6	-41.1	7
Voice 1	60.6	-17.4	-1.6
Voice 2	61.5	-15.1	-1.5
Language	53	-3.5	-15.1
RH			
ToM 1	-52.5	-52.2	21.1
ToM 2	-53.3	-55.2	23.2
Bio Motion	-571	-44.1	14.9
Faces	-54.9	-36.2	7.2
Voice 1	-60.8	-16.3	-1.4
Voice 2	-61	-13.7	-2
Language	-52.5	-0.9	-17.3

2.7 Figures

Figure 1: Individual subject activations to seven different contrasts in two representative example subjects. Analyses were restricted to the bilateral STS mask shown in yellow at the bottom, and were thresholded at a false discovery rate of $q < .01$. The slices displayed are at MNI x-coordinate ±52.

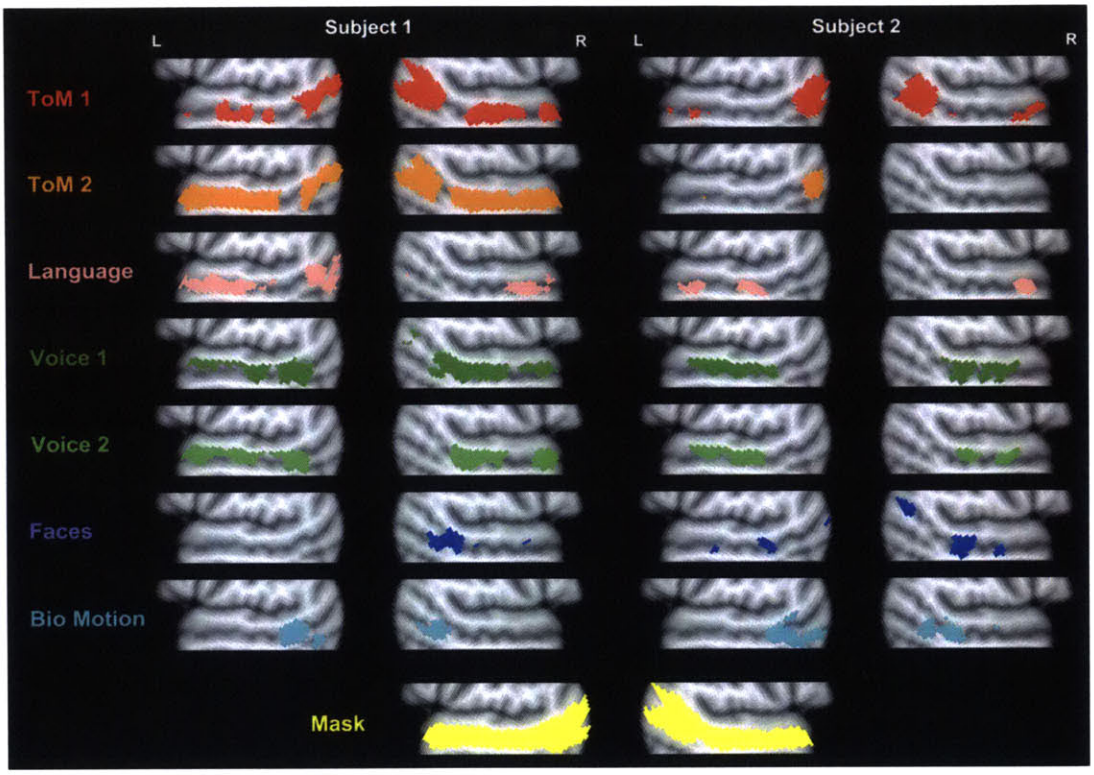
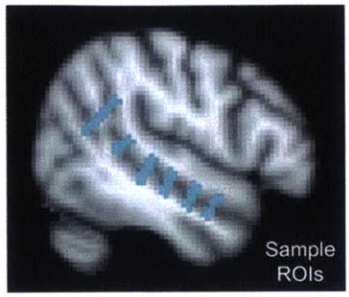
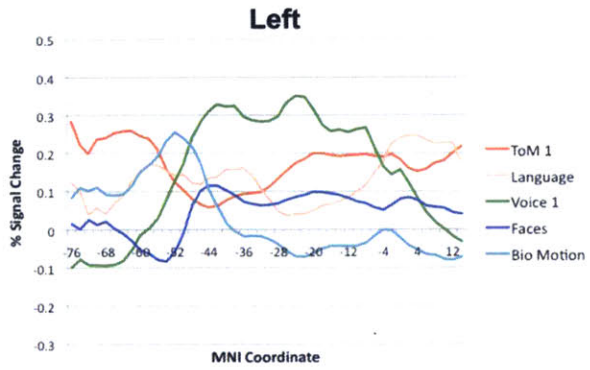


Figure 2: Responses to each task as a function of position along the length of the STS. The upper figure shows the ROIs that were used to extract responses at each position. The lower two graphs show left and right STS responses (percent signal change) for each task, as a function of y-coordinate in MNI space.





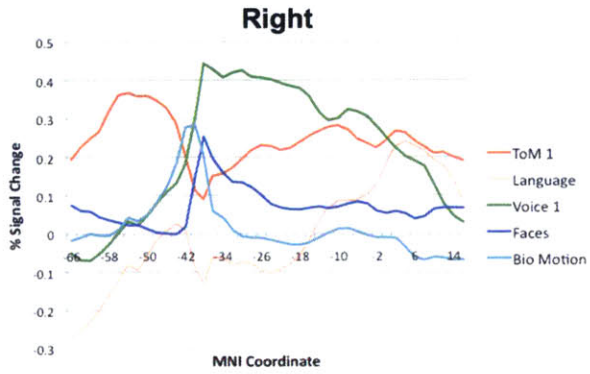


Figure 3: Responses (in percent signal change) of maximally sensitive regions for each contrast, across all conditions. Responses were measured in data independent of those used to define the ROIs.

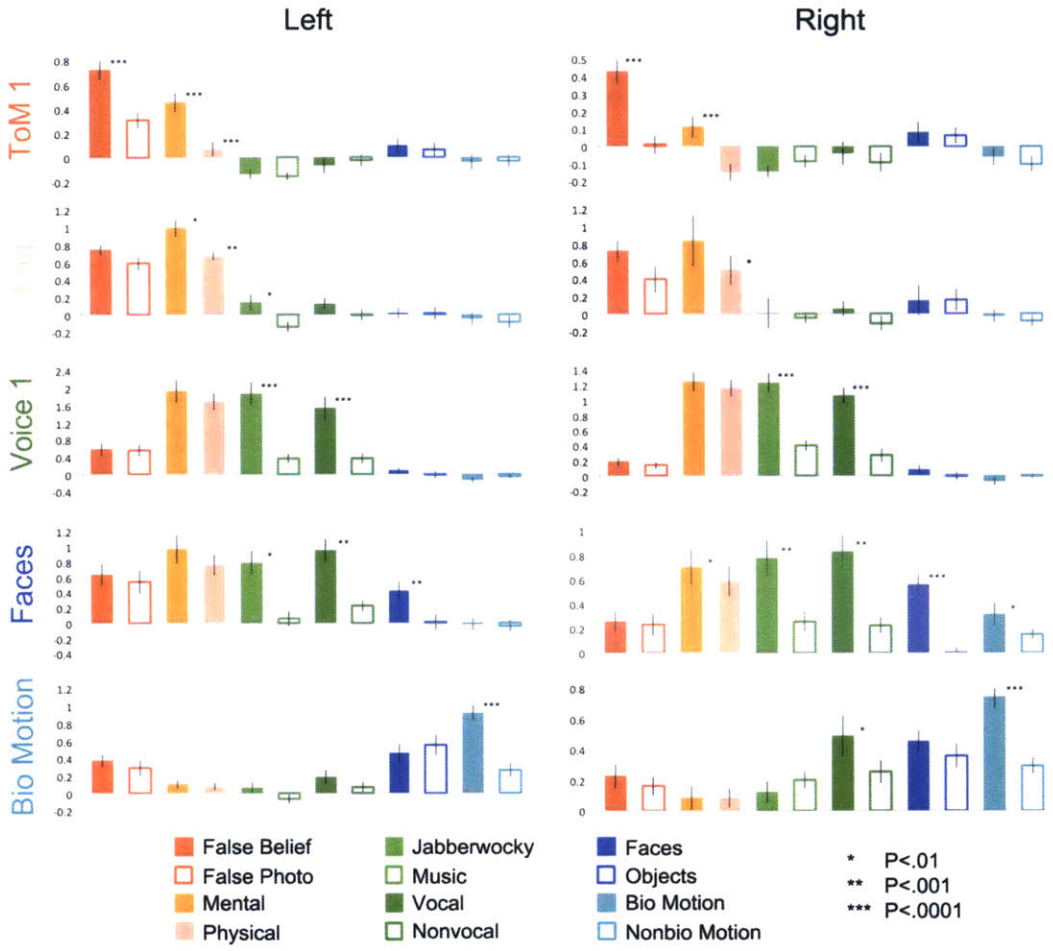


Figure 4: Matrices of functional connectivity similarity (correlations between whole-brain resting-state functional connectivity maps of seed ROIs defined by each contrast) and response similarity (correlations between vectors of task responses from each seed ROI). ROIs are defined to consist of a focal region of maximal activation to a given contrast.

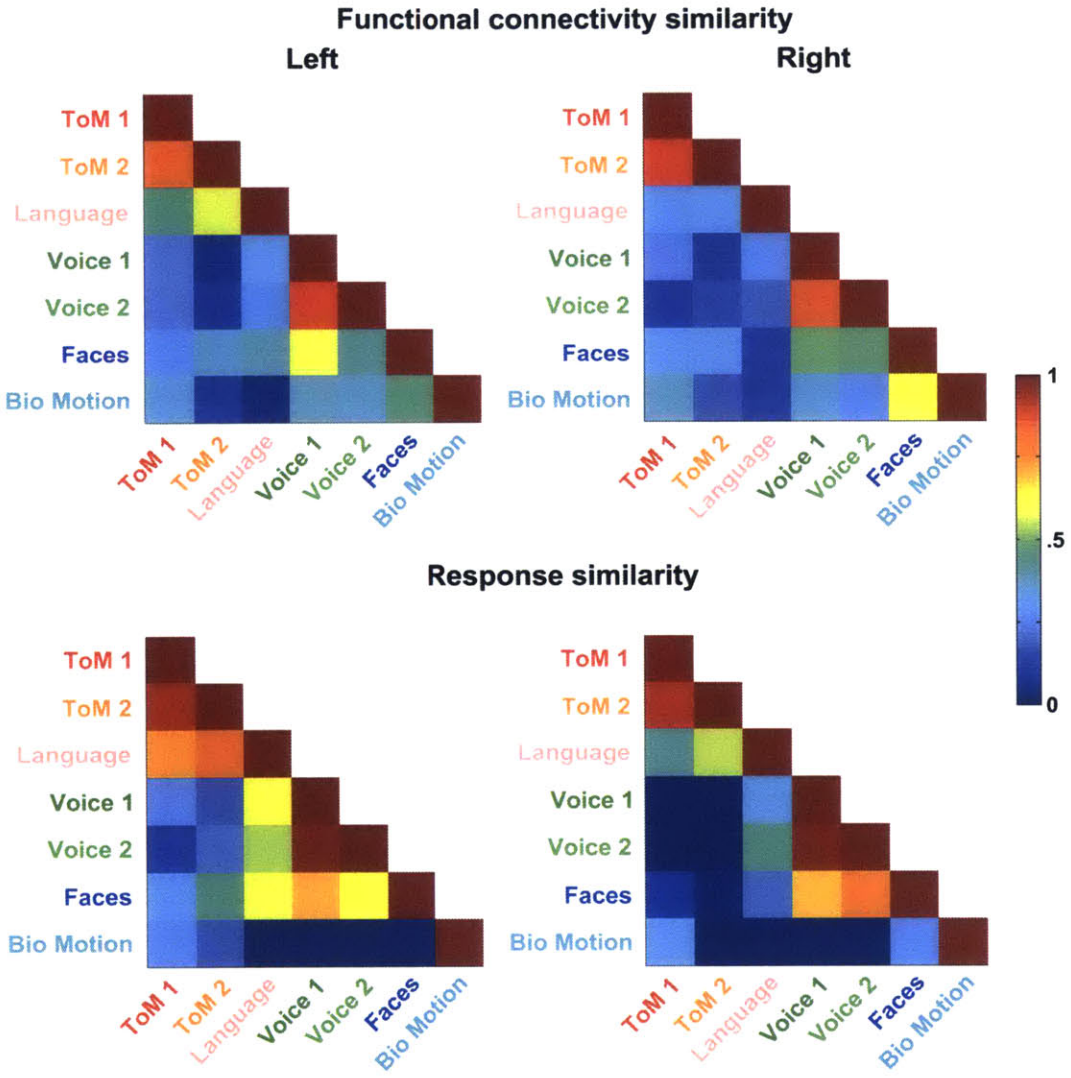


Figure 5: Overlap matrices for regions of activation defined by each task contrast. Each cell in a given overlap matrix is equal to the size of the overlapping region for the tasks on the corresponding row and column, divided by the size of the region of activation for the task on that row, as shown in the graphic on the left-hand side.

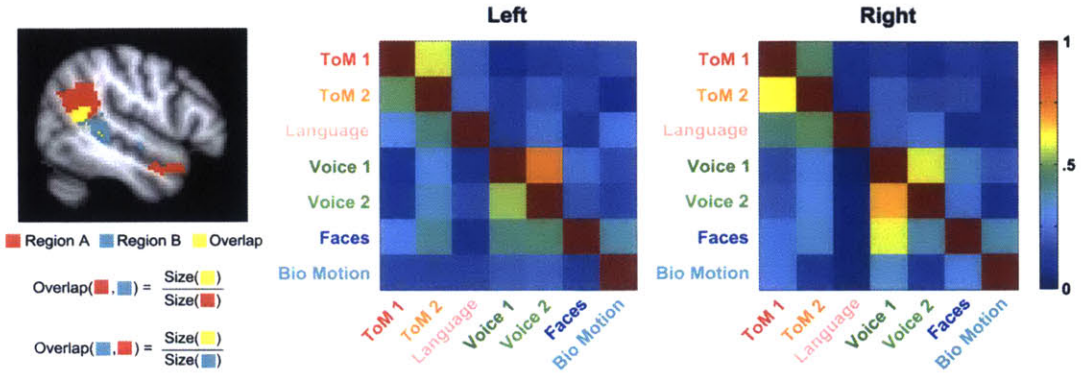
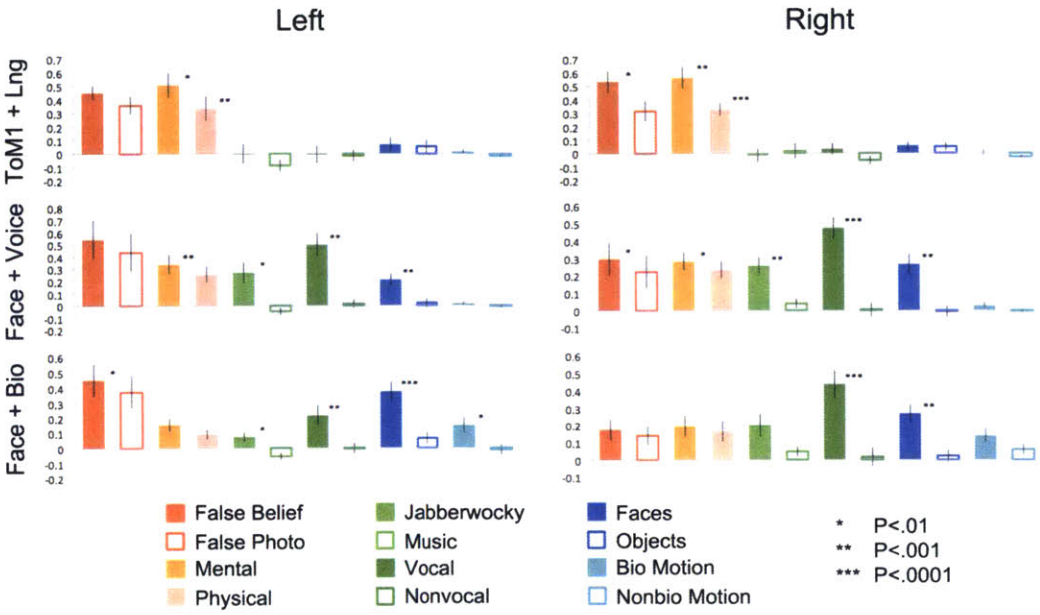
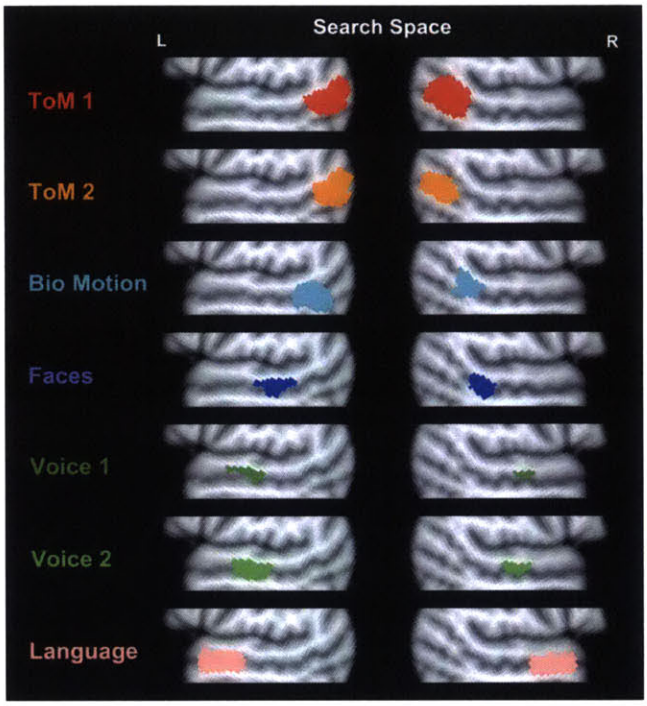


Figure 6: Responses (in percent signal change) of overlapping regions responsive to multiple contrasts, across all conditions. Responses were measured in data independent of those used to define the ROIs.

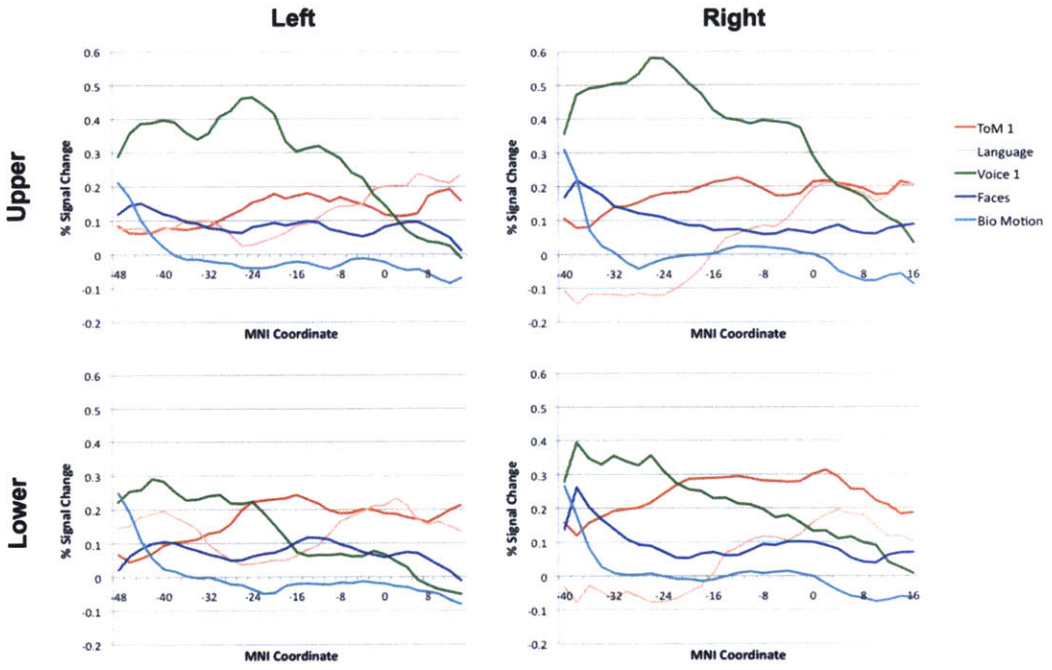


2.8 Supplementary Figures

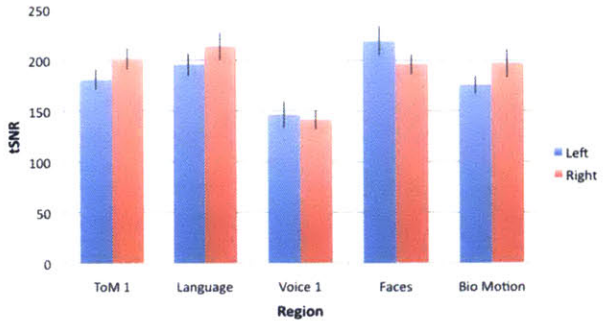
Supplementary Figure 1: Group-level search spaces for each task, placed over the region of maximal response (across individuals) for each contrast. These search spaces are used in the Anterior-posterior organization and Resting-state functional connectivity analysis sections, to constrain the definition of ROIs within individual subjects. The slices displayed at MNI x-coordinate ±52.



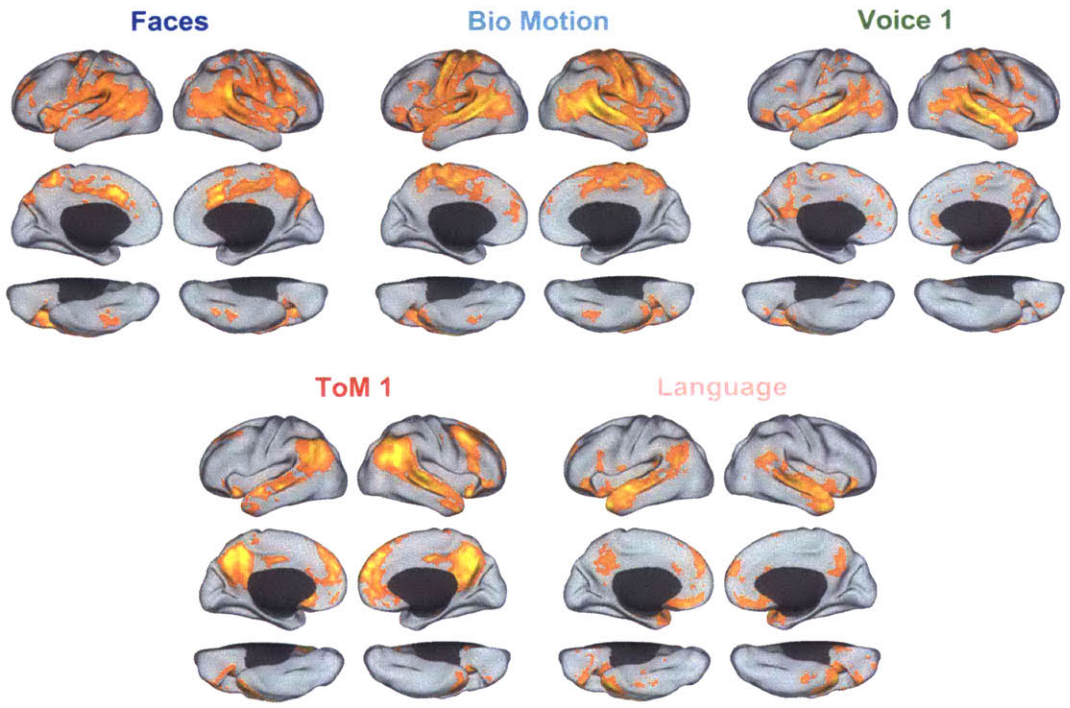
Supplementary Figure 2: Responses to each task (percent signal change) as a function of position along the length of the STS (y-coordinate in MNI space), separately for the upper and lower banks of the STS.



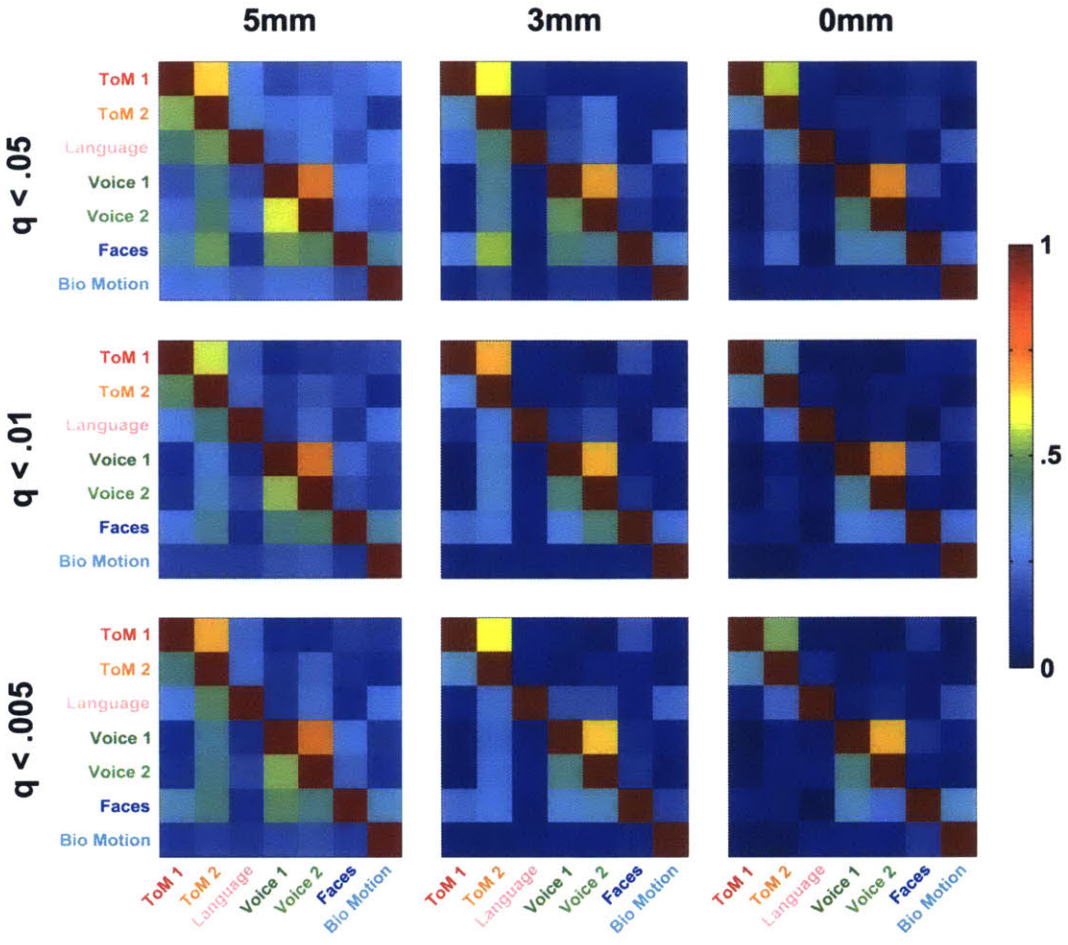
Supplementary Figure 3: Temporal signal-to-noise ratio (tSNR) for ROIs used in the selectivity analysis. Measured by finding the average time course within a region, and computing the ratio of its mean to its standard deviation.



