NeuroImage 221 (2020) 117191

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

Processing communicative facial and vocal cues in the superior temporal sulcus

Ben Deen[a,b,*], Rebecca Saxe[a], Nancy Kanwisher[a]

[a] Department of Brain and Cognitive Sciences and McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge, MA, United States
[b] Laboratory of Neural Systems, The Rockefeller University, New York, NY, United States

A B S T R A C T

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 shown that the face-responsive region of posterior superior temporal sulcus ("fSTS") also responds strongly to vocal sounds. Here, we investigate the functional role of this region and the broader STS by measuring responses 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 action, including both richly social communicative actions, as well as minimally social noncommunicative actions, ruling out hypotheses of specialization for processing speech signals, or communicative signals more generally. 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), indicating sensitivity to an abstract social dimension. These functional properties of the fSTS contrast with a region of middle STS that has a selective, largely unimodal auditory response to speech sounds over both communicative and noncommunicative vocal nonspeech sounds, and nonvocal sounds. Region of interest analyses were corroborated by a data-driven independent component analysis, identifying face-voice and auditory speech responses as dominant sources of voxelwise variance across the STS. These results suggest that the STS contains separate processing streams for the audiovisual analysis of face actions and auditory speech processing.

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. Many prior studies using fMRI and electrocorticography have found responses to human vocal sounds within the middle STS and superior temporal gyrus (Belin et al., 2002, 2000; Binder et al., 2000; Liebenthal et al., 2005; Mesgarani et al., 2014; Norman-Haignere et al., 2015; Overath et al., 2015; Scott et al., 2000; Shultz et al., 2012; Vouloumanos et al., 2001; Wright et al., 2003). These responses have been interpreted either to reflect specialization either for speech processing (Norman-Haignere et al., 2015; Overath et al., 2015; Scott et al., 2000; Vouloumanos et al., 2001), or processing of vocal sounds more generally (Belin et al., 2002, 2000; Deen et al., 2015; Fecteau et al., 2004; Shultz et al., 2012). Within the posterior STS (pSTS), neuroimaging studies have reliably observed visual responses to perceived face movements (Allison et al., 2000; Bernstein et al., 2018; Pelphrey et al., 2005; Pitcher et al., 2011; Puce et al., 1998; Schultz et al., 2013), and spatial patterns of response that discriminate types of face movement (Deen and Saxe, 2019; Said et al., 2010; Srinivasan et al., 2016). These observations have led to the hypothesis that the STS contains a dorsal stream for face processing, specialized for extracting dynamic information from face motion, and distinct from a static form pathway on the ventral surface (Bernstein and Yovel, 2015; Freiwald et al., 2016).

While the face-motion-responsive subregion of pSTS (here termed fSTS) has typically been described as a category-specific visual region (Bernstein and Yovel, 2015; Freiwald et al., 2016; Haxby et al., 2000; O'Toole et al., 2002; Pitcher et al., 2011; Schultz et al., 2013), the broader posterior STS is considered a zone of multimodal association cortex, with responses to both visual and auditory stimuli (Beauchamp et al., 2004, 2008; Hein et al., 2007; Noesselt et al., 2007; Van Atteveldt et al., 2004), and recent studies have found common responses to face movements and vocal sounds within the pSTS in individual human brains (Deen et al., 2015; Watson et al., 2014a; Zhu and Beauchamp, 2017). Our recent work found that fSTS, functionally defined as the maximally face-motion-sensitive subregion of pSTS, has an equally strong response to auditory vocal sounds as to face movements (Deen et al., 2015). These results suggest that fSTS should be considered fundamentally multimodal, and raise questions about the functional role of this face- and voice-specific response.

* Corresponding author at: The Rockefeller University, 1300 York Ave. New York, NY 10065, United States.
E-mail address: benjamin.deen@gmail.com (B. Deen).

https://doi.org/10.1016/j.neuroimage.2020.117191
Received 21 March 2020; Received in revised form 14 July 2020; Accepted 19 July 2020
Available online 23 July 2020

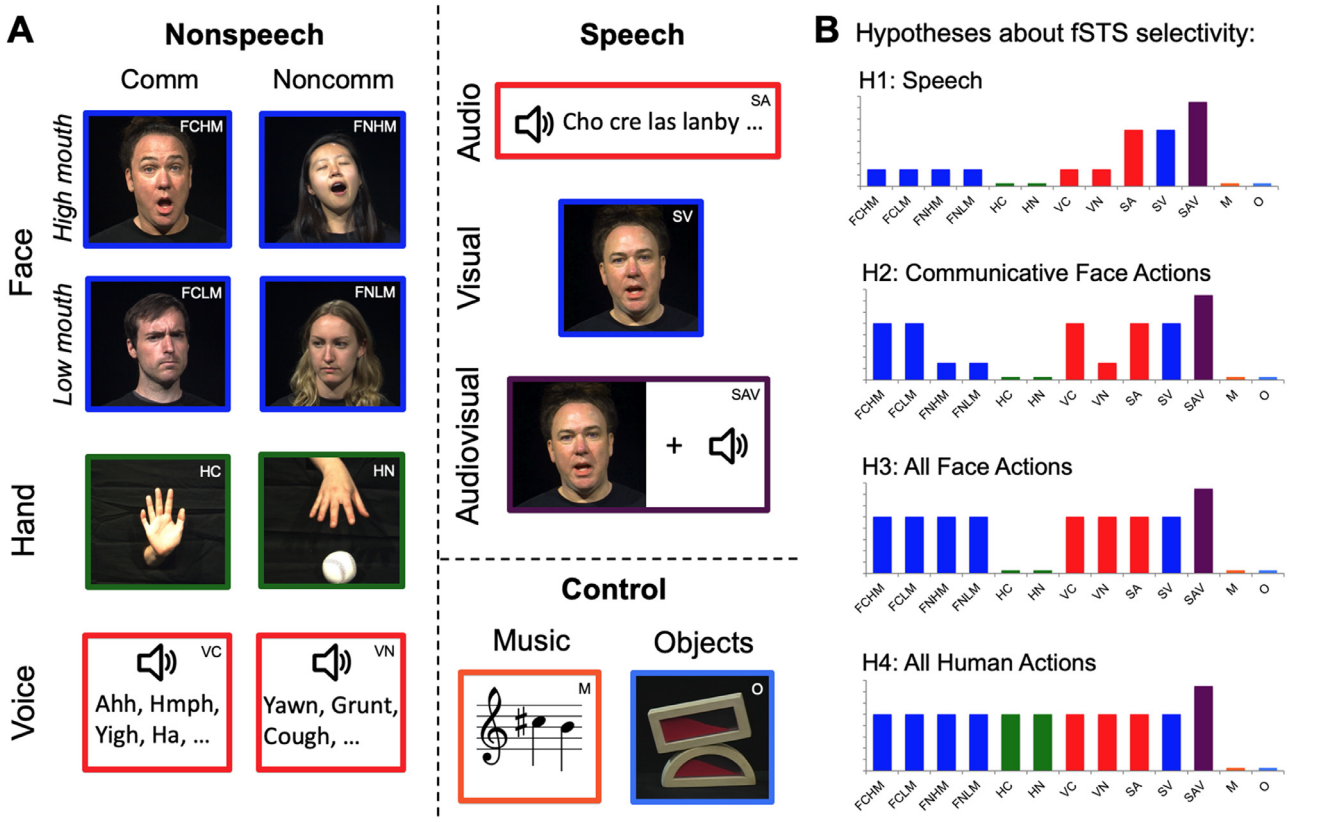

Fig. 1. (A) fMRI condition structure. Thirteen dynamic visual and auditory conditions were used, including face movements and vocal sounds categorized as speech, nonspeech communicative, and noncommunicative, as well as hand and object movement and music as controls. (B) Response profiles predicted by four hypotheses about the selectivity of face-motion-sensitive posterior superior temporal sulcus (fSTS).

Here, we consider four hypotheses regarding the functional role of the fSTS (Fig. 1). *1) The fSTS is specialized for processing audiovisual speech.* Speech is arguably the most ecologically relevant vocal sound we experience, and is well known to be processed audiovisually (McGurk and MacDonald, 1976; Reisberg et al., 1987; Sumby and Pollack, 1954). A face- and voice-responsive area would be well placed to support audiovisual speech processing, and prior studies have found that disrupting pSTS activity using transcranial magnetic or direct current stimulation impairs audiovisual speech perception (Beauchamp et al., 2010; Marques et al., 2014; Riedel et al., 2015). *2) The fSTS is specialized for processing communicative signals produced by faces.* Beyond speech, dynamic facial and vocal signals are used more broadly to communicate social cues via expressions and nonspeech vocalizations. The STS has been argued to play a role in social perception, the inference of abstract social information from perceptual cues (Allison et al., 2000; Brass et al., 2007; Pelphrey et al., 2004; Saxe et al., 2004), and in processing communicative actions in particular (Redcay, 2008; Redcay et al., 2016; Shultz et al., 2012). *3) The fSTS is involved in the perceptual processing of any dynamic audio or visual signal produced by a human face.* On this hypothesis, the fSTS is specialized for processing dynamic facial and vocal cues, but has a broad involvement in processing different actions within this category, including minimally socially relevant actions like a cough or neck stretch. *4) The fSTS is involved in the perceptual processing of any dynamic audio or visual signal produced by a human body.* On this hypothesis, the fSTS not only processes perceptual signals produced by others' faces, but by any body movement, including hand and full body movements. Prior research has found areas within pSTS responsive to both face movements and hand/body movements (Deen et al., 2015; Pelphrey et al., 2005; Thompson et al., 2007), but our recent work found that the strongest pSTS response to naturalistic face movement lies slightly anterior to body movement responses (Deen et al., 2015). The present study aimed to distinguish these hypotheses, and to test how the functional specialization of face-sensitive posterior STS compares to that of voice/speech-responsive middle STS.

