Behavioral/Cognitive

Mouth and Voice: A Relationship between Visual and Auditory Preference in the Human Superior Temporal Sulcus

Lin L. Zhu[1,2] **and** **Michael S. Beauchamp**[2]

[1]Medical Scientist Training Program, Graduate Program in Structural and Computational Biology and Molecular Biophysics, and [2]Department of Neurosurgery and Core for Advanced MRI, Baylor College of Medicine, Houston, Texas 77030

Cortex in and around the human posterior superior temporal sulcus (pSTS) is known to be critical for speech perception. The pSTS responds to both the visual modality (especially biological motion) and the auditory modality (especially human voices). Using fMRI in single subjects with no spatial smoothing, we show that visual and auditory selectivity are linked. Regions of the pSTS were identified that preferred visually presented moving mouths (presented in isolation or as part of a whole face) or moving eyes. Mouth-preferring regions responded strongly to voices and showed a significant preference for vocal compared with nonvocal sounds. In contrast, eye-preferring regions did not respond to either vocal or nonvocal sounds. The converse was also true: regions of the pSTS that showed a significant response to speech or preferred vocal to nonvocal sounds responded more strongly to visually presented mouths than eyes. These findings can be explained by environmental statistics. In natural environments, humans see visual mouth movements at the same time as they hear voices, while there is no auditory accompaniment to visual eye movements. The strength of a voxel's preference for visual mouth movements was strongly correlated with the magnitude of its auditory speech response and its preference for vocal sounds, suggesting that visual and auditory speech features are coded together in small populations of neurons within the pSTS.

Key words: audiovisual; face; multisensory; speech perception

Significance Statement

Humans interacting face to face make use of auditory cues from the talker's voice and visual cues from the talker's mouth to understand speech. The human posterior superior temporal sulcus (pSTS), a brain region known to be important for speech perception, is complex, with some regions responding to specific visual stimuli and others to specific auditory stimuli. Using BOLD fMRI, we show that the natural statistics of human speech, in which voices co-occur with mouth movements, are reflected in the neural architecture of the pSTS. Different pSTS regions prefer visually presented faces containing either a moving mouth or moving eyes, but only mouth-preferring regions respond strongly to voices.

Introduction

It has been known since the time of Wernicke that cortex in and around the human posterior superior temporal sulcus (pSTS) is critical for speech perception. Face-to-face communication is fundamentally multisensory: during conversations, we see the talker's face and hear the talker's voice. Consistent with the multisensory nature of speech perception, the pSTS itself is also multisensory, showing an enhanced response to multisensory audiovisual stimuli (Wright et al., 2003; Beauchamp et al., 2004a; van Atteveldt et al., 2004; Kreifelts et al., 2007; Stevenson et al., 2007) that is subject to cross-modal adaptation (Watson et al., 2014b). These results and others have led to a consensus that the pSTS plays a key role in associating visual and auditory stimuli to create multisensory representations.

In the visual domain, the pSTS is especially responsive to biological motion, the movement of living organisms (Puce et al., 1998; Grossman et al., 2000, 2005; Beauchamp et al., 2002, 2003; Santi et al., 2003; Pitcher et al., 2011). Within the domain of biological motion, the pSTS prefers mouth movements that serve a communicative function (Calvert et al., 1997), responding more to naturally moving speaking faces than still frame images of speech gestures

Received Sept. 13, 2016; revised Jan. 17, 2017; accepted Jan. 24, 2017.

Author contributions: L.L.Z. and M.S.B. designed research; L.L.Z. and M.S.B. performed research; L.L.Z. analyzed data; L.L.Z. and M.S.B. wrote the paper.

This work was supported by the National Institutes of Health (R01NS065395 to M.S.B, F30DC014911 to L.L.Z.) and Baylor Research Advocates for Student Scientists (L.L.Z.). We acknowledge the Core for Advanced MRI and MR technologist Lacey Berry for scanning assistance.

The authors declare no competing financial interests.

Correspondence should be addressed to Michael S. Beauchamp, PhD, Core for Advanced MRI, 1 Baylor Plaza, S104, Houston, TX 77030. E-mail: michael.beauchamp@bcm.edu.

DOI:10.1523/JNEUROSCI.2914-16.2017

Table 1. Stimuli and tasks

Stimulus modality	Stimulus type	Task
Visual only	1. Full-face mouth movements	Identity task (which of the two actors was shown in each video)
	2. Full-face eye movements	
	3. Masked-face mouth movements	
	4. Masked-face eye movements	
Auditory only	1. Short stories	Passive listening
	2. Vocal sounds	One-back same/different
	3. Nonvocal sounds	One-back same/different

(Bernstein et al., 2002; Calvert and Campbell, 2003). The pSTS is also attuned to eye movement, which serves as an important social cue (Hoffman and Haxby, 2000; Hooker et al., 2003).

In the auditory domain, the pSTS also shows strong selectivity. Regions of pSTS are more responsive to human speech and nonspeech vocalizations than to nonhuman environmental or mechanical sounds (Belin et al., 2000) or animal vocalizations (Fecteau et al., 2004). The pSTS demonstrates a preference for communicative (e.g., speech, laughter) over noncommunicative (e.g., coughs, hiccups) auditory signals (Shultz et al., 2012); pseudowords over tones (Binder et al., 2000) or sine-wave analogs (Vouloumanos et al., 2001); and words over signal-correlated noise (Wise et al., 2001).

While we know a great deal about visual and auditory selectivity of the pSTS, little is known about the relationship between these different axes of selectivity. From an early age, humans hear speech while observing mouth movements. Given the importance of the pSTS in associating visual and auditory stimuli, a subset of neurons in the pSTS might code for features of visual speech and auditory speech, forming a neural association between them. If this is true, then some regions of the pSTS with a preference for visual mouth movements might be expected to respond strongly to auditory speech; conversely, regions of the pSTS with a preference for auditory speech should respond to visual mouth movements. To test these hypotheses, we performed a series of fMRI experiments in which subjects watched silent videos containing visual mouth movements or eye movements and listened to continuous speech and isolated vocal and nonvocal sounds.

Materials and Methods

Subjects and stimuli. Twenty healthy subjects participated in the fMRI experiment (seven female; mean age, 26 years). The subjects provided informed written consent under an experimental protocol approved by the Institutional Review Board at Baylor College of Medicine.

All stimuli were presented in Matlab (Mathworks; RRID:SCR_001622) using the Psychophysics Toolbox extensions (RRID:SCR_002881; Brainard, 1997; Pelli, 1997). As shown in Table 1, there were four kinds of visual stimuli, grouped in a two-by-two design. Visual stimuli were presented on a 32 inch 1920 × 1080 MR-compatible BOLDview LCD screen (Cambridge Research Systems) and viewed through a mirror attached to the head coil. The stimuli consisted of videos of actors either moving their mouth (silently mouthing the vowels "aa," "ee," and "oo") or moving their eyes (glancing to the left and glancing to the right). Each video was edited using Final Cut Pro so that the duration was 2 s, and each video started and ended with the speaker in a neutral facial position, with the eye or mouth movement beginning at 500 ms and lasting for 1 s. Videos were manipulated in Psychophysics Toolbox such that they either contained the whole face or were masked so that only the mouth or only the eyes were visible.

Auditory stimuli were played via MR-compatible insert headphones (Sensimetrics). The first auditory stimulus set consisted of a female actor reading 12 different Aesop's fables. The second and third stimuli sets were a subset of those used by Capilla et al. (2013) and consisted of 24 recordings of vocal nonword speech sounds (e.g., "ahh") and 24 recordings of nonvocal nonspeech natural stimuli (e.g., waterfall sound). Each recording lasted 500 ms.

fMRI visual task design. Visual stimuli were presented in a block design. Each 30 s block contained 10 2 s videos of a single type (either mouth or eye movements) with a 1 s response period following each video. An orthogonal task was used in which subjects used a fiber-optic button response pad (Current Designs) to press one of two buttons corresponding to the identity of the actor shown in the video. Each scan series contained six visual stimulation blocks (three each of mouth movements and eye movements) interspersed with 10 s of fixation baseline. Subjects underwent three scan series that contained alternating blocks of full-face stimuli and fixation, and three scan series that contained alternating blocks of masked-face stimuli and fixation.

fMRI auditory task design. In two scan series, subjects listened to auditory short stories presented in 20 s blocks alternated with 20 s of fixation baseline. Subjects were instructed to keep their eyes open and passively listen to the stories but otherwise had no task. In five subjects, only one scan series of auditory short stories was presented; instead, they underwent two additional scan series where they were presented with vocal and nonvocal sounds. In these scan series, there were eight 19.5 s blocks separated by 10 s of fixation baseline. Each block contained 13 clips of either vocal or nonvocal sounds. Subjects performed a one-back same/different task, pressing a response button only for stimuli that repeated.

Statistical fMRI approach. A potential pitfall in fMRI analysis is the use of statistical criteria that are not independent (Baker et al., 2007; Simmons et al., 2007; Vul et al., 2009; Vul and Pashler, 2012). To avoid this problem, we created statistical maps from scan series that contained only visual stimuli and applied them to completely independent data collected in scan series that contained only auditory stimuli. There has also been recent controversy over the use of familywise error correction (Eklund et al., 2016). Our analysis used only voxelwise thresholding without any clustering or familywise error correction.