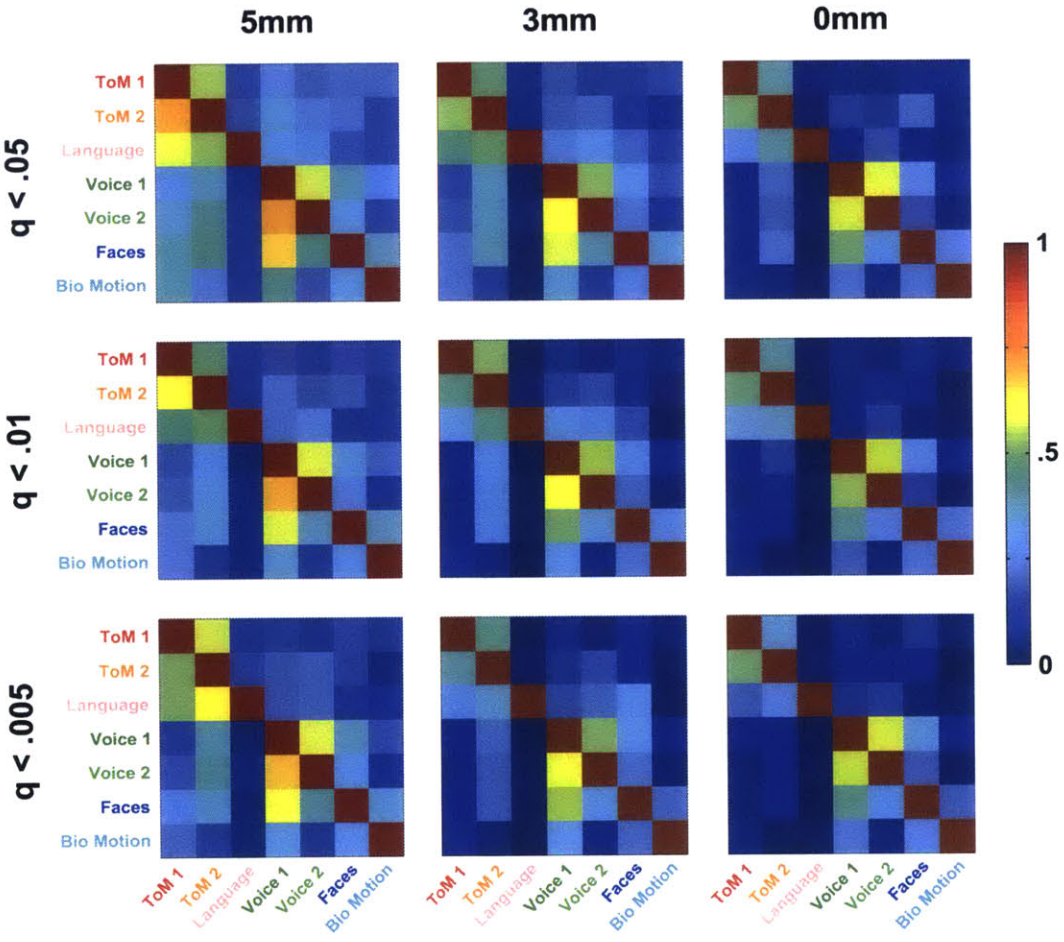
Supplementary Figure 4: Whole-brain functional connectivity maps, with seeds defined by regions of maximal sensitivity to each contrast.



Supplementary Figure 5: Overlap matrices for left-hemisphere regions, at three different thresholds and three different smoothing kernels. Each cell in a given overlap matrix is equal to the size of the overlapping region for the tasks on the corresponding row and column, divided by the size of the region of activation for the task on that row.



Supplementary Figure 6: Overlap matrices for right-hemisphere regions, at three different thresholds and three different smoothing kernels. Each cell in a given overlap matrix is equal to the size of the overlapping region for the tasks on the corresponding row and column, divided by the size of the region of activation for the task on that row.



Chapter 3: Parts-based representations of perceived face movements in the superior temporal sulcus[2]

Facial motion is a primary source of social information about other humans. Prior fMRI studies have identified regions of the superior temporal sulcus that respond specifically to perceived face movements (termed fSTS), but little is known about the nature of motion representations in these regions. Here we use multivoxel pattern analysis to characterize the representational content of the fSTS. Participants viewed a set of specific eye and mouth movements, as well as combined eye and mouth movements. Our results demonstrate that fSTS response patterns contain information about face movements, including subtle distinctions between types of eye and mouth movements. These representations generalize across the actor performing the movement, and across small differences in visual position. Critically, patterns of response to combined movements could be well predicted by linear combinations of responses to individual eye and mouth movements, pointing to a parts-based representation of complex face movements. These results indicate that the fSTS plays an intermediate role in the process of inferring social content from visually perceived face movements, containing a representation that is sufficiently abstract to generalize across low-level visual details, but still tied to the kinematics of face part movements.

[2] The contents of this chapter have been submitted for publication as Deen B, Saxe R. Parts-based representations of perceived face movements in the superior temporal sulcus.

3.1 Introduction

Facial motion provides a critical source of social information about others, regarding their emotional state, direction of attention, and vocal utterances. Among the set of face-responsive regions in the human brain, it has been argued that regions in the superior temporal sulcus (STS) are specialized for processing face motion and changeable aspects of faces (Allison *et al.* 2000; Haxby *et al.* 2000). In contrast with ventral temporal regions, face-responsive regions in the STS respond substantially more strongly to moving than to static faces, and prefer naturalistic motion to videos that are temporally scrambled (Pitcher *et al.* 2011; Schultz *et al.* 2013). Studies using static face images have found that these regions adapt to repeated presentations of the same facial expression, even when facial identity is varied, pointing to an identity-invariant representation of expression (Andrews and Ewbank 2004; Winston *et al.* 2004; Harris *et al.* 2012).

Despite compelling evidence for a role of the STS in processing perceived face motion, very little is known about the nature of face movement representations in this region. Multivoxel pattern analysis (MVPA) provides a powerful technique for characterizing neural representations, by asking which stimulus dimensions can be decoded from subtle variations in spatial patterns of response within a region. Said et al. (2010) found that response patterns to dynamic facial stimuli in anatomically defined anterior and posterior STS regions could be used to classify seven different emotional expressions. Skerry and Saxe (2014) found that responses patterns of a face-responsive STS subregion could classify positively- from negatively-valenced dynamic facial stimuli.

While these studies demonstrate that relevant pattern information can be read out from the STS, many questions remain about the nature of the representations underlying these effects. First, these studies did not attempt to dissociate similarity of facial expression from low-level visual similarity, or similarity of generic motion properties. Does the STS contain representations of face movements that are abstracted from low-level visual properties? Second, these studies used full-face emotional expressions that differed in motions of several face parts. Does the STS represent more subtle distinctions in the motion of individual face parts?

Furthermore, how do the representations of complex face movements relate to the movements of individual parts of the face? Facial expressions typically consist of coordinate movements of different face parts, and there is behavioral evidence that expressions are processed holistically: the expression the top or bottom half of a face influences the perceived expression in the other half (Calder et al. 2000). Does face-responsive STS integrate motion information from multiple face parts to generate a holistic, full-face motion representation? Or are complex movements represented in terms of motion of different parts of the face?

In the current study, we use fMRI and MVPA to address these questions and provide a richer account of the representational content of face-responsive STS. Participants viewed a set of dynamic face movements, including four eye/eyebrow movements, four mouth movements, and combinations of these, performed by one of two actors and presented in one of four visual positions. To test for parts-based versus holistic processing, we asked: can the pattern of response to a combined face movement be predicted from a linear combination of the responses to the eye and mouth component

movements? Or is the response to a combined movement distinct from, and not predictable by, responses to component movements? Our results suggest that 1) face-sensitive regions of the STS represent subtle discriminations in type of face movement; 2) these representations generalize across actors and small differences in visual position; and 3) complex movements are represented in terms of their parts.

3.2 Methods

Methods Preregistration

In order to reduce the risk of false positive results related to researcher degrees of freedom, and thus bolster the reproducibility of our results, we formally preregistered our experimental methods using the Open Science Framework (Deen 2015). Every aspect of our methodology, including the stimuli and task, participants, acquisition parameters, and full analysis pipeline with quantitative parameters fixed, was determined before any data analysis was performed. The analyses reported below are thus all preplanned, unless explicitly labeled otherwise.

Participants

Twenty-four adults participated in the study (age 21-36, ten female). Participants had no history of neurological or psychiatric impairment, and normal or corrected vision. All participants provided written, informed consent. No further exclusion criteria were used.

Paradigm

In the main experiment, participants viewed videos of faces performing a variety of face movements. The videos were generated using Poser 8 character animation software (http://my.smithmicro.com/poser-3d-animation-software.html), allowing tight control over visual properties of the stimuli. The movements included four eye or eyebrow movements (brow raise, eye closing, eye roll, scowl), four mouth movements (smile, frown, mouth opening, snarl), and sixteen combined eye/mouth movements corresponding to all possible combinations of the individual eye and mouth movements (Figure 1). The actions were be performed by two avatars ("actors"), one male and one female. Throughout the scan, participants fixated centrally, while face videos were presented slightly eccentric, at .5° of visual angle in one of four locations: to the upper left, upper right, lower left, and lower right. The faces in these videos subtended 4.2° by 5.9° of visual angle. Across 24 actions, 2 actors, and 4 visual positions, there were 192 distinct stimuli. Throughout a single scan session, four repetitions of each of these stimuli were presented.

Within each run, stimuli were presented in a jittered event-related design. Each clip lasted 2s, with a variable inter-stimulus interval (ISI) of 2, 4, or 6s, each occurring with 1/3 probability. To maintain attention, participants performed a one-back task on the action and actor in the video, irrespective of visual position (i.e., pressed a button with their right pointer finger when the actor and action were repeated across two subsequent trials, regardless of whether visual position repeated or not). In a given run, 48 stimuli were presented (such that all 192 are presented across four runs), as well as 6 repeated stimuli, which did not contribute to pattern analyses. Each run contained one example of each action/actor pair, in one of four visual positions. The order of stimuli in

each run were randomized across runs and participants. With a fixation block of 10s at the beginning of the experiment and 8s at the end of the experiment, each run lasted 5.7 minutes, and 16 runs were acquired throughout a scan session.

In addition to the main experiment, we ran a face localizer to define face-responsive subregions of the STS. Participants passively viewed videos of dynamic faces and videos of dynamic objects. Each video lasted 3s and was presented in a block of 18s. Blocks were presented in palindromic order, with condition order counterbalanced across runs and participants. There were six blocks each of faces and objects, as well as baseline blocks at the beginning, middle, and end of the experiment, in which six uniform color fields were presented for 3s each. Each run lasted 4.5 minutes, and 4 runs were acquired throughout a scan session. Further details about the stimuli have been reported previously (Pitcher *et al.* 2011).

Data acquisition

Data were acquired using a Siemens 3T MAGNETOM Tim Trio scanner (Siemens AG, Healthcare, Erlangen, Germany). High-resolution T1-weighted anatomical images were collected using a MPRAGE pulse sequence (repetition time [TR] = 2.53s; echo time [TE] = 3.48ms, flip angle $\alpha = 7°$, field of view [FOV] = 256mm, matrix = 256x256, slice thickness = 1mm, distance factor = .5, 176 near-axial slices, GRAPPA acceleration factor = 2, 24 reference lines). Functional data were collected using a T2*-weighted echo planar imaging (EPI) pulse sequence sensitive to blood-oxygen-level-dependent (BOLD) contrast (TR = 2s, TE = 30ms, $\alpha = 70°$, FOV = 192mm, matrix =

96x96, slice thickness = 2mm, 42 near-axial slices, multiband acceleration factor = 2, phase partial Fourier = 6/8).

Data Preprocessing and Modeling

Data were processed using the FMRIB Software Library (FSL), version 4.1.8, supplemented by custom MATLAB scripts. Anatomical and functional images were skull-stripped using FSL's brain extraction tool. Functional data were motion corrected using rigid-body transformations to the middle image of each run, and high-pass filtered (Gaussian-weighted least squares fit straight line subtraction, with σ = 50s (Marchini and Ripley 2000)). Localizer data were also spatially smoothed with a 4mm-FWHM isotropic Gaussian kernel, while data from the main task were not smoothed. For the purpose of analyzing group-level data in searchlight analyses, functional data were registered to the Montreal Neurological Institute 152 template brain (MNI space) using the following procedure: functional data were registered to anatomical images using a rigid-body transformation determined by Freesurfer's bbregister (Greve and Fischl 2009), and anatomical images were in turn registered to MNI space using a nonlinear transformation determined by FSL's FNIRT.

Whole-brain general linear model (GLM)-based analyses were performed for each participant, run, and task. Regressors were defined as boxcar functions convolved with a canonical double-gamma hemodynamic response function. All regressors were temporally high-pass filtered in the same way as the data. FSL's FILM was used to correct for residual autocorrelation (Woolrich *et al.* 2001). Data from each run and task was registered to the middle volume of the first run of the main task using a rigid-body

transformation determined by FSL's FLIRT, and further data analysis took place in this space.

For modeling data from the main task, we used the least-squares-single (LSS) method (Mumford et al. 2012). In this approach, a separate model is run for each trial, which consists of one regressor for the trial of interest, and one regressor for all other trials. This provides more accurate and lower variance estimates of response magnitudes for single trials in event-related designs with relatively small ISIs, by reducing collinearity between regressors in each model.

Region-of-interest definition

Analysis of the main task data was conducted using independently defined regions-of-interest (ROIs). We focused on three ROIs: an anatomically defined calcarine sulcus ROI (loosely termed V1), and functionally defined ROIs for motion-sensitive lateral occipitotemporal cortex (loosely termed MT+) and face-sensitive right STS (fSTS). The first two ROIs were intended as controls that were not expected to contain action representations. The V1 ROI was defined for each participant using the bilateral calcarine sulcus parcel from Freesurfer's Desikan-Killiany cortical parcellation. The bilateral MT+ ROI was defined using a group-level map of responses to coherently moving dots over a static baseline, from a separate dataset of 20 participants. Specifically, we placed 15mm-radius spheres around the peak coordinates of activation in lateral occipito-temporal cortex in each hemisphere, and intersected these spheres with the group-level activation map.

The fSTS ROI was defined functionally in individual participants. While face responses are most consistently reported in posterior parts of the STS, middle and anterior STS responses have also been reported (Winston *et al.* 2004; Pitcher *et al.* 2011). Prior studies have not observed clear functional differentiations between these areas, and thus don't suggest hypotheses as to which contain motion representations. For this reason, we chose to simply consider all face-responsive voxels within the STS. The fSTS in each participant was defined as set of voxels within an anatomical right STS mask that respond significantly ($P < .001$ voxelwise) to faces over objects in the localizer task. The anatomical mask was defined by manually drawing STS gray matter on the MNI brain. Any participant who had less than 50 voxels in the resulting ROI was excluded from the fSTS analysis; two participants were excluded based on this criterion.

In addition to predefined ROIs, we performed a hypothesis-neutral search for other brain regions containing action information by using a searchlight analysis across the whole brain. Specifically, we searched for regions whose patterns can discriminate the 24 action conditions, generalizing across position, as described in detail below. We searched across 8mm-radius spheres centered at each voxel in a gray matter mask, with each sphere intersected with the mask. The mask was defined using the MNI gray matter atlas, thresholded at 0%, and intersected with each individual participant's brain mask. Statistical maps within participants were registered to MNI space to perform inference across participants. Because coverage was only near-whole-brain and differed slightly across participants, we only considered voxels in which every participant had data. The resulting statistical map was thresholded at $P < .01$ voxelwise to form contiguous clusters of activation (where two voxels are considered contiguous if they share a vertex). To

correct for multiple comparisons across voxels, we used a permutation test to generate a null distribution for cluster sizes, and used this to threshold clusters of activation at $P <$ .05.

Multivoxel pattern analysis

We next used multivoxel pattern analysis (MVPA) to determine which features of our face motion stimuli could be discriminated by patterns of response within each ROI. In particular, we used the Haxby correlation method (Haxby et al. 2001). In this approach, the data are first split into two halves, and patterns of response to N distinct conditions are computed in each half. Then, a matrix of Fisher-transformed correlations between patterns from the first half and the second half of the data is computed, and for each participant, a difference score or "discrimination index" is computed: the mean within-condition correlation minus the mean between-condition correlation (i.e., the mean of the diagonal elements of this correlation matrix minus the mean of the off-diagonal elements; depicted in Figure 4A). Lastly, a one-tailed *t*-test is performed across participants to determine if these difference scores are significantly greater than zero, indicating that patterns in this region discriminate between the conditions tested. We did not correct for multiple comparisons across the three predefined ROIs, insofar as V1 and MT+ were intended as controls, and fSTS was hypothesized to contain action representations.

As a control measure, we first checked for discrimination of visual position, which we expected to find in V1 and MT+, but not fSTS. For this analysis, we split the data in half by trial number (averaging trial repetitions 1 and 3, and 2 and 4), collapsing

data over actions and actors. For each region, we constructed a 4x4 split-half correlation matrix, treating each position as a distinct condition, and assessed the difference score for this matrix.

To test for the presence of action representations, we performed a hierarchy of analyses, in which we first tested whether a region's patterns could discriminate among the 24 action conditions, and if this was the case, tested several more specific discriminations to detail the nature of action representations. Each of these tests were run in two ways, requiring generalization across either position (left versus right) or actor, by splitting the data across this dimension to compute the split-half correlation matrix. Generalization across position was considered a prerequisite for an abstract action representation, and therefore we only tested further hypotheses if a region's patterns contained action information that generalized across position. For the initial test for action information, we constructed a 24x24 split-half correlation matrix, treating each action as a distinct condition, and tested the difference score for this matrix.

This analysis revealed that fSTS, but not MT+ or V1, contained patterns that discriminated actions across position. Thus for this region, we next performed further specific tests. Three of these assessed the nature of representations of isolated eye and/or mouth movements, termed single movements (as opposed to combined eye and mouth movements). First, we tested for discrimination of eye versus mouth movements, by considering the 8x8 submatrix of correlations between isolated movements, and treating eye to eye and mouth to mouth correlations as within condition, but eye to mouth and mouth to eye correlations as between condition. We also tested for discrimination of

specific eye movements, by computing a difference score from the 4x4 submatrix of eye movements, and did the same for mouth movements (termed eye type and mouth type).

These tests for eye and mouth type information were relatively underpowered, using only 4x4 submatrices of a 24x24 correlation matrix. We thus ran two additional unplanned analyses to test for discrimination of eye and mouth type, taking advantage of the larger amount of data provided by responses to combined movements. First, we tested for discrimination of eye and mouth type within combined movements. Second, we tested for discrimination of eye and mouth type across single and combined movements—i.e., by assessing correlations between patterns of response to single and combined movements.

We next ran two analyses to probe the nature of representations of combined eye/mouth movements. One possibility is that these movements are encoded in a parts-based manner, such that the neural response to a combined movement is roughly the sum of neural responses to eye and mouth movements; this might be expected of a region that encodes the kinematics of face movements. Another possibility is that these representations are holistic, in that the neural response to combined movements cannot be decomposed into responses to individual components; this would be expected, for instance, of a region that encodes the emotion expressed by a face movement. These alternatives are not mutually exclusive: a region could contain neural subpopulations with both types of code.

We tested for the presence of parts-based representations by asking whether patterns of response to combined eye/mouth movements could be discriminated by linear combinations of patterns of response to the isolated movements. Within the first half of

the dataset, we used linear regression to find the linear combination of eye and mouth patterns that best predicted the combined pattern (depicted in Figure 5A). We then computed a 16x16 split-half correlation matrix between these "simulated" combined patterns from the first half, and empirical combined patterns from the second half. To maximize power, we computed two such matrices, where the stimulated patterns were computed from either the first or second half of the dataset, and averaged these together. Finding a significant difference score from this matrix would indicate that combined patterns could be discriminated by linear combinations of eye and mouth patterns.

To test for the presence of holistic representations, we asked whether combined patterns themselves do a better job of discriminating responses to combined movements in left out data than the simulated patterns do. In particular, we computed a difference score for split-half correlations between responses to the 16 combined movements, and asked whether this was significantly greater than the difference score for simulated-to-combined correlations, described above.* Finding a significant difference score in this matrix would indicate the presence of discriminative pattern information in responses to combined movements that isn't captured by the simulated patterns, pointing to a holistic representation.

Univariate analysis

* This approach differed slightly from our planned analysis, which compared within-condition correlations, rather than within/between difference scores. Upon analyzing the data, it became clear that simulated patterns had lower variance than responses to combined movements, which biases toward increased split-half correlations for simulated patterns. Because both within- and between-condition correlations are similarly influenced by differences in variance between simulated and combined patterns, the approach reported here is less influenced by this bias. This difference in analysis did not influence our conclusion regarding holistic processing.

To address whether differences in fSTS patterns across conditions were accompanied by differences in mean response magnitude of the region, we added an unplanned control analysis. We analyzed the mean response of the fSTS to each of the 24 action conditions, by averaging beta values across voxels in the region, as well as trials, actors, and positions. Post-hoc repeated measures ANOVAs were used to assess modulation of mean fSTS responses by action type.

3.3 Results

Investigating face motion representations using fMRI

We used fMRI and multivoxel pattern analysis to investigate cortical representations of perceived face movements. Participants viewed videos of dynamic faces, including 24 specific face movements, performed by two actors, and presented in four visual positions. The movements consisted of 4 eye movements, 4 mouth movements, and 16 combinations of these, allowing us to ask how representations of combined eye and mouth movements relate to representations of their parts. Our analysis focused on face-sensitive subregions of the STS (fSTS), as well as early visual areas V1 and MT+ as control regions. We followed a planned, formally preregistered hierarchy of analyses (Deen 2015). We asked whether patterns of response in each region contained information about perceived face motion, in a way that generalized across visual position. For regions that contained action information, we ran a number of subsequent tests, to detail the nature of this action information and to ask how responses to combined movements relate to responses to component movements.

Position and action decoding

As a control analysis, we first asked whether response patterns in our three ROIs contained information about visual position (Figure 2). As predicted, patterns of fMRI response could discriminate visual position in V1 ($t(23) = 5.99$, $P < 10^{-5}$) and MT+ ($t(23) = 2.12$, $P < .05$), but not fSTS ($t(21) = .94$, $P = .18$). This demonstrates that our approach was sufficiently sensitive to recover a well-established functional property of early visual regions.

We next tested whether these regions' response patterns contained information about action type, generalizing across visual position. Action information was observed in fSTS ($t(21) = 1.90$, $P < .05$), but not in V1 ($t(23) = -2.60$, $P = .99$) or MT+ ($t(23) = .99$, $P = .17$). This indicates that the fSTS contains a position-tolerant representation of face movements.

Could differences in discrimination ability result from differences in ROI size across the three regions? Mean ROI sizes were 549 (V1), 1246 (MT+), and 544 (fSTS) voxels. Because the MT ROI was substantially larger than the other two, we ran an unplanned control analysis using a smaller MT ROI. This ROI was defined in the same way as the planned region, but using 8mm spheres around bilateral peak coordinates, rather than 15mm spheres, and had a mean size of 473 voxels. With this smaller MT+ ROI, we still observed discrimination of position ($t(23) = 2.02$, $P < .05$), but not action ($t(23) = .38$, $P = .35$). This result shows that the observed dissociations between ROIs do not result from differences in ROI size.

Could differences in discrimination ability result from the use of a functional criterion to define fSTS, but not MT+ or V1, which might lead to the inclusion of noisy or unresponsive voxels in the latter ROIs? To ask this, we ran an unplanned control

analysis in which we only included voxels responded significantly to visual stimuli over baseline ($P < .001$ voxelwise) within our planned V1 and MT+ ROIs. These restricted control ROIs yielded qualitatively identical results, with significant discrimination of position (MT+: $t(24) = 2.18$, $P < .05$; V1: $t(24) = 6.78$, $P < 10^{-6}$), but not action (MT+: $t(24) = .57$, $P = .28$; V1: $t(24) = -3.33$, $P \approx 1$). This result shows that the observed dissociations between ROIs do not result from the lack of a functional criterion in defining MT+ and V1.

Do any brain regions other than the fSTS contain pattern information that discriminates perceived face movements? To address this question, we performed a whole-brain searchlight analysis. At our planned threshold ($P < .01$ voxel-wise, $P < .05$ cluster-wise), we did not observe any regions with significant decoding. To check whether any marginal effects could be observed, we additionally applied a threshold of $P < .05$ voxel-wise, $P < .05$ cluster-wise. In this analysis, we observed a single region in the right posterior superior temporal sulcus and middle temporal gyrus (MTG; Figure 3). This region was nearby and overlapping with the location of the posterior STS face response, but was centered slightly posterior and inferior to the face response. Thus, the posterior STS and adjacent cortex may uniquely contain position-tolerant pattern information about perceived face movements.

Action representations in fSTS

We next ran several further MVPA analyses to investigate the nature of action representations in fSTS (Figure 4B). Having demonstrated that patterns in this region discriminate face movements when generalizing across position, we additionally found

that this result held when generalizing across actor ($t(21) = 4.37$, $P < 10^{-3}$). This suggests that fSTS primarily contains representations of face movements themselves, rather than movements tied to specific agents. We also found that patterns of fSTS response could discriminate eye from mouth movements, generalizing across both position ($t(21) = 2.90$, $P < .01$) and actor ($t(21) = 4.12$, $P < 10^{-3}$).

Can fSTS patterns discriminate between different specific eye movements, and specific mouth movements? Within single (eye- or mouth-only) movements, we found no evidence for discrimination of specific movements, either when generalizing across position or actor (P's > .05). However, this negative result could result from a lack of power in these analyses, which focused on 4x4 submatrices of a 24x24 correlation matrix. To address this possibility, we performed an unplanned analysis to ask whether fSTS patterns discriminated type of eye or mouth movement within the combined (eye and mouth) movements, of which there were 16 rather than 4. We found significant discrimination of eye motion type, generalizing across both position ($t(21) = 2.32$, $P <$ .05) and actor ($t(21) = 2.48$, $P < .05$), as well as discrimination of mouth motion type, generalizing across both position ($t(21) = 2.19$, $P < .05$) and actor ($t(21) = 3.03$, $P < .01$). This results indicates that fSTS represents subtle distinctions between types of perceived eye and types of mouth movement.

To bolster this result, we performed a further unplanned analysis, which attempted discriminate specific eye and mouth movements by assessing correlations between patterns of response to single and combined movements. From a machine learning perspective, this corresponds to training a movement type classifier on single movements, and testing on combined movements (and vice versa). In this analysis, both

eye and mouth movements could be discriminated, both generalizing across position (eye: $t(21) = 1.88$, $P < .05$; mouth: $t(21) = 3.78$, $P < 10^{-3}$) and actor (eye: $t(21) = 3.29$, $P < .01$; mouth: $t(21) = 3.81$, $P < 10^{-3}$). This result demonstrates the presence of information about specific movement type even in patterns of response to single movements, suggesting that the negative results for discrimination within single movement responses resulted from a lack of power. Furthermore, this demonstrates that pattern information about eye movement type and mouth movement type generalize from responses to individual eye and mouth movements to combined movements.

Testing parts-based and holistic representations

How do patterns of fSTS response to combined movements relate to patterns of response to single movements? If the fSTS represents face movements in a parts-based fashion, responses to combined movements should reflect a combination of the responses separately evoked by the eye and mouth movements. In contrast, a holistic representation would predict that responses to combined movements cannot simply be decomposed into responses to parts. In order to assess the presence of parts-based and holistic representations in the fSTS, we generated "simulated" patterns of responses to combined movements, by finding an optimal linear combination of evoked responses to the corresponding eye and mouth movements, and asked to what extent these simulations predicted patterns of response to combined movements.