To this end, we used fMRI to measure STS responses to a range of naturalistic face and hand movements, and vocal sounds (Fig. 1). These included speech signals, as well as richly communicative, socially relevant nonspeech signals (e.g., a surprised face, a vocal expression of disgust, a hand gesturing "stop"), and noncommunicative, less socially relevant stimuli (e.g., a chewing face, a throat-clearing sound, and a hand writing with a pen). While many prior fMRI studies have measured responses to a small number of conditions in a given set of participants, directly comparing responses to many stimulus conditions within individual participants can provide stronger constraints on theories of functional specialization (Deen et al., 2015; Fedorenko et al., 2013; Norman-Haignere et al., 2015; Poldrack, 2017). We compare responses across two STS regions-of-interest (ROIs), defined functionally in individual participants: fSTS, defined by a visual dynamic faces > dynamic objects contrast, and vSTS, defined by an auditory voices > music contrast. Additionally, we use a data-driven voxel decomposition method (independent component analysis) to identify dominant sources of variance in responses across the STS.

We find that the fSTS responds broadly to different types of face movements and vocal sounds, including speech, nonspeech communicative, and noncommunicative signals, but does not respond strongly to hand movements or non-social control stimuli (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 response

profile is consistent with a mid-level representation of face actions that is not restricted to socially relevant input, but begins to make abstract social dimensions explicit, and to generalize across input domains. The vSTS, in contrast, responded most strongly to auditory speech signals, over nonspeech vocal sounds, visual stimuli, and nonsocial controls. ROI-based responses were corroborated by a data-driven independent component analysis, demonstrating that voxelwise responses across the STS are well modeled as a linear combination of four component response profiles: responses to visual stimuli, auditory stimuli, faces and voices, and speech. These results suggest that the STS is organized into separate processing streams, one for audiovisual face actions and another for speech sounds.

2. Methods

2.1. 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.

2.2. 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, broadly sampling the space of human social perceptual inputs (Fig. 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. For our purposes, a "communicative" action is defined as one produced to intentionally communicate information to another agent. 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 (Fig. 1A): 1) communicative, high-mouth-motion face movements (FCHM); 2) communicative, low-mouth-motion 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, finger 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 1323 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 5 s, and such that balanced properties of stimuli from each token and actor were included. This resulted in 128 FCHM clips (mean duration 2.23 s), 128 FCLM clips (2.22 s), 128 FNHM clips (2.28 s), 128 FNLM clips (2.31 s), 144 HC clips (1.98 s), 144 HN clips (1.97 s), 157 VC clips (1.32 s), 168 VN clips (1.48 s), and 46 speech clips (5.07 s).

As a nonsocial auditory control condition, we used 150 instrumental music clips from a range of genres (e.g. classical, jazz, rock), cut in duration to 1.5 s to match the length of VN stimuli. Music clips were chosen from a larger set of 724 clips, as the subset of 150 clips that best matched vocal stimuli in frequency spectra (details on the computation of frequency spectra and other acoustic properties are included in Supplementary Information). All audio stimuli were root-mean-square amplitude normed and ramped with a 50 ms linear ramp at the beginning and end of the clip. As a nonsocial visual control condition, we used 60 video clips of dynamic objects, used in a prior experiment (Pitcher et al., 2011), cut to 2.27 s to match the 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 20 s (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 22 s for each block. In each run, 26 blocks (2 per condition) were presented, in palindromic order, with specific block order counterbalanced across runs and participants. Blocks were separated by 6 s of a baseline condition, consisting of a black screen with a white central fixation cross. There was an additional 10 s of baseline at the beginning of the experiment, 16 s in the middle, and 10 s 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). 1-back behavioral performance was high (mean accuracy 93.3%, hit rate 74.1%, false alarm rate 4.3%) and consistent across runs (Supplementary Information, Figure S3).

2.3. Stimulus ratings

To verify that our communicativeness manipulation was effective, we collected behavioral ratings on the stimuli using Amazon Mechanical Turk. For each video or audio clip from the communica-

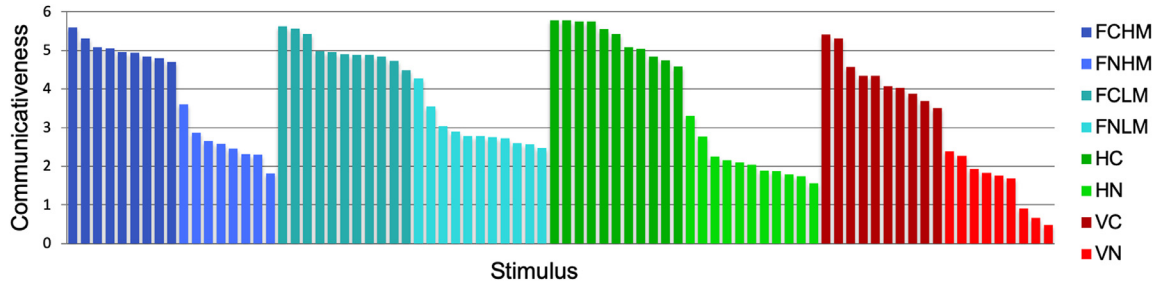

Fig. 2. Behavioral ratings of communicativeness, across the 80 specific actions used in the study, categorized by condition. Condition labels: FC = communicative face movement, FN = noncommunicative face movement, HM = high mouth motion, LM = low mouth motion, HC = communicative hand movement, HN = noncommunicative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

tive/noncommunicative conditions (FCHM, FCLM, FNHM, FNLM, HC, HN, VC, VN), 20 participants viewed or listened to the clip and answered questions in a brief survey. To assess communicativeness, 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 task approval rating of at least 95%, and at least 50 tasks 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 Fig. 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 ($r = 0.99$, $P \approx 0$), 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 ($F(7,72) = 84.14$, $P < 10^{-31}$, $R^2 = 0.89$). 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.

2.4. 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] = 2.53 s; echo time [TE] = 1.64 ms, 3.5 ms, 5.36 ms, 7.22 ms, flip angle $\alpha = 7^{\circ}$, field of view [FOV] = 256 mm, matrix = 256 × 256, slice thickness = 1 mm, 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 = 2 s, TE=30 ms, $\alpha = 90^{\circ}$, FOV = 192 mm, matrix = 64 × 64, slice thickness = 3 mm, slice gap = 0.6 mm, 32 near-axial slices, near-whole-brain coverage).

2.5. 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 (5 mm FWHM), and high-pass filtered (Gaussian-weighted least squares fit straight line subtraction, with $\sigma = 50$ s (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 participant 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 participant using 2nd-level fixed effects analyses, after registering beta maps from each run to a template image in native functional space (the middle image from the first run). Data were also combined across even runs and odd runs, for split-half analyses.

2.6. 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, defining regions with face and voice 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-sensitive region and a middle STS voice-sensitive region (Deen et al., 2015). Search spaces were defined as the set of active voxels (at the group level) within a 15mm-radius sphere around a peak coordinate, and 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

Table 1
Mean coordinates of ROI centers-of-gravity, in MNI space.

ROI	*x*	*y*	*z*
lfSTS	−54.6	−36.9	3.9
rfSTS	54.3	−36.1	5.8
lvSTS	−60.0	−15.8	−0.9
RvSTS	57.5	−15.7	−5.3

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. Mean ROI center-of-gravity coordinates are given in Table 1.

While we used a relatively strict statistical threshold to identify focal regions with particularly strong responses, and for consistency with our prior work (Deen et al., 2015), this method has the disadvantage of excluding participants without ROIs defined. An additional ROI analysis is described in the supplement, which assesses face-responsive regions within posterior/middle/anterior STS search spaces, defining ROIs using a top-N-voxel criterion. This analysis includes all participants by design, and enables us to ask whether significant responses to both faces and voices exist elsewhere along the length of the STS. The results corroborate the presence of face and voice responses in face-motion-sensitive posterior STS observed in the main ROI analysis.

For each ROI in the main analysis (left and right fSTS and vSTS), 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 across conditions and participants, with participant included as a random effect, using MATLAB's fitlme function.

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 analysis 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 by modality (face, voice, hand) by communicativeness ANOVA on all human non-speech 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, to test whether responses to face motion were modulated by the presence of mouth motion, we compared responses to high mouth motion versus low mouth motion videos.

2.7. Multivariate pattern analysis

The ROI analysis revealed that the fSTS responded similarly to communicative and noncommunicative face movements and vocal sounds. We next asked: would spatial patterns of response in these regions 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.

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 2×2 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 each ROI, defined as described above, we performed seven specific comparisons, testing discrimination of communicativeness within and across modalities: 1) within face movements; 2) within vocal sounds; 3) within hand movements; 4) within face movements, generalizing from low to high mouth 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 other regions could discriminate communicative and noncommunicative stimuli. We first tested the vSTS, using the same tests described above. Additionally, we ran a whole-brain searchlight analysis, focusing 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 $P < .05$.

2.8. Independent component analysis

While ROI-based analyses provide a detailed characterization of responses in STS subregions of interest, the STS is a large and functionally diverse area, and response profiles of interest may be missed by restricting focus to specific functional ROIs. We next asked: what are the dominant response profiles to dynamic faces and voices 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 weightings of each profile across voxels are 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.

Methods used for ICA are depicted in Fig. 5. The input data for our implementation of ICA consisted of a condition-by-voxel 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) within each individual participant. Beta values from each of the thirteen conditions were extracted from each selected voxel, to construct a condition-by-voxel 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 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 normal-

ization, and thus doesn't require an assumption that voxels in similar locations across participants are functionally similar, and allows for voxel selection in each participant (Norman-Haignere et al., 2015).