To classify visually responsive voxels, we used a two-step process (Haxby et al., 1999; Ishai et al., 1999, 2000). First, an initial strict threshold based on the full-F (omnibus) test was used to winnow the large number of voxels in the whole brain into a much smaller number of voxels that responded to visual stimuli. For this step, we used a conservative threshold of $F > 7$, $p < 5 \times 10^{-6}$, $q < 1 \times 10^{-4}$, corrected for voxelwise multiple comparisons using the false discovery rate procedure (Genovese et al., 2002). Second, a threshold of $t > 2$, $p < 0.05$ was used to classify visually responsive voxels as preferring mouth movements, preferring eye movements, or showing no preference. This threshold was varied in additional analyses described below.

To classify auditory-responsive voxels, a similar process was used. First, the auditory full-F test ($F > 7$, $p < 5 \times 10^{-6}$, $q < 1 \times 10^{-4}$) was used to identify auditory-responsive voxels. Second, voxels were classified as voice-preferring if they showed a stronger response to vocal sounds than to nonvocal sounds ($t > 2$, $p < 0.05$ uncorrected). Voxels were classified as voice-responsive if they responded more to auditory short stories than baseline ($t > 2$, $p < 0.05$ uncorrected).

General fMRI methods. All MRI data were collected at the Core for Advanced MRI at Baylor College of Medicine using a 3 tesla Siemens Trio MR scanner equipped with a 32-channel head gradient coil. Two T1-weighted MP-RAGE anatomical scans were collected at the beginning of each experimental session (1.0 mm isotropic voxels, 176 slices). The two anatomical scans were aligned to each other and averaged; the resulting average anatomical scan was then used to create a cortical surface model in FreeSurfer (RRID:SCR_001847; Dale et al., 1999; Fischl et al., 1999). T2*-weighted images were collected using gradient-echo echoplanar imaging. Fifteen subjects were scanned with TR = 2.0 s and 2.5 mm isotropic voxels, and five subjects were scanned with TR = 1.5 s and 2.0 mm isotropic voxels.

Single-subject analysis. Data analysis was performed using the Analysis of Functional NeuroImages (AFNI; RRID:SCR_005927) program (Cox, 1996). For each subject, we first performed slice-time correction for each functional scan; we then aligned functional scans to the average anatomical image and performed motion correction. We performed a regression analysis in each voxel using a generalized linear model using the AFNI function 3dDeconvolve. For each experiment, a regressor was created for

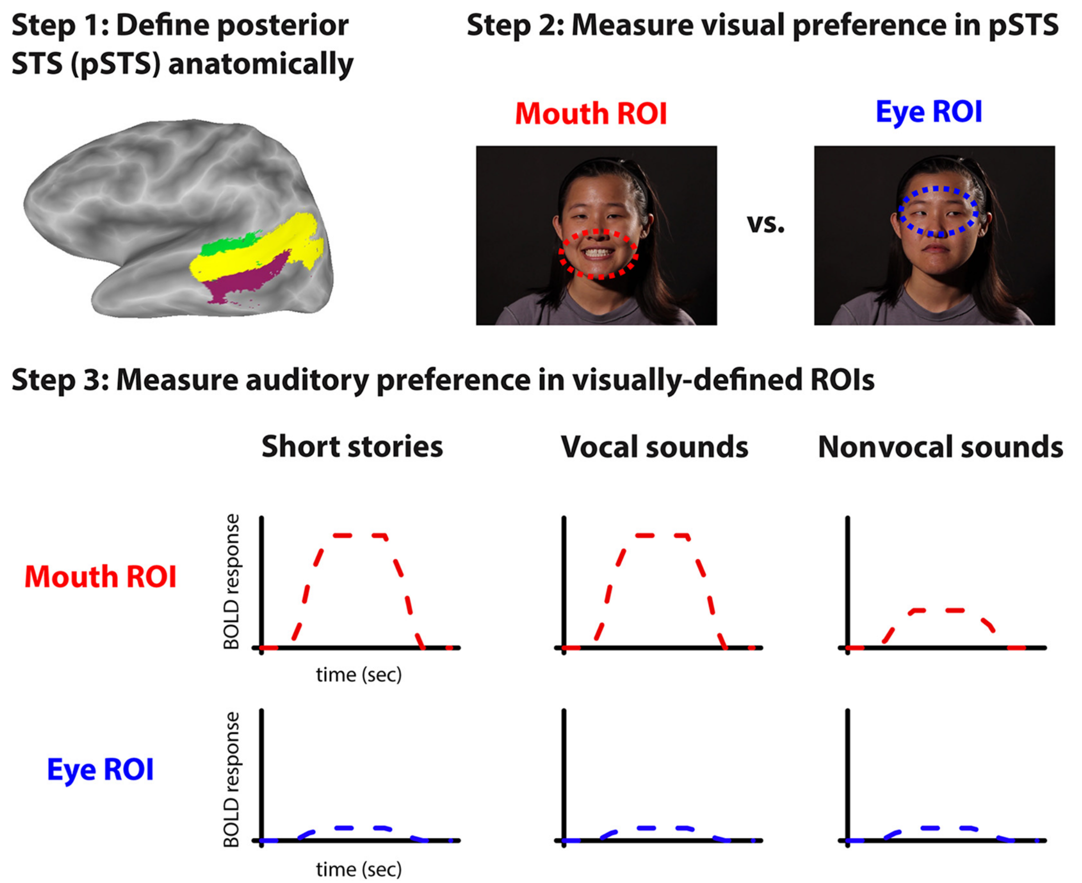

Figure 1. Overview of experimental strategy. Step 1, A cortical surface-based atlas (Destrieux et al., 2010) was used to create an anatomical ROI consisting of the posterior half of the superior temporal sulcus (yellow) and adjacent cortex in the superior temporal gyrus (green) and middle temporal gyrus (purple). Step 2, Visual stimuli consisting of videos of actors making mouth movements or eye movements were presented (location of movements highlighted with colored ellipse, not present in actual video). Within the anatomical ROI, all voxels preferring visually presented mouth movements were grouped into a mouth ROI; voxels preferring eye movements were grouped into an eye ROI. Step 3, Auditory stimuli consisting of short stories, vocal sounds, and nonvocal sounds were presented and the response in the ROIs calculated. The hypothesized responses are shown (see text for details).

each stimulus type by convolving the stimulus onset times with a canonical hemodynamic response function. Baseline drift and six spatial movement parameters were included as regressors of no interest. We first performed deconvolution with block functions to estimate the average percentage signal change to each stimulus type in each voxel. To estimate the impulse response function to each stimulus type in each voxel, a separate deconvolution was performed with tent functions that spanned the time between stimulus onset and 10 s after stimulus offset. This allowed for an unbiased, assumptions-free estimate of the actual BOLD response to each stimulus type (similar to a time-locked average) rather than merely fitting a prespecified function, such as a gamma variate (Glover, 1999).

Analyses were conducted on regions of interest (ROIs) defined individually in each subject. We first combined the Freesurfer-defined superior temporal gyrus, superior temporal sulcus, and middle temporal gyrus into a single ROI (Destrieux et al., 2010). We then selected the posterior half of this ROI using a cutoff placed at the midpoint of the full anterior-to-posterior extent of the ROI. The average location of the cutoff was at $y = -25 \pm 3$ mm (in Talairach space). For reference, the average location of the posterior border of Heschl's gyrus across subjects was $y = -29 \pm 3$ mm.

Group analysis. We used auto_tlrc and adwarp to transform each subject's anatomical and functional datasets into standard space, using the Colin 27 brain as a reference template (Holmes et al., 1998); group analysis was performed using 3dttest++. Analogous to the two-step procedure used for single-subject analysis, we first selected all voxels in the group map that showed a significant response to visual mouth movements or visual eye movements versus fixation ($t > 4$, $q < 0.05$, false discovery rate corrected) and then looked for voxels that showed a significant preference for either mouth or eye movements ($t > 2$, $p < 0.05$).

Results

As summarized in Figure 1, our experimental strategy was to identify regions of posterior superior temporal sulcus, superior temporal gyrus, and middle temporal gyrus (collectively referred to as pSTS for brevity) preferring visually presented eye movements or mouth movements and then measure the responses of these regions to auditory stimuli in independent scan series.

Response to visual face movements

We first identified visually responsive voxels in the pSTS and classified them as preferring visually presented mouth movements or eye movements. Activation maps for two sample subjects are shown in Figure 2*A*. All mouth-preferring voxels were grouped into a mouth ROI (average volume across subjects was 633 ± 110 mm^3 in the left hemisphere and 901 ± 176 mm^3 in the right hemisphere) and all eye-preferring voxels were grouped into an eye ROI (293 ± 70 mm^3 in the left hemisphere and 375 ± 71 mm^3 in the right hemisphere).