We found that patterns of response to combined movements could be discriminated by linear combinations of responses to individual eye and mouth movements (Figure 5), both when requiring generalization across position ($t(21) = 3.63$,

$P < 10^{-3}$) and actor ($t(21) = 4.68$, $P < 10^{-4}$). This provides strong evidence for a parts-based representation of face movements in the fSTS. This result also provides additional evidence for eye and mouth type information in fSTS (described above), which would be necessary to obtain a positive result in this analysis.

Is there action information in fSTS responses to combined movements that cannot be captured by combinations of responses to single movements, pointing to holistic representations? To address this question, we asked whether patterns of response to combined movements do a better job of discriminating between the same patterns in left-out data than simulated patterns do. We found no difference between discrimination ability based on simulated or combined patterns, generalizing across position ($t(21) = -.99$, $P = .83$) or actor ($t(21) = 1.45$, $P = .08$). Thus, our data do not provide evidence for holistic representations of face movements in the fSTS.

Univariate analysis

The above results demonstrate that distinct face movements evoke different patterns of response in the fSTS. Do they also evoke different mean responses, or are these effects specific to patterns? As an unplanned control analysis, we compared mean responses magnitudes to different actions (Figure 6). A one-way, two-level repeated measures ANOVA comparing responses to single and combined face movements revealed significantly stronger responses to combined movements (15% stronger responses to combined; $F(1,505) = 17.11$, $P < 10^{-4}$). Based on this difference, we subsequently looked for effects of action within single and combined movements. One-way repeated measures ANOVAs showed no effect of action condition on response

magnitude, for either single movements ($F(7,147) = .95$, $P = .47$) or combined movements ($F(15,315) = 1.52$, $P = .10$). Thus, in contrast to pattern information, mean responses did not differentiate movement types, apart from the distinction between single and combined movements.

3.4 Discussion

Our results demonstrate that the face-sensitive cortex in the STS (fSTS) represents face movements, in a manner that is robust to changes in actor and small changes in visual position. Such representations were not observed in earlier visual regions (V1 and MT+), where responses are not expected to be position-tolerant. Indeed, a search across the whole brain for position-tolerant action information revealed just a single region of posterior STS/MTG, roughly consistent with the location of fSTS. Action representations in fSTS were sufficiently fine-grained to discriminate subtle differences between types of eye motion and types of mouth motion. Finally, responses to combined eye and mouth movements could be well predicted by responses to the isolated eye and mouth movements, pointing to a parts-based representation of face movements. Taken together, these results indicate that fSTS contains a representation of the kinematics of face movements, which is sufficiently abstract to generalize across actor and across small variations in visual position, but which is nevertheless decomposable into the movements of separate face parts.

These results are consistent with prior findings of STS responses to perceived eye and mouth movements (Puce et al. 1998; Pelphrey *et al.* 2005), and extend these findings by identifying differences in response patterns to distinct types of motion. Strikingly, we find that linear combinations of fSTS responses to individual eye and mouth movements

can be used to discriminate responses to combined movements, and can do so as well as responses to combined movements themselves in independent data. This is consistent with an underlying neural code in which responses to specific eye and mouth movements sum linearly, as has been argued for the coding of an object's shape and its color or material in macaque inferotemporal cortex (Köteles et al. 2008; McMahon and Olson 2009). This novel approach for assessing parts-based versus holistic processing could be equally well applied in other domains of cognitive neuroscience, for characterizing representations in perceptual as well as high-level cognitive domains.

Evidence for parts-based coding was observed in spite of the fact that by design, many of the combined stimuli differed in terms of semantic interpretation from the individual movements they were composed of. For instance: closed eyes and an open mouth appears as a yawn; a scowl and smile appears as an evil grin; and a brow raise with a frown appears as an expression of confusion. A parts-based representation of combined motions is thus more consistent with a kinematic representation than a categorical representation of the interpretation of the movement.

This view of fSTS representations is consistent with an adaptation study using static emotional expressions (Harris *et al.* 2012). This study measured fSTS responses to face images defined along a morph continuum between two emotional expressions. Because emotional expression was perceived categorically, this design provided pairs of stimuli that differed in terms of physical properties of the expression but were perceived as expressing the same emotion, and pairs that differed both in terms of physical properties and perceived emotion. They found that the fSTS response released from adaptation whenever there was a change in physical properties of the expression,

regardless of whether perceived emotion differed; this pattern contrasted with the amygdala, which released from adaptation only for a change in perceived emotion. These findings indicate that the fSTS does not contain a categorical representation of perceived emotion, but a continuous representation of facial expression. Our results suggest a similar conclusion, based on responses to face motion rather than static expression, and further indicate that this representation is parts-based.

In interpreting others' face movements, we begin with a two-dimensional input on the retina, and are ultimately able to infer abstract social properties, such as others' mental or bodily states, from this visual input. The representation characterized in the present study appears to correspond to an intermediate stage in this inferential process: it is sufficiently abstract to generalize over low-level visual details, but still relates more to the properties of face motion itself than to their social interpretation. On this interpretation, where is social information from face movements extracted and represented? Prior evidence indicates that these processes involve both the STS and downstream regions. Watson et al. (2014) observed cross-modal adaptation in the right pSTS to the emotional content of faces and voices, pointing to a representation that generalizes beyond kinematic properties. Similarly, Peelen et al. (2010) found emotion information in a region of left pSTS/STG that generalized across dynamic face, body, and vocal inputs. The STS may thus contain both neural populations that represent face movements in a kinematic format, as well neural populations that encode a more general representation, pooled across multiple input modalities.

Prior evidence also suggests that other regions encode inferred social information with a higher degree of abstraction. The theory of mind network, a set of regions thought

to be involved in the representation of mental states of others, provides a plausible candidate for the substrate of such a downstream representation (Fletcher *et al.* 1995; Saxe and Kanwisher 2003). For instance, Skerry and Saxe (2014) found that the medial prefrontal cortex (mPFC), part of the theory of mind network, contained abstract emotion representations (of positive versus negative valence), which generalized across emotions depicted from dynamic facial expressions to emotions inferred from animations of geometric shapes mimicking social interactions. In contrast, fSTS contained emotion representations within each domain that did not generalize across domains.

Our study has several limitations, which should be noted. First, the present study doesn't directly address whether our results reflect the presence of a generic motion representation in the STS, rather than a representation that is specific to the kinematics of face motion, or human motion more generally. Prior evidence, however, suggests that this interpretation is unlikely. STS subregions that respond to face and body motion have been found to have little or no response to nonhuman motion, including random or coherent motion of dots (Grossman and Blake 2002), radial motion (Puce *et al.* 1998), and object motion (Pelphrey, Mitchell, *et al.* 2003; Pitcher *et al.* 2011). These results indicate that STS subregions are specifically involved in processing different types of human motion, and are not engaged by generic motion stimuli.

Second, although MVPA provides a powerful method for assessing the representational content of human brain regions, the method is intrinsically limited by the spatial resolution of fMRI. MVPA can only detect neural representations that are spatially organized at a scale that can be detected with the 2-3mm resolution of fMRI. There are known representations that lack such a spatial organization, such as

representations of place in the hippocampus or representations of face identity in macaque face patches, which wouldn't be detectable with MVPA (Dombeck et al. 2010; Dubois et al. 2015). Thus, it is not valid to make strong negative claims from MVPA data. In particular, the lack of evidence for a holistic representation in our study does not imply that no such representation exists. Nevertheless, our data do provide positive evidence for the presence of a parts-based representation in fSTS.

Another potential limitation of the current study was the use of animated stimuli. We chose to use animated stimuli to ensure tight visual control over the stimuli, and so that combined movements would be exact combinations of individual eye and mouth motions, which was critical for the logic our analyses. However, the animated stimuli are somewhat nonnaturalistic, and might be less likely to evoke meaningful emotion attributions than real actors would be. Thus, studies using video-recorded stimuli might be better suited for studying emotion representations in fSTS.

Our results point to a number of interesting directions for future research. If the fSTS primarily contains an intermediate representation of face movements, how does this region interact with other areas, such as the amygdala or mPFC, to support social inferences? Research on effective connectivity between these regions, or using combined TMS and fMRI to provide a causal manipulation, may be able to address this question. Beyond the dimensions considered in the present study, is the fSTS representation tolerant to other relevant dimensions, such as size, viewpoint, or larger changes in position? And lastly, if the fSTS representation is largely actor-invariant, corresponding to action type rather than an action-actor pairing, where does action information become

associated with actor to form a representation of a specific agent's motion or implied internal state?

To conclude, the present research provides evidence that the fSTS represents the face movements of others, in a manner that is abstracted from low-level visual details, but tied to the kinematics of face part movements. Future research should further detail the nature of motion representations in the fSTS, and clarify the role of this region in the inferential process that takes us from raw visual input to socially meaningful inferences about other humans.

3.5 References

Allison T, Puce A, McCarthy G (2000) Social perception from visual cues: role of the STS region. Trends in cognitive sciences 4:267-278.

Andrews TJ, Ewbank MP (2004) Distinct representations for facial identity and changeable aspects of faces in the human temporal lobe. Neuroimage 23:905-913.

Calder AJ, Young AW, Keane J, Dean M (2000) Configural information in facial expression perception. J Exp Psychol Hum Percept Perform 26:527-551.

Deen B (2015) FMVPA. Retrieved from osf.io/gqhk9.

Dombeck DA, Harvey CD, Tian L, Looger LL, Tank DW (2010) Functional imaging of hippocampal place cells at cellular resolution during virtual navigation. Nat Neurosci 13:1433-1440.

Dubois J, de Berker AO, Tsao DY (2015) Single-Unit Recordings in the Macaque Face Patch System Reveal Limitations of fMRI MVPA. J Neurosci 35:2791-2802.

Fletcher PC, Happe F, Frith U, Baker SC, Dolan RJ, Frackowiak RS, Frith CD (1995) Other minds in the brain: a functional imaging study of "theory of mind" in story comprehension. Cognition 57:109-128.

Greve DN, Fischl B (2009) Accurate and robust brain image alignment using boundary-based registration. Neuroimage 48:63.

Grossman ED, Blake R (2002) Brain areas active during visual perception of biological motion. Neuron 35:1167-1175.

Harris RJ, Young AW, Andrews TJ (2012) Morphing between expressions dissociates continuous from categorical representations of facial expression in the human brain. Proceedings of the National Academy of Sciences 109:21164-21169.

Haxby J, Gobbini M, Furey M, Ishai A, Shouten J, Pietrini P (2001) Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science 293:2425-2430.

Haxby JV, Hoffman EA, Gobbini MI (2000) The distributed human neural system for face perception. Trends in cognitive sciences 4:223-233.

Köteles K, De Maziere PA, Van Hulle M, Orban GA, Vogels R (2008) Coding of images of materials by macaque inferior temporal cortical neurons. Eur J Neurosci 27:466-482.

Marchini JL, Ripley BD (2000) A new statistical approach to detecting significant activation in functional MRI. Neuroimage 12:366-380.

McMahon DB, Olson CR (2009) Linearly additive shape and color signals in monkey inferotemporal cortex. J Neurophysiol 101:1867-1875.

Mumford JA, Turner BO, Ashby FG, Poldrack RA (2012) Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses. Neuroimage 59:2636-2643.

Peelen MV, Atkinson AP, Vuilleumier P (2010) Supramodal representations of perceived emotions in the human brain. J Neurosci 30:10127-10134.

Pelphrey KA, Morris JP, Michelich CR, Allison T, McCarthy G (2005) Functional anatomy of biological motion perception in posterior temporal cortex: an fMRI study of eye, mouth and hand movements. Cereb Cortex 15:1866-1876.

Pelphrey KA, Mitchell TV, McKeown MJ, Goldstein J, Allison T, McCarthy G (2003) Brain activity evoked by the perception of human walking: Controlling for meaningful coherent motion. J Neurosci 23:6819-6825.

Pitcher D, Dilks DD, Saxe RR, Triantafyllou C, Kanwisher N (2011) Differential selectivity for dynamic versus static information in face-selective cortical regions. Neuroimage 56:2356-2363.

Puce A, Allison T, Bentin S, Gore JC, McCarthy G (1998) Temporal cortex activation in humans viewing eye and mouth movements. J Neurosci 18:2188-2199.

Said CP, Moore CD, Engell AD, Todorov A, Haxby JV (2010) Distributed representations of dynamic facial expressions in the superior temporal sulcus. Journal of Vision 10:11.

Saxe R, Kanwisher N (2003) People thinking about thinking people: The role of the temporo-parietal junction in "theory of mind". Neuroimage 19:1835-1842.

Schultz J, Brockhaus M, Bülthoff HH, Pilz KS (2013) What the human brain likes about facial motion. Cereb Cortex 23:1167-1178.

Skerry AE, Saxe R (2014) A Common Neural Code for Perceived and Inferred Emotion. J Neurosci 34:15997-16008.

Watson R, Latinus M, Noguchi T, Garrod O, Crabbe F, Belin P (2014) Crossmodal Adaptation in Right Posterior Superior Temporal Sulcus during Face-Voice Emotional Integration. J Neurosci 34:6813-6821.

Winston JS, Henson R, Fine-Goulden MR, Dolan RJ (2004) fMRI-adaptation reveals dissociable neural representations of identity and expression in face perception. J Neurophysiol 92:1830-1839.

Woolrich MW, Ripley BD, Brady M, Smith SM (2001) Temporal autocorrelation in univariate linear modeling of FMRI data. Neuroimage 14:1370-1386.

3.6 Figures

Figure 1: Sample frames from video stimuli depicting face movements, from one of two actors. The stimulus set consisted of four eye/eyebrow movements, four mouth movements, and sixteen combined (eye and mouth) movements.

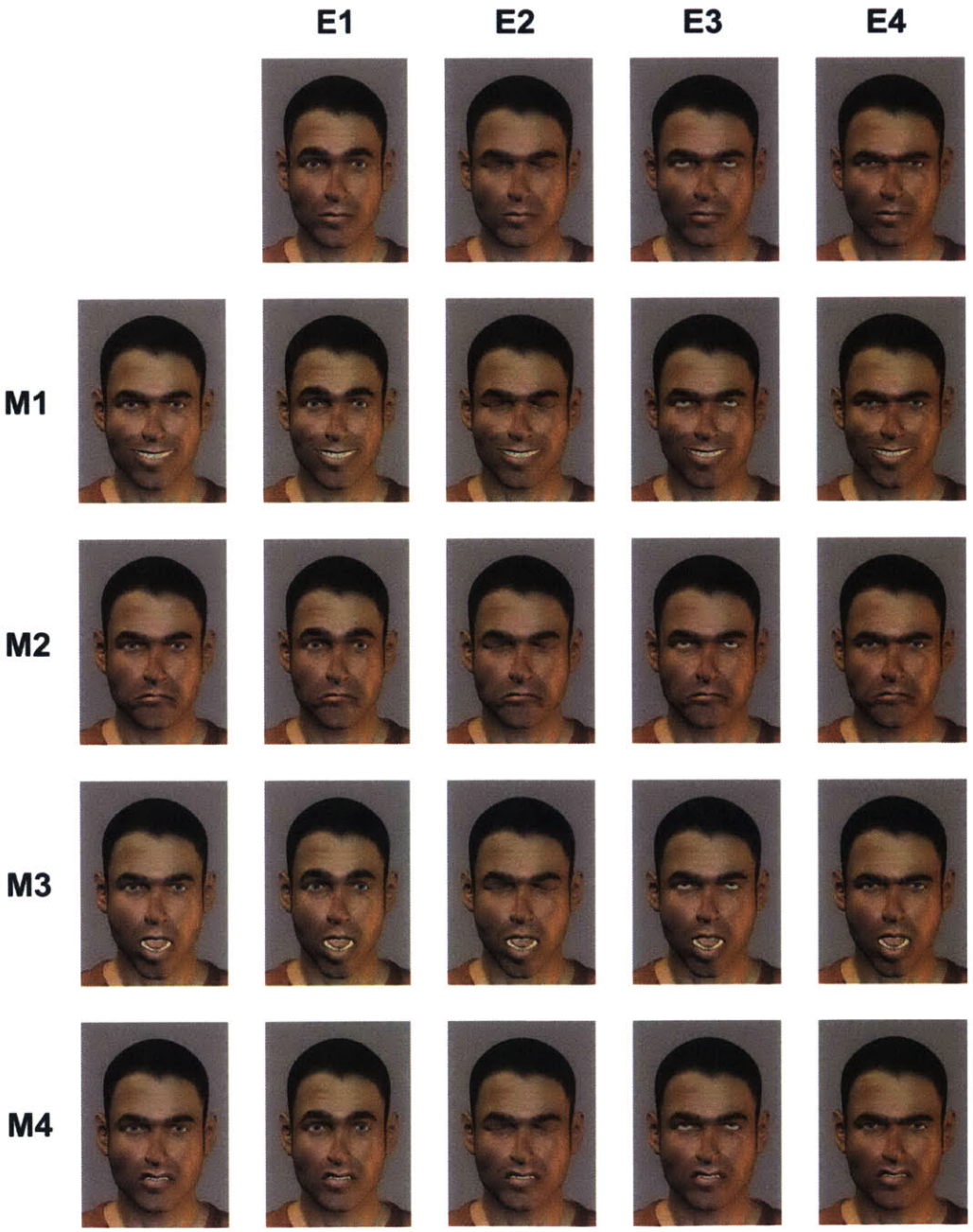


Figure 2: Upper: region of interest (ROI) locations, depicted as maps of the number of participants whose ROI included a given location. Lower: discrimination indices (correlation difference scores) for information about visual position (left), and action, generalizing across position (right). * denotes $P < .05$, and *** denotes $P < 10^{-3}$.

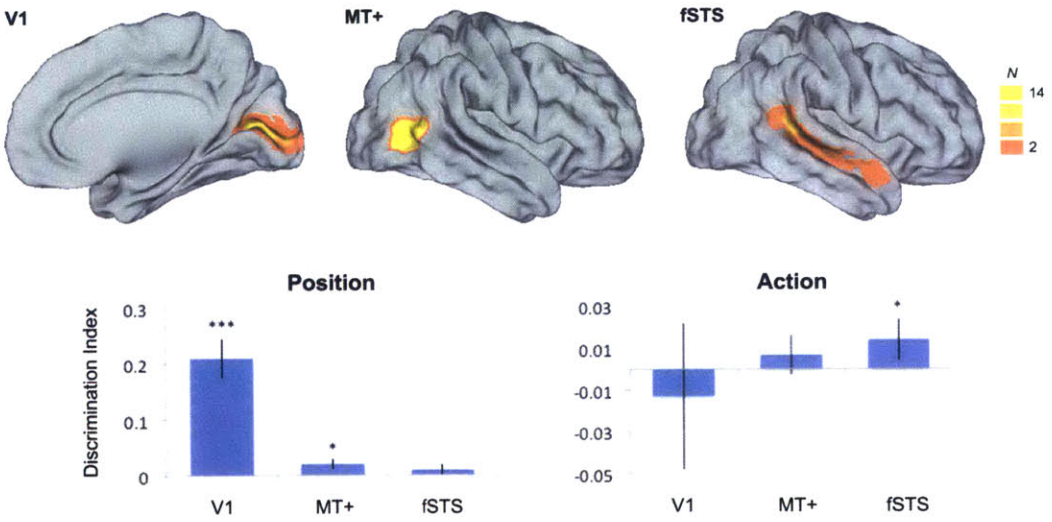


Figure 3: Searchlight analysis for position-tolerant action information. A significant effect indicates that patterns in an 8mm-radius sphere around a given location contain information that discriminates perceived action, in a manner that generalizes across visual position. Thresholded at $P < .05$ voxel-wise, with an additional permutation-based cluster-wise threshold of $P < .05$ to correct for multiple comparisons across voxels.

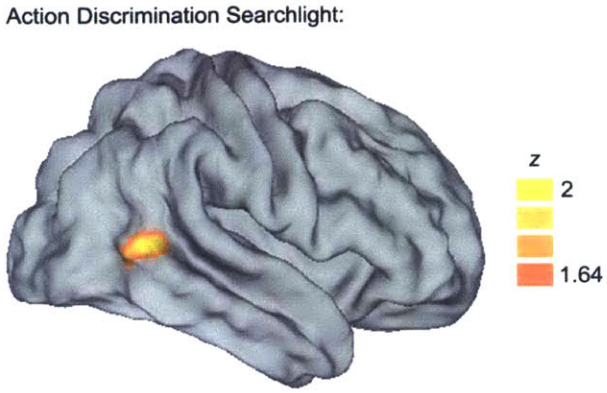


Figure 4: (A) Depiction of correlation difference method used for multivoxel pattern analysis. On the left is a matrix of split-half correlations of patterns of response to each action (where this split is either across visual position or actor). On the right is a matrix indicating which cells are within-condition correlations, and which are between-condition correlations, for action discrimination. Discrimination indices are computed as the difference between within-condition and between-condition correlations (Fisher-transformed). (B) Discrimination indices for various analyses of fSTS patterns. "Mouth type" and "eye type" refer to discrimination of one of four specific mouth (or eye) movements. "Single" refers to individual eye and mouth movements, while "combined" refers to stimuli with both eye and mouth motion. Single-to-combined analyses assessed correlations between patterns of response to single and combined stimuli. * denotes $P < .05$, ** denotes $P < .01$, and *** denotes $P < 10^{-3}$.

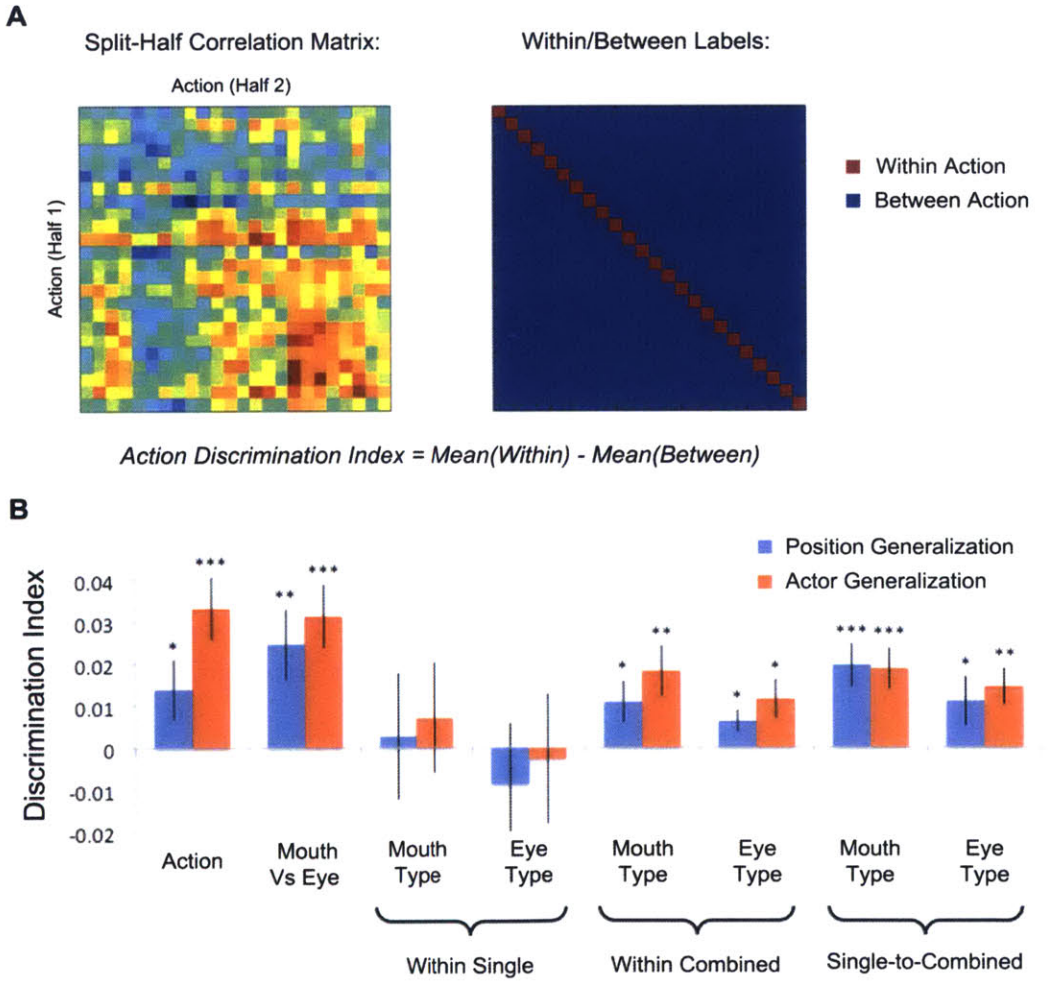


Figure 5: Evidence for a parts-based representation of combined face movements in the fSTS. (A) Method: in one half of the dataset, "simulated patterns" were constructed for each combined movement, as a linear combination of responses to the corresponding individual eye and mouth movements. These simulated patterns were then used to discriminate patterns of response to combined movements in the second half of the dataset. (B) Results from the simulation analysis. *** denotes $P < 10^{-3}$.

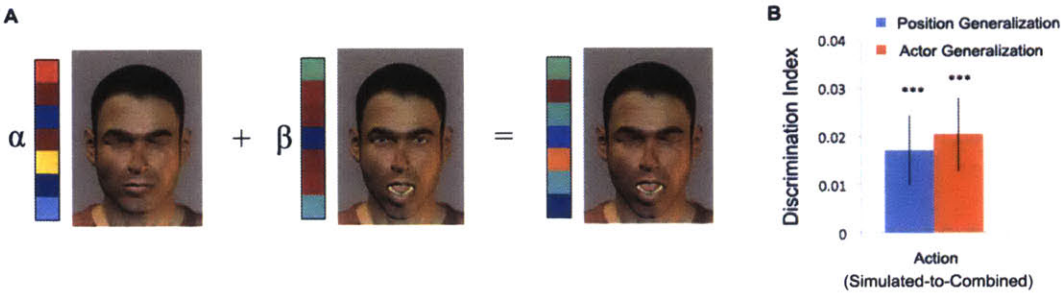
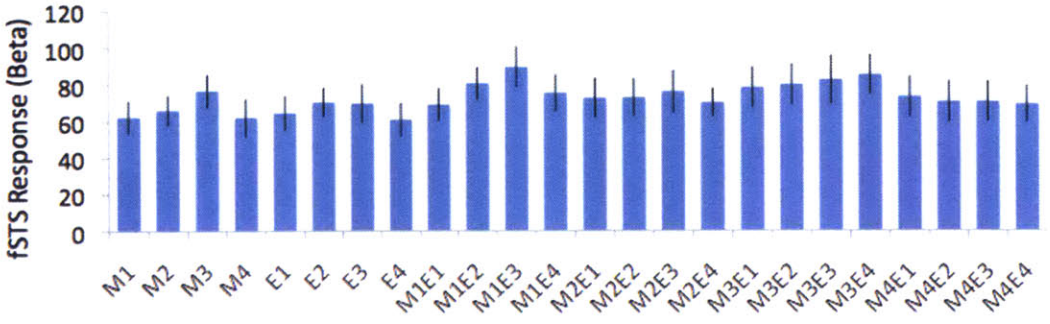


Figure 6: Mean fSTS response magnitudes (beta values) to different actions. The first eight bars correspond to individual eye or mouth movements, while the last sixteen bars correspond to combined movements. Labels M1-4 refer to the four mouth movements, in the following order: smile, frown, mouth opening, snarl. Labels E1-4 refer to the four eye movements, in the following order: brow raise, eye closing, eye roll, scowl.