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 condition-by-voxel data matrix separately for each half. For each potential number of components D (between 1 and 13), we projected the first-half data matrix onto the subspace spanned by the first D 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 (Fig. 5). Identified principal components were highly consistent across left-out participants: the mean normalized dot product between the first four PC vectors across PCA solutions from different left-out participants was 0.99.

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 (Fig. 5). By the Central Limit Theorem, linear combinations of independent random variables will tend toward Gaussian distributions. Thus, identifying underlying independent components from observed linear combinations is equivalent to finding axes with minimally Gaussian data distributions (Hyvärinen and Oja, 2000). We obtained this basis using an algorithm that minimizes entropy along a set of orthogonal axes (Norman-Haignere et al., 2015, nonparametric algorithm, https://github.com/snormanhaignere/nonparametric-ica). For prewhitened data, minimizing entropy is equivalent to minimizing mutual information, a measure of statistical dependence. Minimizing entropy is also equivalent to maximizing non-Gaussianity, because the Gaussian distribution has maximum entropy for a given variance. This procedure yielded 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 spatial maps of voxel weights. Each voxel's response profile was modeled as a linear combination of IC vectors, where the coefficient for each component constituted a weight. These values were normalized to MNI space and averaged across participants to compute spatial maps of voxel weights for each component. To test whether IC weights were lateralized, we computed a laterality index—the difference between the mean voxel weight in left and right hemispheres. This index was tested against the null hypothesis of zero using a one-sample, two-tailed *t*-test across participants.

Our ICA method can only find meaningful independent components if data distributions along these dimensions are non-Gaussian. We tested this assumption by measuring statistical properties of voxel weight distributions—skewness and kurtosis—in each participant. These statistics were tested against the null hypothesis of values from a Gaussian distribution (skewness=0, kurtosis=3) using a nonparametric bootstrap test, resampling from the distribution of statistics across participants (10,000 samples).

Are spatial patterns of IC voxel weights consistent across participants? We next assessed spatial correlations of weight maps from pairs of participants. Correlations were computed between maps in MNI space, restricted to voxels that were used as input for both participants. To assess significance, we compared within-component and between-component correlations using a permutation test. We formed a null distribution for the difference between within- and between-component correlations, by permuting pairs of components (1–1, 1–2, 3–4, etc.), which are exchangeable under the null hypothesis of no difference between within- and between-condition correlations (10 choose 4 = 210 permutations).

Lastly, to evaluate the geometry of IC response profiles in 13-dimensional condition space, we computed normalized dot products between each component's response profile (corresponding to the cosine of the angle between response vectors). For illustration, these were compared to normalized dot products of principal component vectors, which are constrained to be orthogonal.

3. Results

3.1. 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 non-communicative face and hand movements and vocal sounds, as well as nonword speech stimuli. All tests were performed as mixed-effects ANOVAs across conditions and participants, with participant included as a random effect.

Responses in each ROI across all conditions are shown in Fig. 3. We first tested the selectivity profile of face-sensitive posterior STS (fSTS) and voice-sensitive middle STS (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, 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? We tested this using a region by modality by communicativeness ANOVA. Although the regions differed in their overall response (main effect of ROI, $F(3368) = 37.43$, $P < 10^{-20}$) and in their selectivity across modality (ROI by modality interaction, $F(6368) = 4.06$, $P < 10^{-3}$), the communicativeness of the stimuli did not influence the response (main effect and interaction terms involving this factor, all P's > 0.7). 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 performed post-hoc tests comparing responses to communicative versus 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 ($t(42) = 2.56$, $P = .014$); 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 influence on fSTS and vSTS responses, with the exception of an increased response to communicative vocal sounds in the left vSTS.

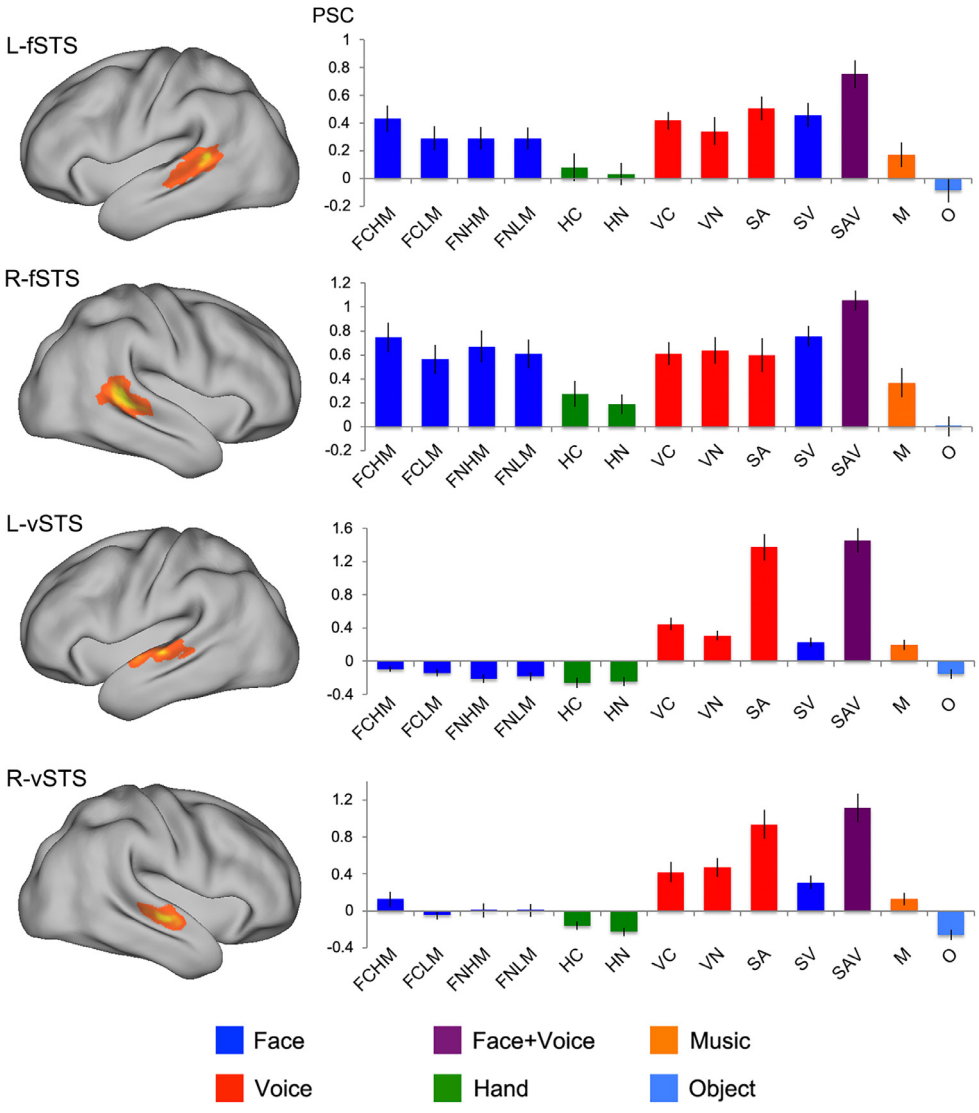

Fig. 3. Face-responsive posterior STS (fSTS) respond strongly to all face movements and vocal sounds, while voice-responsive middle STS (vSTS) responds selectively to speech sounds. Regions were defined using a faces > objects contrast (fSTS) and a voices > music contrast (vSTS). Left: heat maps of region-of-interest locations across participants. Right: responses of these regions (in percent signal change, PSC) 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.

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, $F(3376) = 6.41$, $P < 10^{-3}$), and their relative response to faces and voices (region by modality interaction, $F(3376) = 18.40$, $P < 10^{-10}$). We also observed a region- and modality-specific modulation by speech content (region by modality by speech content interaction, $F(3368) = 4.03$, $P < .01$). 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, we compared the response of each region to nonspeech face movements with and without a mouth motion component (HM versus LM), to ask whether common responses to face movements and vocal sounds are driven by the presence of mouth movement (Zhu and Beauchamp, 2017). While both right and left fSTS responded strongly to face movements with or without a mouth component, responses in the right hemisphere were modulated by the presence of mouth movement (HM > LM, right: $t(58) = 3.06$, $P < .01$; left: $t(38) = 1.49$, $P = .15$). vSTS did not respond strongly to nonspeech face movements, but a marginal effect of mouth movement was observed in the right hemisphere (right: $t(50) = 2.28$, $P < .05$; left: $t(42) = 0.29$, $P = .77$). Thus, face- and voice-sensitive fSTS responded both to movements with and without mouth motion, but had a slight preference for movements with a mouth component in the right hemisphere.

Do face and voice responses, as observed in fSTS, exist in middle and anterior parts of the STS? A supplementary ROI analysis assessed face-motion-responsive ROIs within posterior, middle, and anterior STS, and found that while face and voice responses were most prominent posteriorly, such responses can be found along the length of the STS bilaterally (Fig. S4). This demonstrates that face-motion-responsive regions throughout middle and anterior the STS also have responses to vocal sounds, and shows that the face/voice response observed in posterior STS is robust across multiple methods for defining ROIs.