Response to auditory stimuli

Next, we examined the responses to auditory stimuli within the ROIs constructed from the responses to visual stimulation. Based

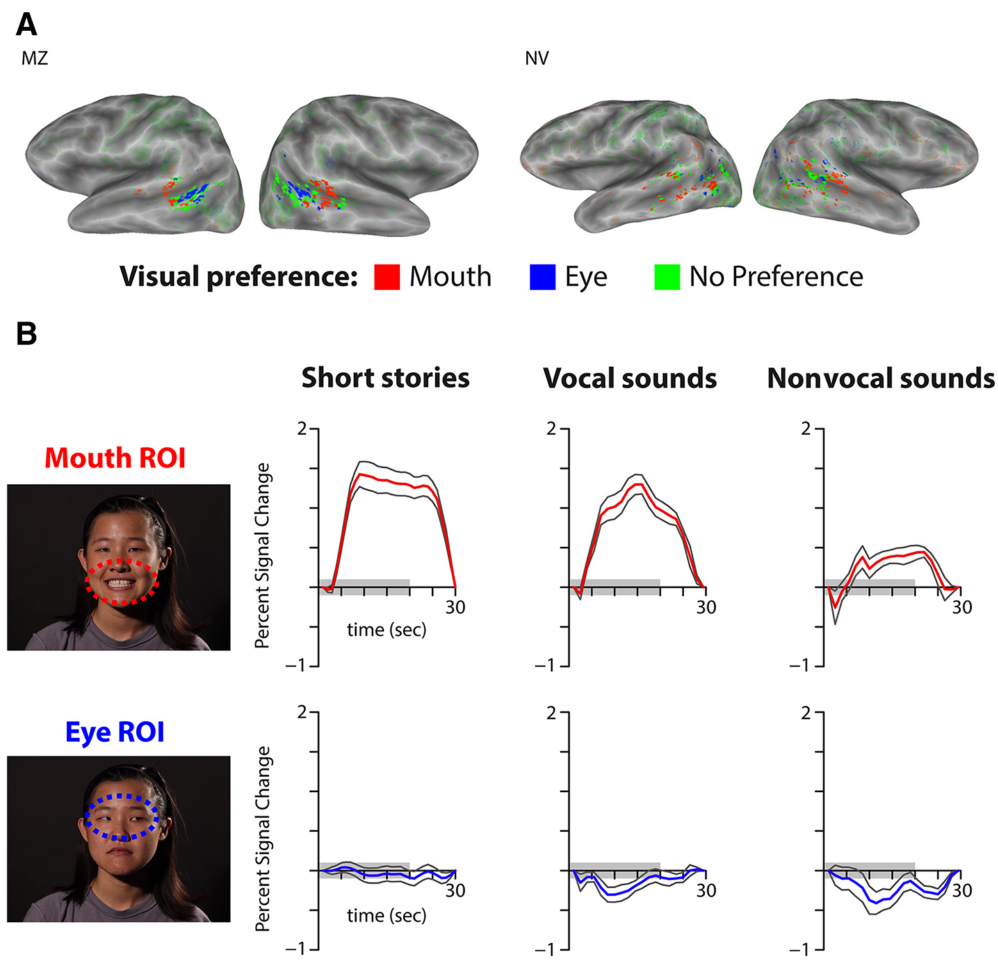

Figure 2. Activation maps and average responses. ***A***, Lateral views of left and right hemisphere cortical surface models showing visual preference maps in two subjects (for maps from the other 18 subjects, see Fig. 5). Mouth-preferring voxels ($t > 2$, $p < 0.05$) are shown in red; eye-preferring voxels ($t < -2$) are shown in blue; voxels with significant visual responses but no preference ($|t| < 2$) are shown in green. Voxels outside the pSTS shown with reduced opacity. ***B***, Average response to auditory stimuli across all subjects, averaged across hemispheres. Top row shows responses in the mouth ROI (all mouth-preferring voxels). Bottom row shows responses in the eye ROI (all eye-preferring voxels). Columns show responses to short stories ($n = 40$), vocal sounds ($n = 10$), and nonvocal sounds ($n = 10$). Gray lines denote SEM, and the gray bar along the x-axis denotes stimulus duration.

on the environmental correspondence between visual mouth movements and auditory speech, our prediction was that the mouth ROI should respond to auditory speech. Because there is no environmental correspondence between visual eye movements and auditory speech, our prediction was that the eye ROI should not respond to auditory speech.

Response to short stories

Subjects listened to blocks of short stories separated by fixation baseline. Across subjects, the mouth ROI responded strongly to auditory stories in both hemispheres (Fig. 2*B*; left hemisphere: 1.6 ± 0.2%, one-sample t test across subjects, $t_{(19)} = 7.0$, $p = 1.3 \times 10^{-6}$; right hemisphere: 1.2 ± 0.2%, $t_{(19)} = 5.9$, $p = 1.1 \times 10^{-5}$). In contrast, the eye ROI showed near-zero responses to auditory stories in both hemispheres (left: 0.1 ± 0.1%, $t_{(19)} = 0.61$, $p = 0.55$; right: −0.2 ± 0.1%, $t_{(19)} = -1.9$, $p = 0.07$). We confirmed the different response patterns in the mouth and eye ROIs using an ANOVA with BOLD response to auditory stories as the dependent measure and ROI (mouth vs eye) and hemisphere (left vs right) as factors. The only significant effect was a main effect of ROI ($F_{(1,76)} = 68$, $p = 3.3 \times 10^{-12}$) driven by a greater response to auditory stories in the mouth ROI.

Response to vocal and nonvocal sounds

Responsiveness to auditory stories in the mouth ROI could reflect either a general responsiveness to any auditory stimuli or a specific response to voices. Because viewing mouth movements and hearing vocal speech are tightly linked, we predicted that the mouth ROI would show a greater response to vocal stimuli compared with nonvocal stimuli. To test this prediction, we scanned five subjects as they listened to blocks of vocal nonword speech sounds (e.g., "aah") or nonvocal environmental sounds (e.g., recording of a waterfall). The mouth ROI responded to both types of sounds, but significantly preferred vocal to nonvocal sounds (Fig. 2*B*; left hemisphere: 1.2 ± 0.2% for vocal vs 0.4 ± 0.1% for nonvocal, $t_{(4)} = 5.2$, $p = 0.007$; right hemisphere: 1.2 ± 0.2% vs 0.6 ± 0.2%, $t_{(4)} = 13$, $p = 2.0 \times 10^{-4}$). In contrast, the eye ROI responded weakly to both types of auditory stimuli (Fig. 2*B*; left hemisphere: −0.2 ± 0.1% vs −0.3 ± 0.2%, $t_{(4)} = 0.6$, $p = 0.58$; right hemisphere: −0.1 ± 0.1% vs −0.2 ± 0.2%, $t_{(4)} = 1.1$, $p = 0.35$).

To quantify the different response patterns in the mouth and eye ROIs, we performed an ANOVA with BOLD response as the dependent measure and ROI (mouth vs eye), auditory stimulus (vocal vs nonvocal), and hemisphere (left vs right) as factors. There was a significant main effect of region ($F_{(1,32)} = 90$, $p =$

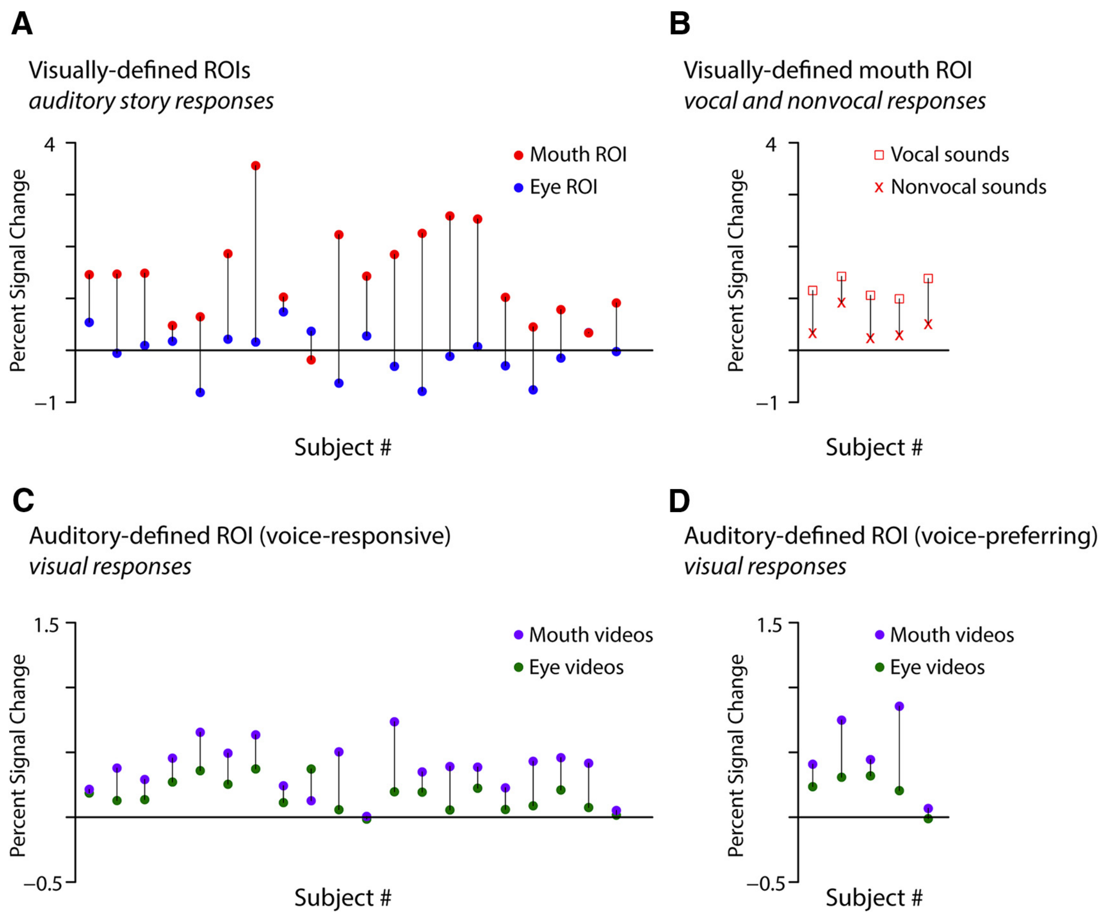

Figure 3. Consistency across subjects. ***A***, BOLD responses in mouth and eye ROIs to auditory short stories for 20 individual subjects, averaged across hemispheres. Each subject is represented by a vertical line, where the red circle represents the auditory response in the mouth ROI and the blue circle represents the auditory response in the eye ROI. ***B***, BOLD responses in mouth ROIs to vocal and nonvocal sounds for five individual subjects, averaged across hemispheres. The red square represents the response to vocal sounds, and the red X represents the response to nonvocal sounds. ***C***, BOLD responses in voice-responsive ROIs to mouth and eye movement videos for 20 individual subjects, averaged across hemispheres. The purple circle represents the response to visual mouth movements and the green circle represents the response to visual eye movements. ***D***, BOLD responses in voice-preferring ROIs to mouth and eye movement videos for five individual subjects, averaged across hemispheres. The purple circle represents the response to visual mouth movements and the green circle represents the response to visual eye movements.