Chapter 4: Dynamic facial and vocal information in the superior temporal sulcus[3]

Facial and vocal cues provide critical social information about other humans, including their emotional and attentional states and the content of their speech. Recent work has identified a region of the human superior temporal sulcus that selectively responds to both face movements and vocal sounds (fSTS). Here, we investigate the functional role of this region by measuring its response profile to a range of face movements, vocal sounds, and hand movements using fMRI. We find that the fSTS responds broadly to different types of audio and visual face actions, including both richly social communicative actions, as well as minimally social noncommunicative actions. Strikingly, however, responses to hand movements were very low, whether communicative or not, indicating a specific role in the analysis of face actions (facial and vocal), not a general role in the perception of any human action. Furthermore, spatial patterns of response in this region were able to decode communicative from noncommunicative face actions, both within and across modality (facial/vocal cues). These functional properties of the fSTS contrast with a region of middle STS that has a selective, largely unimodal response to speech sounds over both communicative and noncommunicative vocal nonspeech sounds. Taken together, these results point to the fSTS as a mid-level stage of social perceptual inference from faces and voices, sensitive to both lower-level action features and more abstract social properties.

[3] The contents of this chapter have been submitted for publication as Deen B, Saxe R, Kanwisher N. Dynamic facial and vocal information in the superior temporal sulcus.

4.1 Introduction

We learn a great deal about the character, thoughts, and emotions of another person by watching their face and listening to their voice. In addition to explicit verbal information, face movements and vocal sounds convey rich nonverbal clues to others' internal states that are essential for normal social interaction. What brain mechanisms underlie the extraction and representation of these communicative signals?

A candidate locus of these processes is the superior temporal sulcus (STS), which is considered a convergence zone for diverse sources of social information. Prior work has identified two subregions of the STS that are prime suspects for the perception of social information from faces and voices. First, a region of the STS that responds strongly to face movements (the fSTS) has recently been shown to respond just as strongly to voices, but not to non-face biological motion. This region thus appears to selectively process visual and audio cues from the face region (Deen et al. 2015). Second, a distinct region of middle STS and superior temporal gyrus (STG) has been argued to respond selectively to voices, but not visual social signals (vSTS; (Belin et al. 2000; Fecteau et al. 2004; Deen *et al.* 2015)).

Do these regions play a role in the perception of communicative signals, and if so what information do they extract? As in the problem of transformation-invariant object recognition (DiCarlo et al. 2012), extracting social meaning from visual and auditory stimuli entails detecting cues that bear a highly nonlinear relationship to raw stimulus features, and thus likely requires multiple stages of processing. Brain regions engaged in early stages in this computation would contain representations tied to lower-level stimulus features, potentially limited to certain domains of social information (face, hand

or body motion, or vocal sounds). In contrast, brain regions situated at the final stages of this inference would contain explicit representations of communicated mental states and/or propositional content, abstracted across a range of stimulus features and input domains (Skerry and Saxe 2014). Between these two extremes lies a broad space of potential mid-level representations, still somewhat tied to stimulus features and/or input domains, but beginning to make explicit more abstract, social dimensions. Determining which of these potential mid-level representations exists in the brain can provide insight into the computations underlying social perceptual inference.

In the present study, we investigate the role of STS subregions in inferring social meaning from perceptual input by measuring their responses to video clips showing a range of different face and hand movements and audio clips of vocal sounds, using fMRI. Within each domain (faces, voices, and hands), we used both richly communicative, socially relevant stimuli (e.g., a surprised face, a vocal expression of disgust, a hand gesturing "stop"), as well as noncommunicative, largely socially irrelevant stimuli (e.g., a chewing face, a throat-clearing sound, and a hand writing with a pen). We also included both speech and nonspeech stimuli, to test whether verbal and nonverbal communicative signals are processed distinctly.

We find that the fSTS responds broadly to different types of face movements and vocal sounds, but does not respond strongly to hand movements or non-social controls (object movements or musical sounds). Although the mean response of the fSTS did not discriminate between communicative and noncommunicative signals, patterns of response in the region could be used to decode this distinction, both within and across input domains (faces and voices). This pattern of a response is consistent with a mid-

level representation of face actions that is not restricted to socially relevant input, but begins to make social dimensions explicit, and to generalize across input domains. The vSTS, in contrast, responded most strongly to audio speech signals, over nonspeech vocal sounds, visual stimuli, and nonsocial controls. These results point to a reconceptualization of the division of labor in the STS: instead of separate streams for facial and vocal inputs, there are separate streams for processing audiovisual face actions and speech sounds.

4.2 Methods

Participants

Fifteen adults participated in the study (age 18-34 years, nine female). Participants had no history of neurological or psychiatric impairment, and normal or corrected vision. All participants provided written, informed consent.

Stimuli and paradigm

Participants viewed a set of video and audio clips depicting various face and hand movements and vocal sounds, as well as nonsocial controls (Figure 1). Among nonspeech stimuli, we included both richly social communicative actions and minimally social noncommunicative actions in each modality, and orthogonally manipulated the presence of mouth motion in face movements. Communicative hand movements consisted of gestures, while noncommunicative hand movements consisted of hand-object interactions. We additionally included audio, visual, and audiovisual speech stimuli, consisting of speakers uttering lists of nonsense words with English phonology. Lastly, we included audio clips of instrumental music as an auditory control, and video

clips of moving objects as a visual control. This led to thirteen total conditions: 1) communicative, high-mouth-motion face movements (FCHM); 2) communicative, low-mouth-mouth face movements (FCLM); 3) noncommunicative, high-mouth-motion face movements (FNHM); 4) noncommunicative, low-mouth-motion face movements (FNLM); 5) communicative hand movements (HC); 6) noncommunicative hand movements (HN); 7) communicative nonspeech vocal sounds (VC); 8) noncommunicative nonspeech vocal sounds (VN); 9) audio nonword speech (SA); 10) visual nonword speech (SV); 11) audiovisual nonword speech (SAV); 12) music (M); 13) objects (O).

Human stimuli were recorded in a television studio using a professional-grade HD video camera and microphone. Face movements, vocal sounds, and speech acts were performed by four actors (two female), wearing black shirts, with a black matte backdrop. Hand movements were performed by three actresses (all female), with their right hand protruding from a black sheet, such that only their hand and upper arm were visible. All actors were unfamiliar to participants in the study.

Among nonspeech stimuli, there were 8-11 specific actions (or tokens) for each condition; each actor performed each action 3-13 times. These tokens were as follows: 1) FCHM: disgusted expression, exhausted exhale, intrigued expression, uncertain expression, uncertain head shake and expression, tongue stick, surprised expression (with mouth open), disapproving head shake and expression ("tsk-tsk"), "yeesh" expression; 2) FCLM: concerned brow raise, confused brow furrow, eye roll, disappointed head hang, head nod ("yes"), head shake ("no"), single head nod ("hi"), skeptical expression, suggestive expression, surprised expression (with mouth closed), wink; 3) FNHM: blow

air, puff cheeks, chew food, cough, move lower jaw left/right, lick lips, pick at teeth with tongue, yawn; 4) FNLM: blink, falling asleep motion (head falling), gaze shift to the lower left, gaze shift to the lower right, gaze shift to the upper left, gaze shift to the upper right, neck stretch (side to side), neck stretch (rotating 180°), shiver, smooth pursuit eye movement, sniff; 5) HC: air quotes, "come here" wave, finger wag, money sign, finger gun gesture, figure point, "so-so" gesture, thumbs down, thumbs up, wave hello, dismissive wave; 6) HN: flip coin, grasp ball (with all fingers), grasp ball (with pointer finger and thumb), shake a bottle, sprinkle seasoning, toss a ball, tug a cord, turn a book page, twist a bottle cap, type on a keyboard, write with a pen; 7) VC: relaxed ahh, sad aww, cute aww, amused ha, hmph, flirtatious rrr, ugh, uh-huh, uh-uh, yigh; 8) VN: ahh (as if opening mouth for a doctor), wretching sound (as if being choked), cough, gargle, grunt, hiccup, throat clear with mouth closed, throat clear with mouth open, yawn. Among speech stimuli, there were 6 tokens (specific lists of nonwords; e.g. "cho cre las lanby caldet raldence cre paments cotlessy ploo"); each actor spoke each list 3-13 times.

From the resulting set of 1,323 video and audio clips of nonspeech actions, we then chose a subset to use for the experiment, such that clip duration was controlled within modality (faces, hands, or voices), and such that balanced proportions of stimuli from each token and actor were included for each condition. Likewise, from the resulting set of 184 speech clips, we chose a subset such that duration of all clips was near 5s, and such that balanced properties of stimuli from each token and actor were included. This resulted in 128 FCHM clips (mean duration 2.23s), 128 FCLM clips (2.22s), 128 FNHM clips (2.28s), 128 FNLM clips (2.31s), 144 HC clips (1.98s), 144 HN clips (1.97s), 157 VC clips (1.32s), 168 VN clips (1.48s), and 46 speech clips (5.07s).

As a high-level auditory control condition, we used 150 instrumental music clips from a range of genres (e.g. classical, jazz, rock), cut in duration to 1.5s to match the length of VN stimuli. All audio stimuli were root-mean-square amplitude normed and ramped with a 50ms linear ramp at the beginning and end of the clip. As a high-level visual control condition, we used 60 video clips of dynamic objects, using in a prior experiment (Pitcher et al. 2011), cut in duration to 2.27s to match the mean duration of face motion clips.

In the fMRI experiment, stimuli were presented in a blocked design, with separate blocks for each of the thirteen conditions. A fixed number of clips were presented in each block; because stimulus durations differed across modalities, this number varied across modalities such that the total stimulus duration for blocks of each condition was roughly 20s (9 stimuli for faces and objects, 10 for hands, 13 for nonspeech vocal sounds and music, and 4 for speech clips). The inter trial interval between clips in a block was chosen such that total block length was 22s for each block. In each run, 26 blocks (2 per condition) were presented, in palindromic order, with specific block order counterbalanced across runs and subjects. Blocks were separated by 6s of a baseline condition, consisting of a black screen with a white central fixation cross. There was an additional 10s of baseline at the beginning of the experiment, 16s in the middle, and 10s at the end, such that each run lasted 12:32 min. Each participant received eight runs of the experiment during a scan session. To maintain attention, participants performed a 1-back task during the experiment, pressing a button when an individual clip within a block repeated itself (one repeat per block).

Stimulus Ratings

To verify that our communicativeness manipulation was effective, we collected ratings on the stimuli using Amazon Mechanical Turk. For each video or audio clip used in the experiment, 20 participants viewed or listened to the clip and answered questions in a brief survey. For the purposes of this study, we asked, "To what extent is this (sound/action) communicative (i.e., produced to intentionally communicate information to another human)?" Participants responded on a scale of 0 (not communicative at all) to 6 (highly communicative). Other questions were asked for separate purposes and are not reported here. Participants were limited to users in the USA, and with a human intelligence task (HIT) approval rating of at least 95%, and at least 50 HITs performed previously. The surveys included a catch question with an objective answer (e.g., "what is the gender of the actor/actress?" for face movement videos). Only responses with a correct answer to the catch question were accepted, to ensure that participants watched or listened to the clip, and weren't responding randomly. Responses were averaged across participants, actors, and specific clips for each token (with an average of 281 responses per token), and statistics were performed across tokens.

Communicativeness ratings across all tokens are shown in Figure 2. To assess the reliability of these responses, we split responses across two subsets of ten participants, and computed the split-half correlation across tokens. This correlation was very high (r = .99, $P \approx 0$), indicating highly reliable responses. We next used a one-way ANOVA to assess the effect of category (treating all eight categories as distinct) on responses, and observed a highly significant effect of category on communicativeness ratings ($F(7,72)$ = 84.14, $P < 10^{-31}$, R^2 = .89). In particular, communicativeness was significantly higher for

FCHM relative to FNHM ($t(15) = 12.42$, $P < 10^{-8}$), FCLM relative to FNLM ($t(20) = 10.84$, $P < 10^{-9}$), HC relative to HN ($t(20) = 15.47$, $P < 10^{-11}$), and VC relative to VN ($t(17) = 9.09$, $P < 10^{-7}$). Within each modality (faces, voices, hands), all tokens in the communicative condition were rated as more communicative than tokens in the noncommunicative condition. All communicative tokens were rated higher than middle score of 3, and all but 5 of the 39 noncommunicative tokens were rated lower than 3. These results demonstrate that our manipulation of communicativeness had the desired effect.

Data acquisition

MRI data were acquired using a Siemens 3T MAGNETOM Tim Trio scanner (Siemens AG, Healthcare, Erlangen, Germany). High-resolution T1-weighted anatomical images were collected using a multi-echo MPRAGE pulse sequence (repetition time [TR] = 2.53s; echo time [TE] = 1.64ms, 3.5ms, 5.36ms, 7.22ms, flip angle $\alpha = 7°$, field of view [FOV] = 256mm, matrix = 256x256, slice thickness = 1mm, 176 near-axial slices, acceleration factor = 3, 32 reference lines). Functional data were collected using a T2*-weighted echo planar imaging (EPI) pulse sequence sensitive to blood-oxygen-level-dependent (BOLD) contrast (TR = 2s, TE=30ms, $\alpha = 90°$, FOV = 192mm, matrix = 64x64, slice thickness = 3mm, slice gap = .6mm, 32 near-axial slices, near-whole-brain coverage).

Data preprocessing and modeling

Data were processed using the FMRIB Software Library (FSL), version 4.1.8, supplemented by custom MATLAB scripts. Anatomical and functional images were skull-stripped using FSL's brain extraction tool. Functional data were motion corrected using rigid-body transformations to the middle image of each run, corrected for interleaved slice acquisition using sinc interpolation, spatially smoothed using an isotropic Gaussian kernel (5mm FWHM), and high-pass filtered (Gaussian-weighted least squares fit straight line subtraction, with σ = 50s (Marchini and Ripley 2000)). Although all analyses were performed in native functional space for each participant, normalization was required for combining results of certain analyses across participants. Functional images were registered to anatomical images using a rigid-body transformation determined by Freesurfer's bbregister (Greve and Fischl 2009). Anatomical images were in turn normalized to the Montreal Neurological Instititute-152 template brain (MNI space), using FMRIB's nonlinear registration tool (FNIRT).

Whole-brain general linear model (GLM)-based analyses were performed for each subject and run. Regressors were defined as boxcar functions including each block from a given condition, convolved with a canonical double-gamma hemodynamic response function. Temporal derivatives of each regressor were included in the models, and all regressors were temporally high-pass filtered. FMRIB's improved linear model (FILM) was used to correct for residual autocorrelation (Woolrich et al. 2001). Lastly, data were combined across runs for each subject using 2nd-level fixed effects analyses, after transforming beta maps to the middle image of the first run for each subject. Data were also combined across even runs and odd runs, for split-half analyses.

Region-of-interest analysis

How do face- and voice-sensitive subregions of the STS respond to communicative and noncommunicative face motions, hand motions, and vocal sounds? To address this question, we performed a region-of-interest (ROI) analysis using two contrasts. The face contrast compared the four face movement conditions to the dynamic object condition. The voice contrast compared the three vocal conditions (communicative/noncommunicative vocal sounds and audio speech) to the music condition. ROIs were defined in individual participants using the face and voice contrasts from the odd runs of the task. To spatially constrain ROI locations, we used search spaces defined based on a prior study, which identified a posterior STS face region and a middle STS voice region (Deen *et al.* 2015). Search spaces were registered from MNI space to each current participant's native functional space. For each participant, hemisphere, and contrast, we defined an ROI as the set of active voxels ($P < 10^{-3}$ voxelwise) within a 7.5mm-radius sphere around the peak coordinate within the search space. Participants with no active voxels were excluded from the corresponding analysis; we identified right fSTS in 15/15 participants, left fSTS in 10, right vSTS in 13, and left vSTS in 11 participants.

For each ROI, we extracted responses (percent signal change) across all thirteen conditions, in independent data from even runs of the experiment. Percent signal change was extracted by averaging beta values across each ROI and dividing by mean BOLD signal in the ROI. We then performed several statistical tests to characterize the response profiles of these regions; all tests were performed as mixed effects ANOVAs, with participant included as a random effect.

We first assessed selectivity profiles by comparing faces to objects, hands to objects, and vocal sounds (including speech) to music, using a separate ANOVA for each contrast and region. This served to confirm that each region had a reliable effect of the contrast used to define it, and to replicate the pattern of selectivity we have observed previously (Deen *et al.* 2015). Second, we tested whether communicativeness modulated ROI responses, using a region (4) by modality (face, voice, hand) by communicativeness ANOVA on all human nonspeech conditions. Third, we tested whether speech content modulated responses, using a region by modality (face, voice) by speech content (speech, non-speech) ANOVA across all face and voice conditions. These ANOVAs were followed up with post-hoc tests to characterize the effects observed. Lastly, for regions with audio and visual speech responses, we compared their response to audiovisual speech with the sum of responses to audio and visual speech, to test for the presence of multimodal interactions manifesting as super- or subadditive responses.

Independent component analysis

While the ROI analysis describes the response profile of several focal STS subregions, the STS is a large and functionally diverse area. We next asked: what are the dominant response profiles to these stimuli across the entire STS? To this end, we analyzed our data using independent component analysis (ICA), which models voxelwise responses as a linear combination of underlying response profiles, such that the weighting of each profile across voxels is maximally statistically independent. This approach complements the ROI analysis in two ways: 1) it is data-driven, allowing the dominant

features of STS functional organization to be revealed by our data; 2) it assesses responses across the full STS, rather than in a set of predefined ROI locations.

The input data for our implementation of ICA consisted of a voxel-by-condition matrix. We first defined an STS mask by manually drawing gray matter in the STS bilaterally in MNI space, and registered this to each participant's native functional space. Within this bilateral STS mask, we selected voxels that responded to a task > rest contrast at a liberal threshold ($P < .01$ voxelwise). Beta values from each of the thirteen conditions were extracted from each selected voxel, to construct a voxel by condition data matrix for a given participant. For each participant, we then removed the mean of this matrix across voxels, and divided by the standard deviation across voxels and conditions, to ensure that each participant contributed similarly to the overall matrix. These within-participant data matrices were then concatenated across participants in the voxel dimension to define a group-level data matrix. This approach to combining data across participants doesn't rely on normalization, and thus doesn't require an assumption that voxels in similar locations across subjects are functionally similar, and allows for voxel selection in each participant.

Prior to performing ICA, we performed dimensionality reduction using principal components analysis (PCA), to restrict our attention to dimensions capturing reliable variance. To this end, we used a leave-one-participant-out approach. For each participant, we ran PCA on a data matrix from the other 14 participants, to obtain a set of 13 principal component vectors in 13-dimensional condition space. We then split the left-out participant's data in half by even and odd runs, and computed a voxel-by-condition data matrix separately for each half. For each potential number of components

N (between 1 and 13), we projected the first-half data matrix onto the subspace spanned by the first N components, and computed the extent to which the resulting projected data could explain the second-half data matrix, by computing explained variance across voxels and conditions. Principal component dimensions capturing reliable variance should increase variance explained in second-half data, while dimensions capturing unreliable variation should decrease it as a result of overfitting the first-half data. Averaging across left-out participants, we found that split-half variance explained was maximized with four components.

Having identified the number of principal component dimensions capturing reliable variance in our data, we next ran PCA on our full data matrix, reduced our data to values along the first four principal component dimensions, and prewhitened the data by dividing by the standard deviation along each dimension. After prewhitening, performing ICA corresponds to finding an orthogonal basis or rotation that minimizes statistical dependence between values along each axis. We obtained this basis using an in-house algorithm that minimizes entropy along a set of orthogonal axes. This procedure provided a set of four 13-dimensional independent component (IC) vectors, corresponding to response profiles capturing maximally independent sources of variance. In addition to reporting these profiles, we assessed the spatial profiles of loadings onto each component (the contribution of each component to a voxel's response). Each voxel's response profile was modeled as a linear combination of the IC vectors, where the coefficient for each component constituted a loading value. These values were normalized to MNI space and averaged across participants to compute spatial maps of loading values for each component.

Multivariate pattern analysis

The ROI analysis revealed that fSTS responded similarly to communicative and noncommunicative face movements and vocal sounds. We next asked: would spatial patterns of response in this region discriminate communicative from noncommunicative stimuli? Multivoxel pattern analysis (MVPA) provides a more sensitive measure of whether a brain region discriminates between two stimulus conditions, indicating that this distinction is represented in the region. We focused on the fSTS for this analysis, because it responded robustly to the nonspeech face and voice conditions, while the vSTS did not.

Specifically, we used the Haxby correlation method (Haxby et al. 2001). For each participant, we first split the data into two halves, and computed patterns of response for communicative and noncommunicative stimuli (for a given modality) in each half. We constructed a 2x2 matrix of Fisher-transformed correlations between patterns from the first and second halves, and used this to compute a difference score or "discrimination index": the mean within-condition correlation minus the mean between-condition correlation (i.e., the diagonal elements minus the off-diagonal elements of this matrix). Lastly, a one-tailed *t*-test was performed across participants, to test whether the discrimination index was significantly greater than zero, indicating that patterns in this region reliably discriminated between communicative and noncommunicative conditions.

In the right and left fSTS, defined as described above, we performed seven specific comparisons, testing discrimination of communicativeness within and across modalities: 1) within face movements; 2) within face movements, generalizing from low

to high mouth movements; 3) within vocal sounds; 4) within hand movements; 5) face movements to vocal sounds; 6) face movements to hand movements; and 7) vocal sounds to hand movements. For the first three analyses, data were split across even and odd runs; for the fourth, across high and low mouth motion conditions; and for the last three, across the relevant modalities.

We next asked whether regions other than the fSTS could discriminate communicative and noncommunicative stimuli, using a whole-brain searchlight analysis. To reduce the number of comparisons, we focused on the crossmodal face-to-voice analysis; using a crossmodal comparison guarantees that decoding is not driven by low-level stimulus confounds. At each voxel in a gray matter mask, we placed an 8mm-radius sphere around the voxel, intersected this with the gray matter mask, and computed a discrimination index for this region. The mask was defined using the MNI gray matter atlas, thresholded at 0%, registered to each participant's native functional space, and intersected with their brain mask. Maps of discrimination indices for each participant were registered to MNI space, and inference was performed across participants, by performing a one-tailed *t*-test on values at each voxel. The resulting statistical maps were thresholded at $P < .01$ voxelwise, to form contiguous clusters of activation (where two voxels are considered contiguous if they share a vertex). To correct for multiple comparisons across voxels, we used a permutation test to generate a null distribution for cluster sizes, and used this to threshold clusters of activation at $P < .05$.

4.3 Results

Region-of-interest analysis

What role do face- and voice-responsive subregions of the STS play in interpreting social communicative signals? Here we ask this question by measuring fMRI responses in these regions to a range of dynamic visual and auditory social stimuli, including communicative and noncommunicative face and hand movements and vocal sounds, as well as nonword speech stimuli.

Responses in each ROI across all conditions are shown in Figure 3. We first tested the selectivity profile of fSTS and vSTS by comparing responses to faces versus objects, hands versus objects, and voices versus music in independent data. The fSTS had a strong response to face versus object movements (left: $t(48) = 6.58$, $P < 10^{-7}$; right: $t(73) = 12.07$, $P < 10^{-18}$) and vocal sounds versus music (left: $t(38) = 4.09$, $P < 10^{-3}$; right: $t(58) = 3.86$, $P < 10^{-3}$), and a small but significant response to hand versus object movements (left: $t(28) = 2.92$, $P < .01$; right: $t(43) = 4.57$, $P < 10^{-4}$). The vSTS bilaterally responded to vocal sounds over music (left: $t(42) = 2.87$, $P < .01$; right: $t(50) = 4.36$, $P < 10^{-4}$). Additionally, there was an effect of faces versus objects in the right vSTS ($t(63) = 4.28$, $P < 10^{-4}$), although this reflected a response below baseline to the object condition, and not a response above baseline to faces. These results indicate that the fSTS responds strongly to both faces and vocal sounds, while the vSTS responds specifically to vocal sounds, consistent with our prior findings (Deen *et al.* 2015).

Are STS responses to social stimuli modulated by communicative content, and does this modulation vary by modality (faces, voices, hands) and region? Although the regions differed in their overall response (main effect of ROI, $F(3,368) = 37.43$, $P < 10^{-20}$) and in their selectivity across modality (ROI by modality interaction, $F(6,368) = 4.06$, $P < 10^{-3}$), the communicativeness of the stimuli did not influence the response (main

effect and interaction terms involving this factor, all P's > .7). This result indicates that communicative content had little influence on mean responses in bilateral fSTS and vSTS.

Because this ANOVA combines data across regions and modalities, it could potentially miss a subtle effect specific to a given region and modality. To address this possibility, we next directly compared responses to communicative and noncommunicative stimuli, within each region and modality. Of these twelve tests, ten yielded null results. We did observe, however, an effect of communicativeness on left vSTS responses for vocal sounds ($t(20) = 3.50$, $P = .002$) and marginally for face movements ($t(42) = 2.56$, $P = .014$); note that the former effect would survive Bonferroni multiple comparisons correction across the twelve tests. These results largely corroborate the above ANOVA, indicating that communicative content has little to no influence on fSTS and vSTS responses, with the exception of an increased response to communicative vocal sounds in the left vSTS.

We next asked whether STS responses to face movements and vocal sounds are modulated by speech content. A region by modality by speech content ANOVA again revealed that regions differed in their overall response (main effect of region, $F(3,376) = 6.41$, $P < 10^{-3}$), and their relative response to faces and voices (region by modality interaction, $F(3,376) = 18.40$, $P < 10^{-10}$). We also observed a region- and modality-specific modulation by speech content (region by modality by speech content interaction, $F(3,368) = 4.03$, $P < .01$). Post-hoc tests revealed that these effects were driven by the presence of modality and speech effects in the vSTS bilaterally, and the absence of these effects in the fSTS. In particular, the vSTS responded more strongly to audio speech

over vocal nonspeech sounds (left: $t(31) = 11.47$, $P < 10^{-11}$; right: $t(37) = 5.05$, $P < 10^{-4}$) and to visual speech over nonspeech face movements (left: $t(53) = 8.94$, $P < 10^{-11}$; right: $t(63) = 5.49$, $P < 10^{-6}$). The vSTS additionally responded more strongly overall to vocal than to face movement stimuli (left: $t(86) = 9.07$, $P < 10^{-13}$; right: $t(102) = 7.88$, $P < 10^{-11}$). In contrast, fSTS responses were not modulated by speech content or modality, with the exception of a marginally stronger response to visual speech over nonspeech in the left fSTS ($t(48) = 2.17$, $P = .035$).

Lastly, in speech-responsive vSTS, we asked whether the response to audiovisual speech differed from the sum of responses to audio and visual speech, which would provide evidence for an audiovisual interaction. In vSTS bilaterally, we found no significant difference between these responses (left: $t(30) = .90$, $P = .37$; right: $t(36) = .74$, $P = .46$). Thus, these data do not provide evidence for super- or subadditivity in the vSTS.

To summarize, we found that fSTS responds strongly to a range of different face movements and vocal sounds, but does not respond strongly to hand movements or nonsocial audio or visual controls. This region responded similarly to various types of face movement and vocal sound, across differences in modality, communicative content, and speech content. In contrast, the response profile of vSTS indicates that this region is largely speech-selective, with a much stronger response to audio speech than to vocal nonspeech sounds and other conditions.