To summarize, we found that face-sensitive posterior STS (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, communica-

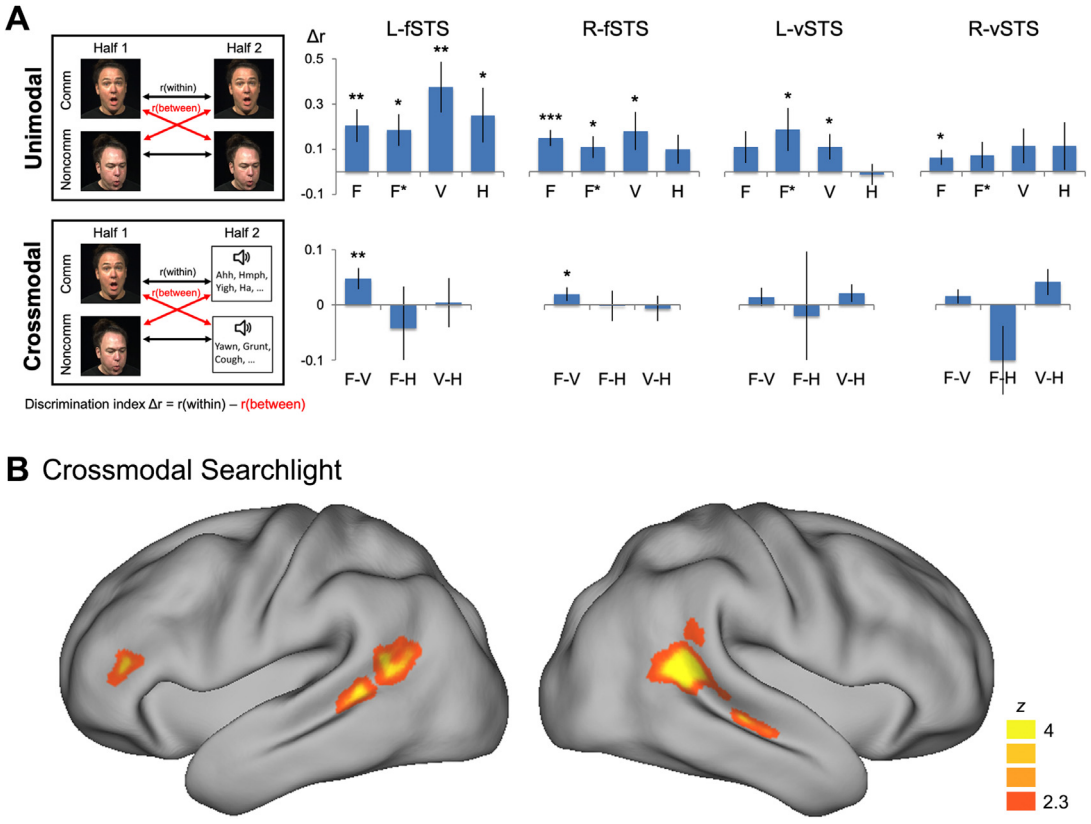

Fig. 4. Multivoxel pattern analysis results: decoding communicativeness from spatial patterns of response, both within and across modality. (A) Region-of-interest-based results, for fSTS and vSTS. Discrimination indices (correlation difference scores) for comparing patterns of response to communicative 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). * denotes $P < .05$, ** $P < .01$, *** $P < .001$. (B) Searchlight results for decoding communicativeness across modality (faces to voices). Whole-brain statistical map thresholded at $P < .01$ voxelwise, followed by a $P < .05$ permutation-based clusterwise threshold to correct for multiple comparisons.

tive content, and speech content. In contrast, the response profile of voice-sensitive middle STS (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.

3.2. Multivoxel pattern analysis

While the ROI analysis showed similar mean responses in fSTS to communicative and noncommunicative face actions, it remains possible that patterns of activity in this region contain information about communicativeness. We next ask whether spatial patterns of response across voxels in fSTS differed between communicative and noncommunicative stimuli, both within and across modalities (faces, voices, hands).

MVPA results are shown in Fig. 4. 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 but not right fSTS was able to discriminate between communicative and noncommunicative hand movements (left: $t(9) = 2.07$, $P < .05$; right: $t(14) = 1.56$, $P = .07$).

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 > 0.45). This result indicates 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. Furthermore, this crossmodal decoding result cannot be explained in terms of low-level visual or acoustic properties that differ across communicative and noncommunicative conditions within either modality.

Can patterns of response differentiating communicative from noncommunicative face actions be observed in other brain regions? We first tested these effects in the vSTS. Patterns in left vSTS were able to discriminate communicativeness for face movements, generalizing across high to low mouth movement conditions ($t(10) = 1.99$, $P < .05$) and for vocal sounds ($t(10) = 1.99$, $P < .05$), while patterns in right vSTS were able to discriminate communicativeness for face movements ($t(12) = 1.88$, $P < .05$). Other unimodal effects, and all crossmodal effects, were not significant (P's > 0.05). Thus, while spatial patterns of response in the vSTS show some sensitivity to communicative content, the effects were relatively weak and inconsistent across hemispheres, and neither region showed evidence for crossmodal decoding.

We next performed a whole-brain searchlight analysis. 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 Fig. 4B. 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 infe-

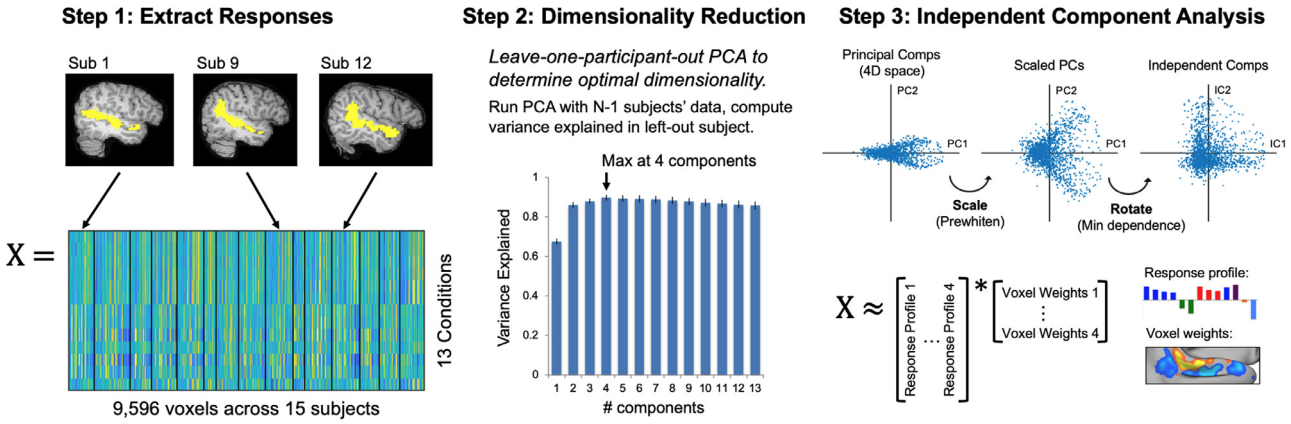

Fig. 5. Independent component analysis methods. Left, step 1: responses (beta values) were extracted across STS voxels and conditions, and concatenated across participants to form a data matrix. Middle, step 2: leave-one-participant-out principal component analysis (PCA) was used to determine the dimensionality for which the PC-spanning subspace explained maximum variance in left-out participants. Right, step 3: independent component analysis (ICA) was performed within the subspace spanned by the first four principal components, by first scaling data to have equal variance along each dimension (prewhitening), and then finding a rotation that minimizes statistical dependence between dimensions. Step 3 is visualized using synthetic data in two dimensions.

rior frontal gyrus. These results indicate that fMRI-decodable 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.

3.3. Independent component analysis

Are parts of the STS beyond functionally-defined fSTS and vSTS involved in processesing dynamic facial and vocal stimuli? While ROI-based analyses provide a detailed characterization of specific functional subregions, they don't assess responses in other parts of the STS, and require a priori assumptions about which regions are involved in 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 (Fig. 5). This subspace captured 95.3% of the total variance across voxels. Running ICA then yielded four response profiles spanning this subspace, with minimal statistical dependence of voxels' responses along each dimension. These response profiles, as well as spatial maps of voxel weights, are shown in Fig. 6A. 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 response to all visual conditions, and roughly zero response to auditory conditions, and thus was termed the visual component. The voxel weights for this component followed a posterior-to-anterior spatial organization, with positive weights posteriorly (adjacent to early visual cortex) and decreasing weights moving anteriorly along the STS. The second component had a positive response to all auditory conditions, and roughly zero response to visual conditions, and thus was termed the auditory component. The voxel weights 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.

The third component had a positive response to all face movement and vocal sound conditions, including communicative and noncommunicative conditions, and speech and nonspeech conditions, but had a negative response to 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 voxel weights for this component were strongest around the posterior STS, with positive weights extending into middle and anterior STS.

The fourth component had a strong response to audio and audiovisual speech conditions, weak response to the visual speech, vocal nonspeech, and music conditions, and a negative response to 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 weights were strongest in the upper bank of the middle STS, and decreased moving ventrally, anteriorly, and posteriorly.

Are the STS response profiles captured by these independent components dominant in a particular hemisphere? We computed a laterality index—the difference between mean voxel weights in the left and right hemispheres—and tested this index across participants (Fig. 6B). This index was only significant for component 3, the face+voice component ($P < .05$, two-tailed t-test), which had stronger weights in the right hemisphere.

Do our data satisfy the key underlying assumption of ICA—that distributions along IC dimensions are non-Gaussian? To assess non-Gaussianity, we measured the skewness and kurtosis of voxel weight distributions (Fig. 6B). We then tested the distributions of these statistics across participants against the null hypothesis of Gaussian values (skewness=0, kurtosis=3) using a nonparametric bootstrap test. Skewness was significantly greater than zero for components 1, 2, and 4 ($P \approx 0$, i.e. no bootstrap samples were less than 0), but not for component 3 ($P \approx 0.09$). Kurtosis was significantly greater than 3 for all components ($P \approx 0$). This demonstrates that components were non-Gaussian, demonstrating sparsity (high kurtosis) and a bias toward positive values (right-skew), which validates the non-Gaussianity assumption of our ICA method. Sparse, right-skewed weight distributions may result from anatomical clustering of neural populations with similar response profiles, yielding a small number of voxels with particularly high weights (Norman-Haignere et al., 2015). Notably, the face+voice component was sparse but not significantly skewed, reflecting the presence of large positive and negative weights across voxels and conditions.