7.9×10^{-11}) driven by a greater response to all auditory stimuli in the mouth ROI, and of stimulus type ($F_{(1,32)} = 16$, $p = 3.8 \times 10^{-4}$) driven by larger overall responses to vocal sounds; there was no main effect of hemisphere ($F_{(1,32)} = 1.5$, $p = 0.23$). The only significant interaction was between region and stimulus type ($F_{(1,32)} = 8.6$, $p = 0.006$) driven by a larger response to vocal sounds in the mouth ROI.

Consistency across subjects and analysis strategies

To determine the consistency of these effects, we examined single-subject data. In 18 of 20 subjects, the response to auditory short stories was greater in the mouth ROI than in the eye ROI (Fig. 3*A*). In five of five subjects, the response in the mouth ROI to vocal sounds was greater than the response to nonvocal sounds (Fig. 3*B*).

Our initial analysis demonstrated that regions of the pSTS defined by their preference for visual mouth movements responded strongly to auditory speech and preferred vocal to nonvocal auditory stimuli. However, if encoding of visual mouth movements and auditory voices occurs in the same population of neurons, we might expect the reverse to be true as well. That is, regions of the pSTS defined by their response to speech (or preference for voices) are predicted to respond more strongly to mouth movements than eye movements. To test this prediction, voice-responsive regions were selected by forming ROIs from all voxels that showed a significant response to auditory short stories (mean volume of voice-responsive ROI in left hemisphere: 4450 ± 327 mm^3; right hemisphere: 3361 ± 290 mm^3). The voice-responsive ROI showed a significantly greater response to visually presented mouth movements than eye movements (left hemisphere: $0.26 \pm 0.05\%$ vs $0.10 \pm 0.03\%$, $t_{(19)} = 4.2$, $p = 4.6 \times 10^{-4}$; right hemisphere: $0.48 \pm 0.05\%$ vs $0.24 \pm 0.03\%$, $t_{(19)} = 5.9$, $p = 1.1 \times 10^{-5}$) and this effect was consistent (19 of 20 subjects; Fig. 3*C*). An ANOVA on BOLD response amplitude with stimulus (mouth vs eye movements) and hemisphere (left vs right) as factors showed a main effect of stimulus ($F_{(1,76)} = 25$, $p = 3.0 \times 10^{-6}$) driven by a greater response to mouth movements than eye movements and a main effect of hemisphere ($F_{(1,76)} = 19$, $p = 3.6 \times 10^{-5}$) driven by greater responses in the right hemisphere. There was no interaction between stimulus and hemisphere.

Voice-preferring regions were selected by forming ROIs from all voxels that preferred vocal sounds to nonvocal sounds (mean volume of voice-preferring ROI in left hemisphere: 1925 ± 318 mm^3; right hemisphere: 1898 ± 358 mm^3). The voice-preferring ROI showed a significantly greater response to visually presented mouth movements than eye movements (left hemisphere: $0.38 \pm 0.11\%$ vs $0.13 \pm 0.08\%$, $t_{(4)} = 3.0$, $p = 0.04$; right hemisphere:

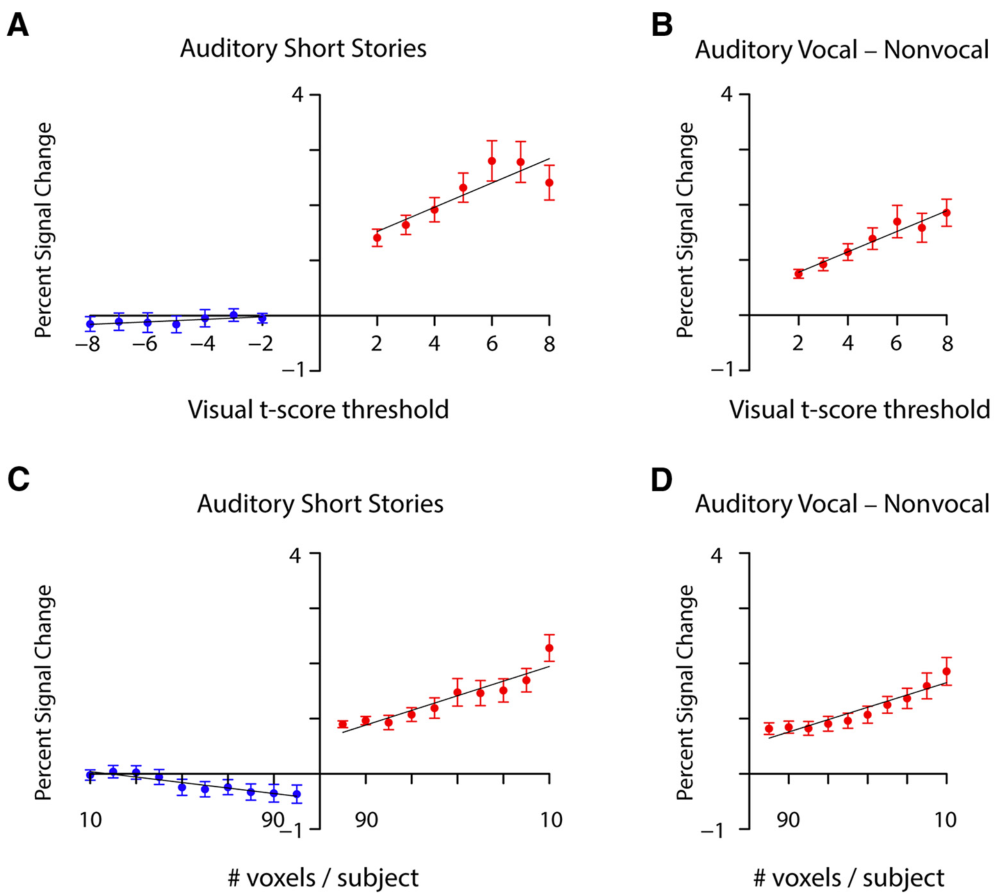

Figure 4. Effect of visual selectivity threshold on auditory responses. ***A***, Mean response to auditory short stories across hemispheres ($n = 40$) plotted as a function of the threshold used to classify voxels as mouth-preferring for inclusion in the mouth ROI (red) or eye-preferring for inclusion in the eye ROI (blue). More extreme *t* scores indicate stronger preference for visual mouth or eye stimuli. Error bars represent SEM. ***B***, Mean Vocal − Nonvocal response across hemispheres plotted as a function of visual *t*-score threshold ($n = 10$). More extreme *t* scores indicate stronger preference for visual mouth stimuli. Error bars represent SEM. ***C***, Mean response to auditory short stories across hemispheres ($n = 40$) plotted as a function of the *x*-most mouth-preferring (blue) or eye-preferring (red) voxels included per subject. Smaller values of *x* indicate stronger overall preference for visual mouth or eye stimuli. Error bars represent SEM. ***D***, Mean Vocal − Nonvocal response across hemispheres plotted as a function of *x* ($n = 10$). Smaller values of *x* indicate stronger overall preference for visual mouth stimuli. Error bars represent SEM.

$0.63 \pm 0.19\%$ vs $0.30 \pm 0.11\%$, $t_{(4)} = 2.4$, $p = 0.07$) and this effect was consistent (five of five subjects; Fig. 3*D*). An ANOVA on BOLD response amplitude including stimulus (mouth vs eye movements) and hemisphere (left vs right) as factors showed a main effect of stimulus ($F_{(1,16)} = 5.0$, $p = 0.04$) driven by a larger response to mouth movements than eye movements. There was no main effect of hemisphere and no interaction between stimulus and hemisphere.

Effect of threshold

Our initial analysis classified visually responsive voxels as mouth-preferring or eye-preferring using a threshold of $t > 2$, $p < 0.05$. Choosing a more stringent threshold of $t > 5$, $p < 10^{-6}$ did not change the pattern of responses: the mouth ROI continued to respond significantly to auditory stories (left hemisphere: $2.5 \pm 0.3\%$, one-sample *t* test $t_{(15)} = 7.6$, $p = 1.5 \times 10^{-6}$; right hemisphere: $2.1 \pm 0.4\%$, $t_{(17)} = 5.3$, $p = 6.4 \times 10^{-5}$) while the eye ROI did not (left hemisphere: $0 \pm 0.3\%$, $t_{(9)} = 0.02$, $p = 0.98$; right hemisphere: $-0.3 \pm 0.2\%$, $t_{(14)} = -1.4$, $p = 0.18$). An ANOVA with BOLD response to auditory stories as the dependent measure and ROI (mouth vs eye) and hemisphere (left vs right) as factors revealed a significant main effect of ROI ($F_{(1,55)} = 52$, $p = 1.5 \times 10^{-9}$) driven by a greater response to auditory stories in the mouth ROI.