Independent component analysis

The ROI analysis provides a detailed characterization of the response profiles of two functional subregions, but doesn't assess responses in other parts of the STS, and requires a priori assumptions about regions relevant to processing our stimuli. We next complemented this approach with a data-driven independent component analysis, to ask more broadly, what are the dominant response profiles to dynamic social stimuli across the STS?

An initial PCA-based dimensionality reduction technique revealed that the split-half reliable sources of variance in response profiles across voxels could be captured by a 4-dimensional subspace of the 13-dimensional space of possible response vectors. This subspace captured 95.3% of the total variance across voxels. Running ICA then yielded four response profiles (or independent component vectors) spanning this subspace, with minimal statistical dependence of voxels' responses along each dimension. These response profiles, as well as spatial maps of their contributions to each voxel's response profile (termed loadings), are shown in Figure 4. Note that they are arbitrarily ordered, and named based on a post-hoc assessment of their response profile.

The first two components had straightforward modality-specific response profiles. The first component had a positive weight for all visual conditions, and roughly zero weight for auditory conditions, and thus was termed the visual component. The voxelwise loadings for this component followed a posterior-to-anterior spatial organization, with positive loadings posteriorly (adjacent to early visual cortex) and decreasing weights moving anteriorly along the STS.

The second component had positive weight for all auditory conditions, and roughly zero weight for visual conditions, and thus was termed the auditory component.

The voxelwise loadings for this component were strongest near the upper bank of the middle STS (near early auditory cortex), and decreased moving ventrally, anteriorly, and posteriorly from this region. Positive loadings for this component were somewhat more extensive in the right than in the left hemisphere, although this reflected the fact that left-hemisphere responses were better captured by the fourth component, described below, and did not reflect a lateralization of auditory responses.

The third component had positive weights on all face movement and vocal sound conditions, including communicative and noncommunicative conditions, and speech and nonspeech conditions, but had negative weights on hand movement, music, and object conditions. Much like the response profile of the fSTS ROI described above, this profile captures the discrimination between facial/vocal and other stimuli, and was thus termed the face+voice component. The voxelwise loadings for this component were strongest around the posterior STS, with positive loadings extending into middle and anterior STS, and were stronger in the right than left hemisphere.

The fourth component had strong positive weights on the audio and audiovisual speech conditions, weak positive weights on the vocal nonspeech, visual speech, and music conditions, and negative weights on the remaining face, hand, and object visual conditions. Similar to the response profile of the vSTS ROI described above, the dominant feature of this profile was audio speech selectivity, with a much stronger weight on audio/audiovisual speech than other conditions, as well as weaker effects of audio over visual stimuli and visual speech over nonspeech face motion. This component was thus termed the speech component. Similar to the auditory component, voxelwise

loadings were strongest in the upper bank of the middle STS, and decreased moving ventrally, anteriorly, and posteriorly.

In sum, a large portion of the voxelwise variance in response to the dynamic visual and auditory stimuli used in this experiment can be captured by four largely independent sources: visual responses, auditory responses, responses to facial and vocal stimuli, and responses to auditory speech. This demonstrates that the face/voice- and speech-related response profiles identified in the ROI analysis are not merely idiosyncratic properties of the focal ROIs we chose, but are dominant profiles that capture variance across the STS and emerge from a data-driven analysis.

Multivoxel pattern analysis

The ROI analysis demonstrated that the fSTS responds strongly to a range of different face movements and vocal sounds, regardless of their communicative content or social relevance. However, if the fSTS plays a role in extracting social meaning from facial and vocal stimuli, we might expect this region to discriminate between these conditions. We next asked this question using a more sensitive measure – by testing whether patterns of response in this region differed for communicative and noncommunicative stimuli, both within and across modalities (faces, voices, hands).

MVPA results are shown in Figure 5. Patterns in the fSTS were able to discriminate communicative from noncommunicative face movements (left: $t(9) = 2.83$, $P < .01$; right: $t(14) = 4.17$, $P < 10^{-3}$), even when requiring generalization across high and low mouth motion conditions (left: $t(9) = 2.64$, $P < .05$; right: $t(14) = 2.27$, $P < .05$). fSTS patterns were also able to discriminate between communicative and

noncommunicative vocal sounds (left: $t(9) = 3.33$, $P < 10^{-3}$; right: $t(14) = 2.17$, $P < .05$), and the left fSTS was able to discriminate between communicative and noncommunicative hand movements ($t(9) = 2.07$, $P < .05$).

Are common patterns of fSTS response evoked by communicative and noncommunicative stimuli from different modalities? Indeed, these patterns could discriminate communicativeness when generalizing across face movements and vocal sounds (left: $t(9) = 2.95$, $P < .01$; right: $t(14) = 2.32$, $P < .05$), but not generalizing across hand movements and face movements or vocal sounds (P's > .45). These results indicate that fSTS responses differentiate communicative and noncommunicative stimuli in a manner that is to some extent consistent across audio and visual face actions, but does not generalize to hand movements.

Does the fSTS uniquely contain patterns of response that differentiate communicative from noncommunicative face actions, or does such information exist in other brain regions as well? We addressed this question using a whole-brain searchlight analysis. To reduce the number of comparisons, we focused on crossmodal decoding of communicativeness from facial to vocal stimuli, because this comparison is impervious to low-level confounds. The results from this searchlight are shown in Figure 6. Regions with significant decoding ability were found in the left posterior STS and right posterior and middle STS, overlapping with but extending posteriorly beyond face-responsive regions. We also observed a region of left inferior frontal gyrus. These results indicate that information about the communicativeness of face movements and vocal sounds is not strictly limited to the fSTS, but circumscribed to a set of focal regions within the STS and frontal cortex.

4.4 Discussion

The present study measured STS responses to a range of visual and auditory social stimuli, in order to characterize the function of face- and voice-responsive STS subregions. We found that the fSTS responded strongly to both face movements and vocal sounds, but weakly to hand movements. This is consistent with our prior results showing strong responses to faces and voices but weak responses to full body movements (Deen *et al.* 2015), and suggests a specific role of this region in processing audio and visual signals from the face region. The fSTS had a similar mean response to a range of types of face movement and vocal sound, including communicative and noncommunicative stimuli, and speech and nonspeech stimuli, in both modalities, pointing to a broad representation of dynamic face actions. However, spatial patterns of response in this region could discriminate communicative and noncommunicative face actions, both within and across modality (faces/voices), demonstrating that this region encodes some abstract social information crossmodally. The fSTS contrasted with adjacent vSTS, which had a strikingly selective response to auditory speech, indicating that this region is largely specialized for speech perception.

Our finding that the STS region defined by a contrast of dynamic faces versus objects had an equally strong response to vocal sounds as to face movements is consistent with our prior results (Deen *et al.* 2015). While prior work has documented overlapping STS responses to faces and voices (Wright *et al.* 2003; Kreifelts *et al.* 2009; Watson, Latinus, Charest, *et al.* 2014), the present result is striking in that the fSTS was defined as the maximally face-selective subregion of posterior STS in individual participants, and

nevertheless had a strong voice response. Rather than a "face region" (Haxby et al. 2000; Pitcher *et al.* 2011), this region is thus better characterized as a "face-voice" region, responsive to dynamic auditory or visual signals from human face, but minimally to nonfacial controls, including hand, body, and object movements, as well as nonvocal music and environmental sounds (see also Deen *et al.* 2015).

The fSTS responded similarly to a range of different categories of face movement and vocal sound, including speech signals, as well as communicative and noncommunicative nonspeech signals. This result rules out the hypothesis that this region is specifically involved in processing audiovisual speech, or communicative signals more generally, instead pointing to a broad role in the perceptual processing of audiovisual face signals. The finding of a strong response to minimally socially relevant, noncommunicative actions (e.g. a neck stretch, a hiccup sound) demonstrates that this region is not limited to processing socially meaningful or significant actions, and points to a representation that is more closely tied to properties of the actions themselves than to their social interpretation.

In spite of this strong mean response to both communicative and noncommunicative actions, spatial patterns of response in the fSTS were able to discriminate these two categories. This result held both within modality for faces and voices, as well as across these two modalities (e.g., training on faces and testing on voices, or vice versa), indicating that this distinction is encoded in an abstract, crossmodal manner. This demonstrates that this region encodes an abstract social dimension, and that representations in this region are to some extent audiovisual, with facial and vocal stimuli organized around a common dimension.

What do these results imply about social perception, the process of inferring abstract social properties from perceptual input? In this inferential process, "low-level" representations are more tied to perceptual properties of the input (even if these are considered high-level from the point of view of the visual or auditory processing streams), while "high-level" representations are more tied to inferred social properties such as mental states or dispositional traits, and generalizing across modality and perceptual details. The properties of the fSTS reported here have some signatures of a low-level representation: the region responds similarly to highly and minimally socially relevant actions, and is specific to facial and vocal signals, not generalizing to socially relevant hand movements. However, other properties are more consistent with a high-level representation: this region responds to stimuli across multiple modalities (visual faces and auditory voices), and pattern analysis indicates that this region represents an abstract social property of these stimuli, in a manner that generalizes across modalities. Taken together, these results suggest that the fSTS constitutes a mid-level stage in social perceptual inference, containing a representation of audiovisual face actions that is not restricted to socially relevant inputs, but which begins to make explicit abstract, social properties across modalities.

In contrast to the broad response profile of the fSTS, the vSTS had a strikingly selective response profile, responding specifically to auditory speech stimuli over all other categories. While prior studies have argued that a similar region of the upper middle STS plays a role in processing speech sounds (Binder et al. 2000; Wright *et al.* 2003; Overath et al. 2015), or vocal sounds more generally (Belin *et al.* 2000; Shultz et al. 2012; Deen *et al.* 2015), the present results strongly suggest that this region is

primarily involved in speech processing. Particularly striking was the strong selectivity of this region for speech sounds over communicative nonspeech sounds, which were somewhat speech-like and typically involved one or multiple English phonemes. A potential explanation for this difference is that this region is sensitive to features of speech at longer timescales than individual phonemes, such as sequences of phonemes or prosodic contours (see also Overath *et al.* 2015). The vSTS also responded more strongly to visually presented speech sounds over other types of face movement, suggesting a potential role in the visual processing of speech signals as well.

Considering the response profiles of the fSTS and vSTS together, our results indicate that the STS contains distinct regions for 1) processing of facial and vocal signals in general, and 2) processing of speech signals. This contrasts with the prevailing view of distinct regions of posterior and middle STS for processing faces and vocal sounds (Belin *et al.* 2000; Haxby *et al.* 2000). This picture of the functional organization of STS was further supported by results from a data-driven ICA analysis, in which face/voice-responsive and speech-selective components emerged as dominant response profiles, contributing largely independent sources of variances in voxelwise responses across the STS.

Several limitations of the present study should be noted. First, while we used a blocked design to maximize power for comparisons across conditions, this precludes extraction of responses to specific types of face movement and vocal sound. It would also be of interest to investigate fSTS responses to specific types of face movement and vocal sound. Such an approach could address whether the strong response to the different categories of action studied here would generalize to different specific actions.

Additionally, this approach could assess what information is contained about specific actions in the spatial pattern of response of the fSTS, to ask more detailed questions about what distinctions between actions this region represents, and to more directly ask whether this region represents actions audiovisually.

Second, a limitation of our MVPA analysis is that because the communicative and noncommunicative categories comprised different actions, which differed in terms of lower-level perceptual properties, it is possible that these categories evoked distinct spatial patterns of response in virtue of differences in these lower-level properties. For the case of face movements, we find that discrimination of communicativeness generalizes across low- and high-mouth-motion stimuli, which suggests against a low-level interpretation, insofar as these categories consist of movements in different parts of the face. Furthermore, crossmodal (face-to-voice) decoding of communicativeness cannot be driven by low-level differences in stimuli across categories, insofar as the stimuli are in different modalities. Nevertheless, it should be noted that low-level differences could contribute to the within-modality MVPA effects reported here.

Lastly, although the nonspeech stimuli in our study are categorized as communicative and noncommunicative, this distinction overlaps with several other distinctions, such as social relevance and emotionality, which are difficult to dissociate. Thus, while we describe our results in terms of effects of communicativeness, they could equally well reflect another of these high-level distinctions. This point is particularly relevant for our MVPA results, where this distinction drives a difference in responses. We thus note that the relevant distinction may not be communicativeness per se, but might be better described as another related high-level social distinction. Importantly,

this does not diminish the claim that the fSTS represents an abstract social dimension crossmodally.

In sum, we find that the face-responsive region of posterior STS responds to a range of face movements and vocal sounds, while the voice-responsive region of middle STS responds selectively to speech sounds. Although the fSTS had a similar mean response to communicative and noncommunicative stimuli, spatial patterns of response in this region differentiated these stimuli across modalities (faces and voices), demonstrating that this region encodes an abstract social feature crossmodally. Future research should further detail the nature of the representations of dynamic facial and vocal signals in these regions.

4.5 References

Belin P, Zatorre RJ, Lafaille P, Ahad P, Pike B. 2000. Voice-selective areas in human auditory cortex. Nature 403:309-312.

Binder JR, Frost JA, Hammeke TA, Bellgowan PS, Springer JA, Kaufman JN, Possing ET. 2000. Human temporal lobe activation by speech and nonspeech sounds. Cereb Cortex 10:512-528.

Deen B, Koldewyn K, Kanwisher N, Saxe R. 2015. Functional organization of social perception and cognition in the superior temporal sulcus. Cereb Cortex.

DiCarlo JJ, Zoccolan D, Rust NC. 2012. How does the brain solve visual object recognition? Neuron 73:415-434.

Fecteau S, Armony JL, Joanette Y, Belin P. 2004. Is voice processing species-specific in human auditory cortex? An fMRI study. Neuroimage 23:840-848.

Greve DN, Fischl B. 2009. Accurate and robust brain image alignment using boundary-based registration. Neuroimage 48:63.

Haxby J, Gobbini M, Furey M, Ishai A, Shouten J, Pietrini P. 2001. Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science 293:2425-2430.

Haxby JV, Hoffman EA, Gobbini MI. 2000. The distributed human neural system for face perception. Trends in cognitive sciences 4:223-233.

Kreifelts B, Ethofer T, Shiozawa T, Grodd W, Wildgruber D. 2009. Cerebral representation of non-verbal emotional perception: fMRI reveals audiovisual integration area between voice-and face-sensitive regions in the superior temporal sulcus. Neuropsychologia 47:3059-3066.

Marchini JL, Ripley BD. 2000. A new statistical approach to detecting significant activation in functional MRI. Neuroimage 12:366-380.

Overath T, McDermott JH, Zarate JM, Poeppel D. 2015. The cortical analysis of speech-specific temporal structure revealed by responses to sound quilts. Nat Neurosci 18:903-911.

Pitcher D, Dilks DD, Saxe RR, Triantafyllou C, Kanwisher N. 2011. Differential selectivity for dynamic versus static information in face-selective cortical regions. Neuroimage 56:2356-2363.

Shultz S, Vouloumanos A, Pelphrey K. 2012. The superior temporal sulcus differentiates communicative and noncommunicative auditory signals. J Cogn Neurosci 24:1224-1232.

Skerry AE, Saxe R. 2014. A Common Neural Code for Perceived and Inferred Emotion. J Neurosci 34:15997-16008.

Watson R, Latinus M, Charest I, Crabbe F, Belin P. 2014. People-selectivity, audiovisual integration and heteromodality in the superior temporal sulcus. Cortex 50:125-136.

Woolrich MW, Ripley BD, Brady M, Smith SM. 2001. Temporal autocorrelation in univariate linear modeling of FMRI data. Neuroimage 14:1370-1386.

Wright TM, Pelphrey KA, Allison T, McKeown MJ, McCarthy G. 2003. Polysensory interactions along lateral temporal regions evoked by audiovisual speech. Cereb Cortex 13:1034-1043.

4.6 Figures

Figure 1: Schematic visualization of the thirteen conditions included in the fMRI experiment.

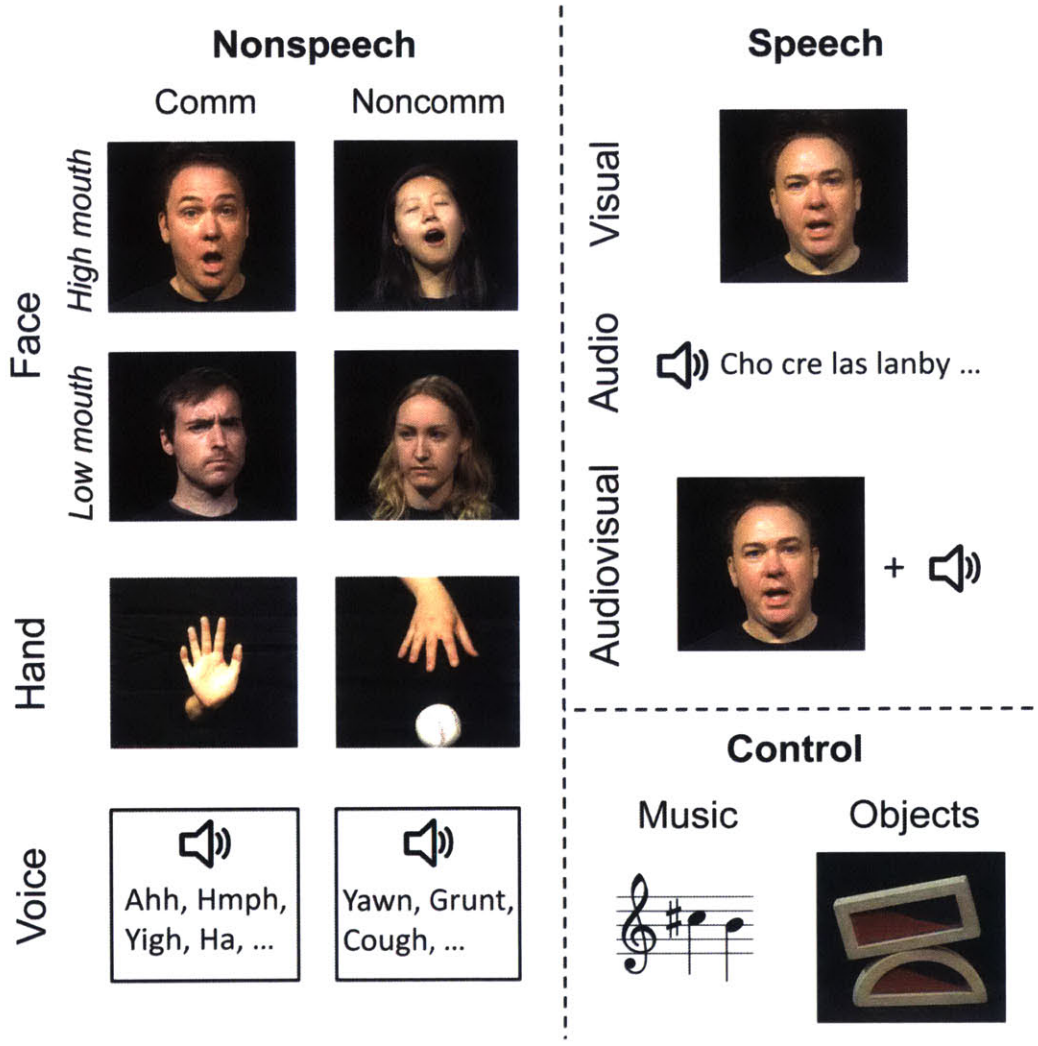


Figure 2: Behavioral ratings of communicativeness, across the 80 specific actions used in the study, categorized by condition. Condition labels: FC = communicative face movement, FN = noncommunicative face movement, HM = high mouth motion, LM = low mouth motion, HC = communicative hand movement, HN = noncommunicative hand movement, VC = communicative vocal sound, VN = noncommunicative vocal sound.

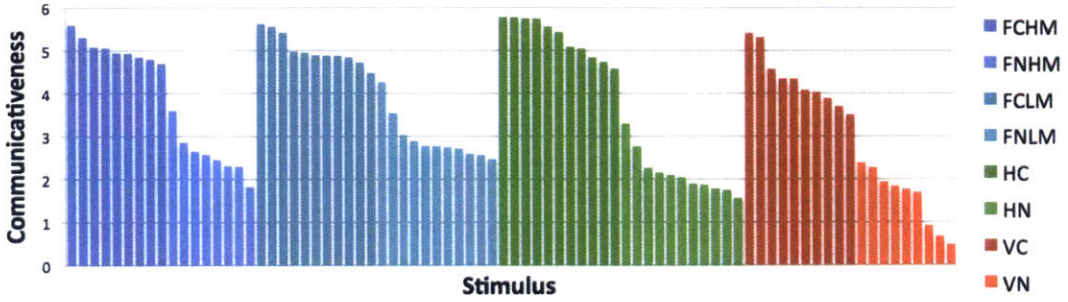


Figure 3: Response profiles of STS subregions. Regions were defined using a face > object contrast (fSTS) and a voice > music contrast (vSTS). Images on the left show heat maps of region-of-interest locations across participants. Images on the right show responses of these regions (in percent signal change) across the thirteen experimental conditions, extracted from data independent from those used to define the regions. Condition labels: FC = communicative face movement, FN = noncommunicative face movement, HM = high mouth motion, LM = low mouth motion, HC = communicative hand movement, HN = noncommunicative hand movement, VC = communicative vocal sound, VN = noncommunicative vocal sound, SA = audio speech, SV = visual speech, SAV = audiovisual speech, M = music, O = objects.

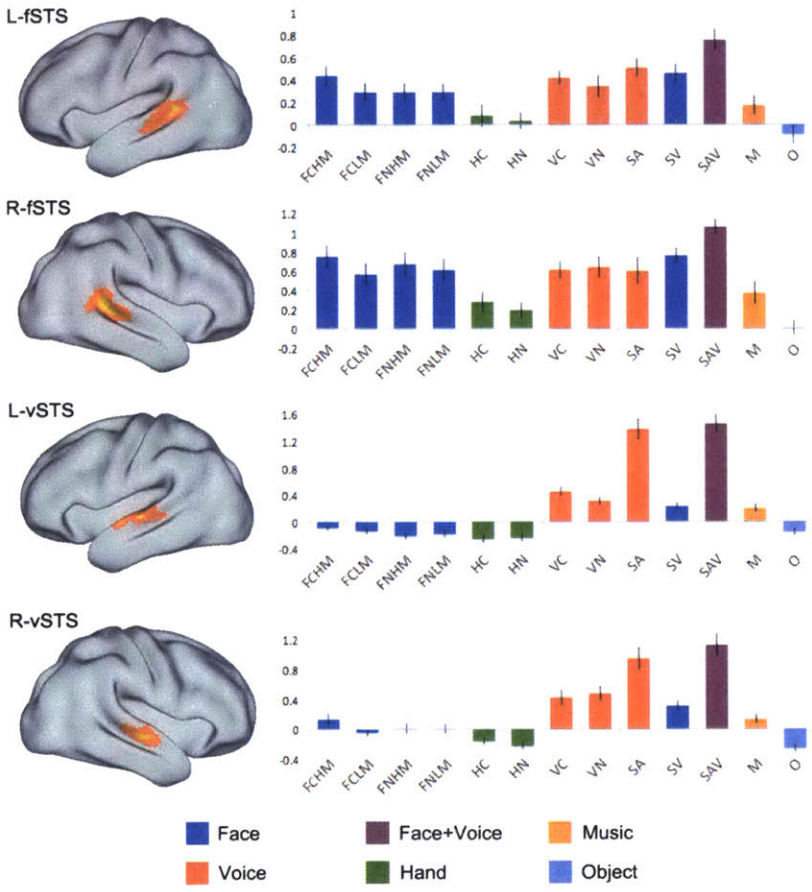


Figure 4: Independent component analysis results. Images on the right show response profiles for four independent component vectors, which together explained ~95% of voxelwise variance in STS responses. Images on the left show spatial patterns of voxelwise loading values, or the contribution of each component to a given voxel's response profile. Components are ordered arbitrarily and named based on post-hoc assessment of their response profiles. Condition labels: FC = communicative face movement, FN = noncommunicative face movement, HM = high mouth motion, LM = low mouth motion, HC = communicative hand movement, HN = noncommunicative hand movement, VC = communicative vocal sound, VN = noncommunicative vocal sound, SA = audio speech, SV = visual speech, SAV = audiovisual speech, M = music, O = objects.

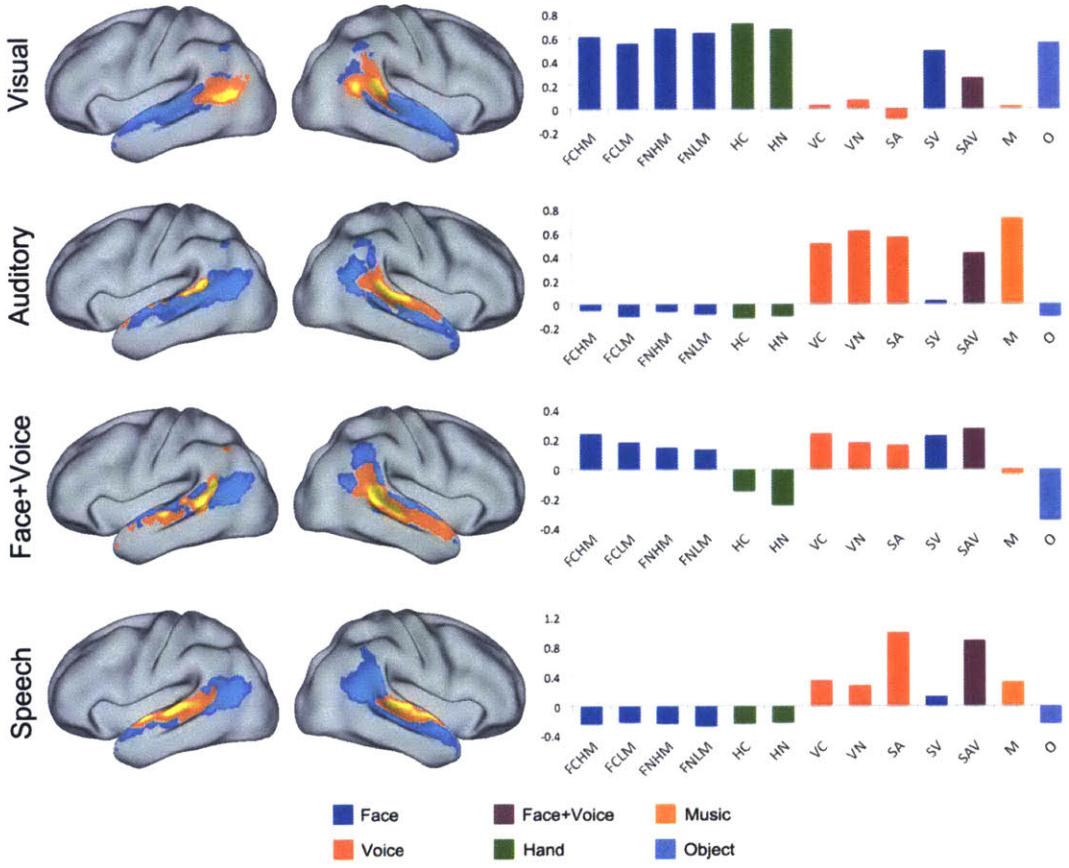


Figure 5: MVPA results: decoding communicativeness from fSTS patterns, both within and across modalities. Discrimination indices (correlation difference scores) for comparing patterns of response to communicative and noncommunicative stimuli. Within modality effects for faces (F); faces, generalizing from high to low mouth motion (F*); voices (V); and hands (H). Crossmodal effects for faces to voices (F-V), faces to hands (F-H), and voices to hands (V-H).