Are spatial maps of voxel weights consistent across individual participants? Maps of voxel weights from a representative set of participants are shown in Fig. 7. These maps showed a consistent spatial structure across participants, despite the IC analysis having no information

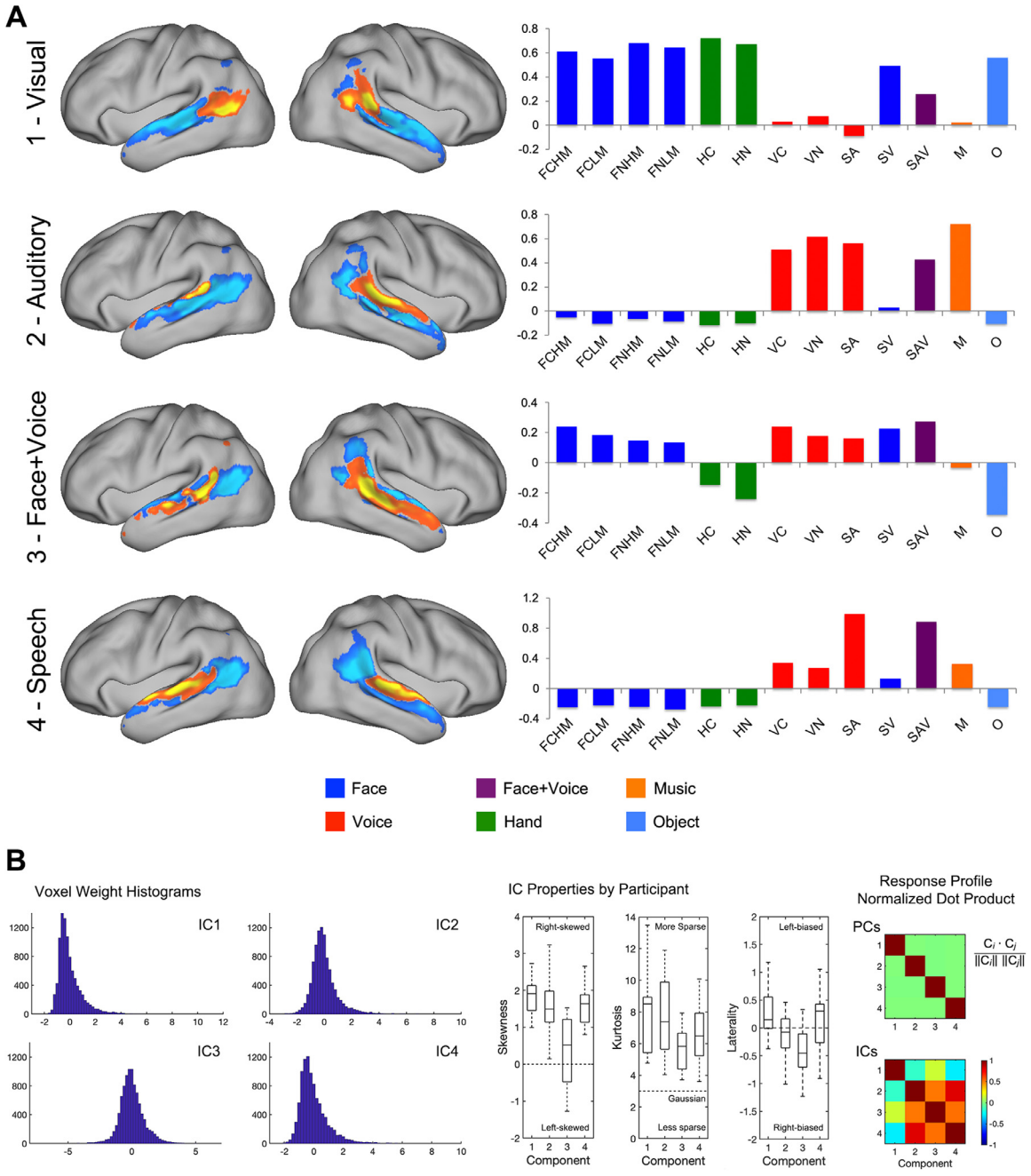

Fig. 6. Independent component analysis identifies face-voice and speech responses as dominant response profiles across the STS. (A) Right: response profiles for four independent components, which together explained ~95% of voxelwise variance in STS responses. Left: maps of voxel weights—the contribution 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. (B) Left: histograms of voxel weights for each component. Middle: properties of voxel weight distributions—skewness, kurtosis, and laterality index—shown as box and whisker plots of the distribution across participants. Boxes show the 25th percentile, median, and 75th percentile of the distribution, and whiskers show the range. Right: matrices showing normalized dot products between pairs of response profiles. Principle components (PCs) are constrained to be orthogonal, while independent components (ICs) are not.

about voxels' spatial location. To quantify this consistency, we compared within- and between-component correlations of spatial maps across participants. Within-component correlations were significantly larger than between-component correlations (mean correlation difference = 0.246, $P < .01$, permutation test).

How does the geometry of response profile vectors differ between principal components and independent components of our data? While PC response profiles are constrained to be orthogonal, IC response profiles do not have this constraint. To compare geometries of PC and IC response profiles, we computed normalized dot products between response profile vectors from each component, equal to cosine of the angle between response profiles in 13-dimensional condition space. These dot products were equal to zero for PCs, but were nonzero for ICs, with normalized dot products >0.5 for components 2, 3, and 4. Thus,

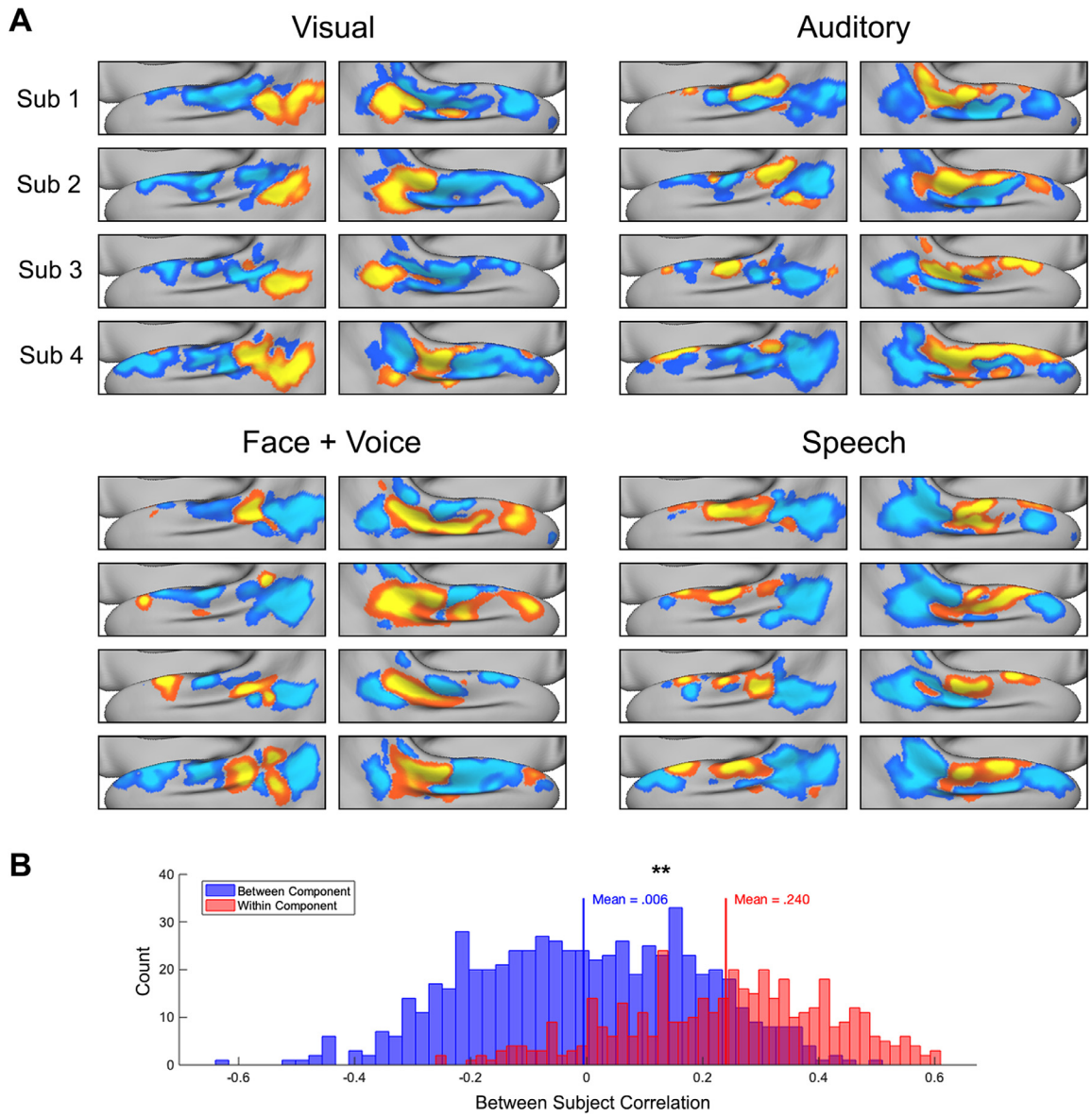

Fig. 7. Independent component voxel weight maps are consistent across participants. (A) Voxel weight maps for four representative participants. (B) Histograms of between-participant correlations in voxel weight maps, either within-component, or between-component. ** denotes $P < .01$.

PCA and ICA yielded components with a rather different geometric structure.

What proportion of unique variance in STS responses is explained by each component? Because IC vectors are nonorthogonal, they do not provide an orthogonal decomposition of voxelwise variance, as PC vectors would. However, we can assess the variance uniquely explained by each component by measuring the increase in explained variance from adding each IC vector to the subspace spanned by the other IC vectors. These measures were: 30% unique variance explained by the visual component, 7% by auditory, 8% by speech, and 5% by face-voice. Thus, each component uniquely explained an appreciable proportion of STS response variance.