More generally, if single neurons that respond to visual mouth movements also respond to auditory vocalizations, then we might expect voxels with greater mouth preference (due to a larger proportion of neurons responding to mouth movements) to also respond more strongly to auditory vocalizations (due to the same high proportion of neurons responding to vocalizations). To determine whether this was the case, we varied the criteria for classifying a voxel as mouth-preferring across a broad range, from $t = 2$ to $t = 8$ (Fig. 4*A*). As the *t*-score threshold for classifying a voxel as mouth-preferring increased, the response to auditory speech also increased, from an auditory speech response of 1.4% at a visual threshold of $t = 2$ (averaged across left hemisphere and right hemisphere) to a maximal response of 2.8% at $t = 6$ before declining slightly to 2.4% at $t = 8$. There was a significant positive correlation between visual mouth preference threshold and auditory response in both hemispheres (left hemisphere: $m = 0.24$, $r^2 = 0.86$, $p = 0.003$; right hemisphere: $m = 0.21$, $r^2 = 0.67$, $p = 0.03$). In contrast, eye-preferring voxels showed little response to auditory-only stories, vocal stimuli, or nonvocal stimuli, regardless of the threshold used to classify them as eye-preferring (Fig. 4*A*).

A similar analysis was performed by subtracting the BOLD response to nonvocal sounds from the response to vocal sounds

Table 2. Group analysis

ROI	Size (mm^3)	Talairach coordinates (mm) x	y	z
Mouth-preferring				
Right hemisphere, primary visual	438	28	−86	3
Left hemisphere, primary visual	200	−24	−90	2
Right hemisphere, STS	70	50	−37	8
Eye-preferring				
Left hemisphere, putamen	17	−25	−2	6

Size and location of mouth-preferring and eye-preferring clusters identified during the whole-brain group analysis.

(Vocal − Nonvocal; Fig. 4B). As the visual category-selectivity threshold was increased to find only the most mouth-preferring voxels, the Vocal − Nonvocal values also increased, from 0.7% at $t = 2$ to a maximum of 1.8% at $t = 8$ (response amplitudes averaged across hemispheres). There was a significant positive correlation between visual category selectivity and Vocal − Nonvocal values in both hemispheres (left: $m = 0.21$, $r^2 = 0.93$, $p = 3.9 \times 10^{-4}$; right: $m = 0.16$, $r^2 = 0.96$, $p = 1.1 \times 10^{-4}$).

On average, the mouth ROI contained more than double the number of voxels as the eye ROI. To control for this possible confound, we performed an analysis in which we equated the number of voxels in the mouth and eye ROIs by selecting the x-most mouth-preferring and eye-preferring voxels in each subject's anatomically defined pSTS ROI. As x decreased for the mouth ROI (selecting only those voxels with the strongest preference for visual mouth movements), the BOLD response to auditory short stories increased from 0.9% at 100 voxels/subject to 2.3% at 10 voxels/subject ($m = 0.013$, $r^2 = 0.86$, $p = 7.9 \times 10^{-5}$; Fig. 4C). Similarly, the Vocal − Nonvocal score increased from 0.8% at 100 voxels/subject to 1.9% at 10 voxels/subject ($m = 0.011$, $r^2 = 0.88$, $p = 5.0 \times 10^{-5}$; Fig. 4D). Eye-preferring voxels showed little response to auditory-only stories, vocal stimuli, or nonvocal stimuli, regardless of the number of voxels included in the eye ROI (Fig. 4C).

Whole-brain analysis

To examine the anatomical consistency of eye versus mouth activations across subjects, we performed a whole-brain group analysis in standard space by examining the results of a voxelwise within-subject paired t test on the difference between mouth and eye movements (Table 2; Fig. 5A).

In the group map, a cluster of voxels in right pSTS preferred visually presented mouth movements. Confirming the ROI analysis conducted in single subjects, the right pSTS region in the group map showed a significant response to short stories (0.54 ± 0.17%, $t_{(19)} = 3.2$, $p = 0.005$) and preferred vocal to nonvocal sounds (0.68 ± 0.16% vs 0.34 ± 0.13%, $t_{(4)} = 3.7$, $p = 0.02$). No regions of pSTS preferred eye movements, likely reflecting inter-individual differences in the location of eye-movement voxels (see Anatomical relationship between eye and mouth regions and variability across subjects).

Additional regions showing differential responses to visual mouth and eye movements were observed, but unlike the right pSTS, these regions did not show positive BOLD responses to auditory stimuli. A bilateral region of early visual cortex preferred visual mouth movements, but showed BOLD deactivations (responses below baseline) to all auditory stimuli (averaged across hemispheres: short stories: −0.16 ± 0.04%; vocal: −0.21 ± 0.15%; nonvocal: −0.45 ± 0.15%; no significant main effects or interactions in an ANOVA with hemisphere and auditory stimulus type as factors). A region of the putamen preferred visual eye movements, but showed near-zero BOLD responses to auditory stimuli (short stories: 0.08 ± 0.08%; vocal sounds: 0.02 ± 0.06%; nonvocal sounds: −0.04 ± 0.1%; no significant difference between vocal and nonvocal sounds, $t_{(4)} = 0.52$, $p = 0.63$).

Anatomical relationship between eye and mouth regions and variability across subjects

To determine whether there were anatomical differences in the locations of the mouth and eye ROIs, we calculated the center of mass of the mouth and eye ROIs in each hemisphere for each subject (Table 3A). In the left hemisphere, the center of mass of the mouth ROI was significantly anterior (6 mm, $t_{(19)} = 2.3$, $p = 0.03$) and lateral (5 mm, $t_{(19)} = 3.7$, $p = 0.002$) to the eye ROI. In the right hemisphere, the mouth ROI was anterior (7 mm, $t_{(19)} = 4.4$, $p = 2.9 \times 10^{-4}$), lateral (3 mm, $t_{(19)} = 2.4$, $p = 0.03$), and inferior (4 mm, $t_{(19)} = 3.2$, $p = 0.005$) to the eye ROI.

To examine the variability of mouth versus eye activations across subjects, we examined each individual subject's activation map. There was significant variability in the volumes of the mouth and eye ROIs across subjects (mouth ROI: left hemisphere range, 109–1969 mm^3; right hemisphere range, 232–3392 mm^3; eye ROI: left, 47–984 mm^3; right, 47–1360 mm^3). There was also substantial variability in the organization of the mouth-preferring and eye-preferring voxels within each subject (Figs. 2A, 5B). In some subjects (Fig. 2A, MZ), the mouth voxels and eye voxels were spatially segregated into a few large clusters. In other subjects (Fig. 2A, NV), the mouth and eye voxels were found in a patchy arrangement, with many small clusters showing mouth or eye preference. To quantify this property, we performed a spatial-cluster analysis and measured the fraction of all voxels found in the largest cluster. On average, 34% of voxels were found in the single largest cluster and 49% were found in the two largest clusters.

Control experiment using masked-face visual stimuli

Our initial experiment examined auditory responses in mouth and eye ROIs defined using visual responses to videos of entire faces. A potential confound is that the visual motion in these full-face stimuli occurred in the lower portion of the display during mouth-movement blocks and in the upper portion of the visual display during eye-movement blocks. Therefore, in a control experiment, we examined auditory responses in mouth and eye ROIs defined using videos of partial faces containing only a moving mouth or moving eyes (Fig. 6). Both types of masked stimuli were presented at the center of the display, eliminating the confound of visual field location. Yet, even controlling for visual field location, we continued to observe a systematic organization of voxels within the pSTS preferring either mouth movements or eye movements. The mouth and eye ROIs constructed from masked-face stimuli did not differ significantly in size from the ROIs constructed from full-face stimuli and their centers of mass were in similar locations (Table 3).

The response pattern to auditory stimuli was similar for ROIs constructed from masked-face and full-face stimuli (compare Figs. 2, 6). The masked-mouth ROI showed a large response to auditory stories (left hemisphere: 1.4 ± 0.2%, one-sample t test across subjects, $t_{(19)} = 6.1$, $p = 8.0 \times 10^{-6}$; right hemisphere: 0.9 ± 0.2%, $t_{(19)} = 6.0$, $p = 9.3 \times 10^{-6}$). An ANOVA with ROI (mouth vs eye) and hemisphere (left vs right) as factors revealed a main effect of ROI ($F_{(1,74)} = 27$, $p = 1.5 \times 10^{-5}$) driven by greater responses to auditory stories in the mouth ROI, as well as a main effect of hemisphere ($F_{(1,74)} = 8.2$, $p = 0.005$) driven by greater responses in the left hemisphere, but no interaction between hemisphere and region.