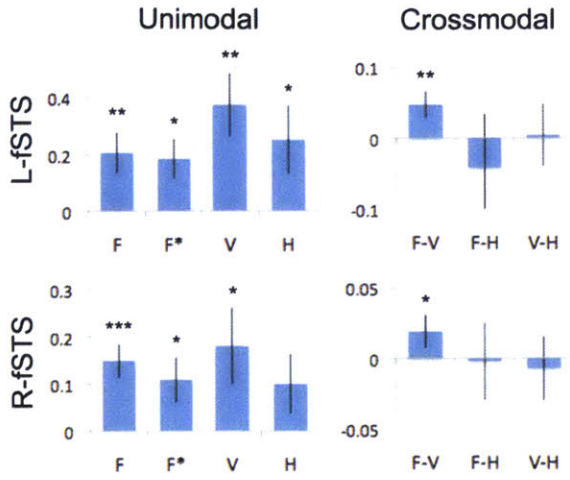
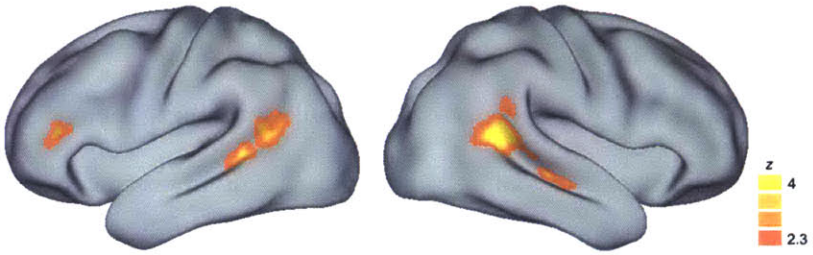


Figure 6: Searchlight MVPA analysis for decoding communicativeness across modality (faces to voices). Whole-brain statistical map thresholded at $P < .01$ voxelwise, followed by a $P < .05$ clusterwise threshold to correct for multiple comparisons.



Chapter 5: Category-sensitive visual regions in human infants[4]

The adult human brain contains numerous functionally specialized regions, dedicated to processing specific types of information (Kanwisher 2010). In visual cortex, category-sensitive regions have been observed, which respond differentially to specific categories of visual input, such as faces, bodies, and scenes (Kanwisher et al. 1997; Epstein and Kanwisher 1998; Downing et al. 2001). A central question about these regions is how they develop: are category-sensitive regions largely innately specified and present from early in development, or do they develop gradually, after extensive visual experience with specific categories (Johnson 2001)? Here we use fMRI in awake infants to demonstrate that category-sensitive regions exist by 4-6 months of age. Comparing responses to faces and scenes in infants, we observe a number of category-sensitive regions in occipitotemporal cortex, consistent with anatomical locations of these regions in adults. However, strongly selective regions, preferring one visual category over all others, were not observed in infants, in contrast to adults. These results demonstrate that while the large-scale functional organization of category preferences in visual cortex is present within a few months after birth, strongly category-selective regions don't emerge until later in development.

[4] The contents of this chapter are in preparation for submission as Deen B, Richardson H, Dilks DD, Takahashi A, Keil B, Wald L, Kanwisher N, Saxe R. Category-sensitive visual regions in human infants.

5.1 Results and Discussion

Many questions about the development of human visual cortex remain open, due to the difficulty of neuroimaging awake infants. While near-infrared spectroscopy can be used to mitigate this difficulty (Grossmann et al. 2008; Lloyd-Fox et al. 2009), this method is limited to superficial regions and has very low spatial resolution, and is thus difficult to relate to the large body of neuroimaging research on adults. Here, we implement novel methods for awake infant fMRI in order to study the early development of high-level visual regions. We employ a number of technical advances to increase participant comfort, optimize signal strength, and minimize head motion artifacts: 1) infant-sized MR head coils; 2) quiet pulse sequences; 3) infant-directed visual stimuli; and 4) a combination of extant and novel data analysis techniques for minimizing motion artifacts.

We obtained low-motion fMRI data from 9 infants (Supplementary Table 1) while viewing engaging, brightly color, infant-friendly videos of dynamic faces, natural scenes, scrambled scenes, human bodies, and objects (Supplementary Figure 1). We first compared responses to faces and scenes, because in adults this comparison yields the most robust effects and delineates a large-scale spatial organization of nonprimary visual cortex (Downing et al. 2006; Nasr et al. 2011). Category-sensitive regions in occipitotemporal cortex were observed in 8 of 9 infants, or 11 of 12 datasets, separated by age (Figure 1, Supplementary Figure 2). In individual infants, face-preferring regions were observed in the fusiform gyrus, lateral occipital cortex, superior temporal sulcus (STS), and medial prefrontal cortex; scene-preferring regions were observed in the parahippocampal gyrus and lateral occipital cortex (see Extended Data Fig. 3 for sample

adult responses). Many of these regions showed reliable responses in a group analysis, combining infants' datasets using a pediatric template (Fig. 1). These results demonstrate that the large-scale functional organization of category preferences in extrastriate visual cortex is similar in 4-6 month old infants and in adults.

Responses in individual infants varied substantially in strength. This variation appeared to be driven largely by variation in the amount of data collected, which ranged from 5 to 50 minutes per dataset. The infants with the most data, shown in Fig. 1, showed extensive category-sensitive responses throughout occipitotemporal cortex. In a multiple regression analysis on activation extent (% of voxels active), effects of both amount of data ($t(56) = 3.48$, $P < 10^{-3}$) and age ($t(56)=2.64$, $P < .05$) were observed (Extended Data Fig. 4). Thus, while observations in individual infants are mainly constrained by data quantity, there is a hint of developmental change in this sample. More or larger regions of cortex may show category-sensitive responses over infants' first year of life.

To assess the magnitude of responses to faces and scenes in category-sensitive regions, and to corroborate whole-brain results with a more sensitive measure, we performed region-of-interest (ROI) analyses. ROIs were defined as the top 5% of voxels with the strongest statistical response to faces over scenes or the reverse, within broad anatomical regions (ventral temporal cortex, lateral occipitotemporal cortex, and STS; Supplementary Figure 5). Responses were extracted from data independent of those used to define ROIs, using leave-one-out cross-validation. ROI locations and responses are shown in Figure 2. Locations of face- and scene-preferring ROIs were highly consistent across infants and adults. All nine infants saw movies of faces, scenes and scrambled scenes (Expt 1); in this experiment, all regions tested showed reliable preferences for

faces or scenes (n = 9; ventral face region, z = 2.85, P < .01; lateral face region, z = 3.27, P < .01; STS face region, z = 4.74, $P < 10^{-5}$; ventral scene region, z = 6.41, $P < 10^{-9}$; lateral scene region, z = 3.43, $P < 10^{-3}$). In six infants, consistent preferences for these categories were independently replicated in independent experiments, using a distinct set of movies, in all of the ROIs (Expts 2-6; ventral face region, z = 2.17, P < .05; lateral face region, z = 2.45, P < .05; STS face region, z = 4.28, $P < 10^{-4}$; ventral scene region, z < 5.35, $P < 10^{-7}$; lateral scene region, z = 5.27, $P < 10^{-6}$). ROI-based category preferences generalized across infants (Extended Data Fig. 6), and as expected, were highly significant in adults (permutation test, n = 3; all P's < 10^{-15}). These results corroborate the whole-brain analyses, providing replicable evidence for face- and scene-preferring visual regions in infants.

In adults, within spatially broad category-sensitive regions of visual cortex, there also exist more spatially focal regions that are highly selective to specific categories, responding substantially more strongly to their preferred category than to any other (Spiridon et al. 2006). Do these focal, strongly selective regions exist in infants? We searched for selective regions by comparing responses to faces and scenes with responses to objects, which in adults yielded spatially focal responses (Supplementary Figure 3). ROIs were again defined as the top 5% of voxels within broad anatomical regions. In adults, these regions showed the expected highly selective response profiles: all regions tested robustly preferred faces or scenes to objects (Figure 3; permutation test, n = 3; ventral face region, z = 5.54, $P < 10^{-7}$; lateral face region, z = 4.35, $P < 10^{-4}$; STS face region, z = 7.13, $P < 10^{-11}$; ventral scene region, z = 6.10, $P < 10^{-8}$; lateral scene region, z = 5.12, $P < 10^{-6}$). In infants, however, no region showed a significant effect of faces or

scenes over objects (permutation test, $n = 6$; ventral face region, $z = -.75$, $P = .45$; lateral face region, $z = .91$, $P = .36$; STS face region, $z = 1.40$, $P = .16$; ventral scene region, $z = -1.36$, $P = .17$; lateral scene region, $z = .81$, $P = .42$). To address the possibility that these results might depend on the specific ROI size used, we next defined ROIs as the top N% of voxels in each anatomical region, varying N from 2% to 30%. In adults, significant effects were observed for all regions and ROI sizes (permutation test, $n = 3$; all P's $< .05$). In contrast, in infants, significant effects were not observed for any region or ROI size (permutation test, $n = 6$; all P's $> .14$). This suggests that regions strongly selective for specific visual categories may not exist in infants. Instead, these regions emerge later in development, possibly requiring more extensive visual experience to tune their selectivity.

Do the category effects reported here reflect truly abstract category preferences, or could they be driven by lower-level visual features that covary with category identity? In particular, high- and low-frequency content and rectilinearity (the presence of 90° angles) have been shown to vary systematically across the categories used here, and to modulate responses in some category-sensitive regions (Rajimehr et al. 2011; Nasr et al. 2014; Yue et al. 2014). This raises the possibility that in development, cortical responses biased by lower-level visual features exist early, while abstract category preferences develop later. Consistent with prior results, we find that these features differ across our stimuli: e.g., faces have strong low-frequency content, while scenes have strong high-frequency content and rectilinearity (Supplementary Figure 7). Are ROI-based responses in infants are better predicted by visual category, or by low-level features? In face-preferring regions, the categorical model did not provide a significantly better fit than a

model based on rectilinearity (Supplementary Figure 7). Because category identity and visual feature values covary in these experiments, this result does not demonstrate that responses in these regions are driven by rectilinearity as opposed to category, but indicates that the current paradigm and data cannot distinguish between these possibilities; more targeted experimental manipulations will be needed to address this question. In scene-preferring regions, however, the category model provided a significantly better fit than models based on these lower-level visual features. This demonstrates that for scene-preferring regions in infants, the category effects observed cannot be explained by responses to high- or low-frequency content or rectilinearity, suggesting against the hypothesis that responses in these regions are primarily driven by lower-level visual features.

In sum, we have shown that infants as early as 4-6 months have category-sensitive visual regions with a similar spatial organization to that observed in adults, but that infants do not have the more focal, highly category-selective regions observed in adults. These results suggest a developmental story in which the large-scale functional organization of category preferences in visual cortex exists early in development, and provides the scaffolding for subsequent refinement of responses in these regions, leading ultimately to the strongly category-selective regions. This conclusion points to a myriad of future questions, including: what is the time course of the development of category-selective visual regions? To what extent does the development of these regions reflect an experience-dependent versus a maturational process? And does a similar principle (an initial preference that is subsequently refined) apply to the development functionally

specific regions in other perceptual and cognitive domains? We hope that the methods introduced here will aid in subsequent investigation of these questions.

5.2 Methods

Participants

We scanned 17 infants (age 2.3-8.6mo, 3 female) and acquired useable, low-motion data from 9 infants (age 3.0-8.0mo, 1 female). We also scanned 3 adults (age 27-34, 1 female) for comparison (Supplementary Table 1). Because low-motion data from infants was relatively rare, whenever possible we scanned infants in multiple sessions (between 1 and 16 scans per infant, for a total of 63 scan sessions). Five infants were scanned at multiple ages, separated by one or more months; datasets from each age were analyzed separately. Adult participants and parents of infant participants provided written, informed consent, in accordance with the Committee on the Use of Humans as Experimental Subjects at MIT.

Paradigm

Stimuli were infant-friendly dynamic video clips depicting faces, objects, bodies, and scenes (Supplementary Figure 1). Participants typically viewed Experiment 1 (Expt 1), a two-condition (face, scene) version, with grid-scrambled scenes included as a baseline. When time permitted, we additionally ran Experiment 2 (Expt 2), a four-condition (face, object, body, scene) version, with distinct face and scene videos, a scrambled scene baseline, and both scene and scrambled scene videos presented at 80% size, to minimize the possibility of a retinotopic confound in the scene vs face comparison. In certain cases, further experiments using different videos of the same

categories were used, to further test generalization of responses across specific videos; these experiments, as well as more detail on stimuli, are further described in the supplement. Stimuli were presented in 18s-long blocks comprising six video clips. Baseline blocks occurred every seven blocks (Expt 1) or five blocks (Expt 2); experimental blocks were ordered pseudorandomly between baseline blocks. During infant functional scans, an experimenter or parent lied in the scanner bore to monitor the infant, and told the experimenters if the infant closed his or her eyes, fell asleep, or fussed out. For infants, individual runs were not fixed in duration, but instead were ended whenever the infant fussed out or fell asleep. For adults, runs lasted 22 blocks (Expt 1) or 21 blocks (Expt 2), with a baseline block at the start and end of each run. Adults received 5 runs each of Expt 1 and Expt 2.

Data acquisition

MRI data were acquired using a Siemens 3T MAGNETOM Tim Trio scanner (Siemens AG, Healthcare, Erlangen, Germany). We used a standard 32-channel head coil for adult participants, and a custom-built infant-sized 32-channel head coil for infants (Keil et al. 2011). The latter was shaped like a reclined car seat to increase comfort, and had coil elements close to the infant's head, to reduce head motion and increase signal-to-noise ratio. For infants whose head did not fit in this coil, a head coil designed for 5-year-olds was used instead. In order to further increase infant comfort, we acquired data using a quiet (70-72dB SPL) T2*-weighted pulse sequence (Zapp et al. 2012), sensitive to blood-oxygen-level-dependent (BOLD) contrast (repetition time [TR] = 3s, echo time [TE]=43ms, $\alpha = 90°$, field of view [FOV] = 192mm, matrix = 64x64, slice thickness =

3mm, slice gap = .6mm). For infants, we used 18-24 near-axial slices, using the minimum number of slices required to cover occipitotemporal cortex for a given head size, because pulse sequence volume scaled with number of slices; for adults, we used 36 near-axial slices for whole-brain coverage.

Anatomical images were only collected in certain cases, because our focus was normally to collect as much awake functional data as possible, and because collecting a high quality anatomical typically required the infant to be asleep to reduce motion. When anatomicals were collected, we used one of three T1-weighted pulse sequences of varying length, using briefer, lower-quality sequences when an infant would only stay still for a briefer duration. These included a 24s sequence (TR = 283ms, TE = 1.8ms, flip angle $\alpha = 9°$, FOV = 159mm, matrix = 106x106, slice thickness = 1.5mm, 96 sagittal slices), a 2.2-min sequence (TR = 800ms, TE = 3.43ms, flip angle $\alpha = 9°$, FOV = 160mm, matrix = 160x160, slice thickness = 1mm, 144 sagittal slices), and a 6.5-min sequence (TR = 2530ms, TE = 1.64ms, flip angle $\alpha = 7°$, FOV = 256mm, matrix = 256x256, slice thickness = 1mm, 176 sagittal slices, acceleration factor = 2, 24 reference lines). In adults anatomicals were acquired using the 6-min sequence.

Data selection

Data were processed primarily using custom scripts, with tools from the FMRIB Software Library (FSL) version 4.1.8 and Freesurfer additionally used for registration and motion correction. Because some of our infant data contained a substantial amount of head movement, and because head motion causes highly deleterious artifacts in fMRI data (Friston et al. 1996), we first aimed to discard high-motion data that could corrupt

our results and lead to false negatives. Each run was first motion corrected by registering each volume to the middle volume, using rigid transformations determined by FSL's MCFLIRT. Using the motion parameters estimated by this correction, we applied a technique known as scrubbing (Power et al. 2014), removing pairs of adjacent volumes with more than .5mm of total translation or .5° of total rotation between them. We also removed volumes where the participant's eyes were closed, and the first three volumes of each run (to allow the MR system to equilibrate).

While this technique is effective in removing artifactual spikes of response that occur at high-motion time points, it can still leave large baseline shifts in voxels' time courses that occur when a participant's head moves substantially and remains in a new location relative to the head coil and external magnetic field. We thus instituted a second cutoff on scrubbed data, at pairs of adjacent volumes with more than 2mm of total translation or 2° of total rotation between them. At these cutoff points, we temporally split runs to form "pseudoruns" of scrubbed data where the head was in a relatively consistent position. These pseudoruns were subsequently analyzed as one would normally analyze a full run. Pseudoruns were kept for analysis if they contained at least 24 time points, as well as six time points per condition for all conditions (where condition timing was lagged by 6s to account for hemodynamic delay), such that responses to each condition could be estimated. Lastly, participants were included in analyses if they had at least five minutes of data saved after this procedure, across experiments.

Supplementary Table 1 shows the amount of data acquired and saved, across participants. We initially acquired 23.06 hours of data across 17 infants, and were left with 4.26 hours of data across 9 infants after motion screening. Resulting pseudoruns in

infants ranged in length from 1.2-17.5 minutes (mean 4.3). While this procedure led to a substantial reduction in data quantity, it drastically reduced the amount of head motion present in the resulting data, reducing mean volume-to-volume translation from 1.11mm to .13mm, and mean rotation from 1.69° to .17°. In adults, no volumes were removed either by scrubbing or pseudorun selection, such that pseudoruns were equivalent to the original runs. Adult data had mean volume-to-volume translation of .04mm, and mean rotation of .02°.

Data preprocessing

The resulting pseudoruns were first motion-corrected by registering each volume to the middle volume, using rigid transformations determined by FSL's MCFLIRT. Data were skull-stripped using FSL's Brain Extraction Tool, and spatially smoothed using a 3mm-FWHM Gaussian kernel.

Data registration

In order to combine data across pseudoruns, middle volumes from each pseudorun for a given participant were all registered to a specific "target" middle volume, chosen to have minimal distortion. All registration was performed using FSL's FLIRT, unless otherwise noted. In infant data, head motion across pseudoruns posed challenges for this registration: different volumes could have different positions within the bounding box, and different types of nonrigid distortion. To optimize registration, we thus adopted the following procedure: 1) middle volumes were algorithmically registered to target volumes using both a rigid transformation and a general affine transformation; 2)

translation and rotation parameters for both of these transformations were hand-tuned to improve registration quality; 3) we selected whichever resulting transformation (hand-tuned rigid or hand-tuned affine) provided a more accurate registration based on visual inspection of anatomical landmarks. For adult data, middle volumes were registered to the target using a rigid transformation.

For infant data, in cases where anatomicals were collected, target functional volumes were registered to anatomical images using a rigid transformation, with translation and rotation parameters subsequently hand-tuned. For adult data, because surface reconstructions could be obtained, target functionals were registered to anatomicals with a rigid transformation determined by Freesurfer's bbregister. Anatomical images in adults were in turn registered to the MNI template brain using a nonlinear transformation determined by FSL's FNIRT.

Lastly, we aimed to register data across infants, for the purposes of registering search spaces for region-of-interest analyses (described below), and to compute group-level whole brain statistical maps. To this end, target functional volumes from each infant were registered to the target functional of Infant1, dataset 3 (the infant and dataset with the most useable data) using an affine transformation, with translation and rotation parameters subsequently hand-tuned. While these transformations were not perfect, insofar as linear registration cannot perfectly align different brains, they were primarily used for the registration of large search spaces, which should be tolerant to minor inaccuracies in registration. Lastly, to transform search spaces across infants and adults, this target functional was registered to the MNI brain using an affine transformation, with translation and rotation parameters subsequently hand-tuned.

Data modeling

For each pseudorun, whole-brain voxelwise linear models were performed to estimate the BOLD response to visual stimuli. Regressors for each condition (excluding the baseline) were defined as boxcar functions with value 1 during blocks of that condition, convolved with a canonical double-gamma hemodynamic response function. Twelve nuisance regressors were additionally included to reduce the influence of potential artifacts. A linear trend regressor was included to account for signal drift. Motion parameter regressors (3 translation parameters and 3 rotation parameters determined by motion correction) were used to minimize effects of head motion. Lastly, 5 principal component analysis (PCA)-based noise regressors were used to account for other noise sources (a method similar to GLMDenoise (Kay et al. 2013)). PCA-based regressors were defined by: 1) choosing a "noise pool" of voxels with <1% of variance explained by the task regressors; 2) running PCA on time series from these voxels; and 3) choosing the top 5 principal components as regressors. For both task and nuisance regressors, time points that were scrubbed in data selection were removed after the regressors were defined (with the exception of PCA-based regressors, which were defined using scrubbed data).

This analysis provided beta values for task regressors corresponding to the magnitude of response to each condition, and contrast values corresponding differences across conditions. To combine the resulting contrast values across pseudoruns for a given participant and dataset, we computed a weighted average of contrast maps registered to a common functional space, using weights corresponding to the amount of

data contributed by each pseudorun. Weights were proportional to $(c^T(X^TX)^{-1}c)^{-1}$, where c is the contrast vector, and X is the design matrix for a given pseudorun. For a given contrast (e.g., faces vs scenes), we combined data across all experiments containing that contrast.

We next statistically assessed these average contrast values for each participant. Because fMRI time series are temporally autocorrelated, within-participant statistics are typically computed using feasible generalized least squares, with an empirical estimate of the autocorrelation structure. However, the validity of extant methods for estimating the autocorrelation of fMRI data is not well established (Eklund et al. 2012), and these methods have not been validated in infant data. To obviate the need for any assumptions about the autocorrelation structure in our data, we instead used a nonparametric permutation test (Nichols and Holmes 2002). Specifically, on each of 5,000 iterations, we randomly permuted the block order for each pseudorun, and computed a contrast value for each voxel. This provided a null distribution that was used to threshold voxelwise contrast values at $P < .01$. Estimated null distributions were fit with a Gaussian distribution, allowing us to estimate small p-values that wouldn't be possible to estimate from the fraction of samples from the null distribution exceeding the observed statistic; for statistics with larger p-values, the Gaussian fit gave very similar p-values to those computed using the raw null distribution. For visualization and reporting purposes, voxelwise statistics were converted to z-values based on their computed p-value. To correct for multiple comparisons across voxels, we additionally used a permutation test to build a null distribution for sizes of contiguous clusters of activation, and thresholded cluster sizes at $P < .05$.

We additionally computed a group-level statistical map to perform inference across infants. Average contrast maps for each infant were registered to the target functional space of Infant1, dataset 3, and voxelwise *t*-tests were performed across infants, comparing contrast values to zero, thresholded at $P < .01$. For infants with multiple datasets acquired at different ages, we only used the dataset with the largest amount of saved data. As above, voxelwise *t*-statistics were converted to *z*-values based on their computed *p*-value for visualization purposes. To correct for multiple comparisons across voxels, a permutation test was used to build a null distribution for sizes of contiguous clusters of activation (where on each iteration, signs of contrast values for each infant were randomly flipped), and thresholded cluster sizes at $P < .05$.

Lastly, we asked which factors contributed to differences in the extent of activations across datasets. In particular, we hoped to demonstrate that datasets with larger amounts of data yielded more extensive activations, and to ask whether there was an additional effect of age. Statistics were performed across pseudoruns rather than infants, to increase power for correlational analyses. For each pseudorun, activation extent was quantified as the proportion of voxels across the whole brain reaching $P < .01$ significance using ordinary-least-squares statistics on time series (permutation-based statistics were not used because individual pseudoruns typically did not contain enough data to yield significant results with this more stringent method). We then performed a multiple regression analysis on activation extent, with amount of data (number of volumes per subrun) and age as independent variables.

Region-of-interest analysis

To assess response profiles of brain regions identified in the whole-brain analysis, we performed region-of-interest (ROI) analyses. In order to maximize the amount of data used to define regions, but still extract responses from data independent of those used to define the ROI (Vul et al. 2009), we used a leave-one-pseudorun-out analysis: ROIs were defined using data from all but one pseudorun, responses were extracted from the remaining pseudorun, and after iterating this process across all pseudoruns and participants, the resulting beta values and contrasts were combined using the weighted average described above (Data modeling). Beta values and contrasts were converted to percent-signal-change (PSC) values by dividing by mean signal strength within the ROI. These values were statistically assessed using a permutation test, analogous to the procedure described above (Data modeling). For a given iteration, ROIs were defined as the set of voxels within a broad anatomical search space with the top *N%* of statistical values for a specific contrast, such as faces vs scenes or faces vs objects. The value *N* was either chosen to be 5% to assess responses in a maximally selective region, or varied from 2-30% to measure selectivity as a function of ROI size. Search spaces were hand-drawn on the anatomical image of a specific participant (Infant1, dataset 3), and registered to other participants' functional images as described above (Data registration). They included: 1) lateral occipitotemporal cortex, covering the expected locations of the occipital face area and occipital place area; 2) ventral temporal cortex, covering the expected locations of the fusiform face area and parahippocampal place area; and 3) superior temporal sulcus (STS), covering the expected location of the posterior STS face region (Supplementary Figure 5).

Because the above statistical analyses combine data from different infants without explicitly testing whether effects generalize infants, we next tested generalization by performing statistics across infants. We focused on responses to faces and scenes, because this data was acquired in all infants, and combined data across two- and four-category experiments (Expts 1-6) to increase power within each participant. For each ROI (defined as described above, using the face vs scene contrast), mean PSC values were computed for each participant and averaged across participants, and the difference between responses to faces and scenes was statistically compared to zero using a *t*-test across infants. As with the whole-brain group-level analysis, when infants yielded multiple datasets acquired at different ages, we only used the dataset with the largest amount of saved data.

Visual feature analysis

We next asked whether responses in category-sensitive visual regions could be explained in terms of lower-level visual features. In particular, we focused on high- and low-frequency content and rectilinearity (the presence of 90° angles in an image), which have been argued previously to modulate responses in category-sensitive visual regions (Rajimehr *et al.* 2011; Nasr *et al.* 2014; Yue *et al.* 2014). Frequency content and rectilinearity measures were computed on individual frames from each video clip, and averaged across frames for a given clip. Frames were first converted to grayscale and normalized to have zero mean and unit standard deviation, to remove effects of overall luminance and contrast. We then computed the discrete Fourier transform of each frame, and defined low-frequency content as total power at frequencies less than 1 cycle per

degree of visual angle, and high-frequency content as total power at frequencies greater than 5 cycles per degree of visual angle, following the cutoffs used by Razimehr et al. (2011). Rectilinearity was computed using a procedure described by Nasr et al. (2014): frames were convolved with a bank of 90° angle Gabor filters at different scales and orientations, and magnitudes of convolved images were averaged across spatial position and filter to yield a single measure (see Supplementary Figure 7).