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 a linear combination of four components: visual responses, auditory responses, responses to facial and vocal stimuli, and responses to auditory speech. Thus, 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 in responses across the STS and emerge from a data-driven analysis.

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, fSTS and vSTS. We found that the fSTS responded strongly to both face movements and vocal sounds, but weakly to hand movements or nonsocial control stimuli. These findings are consistent with our prior results showing strong responses to faces and voices but weak responses to whole-body movements (Deen et al., 2015), and suggest a specific role of this region in processing audio and visual signals from the face. 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. These findings argue against hypotheses that fSTS is specialized for processing audiovisual speech, or communicative sig-

nals more generally. 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 an abstract social feature crossmodally. The response profile of the fSTS contrasted with that of the adjacent vSTS, which had a selective response to auditory speech.

While prior work has documented overlapping posterior STS responses to faces and voices (Kreifelts et al., 2009; Watson et al., 2014a; Wright et al., 2003), the present result is striking in that the fSTS was defined as the maximally face-sensitive subregion of posterior STS in individual participants, and nevertheless it responded as strongly to vocal sounds as to faces. Face-responsive regions of middle and anterior STS were also found to have voice responses (Figure S3). Furthermore, a data-driven independent component analysis identified responses to faces and voices as a dominant source of voxelwise variance across the STS, with strongest voxel weights in posterior STS, and positive weights extending along the length of the STS in some participants. These results argue that "face regions" of the human STS (Haxby et al., 2000; Pitcher et al., 2011) are better characterized as "face-voice" regions, responsive to dynamic visual or auditory 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). This conclusion suggests a straightforward update to existing models of the human brain's face perception system (Bernstein and Yovel, 2015): the dorsal (STS) face processing stream is specialized not just for dynamic visual information from faces, but also dynamic auditory information from faces (see also Yovel and O'Toole, 2016).

What does the response profile of fSTS across multiple types of face and hand movement and vocal sound tell us about the functional role of this region? This region responded weakly to hand movements, even when communicative, suggesting against a role in processing any body movement. Among dynamic facial and vocal stimuli, however, the fSTS responded strongly to all stimulus categories presented—including speech and nonspeech, communicative and noncommunicative—and a similar pattern of response was observed for the face+voice component identified by ICA. This result argues against a strict specialization of this region for speech processing or social perceptual inference, instead pointing to a more general role in the multimodal perceptual processing of signals from faces. Such a region could plausibly contribute to a range of functions relying on audiovisual perceptual representations of face actions, including speech perception, social perception, and person identification. The broad response profile observed also suggests against the claim that voice responses within pSTS are specifically linked to mouth movement responses (Zhu and Beauchamp, 2017). While a small preference for stimuli with mouth movement was observed in the right hemisphere, the fSTS bilaterally responded strongly to both movements with and without a mouth component, and our prior work has found that a similarly defined region contains information about both eye and mouth movement type (Deen and Saxe, 2019). While our results don't contradict prior findings of subregions within posterior STS with preferences for eye or mouth movements (Pelphrey et al., 2005; Zhu and Beauchamp, 2017), they demonstrate that activations to any face movement or vocal sound constitute a dominant response profile across the STS.

While the fSTS responded strongly 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 finding 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. In providing evidence for crossmodal coding of a socially relevant dimension, these results are broadly consistent with findings of crossmodal emotional state decoding in a region of pSTS/middle temporal gyrus (Peelen et al., 2010), and crossmodal adaptation for emotional state information in a region of pSTS (Watson et al., 2014b). However, we note that the regions assessed in these two studies likely differ slightly from the area studied here: e.g., the region found by Peelen et al. was not face-selective, and the region found by Watson et al. was not voice-selective.

What do these results tell us about the role of the fSTS in social perception, the process of inferring abstract social properties from perceptual input? 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 might benefit from a hierarchical processing architecture. Brain regions positioned "lower" in the hierarchy 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 "higher" in the hierarchy 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). On this view, social perceptual inference involves an interplay of regions across the hierarchy, with feedforward connections transmitting updated sensory input, and feedback connections conveying predictions driven by high-level representations (Koster-Hale and Saxe, 2013).

Where is the fSTS situated in this putative hierarchy? The properties 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: fSTS responds to stimuli across multiple modalities (visual faces and auditory voices), and pattern analysis indicates that this region represents an abstract social property, in a manner that generalizes across modalities. Taken together, these results suggest that the fSTS plays a mid-level role 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 features across modalities.

What parts of the brain constitute "higher" regions in this hierarchy, with more explicit representations of abstract social information? Areas within higher-order association cortex implicated in high-level social cognition and theory of mind provide a plausible candidate network (Fletcher et al., 1995; Saxe and Kanwisher, 2003). These regions fall within the default mode network or apex network, situated at the top of the cortical sensory/motor processing hierarchy (DiNicola et al., 2020; Margulies et al., 2016), and have been found to contain abstract representations of features of others' internal states, including emotional states (Skerry and Saxe, 2014, 2015) and beliefs (Koster-Hale et al., 2014, 2017). This network contains a component in the anterior STS, and our prior work has found partial overlap between face movement and theory of mind responses within anterior STS (Deen et al., 2015). Thus, socially-sensitive subregions of the anterior STS could plausibly constitute a route through which information about dynamic face/voice signals is relayed from fSTS to areas involved in high-level social cognition. Future work should explore this possibility.

Beyond social perception, our results are consistent with prior studies implicating the posterior STS in the use of audiovisual information for speech perception and person identification. Studies using transcranial magnetic or direct current stimulation have found that disrupting the pSTS can disrupt audiovisual processing of speech content (Beauchamp et al., 2010; Marques et al., 2014; Riedel et al., 2015). fMRI studies have found sensitivity of pSTS responses to vocal identity (von Kriegstein et al., 2007, 2010), sensitivity to dynamic facial information relevant to identity has been hypothesized (Bernstein and Yovel, 2015; O'Toole et al., 2002), and recent studies have found ev-

idence for crossmodal identity representations (Anzellotti and Caramazza, 2017; Hasan et al., 2016; Tsantani et al., 2019). Furthermore, pSTS responses have been linked to benefits in auditory speech processing resulting from face-voice learning (Blank and von Kriegstein, 2013; von Kriegstein et al., 2008). However, we note that it is difficult to establish whether the studies mentioned above are in fact studying a common region of pSTS, or nearby but functionally distinct regions. Given differences in the precise anatomical location of functional regions across human participants, and differences in analysis and registration strategies across studies, finding responses in similar stereotaxic coordinates across studies does not demonstrate that these studies are assessing the same region (Brett et al., 2002); in fact, our prior work has demonstrated that nearby and even overlapping pSTS areas can have rather different response profiles (Deen et al., 2015). Using functional criteria to define regions in a consistent manner across studies provides one way to resolve this issue (Saxe et al., 2006).

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; Liebenthal et al., 2005; Scott et al., 2000; Vouloumanos et al., 2001; Wright et al., 2003), or vocal sounds more generally (Belin et al., 2002, 2000; Deen et al., 2015; Fecteau et al., 2004; Shultz et al., 2012), the present results suggest that this region is primarily specialized for speech processing. This result is consistent with a recent study assessing responses to a broad set of natural sounds, which found a response component localized to middle STS/STG with a much stronger response to speech than a variety of other sound categories, including nonspeech vocal sounds (Norman-Haignere et al., 2015; see also Pernet et al., 2015). Particularly striking here was the strong selectivity of vSTS 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 (Overath et al., 2015). Although our results suggest that vSTS is specialized for processing speech over arbitrary vocal sounds, this doesn't rule out a potential role for this region for voice identification, given that speech sounds are the primary cue humans use to determine voice identity (Latinus et al., 2013).

The vSTS also responded more strongly to visually presented speech over other types of face movement, suggesting a potential role in the visual processing of speech signals as well. This finding is consistent with prior studies finding mid-STS responses to visual speech (Callan et al., 2004; Calvert et al., 1997; Capek et al., 2008), and extends these studies by including a number of meaningful face movement controls, including communicative nonspeech mouth movements.

Considering the response profiles of the fSTS and vSTS together, our results indicate that the STS contains distinct pathways for 1) processing of facial and vocal signals in general (corresponding to the dorsal face processing pathway), and 2) processing of speech signals. This conclusion contrasts with the common notion that the STS is subdivided into areas for processing faces (Haxby et al., 2000) and vocal sounds (Belin et al., 2000). This view of STS functional organization was further supported by data-driven ICA results, in which face/voice-responsive and speech-selective components emerged as dominant response profiles, contributing largely independent sources of variance in voxelwise responses across the STS. While we designate the regions studied here as fSTS and vSTS based on the functional criteria used to define them (face and voice responses), these results suggest that fvSTS and spSTS would be more appropriate names.

How do these findings relate to our understanding of systems for face and voice processing in nonhuman primates? The dorsal face processing stream in humans has been argued to relate to a dorsal stream within the upper bank of the macaque STS, which contains regions that respond selectively to face motion (Fisher and Freiwald, 2015; Freiwald et al., 2016). The upper bank of the macaque STS primarily comprises a polysensory region, the superior temporal polysensory area (STP, also termed TPO; Bruce et al., 1981; Seltzer and Pandya, 1978), which contains neurons responsive to faces and vocal sounds, some of which show multimodal interactions (Barraclough et al., 2005; Ghazanfar et al., 2008; Perrodin et al., 2014). Thus, the claim that the dorsal face processing stream is multimodal is generally consistent with the anatomical positioning of macaque face motion areas. However, macaque fMRI studies on responses to vocal sounds have yielded mixed results within the STP (Gil-da-Costa et al., 2006; Joly et al., 2012; Petkov et al., 2008), with responses observed primarily within the superior temporal plane and posterior STP, not consistent in location with face-motion responses. Thus, while an evolutionary relationship between face-motion-sensitive areas of macaque STP and human STS remains plausible, it is not clear whether the macaque STP contains subregions with selective, fMRI-detectable responses to both face motion and vocal sounds, as we observe here in humans. Future studies should test this by directly measuring responses to face movements and vocal sounds within individual macaques.