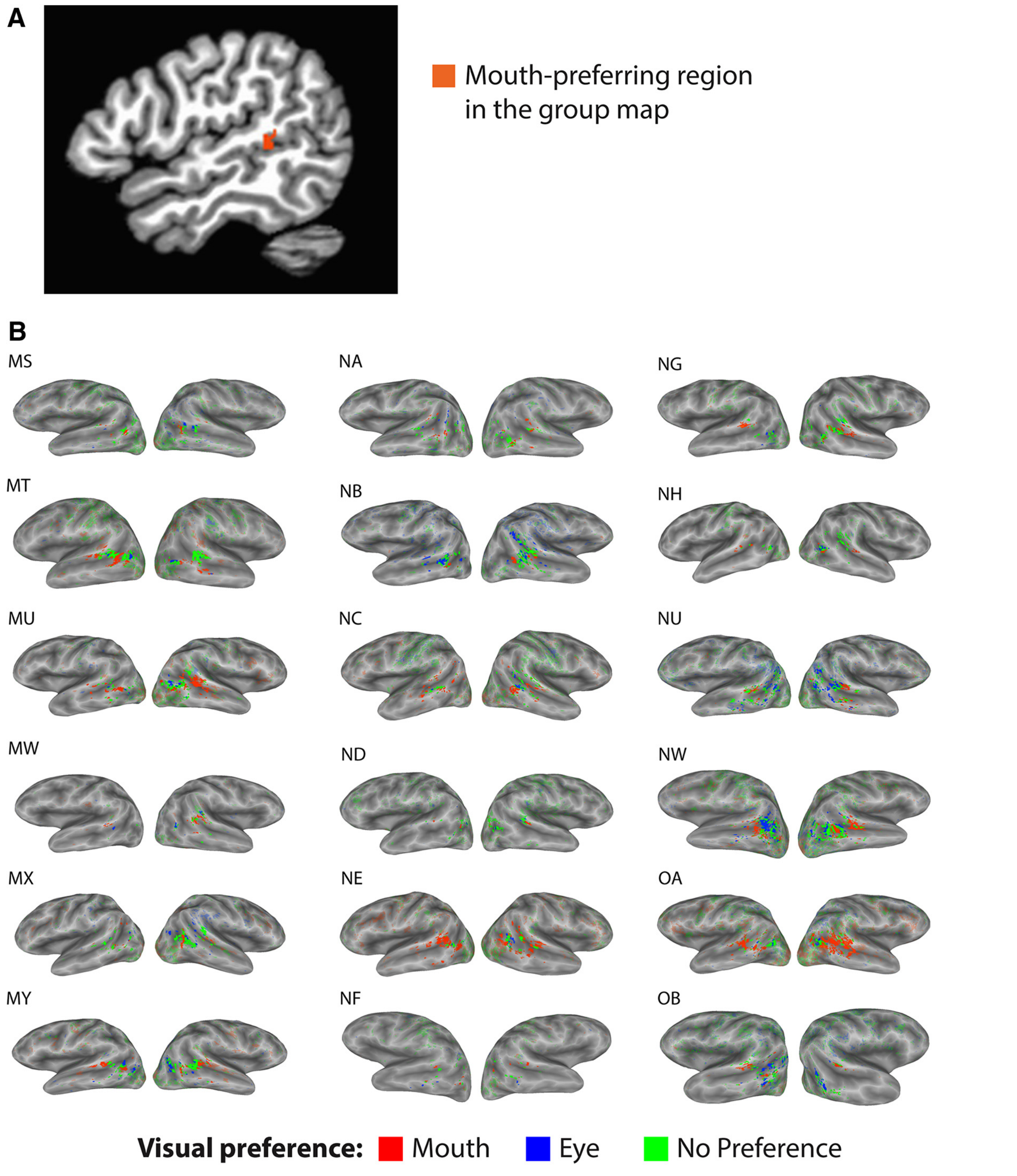

Figure 5. Anatomical consistency and variability across subjects. ***A***, Sagittal slice through the group average dataset showing mouth-preferring region (orange) in pSTS (center of mass at 50, −37, 8). ***B***, Lateral views of left and right hemisphere cortical surface models showing visual preference for mouth and eye movements in 18 subjects (remaining 2 subjects shown in Fig. 2). Mouth-preferring voxels ($t > 2$, $p < 0.05$) shown in red; eye-preferring voxels ($t < -2$) shown in blue; voxels with significant visual responses but no preference ($|t| < 2$) in green. Voxels outside the pSTS shown with reduced opacity.

As with the full-face mouth ROI, the masked-mouth ROI preferred vocal to nonvocal sounds (left hemisphere: 1.3 ± 0.3% vs 0.4 ± 0.2%, $t_{(4)} = 3.7$, $p = 0.02$; right hemisphere: 1.1 ± 0.2% vs 0.4 ± 0.2%, $t_{(4)} = 48.7$, $p = 1.1 \times 10^{-6}$). An ANOVA with region (mouth vs eye), stimulus (vocal vs nonvocal), and hemisphere (left vs right) as factors showed a main effect of region ($F_{(1,32)} = 44$, $p = 1.6 \times 10^{-7}$) driven by a greater response to all auditory stimuli in the mouth ROI) and a main effect of stimulus type ($F_{(1,32)} = 14$, $p = 6.0 \times 10^{-4}$) driven by greater overall responses to vocal sounds; there was no main effect of hemisphere ($F_{(1,32)} = 0.15$, $p = 0.70$). The only significant interaction was between region and stimulus type ($F_{(1,32)} = 4.4$, $p = 0.04$) driven by a greater response to vocal sounds in mouth regions.

Discussion

The key finding of our experiments is that visual preference for mouth movements predicted auditory responses within single voxels in the pSTS. Populations of pSTS neurons that showed a

Table 3. ROI sizes and locations

ROI	Size (mm^3)	Talairach coordinates (mm) x	y	z
Full-face stimuli				
Left hemisphere pSTS				
Mouth	633 ± 110	−54 ± 3	−46 ± 6	9 ± 4
Eye	293 ± 70	−50 ± 5	−52 ± 7	10 ± 5
Right hemisphere pSTS				
Mouth	901 ± 176	53 ± 3	−43 ± 4	7 ± 4
Eye	375 ± 71	50 ± 4	−50 ± 5	11 ± 4
Masked-face stimuli				
Left hemisphere pSTS				
Mouth	525 ± 65	−54 ± 4	−49 ± 6	9 ± 4
Eye	393 ± 78	−49 ± 5	−51 ± 6	12 ± 4
Right hemisphere pSTS				
Mouth	829 ± 123	53 ± 3	−44 ± 3	8 ± 4
Eye	357 ± 69	51 ± 5	−49 ± 6	11 ± 5

Average size and location of ROIs created from full-face and masked visual stimuli averaged across all subjects (mean ± SD).

preference for visually presented mouth movements responded significantly to auditory stimuli, with a preference for vocal compared with nonvocal sounds. In contrast, pSTS regions that preferred visually presented eye movements did not respond to auditory stimuli, either vocal or nonvocal.

Our findings build upon those of previous studies that have examined visual or auditory selectivity within the pSTS in isolation. The pSTS responds strongly to visual stimuli containing biological motion (Grossman et al., 2000, 2005; Beauchamp et al., 2002, 2003; Santi et al., 2003; Pitcher et al., 2011), including facial movements, such as mouth movements and eye movements (Puce et al., 1998; Pelphrey et al., 2005). The pSTS is also known to respond strongly to auditory stimuli, especially speech sounds (Belin et al., 2000; Binder et al., 2000; Vouloumanos et al., 2001; Wise et al., 2001; Fecteau et al., 2004; Shultz et al., 2012). Our study links these two previously demonstrated axes of visual and auditory selectivity by showing that at a fine anatomical scale, pSTS regions that prefer a certain category of visual stimuli (*i.*e., moving mouths) respond strongly to auditory speech.

While previous studies have shown that faces and voices activate overlapping regions of the pSTS (Kreifelts et al., 2009; Watson et al., 2014a; Deen et al., 2015), our study examines this overlap in individual voxels without spatial smoothing. Smoothing produces large "blobs" of activation; interpreting overlap between these blobs is problematic and results in an artificially large overlap (Deen et al., 2015). We found that single voxels within close spatial proximity showed different preferences for different visual stimuli—preferring either mouths or eyes—and different preferences for auditory stimuli—responding strongly or not at all to vocal stimuli.

A logical explanation for these findings is that when humans communicate face to face with conspecifics, they see the movements of the mouth at the same time they hear speech. Given this environmental correspondence, it would be parsimonious for the same populations of neurons in the pSTS to represent both the visual features of mouth movements (Bernstein and Liebenthal, 2014; Irwin et al., 2015) and the auditory features of speech (Mesgarani et al., 2014; Arsenault and Buchsbaum, 2015; Correia et al., 2015; Norman-Haignere et al., 2015; Brewer and Barton, 2016). Coding information in this way is advantageous because it allows the organism to benefit from the independent sources of information available about the environment in the two modalities. For instance, humans frequently encounter environments with a high degree of auditory noise. Visual speech information is unaffected by auditory noise, allowing it to compensate for the degraded auditory information and restore comprehension to near-normal levels. Neurons in the pSTS that code for both auditory and visual speech features provide a likely neural mechanism for this process.

This idea is supported by studies in nonhuman primates demonstrating that single neurons in macaque STS respond to both auditory and visual stimuli (Bruce et al., 1981) and prefer face and voice stimuli in which the seen mouth movement matches the heard vocalization (Barraclough et al., 2005; Ghazanfar et al., 2005; Perrodin et al., 2014). While monkeys do not have as large a vocal repertoire as humans, it seems reasonable that single neurons in the human pSTS have a similar learned correspondence between vocal sounds and the mouth movements that produce them.