We then assessed whether responses in ROIs defined using face vs scene and scene vs face contrasts were better predicted by category identity or by visual features. Regressors for visual features were defined by constructing time series of feature values for each individual video in a given pseudorun, convolved with a canonical double-gamma hemodynamic response function. Categorical regressors were defined as described above (Data modeling). We compared five models: category (containing regressors for each visual category in an experiment), low-frequency content, high-frequency content, rectilinearity, and a model containing low-frequency, high-frequency, and rectilinearity regressors. To eliminate the possibility that differences in model fit resulted from different degrees of freedom across models, model fit was assessed using leave-one-pseudorun-out cross-validation. For a given pseudorun, models were fit using data from all other pseudoruns with the same set of conditions from that participant and dataset (ROIs were also defined using data independent from the left-out pseudorun, as described in the Region-of-interest analysis section above). This provided a set of beta values that was used to define a predicted response for the left-out pseudorun, for each model. Model fit was assessed by computing the Fisher-transformed correlation (z'-value) between the time series in the left-out pseudorun and the predicted response.

Linear trend and motion parameter nuisance regressors were included in all models. Model fit estimates were compared across models using paired *t*-tests across pseudoruns.

5.3 References

Downing P, Chan A-Y, Peelen M, Dodds C, Kanwisher N. 2006. Domain specificity in visual cortex. Cereb Cortex 16:1453-1461.

Downing PE, Jiang Y, Shuman M, Kanwisher N. 2001. A cortical area selective for visual processing of the human body. Science 293:2470-2473.

Eklund A, Andersson M, Josephson C, Johannesson M, Knutsson H. 2012. Does parametric fMRI analysis with SPM yield valid results?,ÄîAn empirical study of 1484 rest datasets. Neuroimage 61:565-578.

Epstein R, Kanwisher N. 1998. A cortical representation of the local visual environment. Nature 392:598-601.

Friston KJ, Williams S, Howard R, Frackowiak RS, Turner R. 1996. Movement-related effects in fMRI time-series. Magn Reson Med 35:346-355.

Grossmann T, Johnson MH, Lloyd-Fox S, Blasi A, Deligianni F, Elwell C, Csibra G. 2008. Early cortical specialization for face-to-face communication in human infants. Proceedings of the Royal Society B: Biological Sciences 275:2803-2811.

Johnson MH. 2001. Functional brain development in humans. Nature reviews neuroscience 2:475-483.

Kanwisher N. 2010. Functional specificity in the human brain: a window into the functional architecture of the mind. Proceedings of the National Academy of Sciences 107:11163-11170.

Kanwisher N, McDermott J, Chun MM. 1997. The fusiform face area: a module in human extrastriate cortex specialized for face perception. J Neurosci 17:4302-4311.

Kay KN, Rokem A, Winawer J, Dougherty RF, Wandell BA. 2013. GLMdenoise: a fast, automated technique for denoising task-based fMRI data. Frontiers in neuroscience 7.

Keil B, Alagappan V, Mareyam A, McNab JA, Fujimoto K, Tountcheva V, Triantafyllou C, Dilks DD, Kanwisher N, Lin W. 2011. Size-optimized 32-channel brain arrays for 3 T pediatric imaging. Magn Reson Med 66:1777-1787.

Lloyd-Fox S, Blasi A, Volein A, Everdell N, Elwell CE, Johnson MH. 2009. Social perception in infancy -- a near infrared spectroscopy study. Child Dev 80:986-999.

Nasr S, Echavarria CE, Tootell RB. 2014. Thinking outside the box: rectilinear shapes selectively activate scene-selective cortex. J Neurosci 34:6721-6735.

Nasr S, Liu N, Devaney KJ, Yue X, Rajimehr R, Ungerleider LG, Tootell RB. 2011. Scene-selective cortical regions in human and nonhuman primates. J Neurosci 31:13771-13785.

Nichols TE, Holmes AP. 2002. Nonparametric permutation tests for functional neuroimaging: a primer with examples. Hum Brain Mapp 15:1-25.

Power JD, Mitra A, Laumann TO, Snyder AZ, Schlaggar BL, Petersen SE. 2014. Methods to detect, characterize, and remove motion artifact in resting state fMRI. Neuroimage 84:320-341.

Rajimehr R, Devaney KJ, Bilenko NY, Young JC, Tootell RB. 2011. The "parahippocampal place area" responds preferentially to high spatial frequencies in humans and monkeys.

Spiridon M, Fischl B, Kanwisher N. 2006. Location and spatial profile of category-specific regions in human extrastriate cortex. Hum Brain Mapp 27:77-89.

Vul E, Harris C, Winkielman P, Pashler H. 2009. Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. Perspectives on psychological science 4:274-290.

Yue X, Pourladian IS, Tootell RB, Ungerleider LG. 2014. Curvature-processing network in macaque visual cortex. Proceedings of the National Academy of Sciences 111:E3467-E3475.

Zapp J, Schmitter S, Schad LR. 2012. Sinusoidal echo,Äêplanar imaging with parallel acquisition technique for reduced acoustic noise in auditory fMRI. J Magn Reson Imaging 36:581-588.

5.4 Figures

Figure 1: Whole-brain activation maps, comparing faces to scenes. The top two rows of images show results from the two infants with the largest amount of usable data, while the third shows a group map with statistics across infants. Maps are thresholded at $P < .01$ voxelwise, and corrected for multiple comparisons using a clusterwise threshold of $P < .05$.

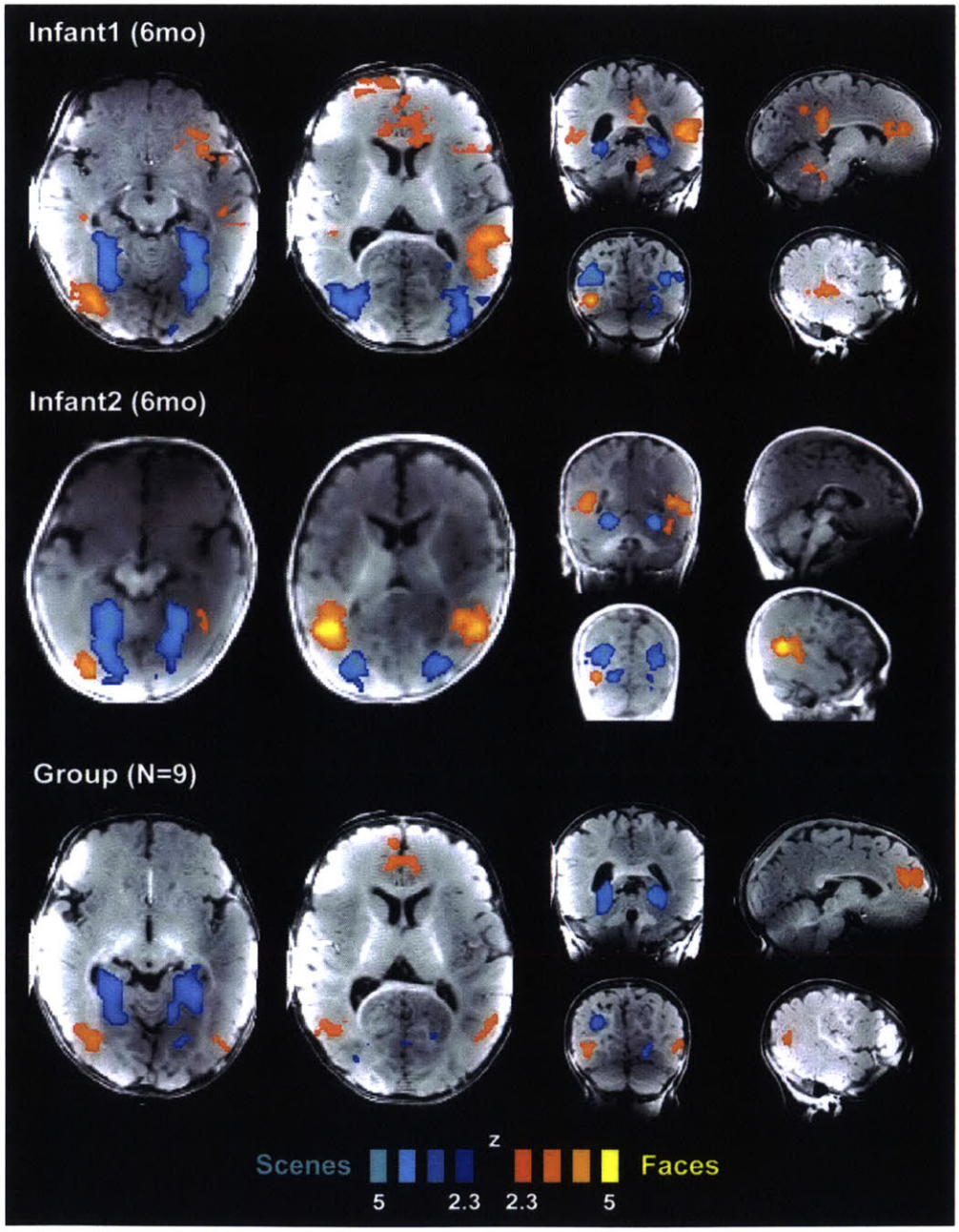


Figure 2: Region-of-interest (ROI) responses (percent signal change, PSC) in regions defined by a face vs scene contrast, in infants and adults. Responses were extracted from data independent from those used to define ROIs, using leave-one-pseudorun-out cross validation. Brain images show heat maps of ROI location across participant and left-out pseudorun. Responses are shown separately for Expt 1 and Expts 2-6. Error bars show the standard deviation of a permutation-based null distribution for the corresponding PSC value.

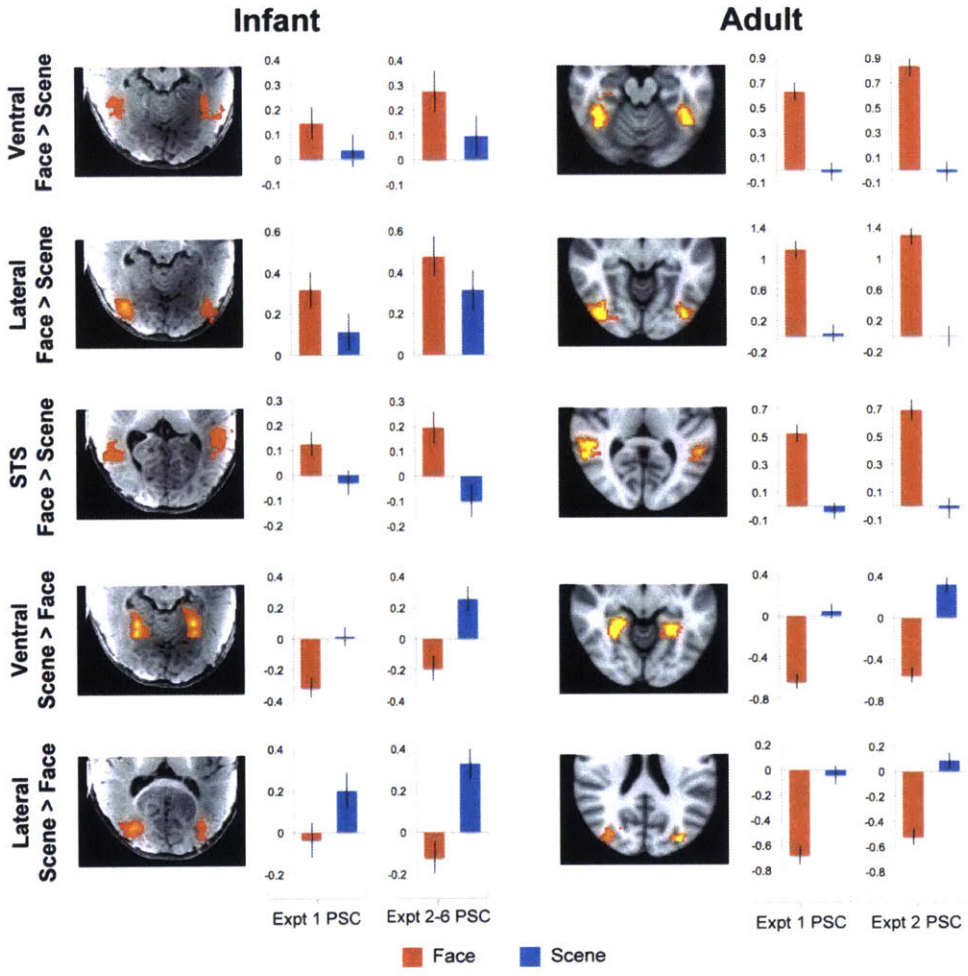
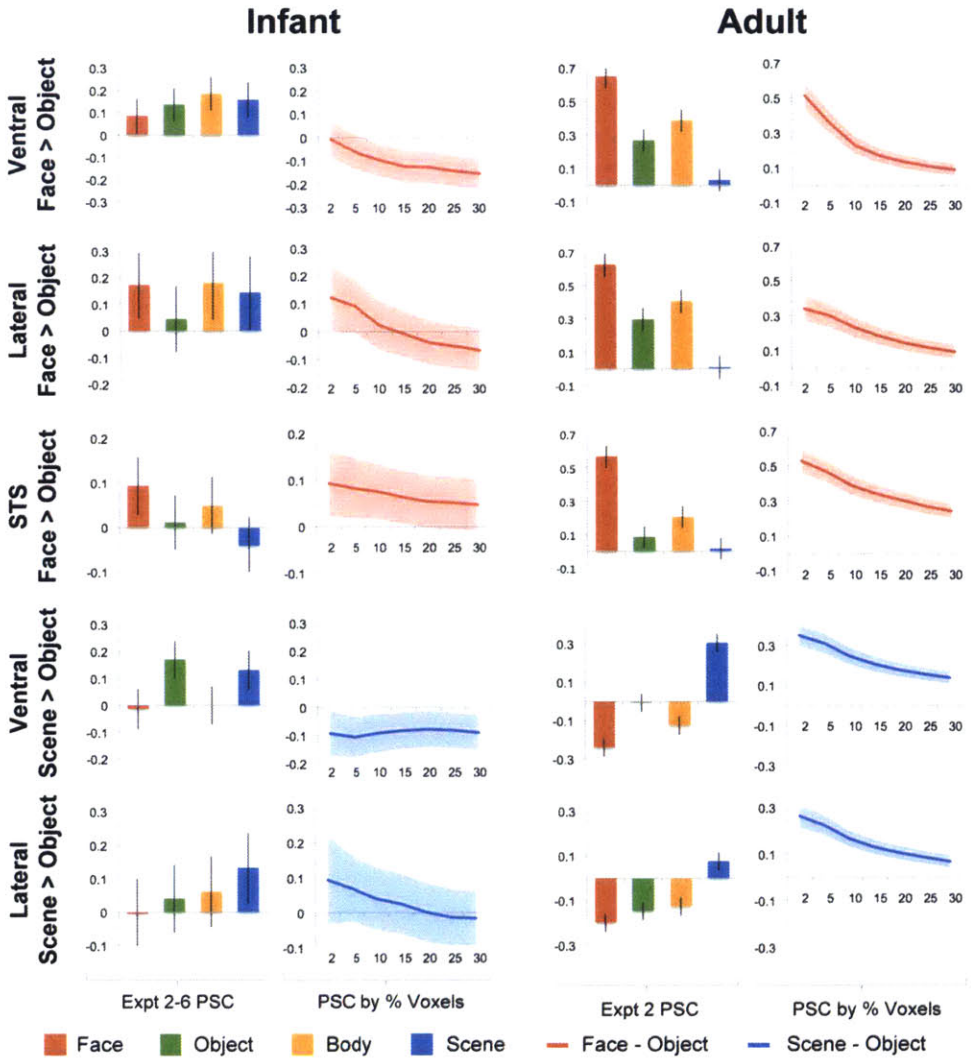


Figure 3: Region-of-interest (ROI) responses (percent signal change, PSC) in regions defined by a face vs object and scene versus object contrasts, in infants and adults. Responses were extracted from data independent from those used to define ROIs, using leave-one-pseudorun-out cross validation. Bar plots show responses of ROIs defined as the top 5% of voxels within an anatomical region, while line graphs show how the difference between face and scene responses varies as a function of % of voxels used to define the ROI. Error bars show the standard deviation of a permutation-based null distribution for the corresponding PSC value or PSC difference.



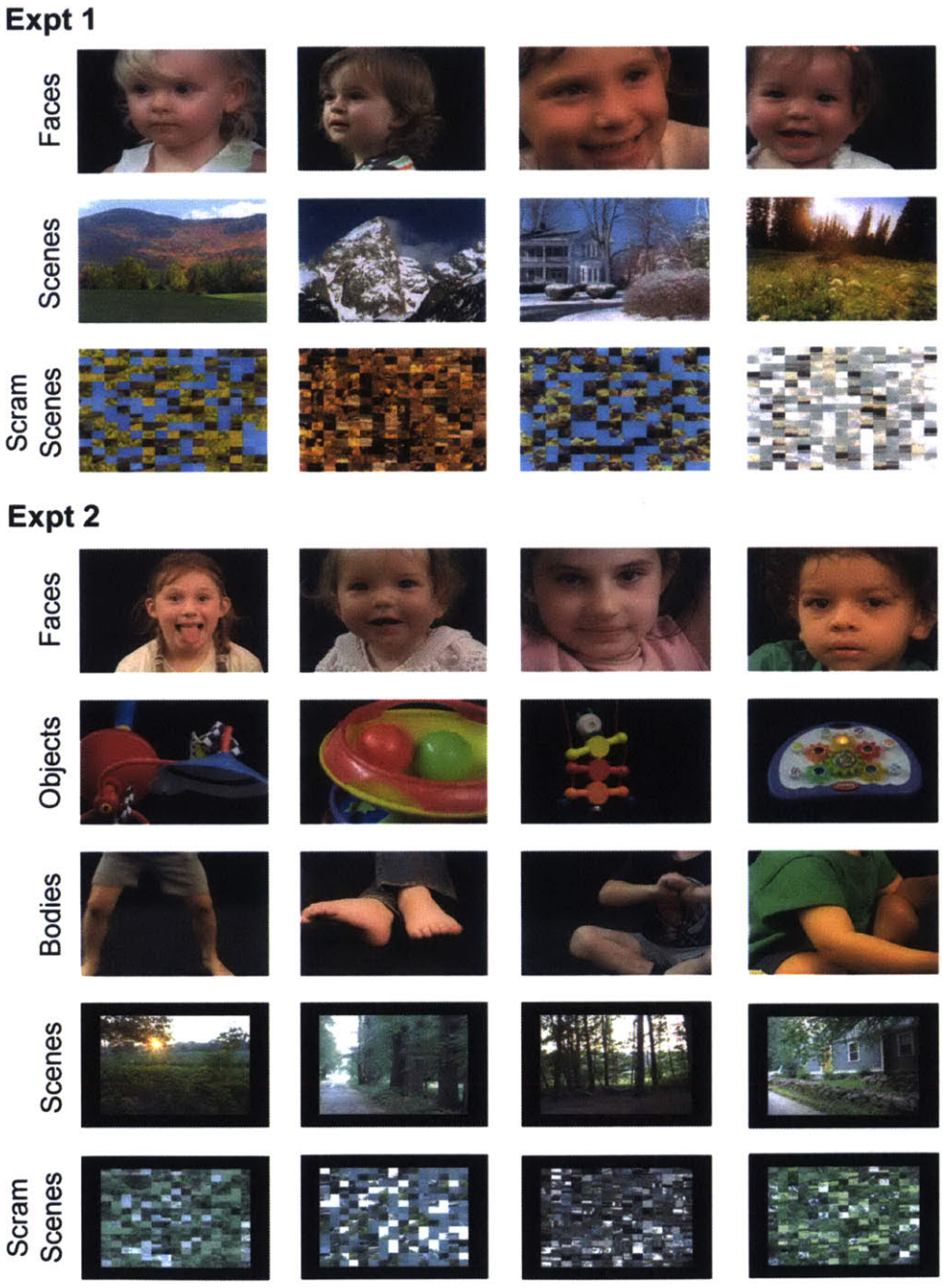
5.6 Supplementary Tables

Supplementary Table 1: Participant demographic information. In cases were infants were scanned at different ages, entries are separated by dataset. Infants are labeled by amount of data kept in their largest dataset, such that Infant1 had the largest amount of data in a single dataset.

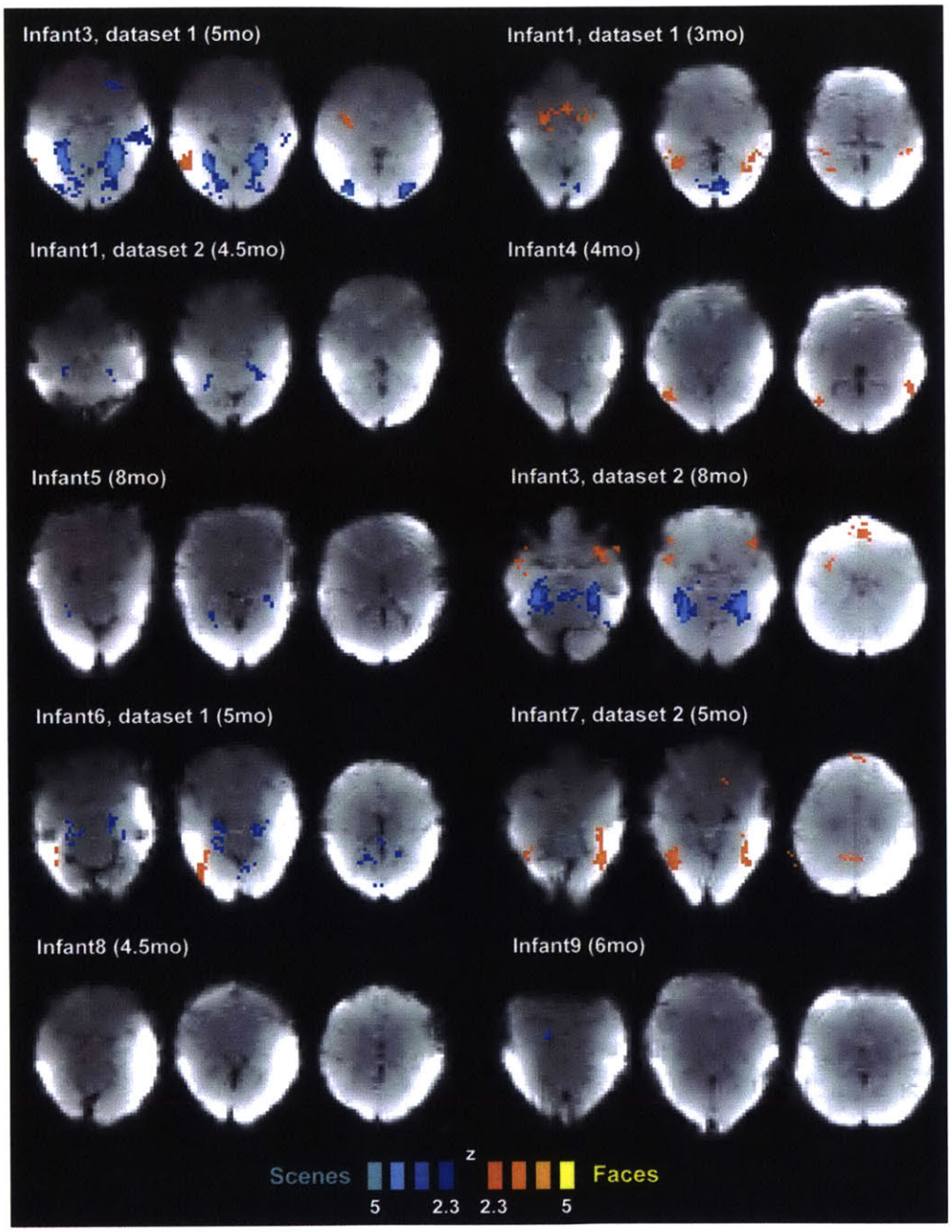
ID	Dataset	Age (mo)	Sex	Scans	Expts	Participant Included	Volumes Acquired	Volumes Kept
Infant1	3/3	5.9	M	4	1, 2	1	2130	1173
Infant2	3/3	6.2	M	4	1, 2, 3	1	3140	908
Infant3	1/2	5.2	M	6	5, 6, 7	1	2748	600
Infant1	1/3	3	M	7	1	1	2626	530
Infant1	2/3	4.5	M	5	1	1	2111	479
Infant4	1/1	4.2	M	4	1, 2	1	2216	292
Infant5	1/1	7.8	M	2	1, 2	1	1136	256
Infant3	2/2	8	M	1	1, 8	1	854	234
Infant6	1/2	5.2	F	1	1	1	455	212
Infant7	2/3	4.9	M	2	1, 4	1	550	211
Infant8	1/1	4.5	M	1	1	1	209	111
Infant9	1/1	5.8	M	2	1	1	834	100
Infant6	2/2	8.6	F	1	1	0	1080	86
Infant7	3/3	6.5	M	1	4	0	320	76
Infant10	1/1	4.1	F	7	1	0	2032	31
Infant7	1/3	2.8	M	1	1	0	358	30
Infant11	1/1	3.4	M	2	1	0	820	0
Infant2	2/3	4.7	M	1	1	0	787	0
Infant2	1/3	3.3	M	1	1	0	707	0
Infant12	1/1	3.7	M	2	1	0	629	0
Infant13	1/1	2.5	M	1	1	0	566	0
Infant14	1/1	2.3	F	1	1	0	560	0
Infant15	1/1	3.9	M	2	1	0	326	0
Infant16	1/1	4.3	M	1	1	0	259	0
Infant17	1/1	4.9	M	3	1	0	222	0
Adult1	1/1	27	M	1	1, 2	1	1260	1260
Adult2	1/1	34	M	1	1, 2	1	1260	1260
Adult3	1/1	27	F	1	1, 2	1	1260	1260

5.7 Supplementary Figures

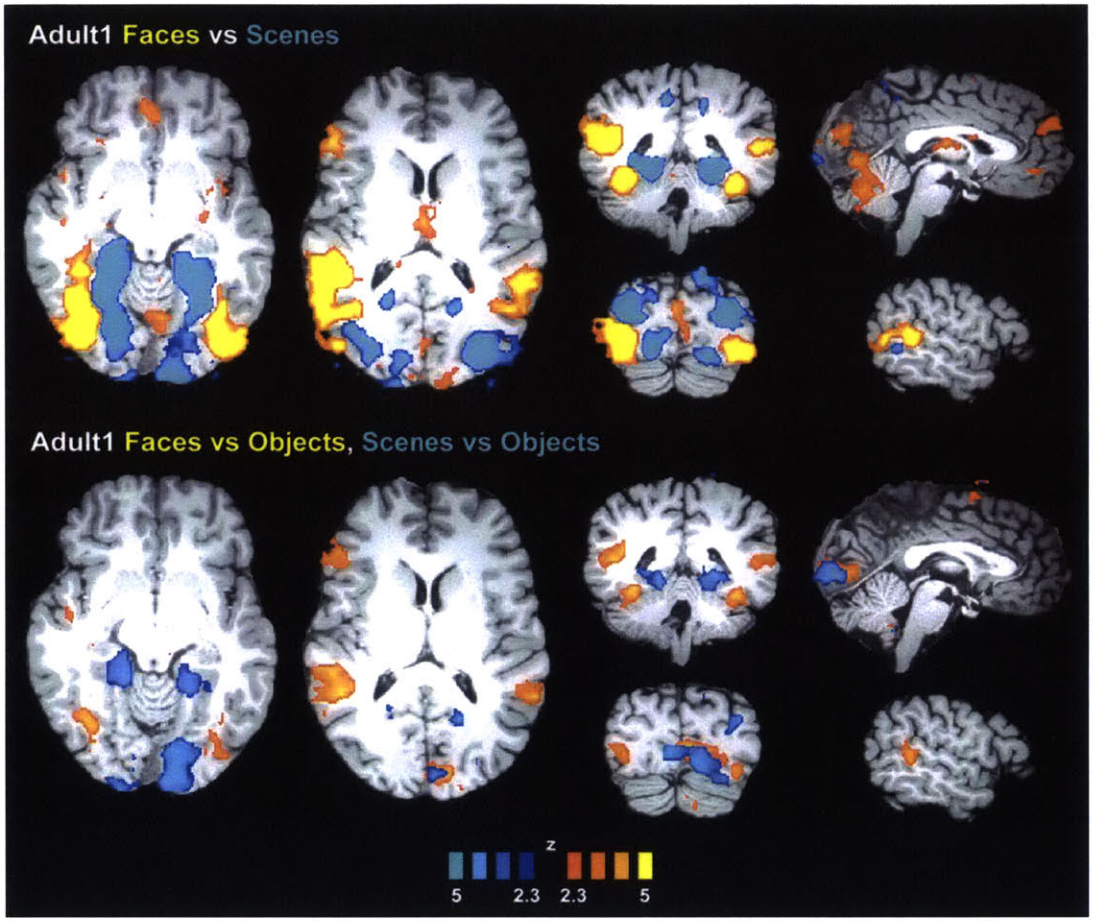
Supplementary Figure 1: Sample frames from video stimuli used in Expts 1-2.