Can the response profiles reported here be accounted for by differences across categories in low-level visual or acoustic features? Face motion videos had lower motion energy than hand movement or object videos, suggesting against the possibility that face responses were driven by motion per se (Fig. S1). While different categories of auditory stimuli were reasonably well matched on frequency content, categories differed somewhat in spectrotemporal modulation, with stronger 2–4 Hz modulation for speech stimuli (Fig. S2). Thus, we can't rule out the possibility that responses were influenced by differences in acoustic properties. However, the response profile of the fSTS across multiple categories—a strong response to speech, nonspeech communicative, and noncommunicative vocal sounds, and weak response to music and nonvocal environmental sounds (Deen et al., 2015)—is not easily accounted for in terms of responses to spectral or temporal modulation. Furthermore, decoding of communicativeness from fSTS patterns generalized across auditory and visual modalities, and thus can't be explained by low-level features. Could the heightened vSTS response to speech over nonspeech vocal sounds simply reflect the spectrotemporal complexity of speech? Recent work has found that speech responses in middle STS/STG are substantially reduced to synthetic sounds matched in spectrotemporal modulation statistics, suggesting against this explanation (Norman-Haignere and McDermott, 2018).

Could effects attributed here to communicativeness relate to a different high-level factor? The distinction between communicative and noncommunicative stimuli 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 the distinction drives a difference in responses. Importantly, this does not diminish the claim that the fSTS represents an abstract social dimension crossmodally.

Are the fSTS responses reported here contingent on the behavioral task used in the scanner? Here, we used a task that is unrelated to the stimulus distinctions of interest—a 1-back task on individual video/audio clips—to ensure that differences in response across categories cannot be explained by task effects. However, prior studies have found a modest influence of task on pSTS responses to visually presented faces, with stronger responses when participants attend to gaze direction or facial expression than to identity (Bernstein et al., 2018; Hoffman and Haxby, 2000). Future studies should investigate fSTS responses to audiovisual social stimuli in during tasks involving social perceptual inference.

Lastly, we note that while our ICA results show that face/voice and speech responses constitute dominant response profiles across the STS, they of course don't rule out the possibility that other meaningful response profiles exist within this large region. Response profiles that ac-

count for a small amount of variance in STS-wide responses, or that don't satisfy the model's assumption of spatial orthogonality of voxel weights among components, could have been missed by this method. Furthermore, our ability to identify dominant sources of response variance is intrinsically constrained by the stimulus set chosen: there could be features driving STS variance that don't vary across the particular stimuli used here. Thus, the current results shouldn't be considered a full characterization of response variability to audiovisual face actions within the STS, but rather an assessment of dominant response profiles to a set of broad categories that capture multiple theoretically relevant dimensions.

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. Spatial patterns of response in the fSTS differentiated communicative and noncommunicative stimuli across modalities (faces and voices), demonstrating that this region encodes an abstract social feature cross-modally. Future research should further detail the nature of representations of dynamic facial and vocal signals in these regions.

CRediT authorship contribution statement

Ben Deen: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Writing - original draft, Writing - review & editing, Visualization, Project administration. **Rebecca Saxe:** Conceptualization, Methodology, Writing - review & editing, Supervision. **Nancy Kanwisher:** Conceptualization, Methodology, Resources, Writing - review & editing, Supervision, Funding acquisition.

Acknowledgements

This research was funded by the NSF Center for Brains, Minds, and Machines (CCF- 1231216). B.D. was supported by an NSF graduate research fellowship and the Helen Hay Whitney Fellowship. The authors declare no competing financial interests.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2020.117191.

References

Allison, T., Puce, A., McCarthy, G., 2000. Social perception from visual cues: role of the STS region. Trends Cogn. Sci. (Regul. Ed.) 4, 267–278.
Anzellotti, S., Caramazza, A., 2017. Multimodal representations of person identity individuated with fMRI. Cortex 89, 85–97.
Barraclough, N.E., Xiao, D., Baker, C.I., Oram, M.W., Perrett, D.I., 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, M.S., Lee, K.E., Argall, B.D., Martin, A., 2004. Integration of auditory and visual information about objects in superior temporal sulcus. Neuron 41, 809–824.
Beauchamp, M.S., Nath, A.R., Pasalar, S., 2010. fMRI-guided transcranial magnetic stimulation reveals that the superior temporal sulcus is a cortical locus of the McGurk effect. J. Neurosci. 30, 2414–2417.
Beauchamp, M.S., Yasar, N.E., Frye, R.E., Ro, T., 2008. Touch, sound and vision in human superior temporal sulcus. Neuroimage 41, 1011–1020.
Belin, P., Zatorre, R.J., Ahad, P., 2002. Human temporal-lobe response to vocal sounds. Cognitive Brain Res. 13, 17–26.
Belin, P., Zatorre, R.J., Lafaille, P., Ahad, P., Pike, B., 2000. Voice-selective areas in human auditory cortex. Nature 403, 309–312.
Bernstein, M., Erez, Y., Blank, I., Yovel, G., 2018. An integrated neural framework for dynamic and static face processing. Sci. Rep. 8, 1–10.
Bernstein, M., Yovel, G., 2015. Two neural pathways of face processing: a critical evaluation of current models. Neurosci. Biobehav. Rev. 55, 536–546.
Binder, J.R., Frost, J.A., Hammeke, T.A., Bellgowan, P.S., Springer, J.A., Kaufman, J.N., Possing, E.T., 2000. Human temporal lobe activation by speech and nonspeech sounds. Cereb. Cortex 10, 512–528.
Blank, H., von Kriegstein, K., 2013. Mechanisms of enhancing visual–speech recognition by prior auditory information. Neuroimage 65, 109–118.
Brass, M., Schmitt, R.M., Spengler, S., Gergely, G., 2007. Investigating action understanding: inferential processes versus action simulation. Curr. Biol. 17, 2117–2121.
Brett, M., Johnsrude, I.S., Owen, A.M., 2002. The problem of functional localization in the human brain. Nat. Rev. Neurosci. 3, 243–249.
Bruce, C., Desimone, R., Gross, C.G., 1981. Visual properties of neurons in a polysensory area in superior temporal sulcus of the macaque. J. Neurophysiol. 46, 369–384.
Callan, D.E., Jones, J.A., Munhall, K., Kroos, C., Callan, A.M., Vatikiotis-Bateson, E., 2004. Multisensory integration sites identified by perception of spatial wavelet filtered visual speech gesture information. J. Cogn. Neurosci. 16, 805–816.
Calvert, G.A., Bullmore, E.T., Brammer, M.J., Campbell, R., Williams, S.C., McGuire, P.K., Woodruff, P.W., Iversen, S.D., David, A.S., 1997. Activation of auditory cortex during silent lipreading. Science 276, 593–596.
Capek, C.M., MacSweeney, M., Woll, B., Waters, D., McGuire, P.K., David, A.S., Brammer, M.J., Campbell, R., 2008. Cortical circuits for silent speechreading in deaf and hearing people. Neuropsychologia 46, 1233–1241.
Deen, B., Koldewyn, K., Kanwisher, N., Saxe, R., 2015. Functional organization of social perception and cognition in the superior temporal sulcus. Cereb. Cortex 25, 4596–4609.
Deen, B., Saxe, R., 2019. Parts-based representations of perceived face movements in the superior temporal sulcus. Hum. Brain Mapp. 40, 2499–2510.
DiCarlo, J.J., Zoccolan, D., Rust, N.C., 2012. How does the brain solve visual object recognition? Neuron 73, 415–434.
DiNicola, L.M., Braga, R.M., Buckner, R.L., 2020. Parallel distributed networks dissociate episodic and social functions within the individual. J. Neurophysiol. 123, 1144–1179.
Fecteau, S., Armony, J.L., Joanette, Y., Belin, P., 2004. Is voice processing species-specific in human auditory cortex? An fMRI study. Neuroimage 23, 840–848.
Fedorenko, E., Duncan, J., Kanwisher, N., 2013. Broad domain generality in focal regions of frontal and parietal cortex. Proc. Natl. Acad. Sci. 110, 16616–16621.
Fisher, C., Freiwald, W.A., 2015. Contrasting specializations for facial motion within the macaque face-processing system. Curr. Biol. 25, 261–266.
Fletcher, P.C., Happe, F., Frith, U., Baker, S.C., Dolan, R.J., Frackowiak, R.S., Frith, C.D., 1995. Other minds in the brain: a functional imaging study of "theory of mind" in story comprehension. Cognition 57, 109–128.
Freiwald, W., Duchaine, B., Yovel, G., 2016. Face processing systems: from neurons to real-world social perception. Annu. Rev. Neurosci. 39, 325–346.
Ghazanfar, A.A., Chandrasekaran, C., Logothetis, N.K., 2008. Interactions between the superior temporal sulcus and auditory cortex mediate dynamic face/voice integration in rhesus monkeys. J. Neurosci. 28, 4457–4469.
Gil-da-Costa, R., Martin, A., Lopes, M.A., Munoz, M., Fritz, J.B., Braun, A.R., 2006. Species-specific calls activate homologs of Broca's and Wernicke's areas in the macaque. Nat. Neurosci. 9, 1064–1070.
Greve, D.N., Fischl, B., 2009. Accurate and robust brain image alignment using boundary-based registration. Neuroimage 48, 63.
Hasan, B.A.S., Valdes-Sosa, M., Gross, J., Belin, P., 2016. "Hearing faces and seeing voices": amodal coding of person identity in the human brain. Sci Rep 6, 37494.
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, J.V., Hoffman, E.A., Gobbini, M.I., 2000. The distributed human neural system for face perception. Trends Cogn. Sci. (Regul. Ed.) 4, 223–233.
Hein, G., Doehrmann, O., Müller, N.G., Kaiser, J., Muckli, L., Naumer, M.J., 2007. Object familiarity and semantic congruency modulate responses in cortical audiovisual integration areas. J. Neurosci. 27, 7881–7887.
Hoffman, E.A., Haxby, J.V., 2000. Distinct representations of eye gaze and identity in the distributed human neural system for face perception. Nat. Neurosci. 3, 80–84.
Hyvärinen, A., Oja, E., 2000. Independent component analysis: algorithms and applications. Neural Netw. 13, 411–430.
Joly, O., Ramus, F., Pressnitzer, D., Vanduffel, W., Orban, G.A., 2012. Interhemispheric differences in auditory processing revealed by fMRI in awake rhesus monkeys. Cereb. Cortex 22, 838–853.
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., Richardson, H., Velez, N., Asaba, M., Young, L., Saxe, R., 2017. Mentalizing regions represent distributed, continuous, and abstract dimensions of others' beliefs. Neuroimage 161, 9–18.
Koster-Hale, J., Saxe, R., 2013. Theory of mind: a neural prediction problem. Neuron 79, 836–848.
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.
Latinus, M., McAleer, P., Bestelmeyer, P.E., Belin, P., 2013. Norm-based coding of voice identity in human auditory cortex. Curr. Biol. 23, 1075–1080.
Liebenthal, E., Binder, J.R., Spitzer, S.M., Possing, E.T., Medler, D.A., 2005. Neural substrates of phonemic perception. Cereb. Cortex 15, 1621–1631.
Marchini, J.L., Ripley, B.D., 2000. A new statistical approach to detecting significant activation in functional MRI. Neuroimage 12, 366–380.
Margulies, D.S., Ghosh, S.S., Goulas, A., Falkiewicz, M., Huntenburg, J.M., Langs, G., Bezgin, G., Eickhoff, S.B., Castellanos, F.X., Petrides, M., 2016. Situating the default-mode network along a principal gradient of macroscale cortical organization. Proc. Natl. Acad. Sci. 113, 12574–12579.
Marques, L.M., Lapenta, O.M., Merabet, L.B., Bolognini, N., Boggio, P.S., 2014. Tuning and disrupting the brain—Modulating the McGurk illusion with electrical stimulation. Front. Hum. Neurosci. 8, 533.
McGurk, H., MacDonald, J., 1976. Hearing lips and seeing voices. Nature 264, 746–748.
Mesgarani, N., Cheung, C., Johnson, K., Chang, E.F., 2014. Phonetic feature encoding in human superior temporal gyrus. Science 343, 1006–1010.