The present study suggests a link between the organization of the pSTS into auditory, visual, and audiovisual patches and the organization of sensory cortex into regions preferring different stimulus categories (Grill-Spector and Malach, 2004). Neuroanatomical studies using anterograde tracers have shown that the macaque STS receives projections from both auditory and visual areas in a patchy fashion (Seltzer and Pandya, 1994; Seltzer et al., 1996). In humans, high-resolution fMRI has been used to show that small patches of cortex in pSTS preferentially respond to auditory, visual, or audiovisual stimuli (Beauchamp et al., 2004b), an organization confirmed with single-unit studies in macaques (Dahl et al., 2009). We propose that the mouth-preferring and voice-preferring regions identified in the present study correspond to the audiovisual patches described by Beauchamp et al. (2004b) and Dahl et al. (2009). Neurons in audiovisual patches that receive both auditory and visual inputs are put into service to recognize the correspondence between visual mouth movements and auditory speech. Neurons in visual patches become specialized for recognizing eye movements, as there is no auditory stimulus associated with motion of the eyes. Finally, neurons in auditory patches become specialized for processing sounds that do not have a strong visual association, such as a beeping alarm (Belin et al., 2000; Kreifelts et al., 2009).

The idea of co-occurrence of auditory and visual selectivity in the same population of neurons is supported by the analysis in which we adjusted the threshold for classifying voxels as mouth-preferring. Under the null hypothesis of no link between auditory and visual selectivity, how strongly a voxel prefers mouth movements should have no influence on its response to auditory voices. Instead, we found a strong positive correlation: voxels with the strongest preference for visual mouth movements showed the greatest response to auditory voice stimuli and the greatest preference for vocal sounds. An obvious explanation for this finding is that these voxels contain a high proportion of neurons that respond to both visual mouth movements and auditory voices.

Laterality

The visually defined mouth and eye ROIs were qualitatively similar between hemispheres (Fig. 5) and ANOVAs showed no significant interhemispheric difference in the amplitudes of the BOLD responses to auditory stimuli. The visually defined ROI volumes were larger in the right hemisphere, consistent with previous studies of responses to visually presented biological motion (Beauchamp et al., 2002, 2003). For the auditory-defined voice-responsive ROIs, the ANOVA revealed a significantly greater response to visual stimuli (both mouth and eye movements) in the

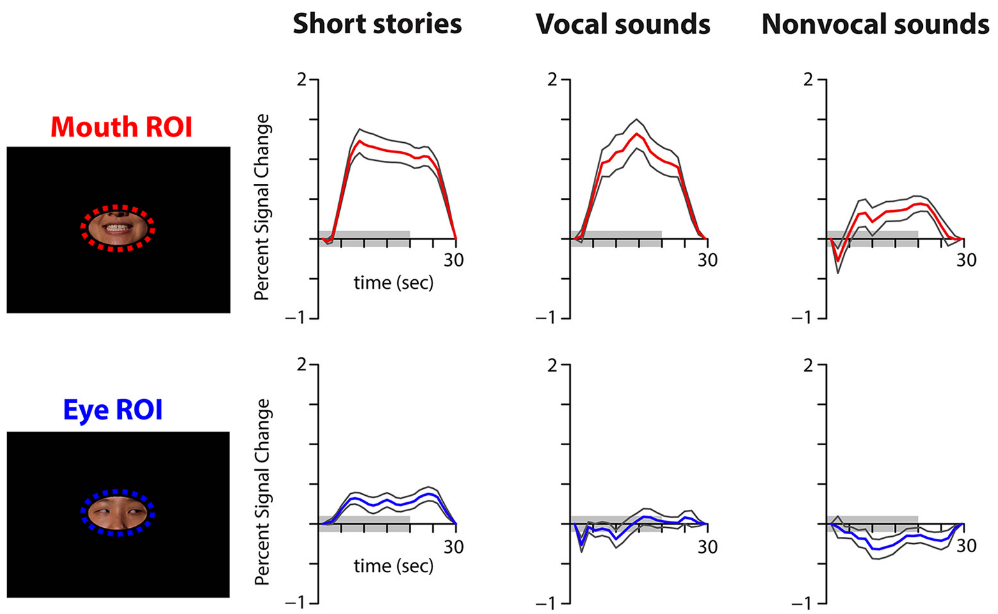

Figure 6. Auditory responses in visual regions defined using masked stimuli. Visual stimuli consisting of videos of actors making mouth movements or eye movements were presented, masked so that only the mouth or the eyes were visible, and presented at the center of display (location of movements highlighted with colored ellipse, not present in actual video). Top row shows responses to auditory stimuli in the mouth ROI (all masked mouth-preferring voxels). Bottom row shows responses in the eye ROI (all masked eye-preferring voxels). Columns show responses to short stories ($n = 40$), vocal sounds ($n = 10$), and nonvocal sounds ($n = 10$). Gray lines denote SEM, and the gray bar along the x-axis denotes the length of stimulus presentation.

right hemisphere compared with the left hemisphere, consistent with previous findings of stronger multisensory responses in right compared with left STS (Beauchamp et al., 2004a).

Control experiment using masked-face visual stimuli

Our main experiment used full-face stimuli, confounding the type of motion (mouth vs eyes) with the location of the motion (lower vs upper visual field). However, a control experiment using masked-face stimuli presented at the center of gaze also observed voice-selectivity in mouth-preferring regions, ruling out visual field location as the sole explanation for our results. This is consistent with the known properties of the face-processing network in monkey inferotemporal cortex, in which both the posterior lateral and the medial fundus face patches show responses strongly modulated by the presence or absence of specific face parts, especially eyes, regardless of where they are presented in the visual field (Freiwald et al., 2009; Issa and DiCarlo, 2012).

Anatomical relationship between eye and mouth regions

We found that across subjects, the average center of mass of the mouth ROI was located slightly lateral and anterior to the eye ROI. This is consistent with previous studies showing lateral and anterior activation for moving mouths compared with moving eyes (Pelphrey et al., 2005) or static images of mouths compared with static images of eyes (Orlov et al., 2010). As proposed by Pelphrey et al. (2005), an attractive explanation for these anatomical differences is efficient connectivity: the more anterior location of mouth-preferring regions decreases the distance axons must travel from auditory cortex, allowing for efficient integration of mouth movement information with auditory speech information during multisensory speech processing. In contrast, the more posterior location of eye-preferring regions creates anatomical proximity with the intraparietal sulcus, an area critical for shifting spatial attention in response to the spatial cues provided by social gaze signals (Hoffman and Haxby, 2000; Nummenmaa et al., 2010).

Future directions

Our study raises a number of important questions for future research. First, our study did not investigate the role of the pSTS in processing vocal nonspeech, such as yawns or coughs. While these stimuli are less frequent and less important for human communication than speech, it is reasonable to assume that the pSTS also forms auditory–visual associations between mouth movements and vocal nonspeech sounds. This is especially likely to be true for emotional audiovisual stimuli, such as crying or laughing, given that the pSTS is important for processing the content of emotional stimuli (Grandjean et al., 2005; Kreifelts et al., 2007, 2009; Collignon et al., 2008; Charbonneau et al., 2013; Watson et al., 2014b). Second, our study did not investigate the role of auditory imagery: subjects might have imagined the auditory speech that would normally accompany the mouth movements shown in our silent videos, leading to activity in auditory regions. However, careful studies have shown no activation in auditory areas for silent videos with implied sound (Hsieh et al., 2012) or deactivation below baseline in auditory areas for auditory imagery (Daselaar et al., 2010), rendering it unlikely that imagery is the sole explanation of our findings. Third, our study did not investigate how the observed intersection of visual and auditory stimulus selectivity is related to multisensory responses in the pSTS, given that the pSTS is known to be a hub for audiovisual integration (Wright et al., 2003; Beauchamp et al., 2004a; van Atteveldt et al., 2004; Kreifelts et al., 2007; Stevenson et al., 2007): associating individual faces and voices leads to increased pSTS activity and improved performance on auditory-only speech-recognition tasks (von Kriegstein et al., 2008, 2010; Schall and von Kriegstein, 2014), while audiovisual integration can be disrupted by temporary lesions of the pSTS induced by transcranial magnetic stimulation (Beauchamp et al., 2010; Riedel et al., 2015).

References

Arsenault JS, Buchsbaum BR (2015) Distributed neural representations of phonological features during speech perception. J Neurosci 35:634–642. CrossRef Medline

Baker CI, Hutchison TL, Kanwisher N (2007) Does the fusiform face area contain subregions highly selective for nonfaces? Nat Neurosci 10:3–4. CrossRef Medline

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

Beauchamp MS, Lee KE, Haxby JV, Martin A (2002) Parallel visual motion processing streams for manipulable objects and human movements. Neuron 34:149–159. CrossRef Medline

Beauchamp MS, Lee KE, Haxby JV, Martin A (2003) FMRI responses to video and point-light displays of moving humans and manipulable objects. J Cogn Neurosci 15:991–1001. CrossRef Medline

Beauchamp MS, Lee KE, Argall BD, Martin A (2004a) Integration of auditory and visual information about objects in superior temporal sulcus. Neuron 41:809–823. CrossRef Medline

Beauchamp MS, Argall BD, Bodurka J, Duyn JH, Martin A (2004b) Unraveling multisensory integration: patchy organization within human STS multisensory cortex. Nat Neurosci 7:1190–1192. CrossRef Medline

Beauchamp MS, Nath AR, 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. CrossRef Medline

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

Bernstein LE, Liebenthal E (2014) Neural pathways for visual speech perception. Front Neurosci 8:386. CrossRef Medline

Bernstein LE, Auer ET Jr, Moore JK, Ponton CW, Don M, Singh M (2002) Visual speech perception without primary auditory cortex activation. Neuroreport 13:311–315. CrossRef Medline

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

Brainard DH (1997) The psychophysics toolbox. Spat Vis 10:433–436. CrossRef Medline

Brewer AA, Barton B (2016) Maps of the auditory cortex. Annu Rev Neurosci 39:385–407. CrossRef Medline