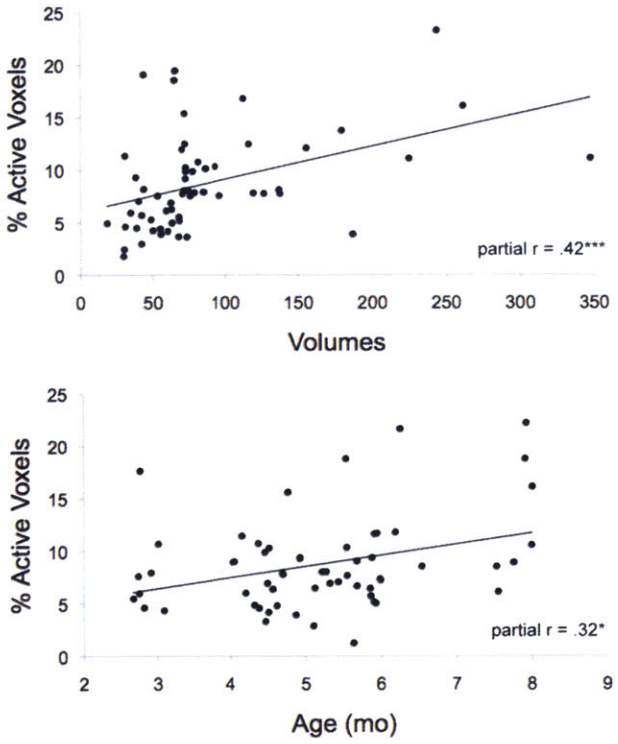
Supplementary Figure 2: Whole-brain activation maps, comparing faces to scenes, in all infants and datasets not included in Figure 1. Maps are thresholded at $P < .01$ voxelwise, and corrected for multiple comparisons using a clusterwise threshold of $P < .05$.



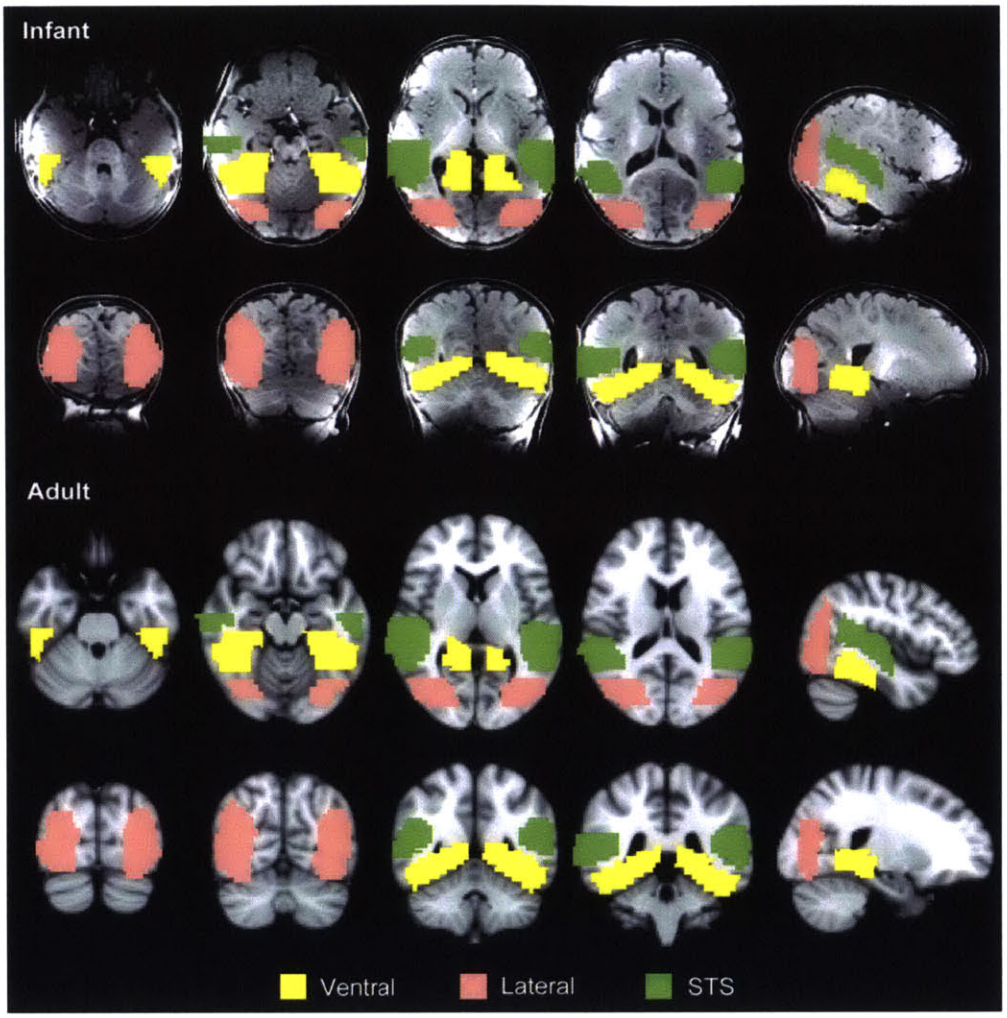
Supplementary Figure 3: Whole-brain activation maps in a representative adult participant. The top row of images shows results from a faces vs scene scontrast, while the bottom row shows faces vs objects and scenes vs objects. Maps are thresholded at $P < .01$ voxelwise, and corrected for multiple comparisons using a clusterwise threshold of $P < .05$.



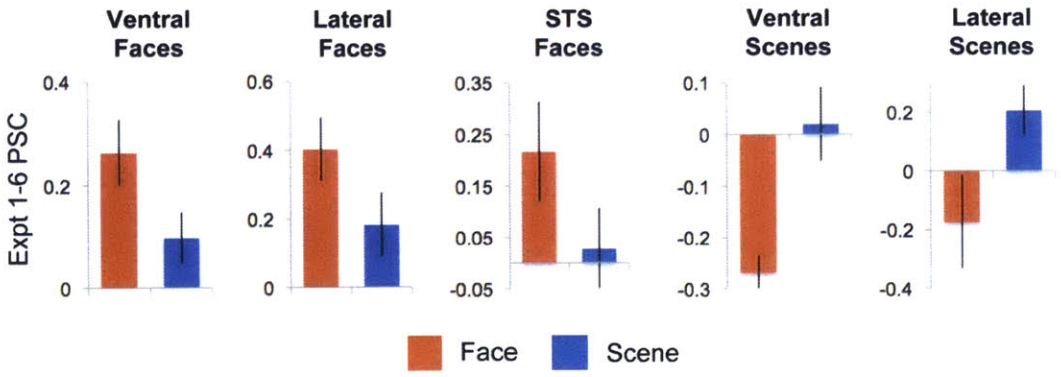
Supplementary Figure 4: Partial regression plots showing the effect of number of volumes and age on activation extent (% of active voxels), across pseudoruns. Both plots show data that has been residualized with respect to the other of the two independent variables. * indicates $P < .05$, *** indicates $P < .001$.



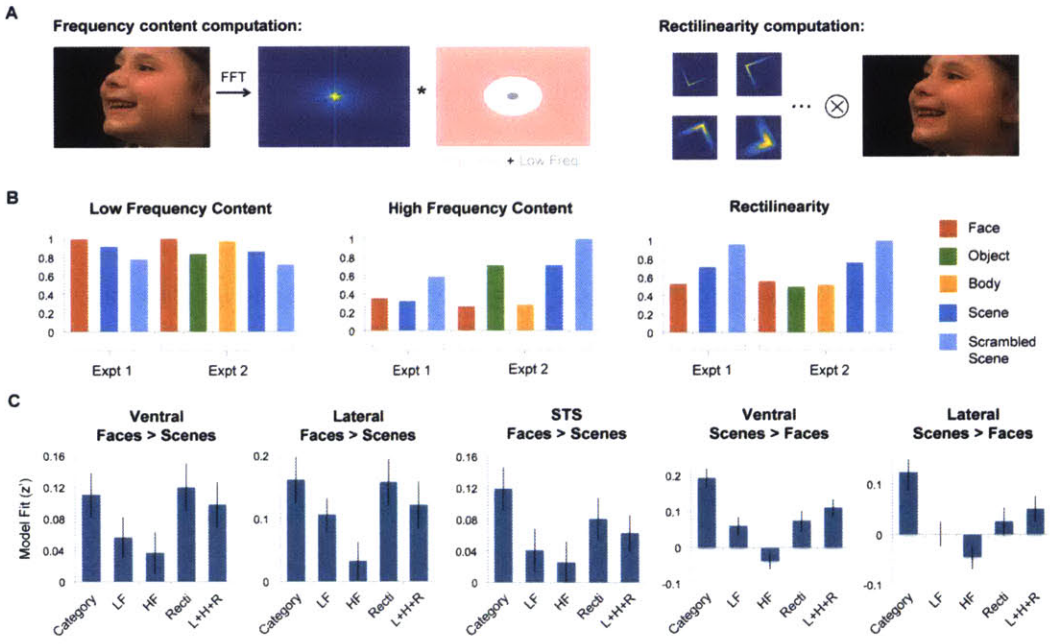
Supplementary Figure 5: Anatomical search spaces used to define ROIs, in infants and adults.



Supplementary Figure 6: Region-of-interest (ROI) responses (percent signal change, PSC) in regions defined by a face vs scene contrast in infants, showing statistics across participants. Responses were extracted from data independent from those used to define ROIs, using leave-one-pseudorun-out cross validation. Error bars show standard error of PSC values across subjects. The effect of category was significant across participants in four of the five regions ($n = 9$; ventral face region, $z = 2.62$, $P < .05$; lateral face region, $z = 2.84$, $P < .05$; STS face region, $z = 2.76$, $P < .05$; ventral scene region, $z = 2.92$, $P < .05$), and near significant in the lateral scene region ($z = 1.87$, $P = .098$).



Supplementary Figure 7: Assessing the relationship between low-level visual features and region-of-interest (ROI) responses. A. Schematic showing how high- and low-frequency content and rectilinearity were computed from video frames. B. Mean values of these visual features across the stimuli used in Expts 1-2, normalized such that the maximum value across categories is set to 1. C. Model fits of category and visual feature models to ROI responses. In all three face-preferring regions, there was no significant difference between the category model and the best-performing visual feature model (ventral face region, $t(54) = -.48$, $P = .64$; lateral face region, $t(54) = .25$, $P = .80$; STS face region, $t(54) = 1.55$, $P = .13$). In contrast, in the two scene-preferring regions, the category model significantly outperformed all visual feature models; for brevity, we report statistics only for the comparison with the best-performing model (ventral scene region, $t(54) = 3.56$, $P < 10^{-3}$; lateral scene region, $t(54) = 2.56$, $P < .05$). Abbreviations: LF = low-frequency content; HF = high-frequency content; Recti = rectilinearity; L+H+R = low-frequency content, high-frequency content, and rectilinearity.



5.7 Supplementary Methods

Paradigm: further details

Across infants, eight slightly different experiments were run. Experiment 1 contained two categories (face, scene) and was run in every infant. Experiment 2 contained four categories (face, body, object, scene) and was run in a subset of $n = 4$ infants. Experiments 3-8 contained 3-4 categories and were each only run in a single infant. Experiments 3-7 use stimuli that are very similar to those used in experiment 2, and were used in early scanning sessions before switching to experiment 2. Experiment 8 contains distinct stimuli and was intended to provide additional evidence for generalization of category preferences across different specific videos. Because we did not acquire enough usable data with experiments 3-8 to analyze them in isolation, they were ultimately only used in combination with other experiments, to increase power for various analyses. In particular, because all experiments contained face and scene categories, all were used for whole-brain face vs scene comparisons, and to define ROIs based on this contrast. Because experiments 3-6 contained four categories, they additionally contributed to four-condition ROI responses.

Experiment 1 consisted of Filmed Faces and Baby Einstein Scenes conditions, as well as a baseline condition of spatially scrambled scenes (using 15x15 grid scrambling, as is the case for all scrambled stimuli). The Filmed Faces were 60 3s-long close-up videos of children's faces on a black background, filmed by the experimenters, as used in a previous experiment in adults(Pitcher et al. 2011). These videos did not contain parts of the body below the neck. The Baby Einstein Scenes were 36 3s-long videos of scenes

taken from the Baby Einstein video collection, which all depicted a three-dimensional spatial layout, and did not contain humans or animals.

Experiment 2 consisted of Filmed Front Faces, Filmed Objects*, Filmed Bodies, Filmed Scenes (presented at 80% size), and a baseline condition of spatially scrambled scenes (also presented at 80% size). The Filmed Front Faces were 30 3s-long videos of front-view faces, similar to the Filmed Faces condition but containing distinct specific videos. The Filmed Objects* were a set of 20 3s-long close-up videos of children's toys on a black background (e.g. rolling balls, moving gear toys), filmed by the experimenters. These 20 clips were selected from a larger set of 60 clips used in a previous experiment(Pitcher *et al.* 2011) (where the * denotes the subset), which were chosen to have virtually no information about three-dimensional scene layout (e.g., corners between walls or between a wall and a floor). The Filmed Bodies were a set of 60 3s-long close-up videos of children's bodies or body parts (not showing faces) on a black background, as used in a previous experiment(Pitcher *et al.* 2011). The Filmed Scenes were a set of 60 3s-long videos filmed by the experimenters from a camera moving through an outdoor scene (e.g. a road, a field), as used in a previous experiment(Pitcher *et al.* 2011). These all depicted a three-dimensional spatial layout, and did not contain humans or animals.

Experiment 3 consisted of Filmed Faces, Filmed Objects*, Filmed Bodies, Baby Einstein Scenes, and a baseline condition of spatially scrambled scenes.

Experiment 4 consisted of Filmed Front Faces, Filmed Objects, Filmed Bodies, Filmed Scenes, and a baseline condition of spatially scrambled objects. Filmed Objects were the full set of 60 filmed object videos from which the Filmed Objects* videos were selected.

Experiment 5 consisted of Filmed Faces, Filmed Objects, Filmed Bodies, Filmed Scenes, and a baseline condition of spatially scrambled objects.

Experiment 6 consisted of Filmed Faces, Filmed Objects, Filmed Bodies, Baby Einstein Scenes, and a baseline condition of spatially scrambled objects.

Experiment 7 consisted of Filmed Front Faces, Filmed Side Faces, Filmed Objects*, Baby Einstein Scenes (presented at 80% size), and a baseline condition of spatially scrambled scenes. The Filmed Side Faces were 35 3s-long videos of side-view faces, similar to the Filmed Faces and Filmed Front Faces conditions but containing distinct specific videos.

Experiment 8 consisted of Baby Einstein Faces, Baby Einstein Objects, Animated Scenes, and a baseline condition of spatially scrambled scenes. Baby Einstein Faces were 3 18s-long videos (containing multiple clips) of children's faces, taken from the Baby Einstein video collection. While these videos typically only contained faces, hands were occasionally presented in the vicinity of the face. Baby Einstein Objects were 3 18s-long videos (containing multiple clips) of children's toys and other objects in motion, taken from the Baby Einstein video collection. Animated Scenes were 18 6s videos designed by having a camera move through an animated scene created using Blender 3D animation software. These all depicted a three-dimensional spatial layout, and did not contain humans or animals.

5.8 Supplementary References

Pitcher D, Dilks DD, Saxe RR, Triantafyllou C, Kanwisher N. 2011. Differential selectivity for dynamic versus static information in face-selective cortical regions. Neuroimage 56:2356-2363.

Chapter 6: Conclusion

6.1 Summary

The studies in this thesis detail the functional organization of STS responses to varied social stimuli, the functional properties of a face- and voice-responsive STS subregion, and the nature of functionally specific visual responses in infants. Taken together, these results inform the cognitive architecture of social perception and the role of the STS therein, as well as the development of functionally specialized brain regions more broadly.

Chapter 2 investigates the large-scale functional organization of STS responses to different social stimuli, and identifies STS subregions selective for processing certain types of stimuli, such as body movements, vocal/speech sounds, and mental states. This was true even for relatively similar domains of input processed by spatially neighboring regions, such as face movement and body movement, suggesting relatively fine-grained dissociations in function. The method of testing responses to many different types of stimuli in the same set of brain regions allows this study to make stronger claims about selective responses than prior work, which has largely tested restricted domains of social perceptual processing in isolation. This result points to a cognitive division of labor in social perception, in which different domains of social input are processed separately, and in which abstract social properties such as mental states are largely represented in distinct regions from those involved in perceptual analysis.

A surprising and intriguing result from this work was that the fSTS, defined as the STS subregion with the maximal response to face movement over object movement, had

an equally strong response to vocal sounds. This contradicts the long-standing view of this region as "face region," instead suggesting a multimodal role. The fSTS did not respond strongly to body movements (Chapter 2) or hand movements (Chapter 4), indicating responses selective to signals from the face, whether from the visual or auditory modality. Furthermore, the fSTS responded similarly to a range of face movements and vocal sounds, irrespective of linguistic or speech content, or perceived communicativeness or social relevance, indicating that this region plays a general role in the perceptual analysis of a broad range of face movements and vocal sounds.

With regard to the role of this region in social perception, a number of observations from Chapters 3-4 indicate that the fSTS contains a mid-level representation, which abstracts across low-level visual properties and begins to show sensitivity to high-level features, but which is still more tied to perceptual input than would be an abstract representation of inferred social properties. First, the strong response of this region to noncommunicative, minimally socially relevant facial and vocal signals is more consistent with a representation of audiovisual features of face actions than with a representation of inferred social properties, insofar as these stimuli may not strongly elicit inferences about social properties. Second, the low response of this region to socially significant hand gestures argues against a generic representation of inferred social properties, as these stimuli have rich social interpretations. Finally, the finding of a parts-based, rather than holistic, representation of face movements (Chapter 3) is more consistent with a kinematic representation of face movement type than with a representation of the mental state implied by a face movement, insofar as the latter are

tied to full face movement patterns (e.g., a smile with a brow raise is perceived as happy or excited, while a smile with a scowl is perceived as sinister).

On the other hand, several properties of this region distinguish it from lower-level perceptual regions that may serve as inputs in the process of social perceptual inference. Face movement representations in this region generalize across visual features such as position and actor, indicating some degree of abstraction (Chapter 3). Furthermore, the fact that this region responds to both dynamic faces and voices is suggestive of representations that abstract across modality to some extent. And indeed, I find that this region is sensitive to an abstract social property, with distinct evoked spatial patterns of activity for communicative and noncommunicative stimuli, in a manner that generalizes across visual and auditory inputs. Hence representations in this region begin to make abstract social properties explicit.

Taken together, these considerations point to the fSTS supporting a mid-level representation in social perception, between lower-level visual and auditory regions and higher-level regions (explicitly encoding inferred social properties) in a computational hierarchy. As such, studying this region may further inform the "black box" of mid-level representations and computations in social perceptual inference.

One of the striking results from the first four chapters is that STS subregions respond robustly to specific types of social stimuli. Are such functionally specialized brain regions largely innate and present early in development, or do they emerge as a result of extensive experience with the relevant input? Chapter 5 of this thesis investigates the nature of functionally specialized brain regions in infants, and finds regions of visual cortex that are sensitive to different categories of input (such as faces

and scenes), but not regions that are strongly selective for specific categories, as found in adults. This supports a view in which the development of functionally specialized brain regions relies both on innate or early-developing constraints, as well as subsequent tuning of response profiles through development, perhaps in virtue of extensive experience with certain domains of input. Specifically, for high-level visual cortex, the large-scale functional organization of category preferences is present in infants, and may provide a scaffold to guide and support subsequent development of increasingly specialized brain regions.

6.2 Future Directions

This thesis work raises myriad directions for future research. Having a catalog of functional subregions within the STS, and a procedure for defining these regions using independent localizer experiments, we can next probe individual subregions in more detail. Specifically, future work should investigate their responses to broad sets of stimuli tailored for each region, and use more sensitive techniques such as MVPA and adaptation, in order to further assess their functional role in social perception. While some prior studies have begun this line of work (Koster-Hale et al. 2013; Koster-Hale et al. 2014; Vangeneugden et al. 2014), much further progress can be made on these problems. This research should further our understanding of both the cognitive architecture of social perception, and detail the role of the STS therein.

Future work should also continue to probe the functional role of the fSTS in particular. Results of Chapters 2 and 4 of this thesis are suggestive of audiovisual representations of face actions in this region, but these studies do not directly test this (see Peelen et al., 2010 and Watson et al., 2014 for evidence from nearby, possibly

overlapping regions). Future work using crossmodal stimulus decoding and adaptation techniques should determine whether representations in this region are audiovisual, and further detail the format of these representations. Additionally, studying the nature of connectivity and interactions between the fSTS and brain regions thought to encode more abstract features of others' mental states will help address whether this set of regions does in fact constitute a computational hierarchy.

While human fMRI has proven remarkably useful for studying the nature of representations in human brain regions, this technique remains fundamentally limited in spatial and temporal resolution, particularly for questions whether the ideal answer would come in the form of neuronal population responses to different stimuli. Thus, a deeper level of understanding of the function of these regions may come from studying similar regions in nonhuman primates such as macaques, using single-unit and population recordings. While the social repertoire and understanding of nonhuman primates is clearly limited relative to humans, some species of apes and monkeys have been shown possess basic aspects of social perception: they can perceive other beings and animate things with self-driven behavior, and appear to understand basic mental states such as intentions and percepts to some degree (Flombaum and Santos 2005; Phillips et al. 2009; Marticorena et al. 2011). This suggests that studying the neural basis of representations of basic mental states and the process of inferring these states from perceptual input should be possible in nonhuman primates.

A critical component of the effort to relate work in human and nonhuman primates will be to establish homologies or analogies between social brain regions across species. Early physiological work showing responses to faces and biological motion in

the macaque STS (Perrett et al. 1982; Desimone et al. 1984; Perrett et al. 1985; Perrett et al. 1989) has been taken as an indication of similar functional roles of the STS in the macaque and humans (Allison et al. 2000). However, several considerations suggest that the relationship between the STS in macaques and in humans is more complicated, and not yet well understood. First, many parts of the human STS are sensitive to linguistic and speech inputs; these regions are unlikely to have monkey homologues, although they could plausibly have precursors of some form. Voice responses have been observed within the STS in humans, but only in regions outside of the STS in macaques, at least with imaging methods (Petkov et al. 2008). And while a number of face-selective regions have been observed in the macaque STS (Tsao et al. 2003), some of these regions bear a systematic spatial relationship with color- and scene-selective responses in a way that mimics spatial relationships observed in the ventral visual pathway in humans (Lafer-Sousa et al. submitted), indicating that at least some face regions in the macaque STS correspond better to regions outside of the STS in humans. These considerations suggest that more work is needed to better understand the relationship between regions in macaque and human STS, such that insights from physiological work in macaques can be related to humans.

Finally, my work on the nature of category-sensitive visual regions in human infants also suggests a number of important questions for future research. What is the developmental time course of the development of increasingly specialized regions, and to what extent does this process rely on experience? To what extent does the finding of regions that are sensitive to high-level distinctions, but not strongly selective for certain inputs, generalize to other domains of input, such as auditory signals, language, and other

types of social signals? And lastly, how does the development of functionally specialized brain regions for vision, social perception, and other processes relate to behavior in these domains? The methods introduced in this thesis should be useful for address these questions in future work.

6.3 References

Allison T, Puce A, McCarthy G. 2000. Social perception from visual cues: role of the STS region. Trends in cognitive sciences 4:267-278.

Desimone R, Albright TD, Gross CG, Bruce C. 1984. Stimulus-selective properties of inferior temporal neurons in the macaque. J Neurosci 4:2051-2062.

Flombaum JI, Santos LR. 2005. Rhesus monkeys attribute perceptions to others. Curr Biol 15:447-452.

Koster-Hale J, Bedny M, Saxe R. 2014. Thinking about seeing: Perceptual sources of knowledge are encoded in the theory of mind brain regions of sighted and blind adults. Cognition 133:65-78.

Koster-Hale J, Saxe R, Dungan J, Young LL. 2013. Decoding moral judgments from neural representations of intentions. Proceedings of the National Academy of Sciences 110:5648-5653.

Lafer-Sousa R, Conway B, Kanwisher N. submitted. Color-biased regions of the ventral visual pathway lie between face- and place-selective regions in humans, as in macaques.

Marticorena DC, Ruiz AM, Mukerji C, Goddu A, Santos LR. 2011. Monkeys represent others' knowledge but not their beliefs. Developmental science 14:1406-1416.

Peelen MV, Atkinson AP, Vuilleumier P. 2010. Supramodal representations of perceived emotions in the human brain. J Neurosci 30:10127-10134.

Perrett D, Rolls E, Caan W. 1982. Visual neurones responsive to faces in the monkey temporal cortex. Exp Brain Res 47:329-342.

Perrett D, Smith P, Mistlin A, Chitty A, Head A, Potter D, Broennimann R, Milner A, Jeeves M. 1985. Visual analysis of body movements by neurones in the temporal cortex of the macaque monkey: a preliminary report. Behav Brain Res 16:153-170.

Perrett DI, Harries MH, Bevan R, Thomas S, Benson P, Mistlin AJ, Chitty AJ, Hietanen JK, Ortega J. 1989. Frameworks of analysis for the neural representation of animate objects and actions. J Exp Biol 146:87-113.

Petkov CI, Kayser C, Steudel T, Whittingstall K, Augath M, Logothetis NK. 2008. A voice region in the monkey brain. Nat Neurosci 11:367-374.

Phillips W, Barnes JL, Mahajan N, Yamaguchi M, Santos LR. 2009. 'Unwilling' versus 'unable': capuchin monkeys' (Cebus apella) understanding of human intentional action. Developmental science 12:938-945.

Tsao DY, Freiwald WA, Knutsen TA, Mandeville JB, Tootell RB. 2003. Faces and objects in macaque cerebral cortex. Nat Neurosci 6:989-995.

Vangeneugden J, Peelen MV, Tadin D, Battelli L. 2014. Distinct neural mechanisms for body form and body motion discriminations. J Neurosci 34:574-585.

Watson R, Latinus M, Noguchi T, Garrod O, Crabbe F, Belin P. 2014. Crossmodal Adaptation in Right Posterior Superior Temporal Sulcus during Face-Voice Emotional Integration. J Neurosci 34:6813-6821.