Noesselt, T., Rieger, J.W., Schoenfeld, M.A., Kanowski, M., Hinrichs, H., Heinze, H.-J., Driver, J., 2007. Audiovisual temporal correspondence modulates human multisensory superior temporal sulcus plus primary sensory cortices. J. Neurosci. 27, 11431–11441.

Norman-Haignere, S., Kanwisher, N.G., McDermott, J.H., 2015. Distinct cortical pathways for music and speech revealed by hypothesis-free voxel decomposition. Neuron 88, 1281–1296.

Norman-Haignere, S.V., McDermott, J.H., 2018. Neural responses to natural and model-matched stimuli reveal distinct computations in primary and nonprimary auditory cortex. PLoS Biol. 16, e2005127.

O'Toole, A.J., Roark, D.A., Abdi, H., 2002. Recognizing moving faces: a psychological and neural synthesis. Trends Cogn. Sci. (Regul. Ed.) 6, 261–266.

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

Peelen, M.V., Atkinson, A.P., Vuilleumier, P., 2010. Supramodal representations of perceived emotions in the human brain. J. Neurosci. 30, 10127–10134.

Pelphrey, K.A., Morris, J.P., 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, K.A., Morris, J.P., Michelich, C.R., 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.

Pernet, C.R., McAleer, P., Latinus, M., Gorgolewski, K.J., Charest, I., Bestelmeyer, P.E., Watson, R.H., Fleming, D., Crabbe, F., Valdes-Sosa, M., Belin, P., 2015. The human voice areas: spatial organization and inter-individual variability in temporal and extra-temporal cortices. Neuroimage 119, 164–174.

Perrodin, C., Kayser, C., Logothetis, N.K., Petkov, C.I., 2014. Auditory and visual modulation of temporal lobe neurons in voice-sensitive and association cortices. J. Neurosci. 34, 2524–2537.

Petkov, C.I., Kayser, C., Steudel, T., Whittingstall, K., Augath, M., Logothetis, N.K., 2008. A voice region in the monkey brain. Nat. Neurosci. 11, 367–374.

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

Poldrack, R.A., 2017. Precision neuroscience: dense sampling of individual brains. Neuron 95, 727–729.

Puce, A., Allison, T., Bentin, S., Gore, J.C., McCarthy, G., 1998. Temporal cortex activation in humans viewing eye and mouth movements. J. Neurosci. 18, 2188–2199.

Redcay, E., 2008. The superior temporal sulcus performs a common function for social and speech perception: implications for the emergence of autism. Neurosci. Biobehav. Rev. 32, 123–142.

Redcay, E., Velnoskey, K.R., Rowe, M.L., 2016. Perceived communicative intent in gesture and language modulates the superior temporal sulcus. Hum Brain Mapp 37, 3444–3461.

Reisberg, D., Mclean, J., Goldfield, A., 1987. Easy to hear but hard to understand: a lip-reading advantage with intact auditory stimuli. In: Dodd, B., Campbell, R. (Eds.), Hearing By eye: The psychology of Lip-Reading. Lawrence Erlbaum Associates, Inc., Hillsdale, NJ, pp. 97–113.

Riedel, P., Ragert, P., Schelinski, S., Kiebel, S.J., von Kriegstein, K., 2015. Visual face–movement sensitive cortex is relevant for auditory-only speech recognition. Cortex 68, 86–99.

Said, C.P., Moore, C.D., Engell, A.D., Todorov, A., Haxby, J.V., 2010. Distributed representations of dynamic facial expressions in the superior temporal sulcus. J Vis 10, 11.

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., Xiao, D.-K., Kovacs, G., Perrett, D., Kanwisher, N., 2004. A region of right posterior superior temporal sulcus responds to observed intentional actions. NeuropsychologiaNeuropsychologia 42, 1435–1446.

Schultz, J., Brockhaus, M., Bülthoff, H.H., Pilz, K.S., 2013. What the human brain likes about facial motion. Cereb. Cortex 23, 1167–1178.

Scott, S.K., Blank, C.C., Rosen, S., Wise, R.J., 2000. Identification of a pathway for intelligible speech in the left temporal lobe. Brain 123, 2400–2406.

Seltzer, B., Pandya, D.N., 1978. Afferent cortical connections and architectonics of the superior temporal sulcus and surrounding cortex in the rhesus monkey. Brain Res. 149, 1–24.

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

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

Skerry, A.E., Saxe, R., 2015. Neural representations of emotion are organized around abstract event features. Curr. Biol. 25, 1945–1954.

Srinivasan, R., Golomb, J.D., Martinez, A.M., 2016. A neural basis of facial action recognition in humans. J. Neurosci. 36, 4434–4442.

Sumby, W.H., Pollack, I., 1954. Visual contribution to speech intelligibility in noise. J. Acoust. Soc. Am. 26, 212–215.

Thompson, J.C., Hardee, J.E., Panayiotou, A., Crewther, D., Puce, A., 2007. Common and distinct brain activation to viewing dynamic sequences of face and hand movements. Neuroimage 37, 966–973.

Tsantani, M., Kriegeskorte, N., McGettigan, C., Garrido, L., 2019. Faces and voices in the brain: a modality-general person-identity representation in superior temporal sulcus. Neuroimage 201, 116004.

Van Atteveldt, N., Formisano, E., Goebel, R., Blomert, L., 2004. Integration of letters and speech sounds in the human brain. Neuron 43, 271–282.

von Kriegstein, K., Dogan, Ö., Grüter, M., Giraud, A.-L., Kell, C.A., Grüter, T., Kleinschmidt, A., Kiebel, S.J., 2008. Simulation of talking faces in the human brain improves auditory speech recognition. Proc. Natl. Acad. Sci. 105, 6747–6752.

von Kriegstein, K., Smith, D.R., Patterson, R.D., Ives, D.T., Griffiths, T.D., 2007. Neural representation of auditory size in the human voice and in sounds from other resonant sources. Curr. Biol. 17, 1123–1128.

von Kriegstein, K., Smith, D.R., Patterson, R.D., Kiebel, S.J., Griffiths, T.D., 2010. How the human brain recognizes speech in the context of changing speakers. J. Neurosci. 30, 629–638.

Vouloumanos, A., Kiehl, K.A., Werker, J.F., Liddle, P.F., 2001. Detection of sounds in the auditory stream: event-related fMRI evidence for differential activation to speech and nonspeech. J. Cogn. Neurosci. 13, 994–1005.

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

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

Woolrich, M.W., Ripley, B.D., Brady, M., Smith, S.M., 2001. Temporal autocorrelation in univariate linear modeling of FMRI data. Neuroimage 14, 1370–1386.

Wright, T.M., Pelphrey, K.A., Allison, T., McKeown, M.J., McCarthy, G., 2003. Polysensory interactions along lateral temporal regions evoked by audiovisual speech. Cereb. Cortex 13, 1034–1043.

Yovel, G., O'Toole, A.J., 2016. Recognizing people in motion. Trends Cogn. Sci. (Regul. Ed.) 20, 383–395.

Zhu, L.L., Beauchamp, M.S., 2017. Mouth and voice: a relationship between visual and auditory preference in the human superior temporal sulcus. J. Neurosci. 37, 2697–2708.