Bruce C, Desimone R, Gross CG (1981) Visual properties of neurons in a polysensory area in superior temporal sulcus of the macaque. J Neurophysiol 46:369–384. Medline

Calvert GA, Campbell R (2003) Reading speech from still and moving faces: the neural substrates of visible speech. J Cogn Neurosci 15:57–70. CrossRef Medline

Calvert GA, Bullmore ET, Brammer MJ, Campbell R, Williams SC, McGuire PK, Woodruff PW, Iversen SD, David AS (1997) Activation of auditory cortex during silent lipreading. Science 276:593–596. CrossRef Medline

Capilla A, Belin P, Gross J (2013) The early spatio-temporal correlates and task independence of cerebral voice processing studied with MEG. Cereb Cortex 23:1388–1395. CrossRef Medline

Charbonneau G, Bertone A, Lepore F, Nassim M, Lassonde M, Mottron L, Collignon O (2013) Multilevel alterations in the processing of audio-visual emotion expressions in autism spectrum disorders. Neuropsychologia 51:1002–1010. CrossRef Medline

Collignon O, Girard S, Gosselin F, Roy S, Saint-Amour D, Lassonde M, Lepore F (2008) Audio-visual integration of emotion expression. Brain Res 1242:126–135. CrossRef Medline

Correia JM, Jansma BM, Bonte M (2015) Decoding articulatory features from fMRI responses in dorsal speech regions. J Neurosci 35:15015–15025. CrossRef Medline

Cox RW (1996) AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. Comput Biomed Res 29:162–173. CrossRef Medline

Dahl CD, Logothetis NK, Kayser C (2009) Spatial organization of multisensory responses in temporal association cortex. J Neurosci 29:11924–11932. CrossRef Medline

Dale AM, Fischl B, Sereno MI (1999) Cortical surface-based analysis. I. Segmentation and surface reconstruction. Neuroimage 9:179–194. CrossRef Medline

Daselaar SM, Porat Y, Huijbers W, Pennartz CM (2010) Modality-specific and modality-independent components of the human imagery system. Neuroimage 52:677–685. CrossRef Medline

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. CrossRef Medline

Destrieux C, Fischl B, Dale A, Halgren E (2010) Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. Neuroimage 53:1–15. CrossRef Medline

Eklund A, Nichols TE, Knutsson H (2016) Cluster failure: why fMRI inferences for spatial extent have inflated false-positive rates. Proc Natl Acad Sci U S A 113:7900–7905. CrossRef Medline

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

Fischl B, Sereno MI, Dale AM (1999) Cortical surface-based analysis. II: Inflation, flattening, and a surface-based coordinate system. Neuroimage 9:195–207. CrossRef Medline

Freiwald WA, Tsao DY, Livingstone MS (2009) A face feature space in the macaque temporal lobe. Nat Neurosci 12:1187–1196. CrossRef Medline

Genovese CR, Lazar NA, Nichols T (2002) Thresholding of statistical maps in functional neuroimaging using the false discovery rate. Neuroimage 15:870–878. CrossRef Medline

Ghazanfar AA, Maier JX, Hoffman KL, Logothetis NK (2005) Multisensory integration of dynamic faces and voices in rhesus monkey auditory cortex. J Neurosci 25:5004–5012. CrossRef Medline

Glover GH (1999) Deconvolution of impulse response in event-related BOLD fMRI. Neuroimage 9:416–429. CrossRef Medline

Grandjean D, Sander D, Pourtois G, Schwartz S, Seghier ML, Scherer KR, Vuilleumier P (2005) The voices of wrath: brain responses to angry prosody in meaningless speech. Nat Neurosci 8:145–146. CrossRef Medline

Grill-Spector K, Malach R (2004) The human visual cortex. Annu Rev Neurosci 27:649–677. CrossRef Medline

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

Grossman ED, Battelli L, Pascual-Leone A (2005) Repetitive TMS over posterior STS disrupts perception of biological motion. Vis Res 45:2847–2853. CrossRef Medline

Haxby JV, Ungerleider LG, Clark VP, Schouten JL, Hoffman EA, Martin A (1999) The effect of face inversion on activity in human neural systems for face and object perception. Neuron 22:189–199. CrossRef Medline

Hoffman EA, Haxby JV (2000) Distinct representations of eye gaze and identity in the distributed human neural system for face perception. Nat Neurosci 3:80–84. CrossRef Medline

Holmes CJ, Hoge R, Collins L, Woods R, Toga AW, Evans AC (1998) Enhancement of MR images using registration for signal averaging. J Comput Assist Tomogr 22:324–333. CrossRef Medline

Hooker CI, Paller KA, Gitelman DR, Parrish TB, Mesulam MM, Reber PJ (2003) Brain networks for analyzing eye gaze. Brain Res Cogn Brain Res 17:406–418. CrossRef Medline

Hsieh PJ, Colas JT, Kanwisher N (2012) Spatial pattern of BOLD fMRI activation reveals cross-modal information in auditory cortex. J Neurophysiol 107:3428–3432. CrossRef Medline

Irwin J, Preston J, Brancazio L, D'angelo M, Turcios J (2015) Development of an audiovisual speech perception app for children with autism spectrum disorders. Clin Linguist Phon 29:76–83. CrossRef Medline

Ishai A, Ungerleider LG, Martin A, Schouten JL, Haxby JV (1999) Distributed representation of objects in the human ventral visual pathway. Proc Natl Acad Sci U S A 96:9379–9384. CrossRef Medline

Ishai A, Ungerleider LG, Haxby JV (2000) Distributed neural systems for the generation of visual images. Neuron 28:979–990. CrossRef Medline

Issa EB, DiCarlo JJ (2012) Precedence of the eye region in neural processing of faces. J Neurosci 32:16666–16682. CrossRef Medline

Kreifelts B, Ethofer T, Grodd W, Erb M, Wildgruber D (2007) Audiovisual integration of emotional signals in voice and face: an event-related fMRI study. Neuroimage 37:1445–1456. CrossRef Medline

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

Mesgarani N, Cheung C, Johnson K, Chang EF (2014) Phonetic feature encoding in human superior temporal gyrus. Science 343:1006–1010. CrossRef Medline

Norman-Haignere S, Kanwisher NG, McDermott JH (2015) Distinct cortical pathways for music and speech revealed by hypothesis-free voxel decomposition. Neuron 88:1281–1296. CrossRef Medline

Nummenmaa L, Passamonti L, Rowe J, Engell AD, Calder AJ (2010) Connectivity analysis reveals a cortical network for eye gaze perception. Cereb Cortex 20:1780–1787. CrossRef Medline

Orlov T, Makin TR, Zohary E (2010) Topographic representation of the human body in the occipitotemporal cortex. Neuron 68:586–600. CrossRef Medline

Pelli DG (1997) The VideoToolbox software for visual psychophysics: transforming numbers into movies. Spat Vis 10:437–442. CrossRef Medline

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

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

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

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

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

Santi A, Servos P, Vatikiotis-Bateson E, Kuratate T, Munhall K (2003) Perceiving biological motion: dissociating visible speech from walking. J Cogn Neurosci 15:800–809. CrossRef Medline

Schall S, von Kriegstein K (2014) Functional connectivity between face-movement and speech-intelligibility areas during auditory-only speech perception. PloS One 9:e86325. CrossRef Medline

Seltzer B, Pandya DN (1994) Parietal, temporal, and occipital projections to cortex of the superior temporal sulcus in the rhesus monkey: a retrograde tracer study. J Comp Neurol 343:445–463. CrossRef Medline

Seltzer B, Cola MG, Gutierrez C, Massee M, Weldon C, Cusick CG (1996) Overlapping and nonoverlapping cortical projections to cortex of the superior temporal sulcus in the rhesus monkey: double anterograde tracer studies. J Comp Neurol 370:173–190. CrossRef Medline

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

Simmons WK, Bellgowan PS, Martin A (2007) Measuring selectivity in fMRI data. Nat Neurosci 10:4–5. CrossRef Medline

Stevenson RA, Geoghegan ML, James TW (2007) Superadditive BOLD activation in superior temporal sulcus with threshold non-speech objects. Exp Brain Res 179:85–95. CrossRef Medline

van Atteveldt N, Formisano E, Goebel R, Blomert L (2004) Integration of letters and speech sounds in the human brain. Neuron 43:271–282. CrossRef Medline

von Kriegstein K, Dogan O, Grüter M, Giraud AL, Kell CA, Grüter T, Kleinschmidt A, Kiebel SJ (2008) Simulation of talking faces in the human brain improves auditory speech recognition. Proc Natl Acad Sci U S A 105:6747–6752. CrossRef Medline

von Kriegstein K, Smith DR, Patterson RD, Kiebel SJ, Griffiths TD (2010) How the human brain recognizes speech in the context of changing speakers. J Neurosci 30:629–638. CrossRef Medline

Vouloumanos A, Kiehl KA, Werker JF, Liddle PF (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. CrossRef Medline

Vul E, Pashler H (2012) Voodoo and circularity errors. Neuroimage 62: 945–948. CrossRef Medline

Vul E, Harris C, Winkielman P, Pashler H (2009) Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. Perspect Psychol Sci 4:274–290. CrossRef Medline

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. CrossRef Medline

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. CrossRef Medline

Wise RJ, Scott SK, Blank SC, Mummery CJ, Murphy K, Warburton EA (2001) Separate neural subsystems within 'Wernicke's area'. Brain 124: 83–95. CrossRef Medline

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