HHS Public Access

Author manuscript

Neuroimage. Author manuscript; available in PMC 2018 December 01.

Published in final edited form as:

Neuroimage. 2018 December ; 183: 25–36. doi:10.1016/j.neuroimage.2018.08.008.

Free viewing of talking faces reveals mouth and eye preferring regions of the human superior temporal sulcus

Johannes Rennig and **Michael S Beauchamp**

Department of Neurosurgery and Core for Advanced MRI, Baylor College of Medicine, Houston TX, USA

Abstract

During face-to-face communication, the mouth of the talker is informative about speech content, while the eyes of the talker convey other information, such as gaze location. Viewers most often fixate either the mouth or the eyes of the talker's face, presumably allowing them to sample these different sources of information. To study the neural correlates of this process, healthy humans freely viewed talking faces while brain activity was measured with BOLD fMRI and eye movements were recorded with a video-based eye tracker. *Post hoc* trial sorting was used to divide the data into trials in which participants fixated the mouth of the talker and trials in which they fixated the eyes. Although the audiovisual stimulus was identical, the two trials types evoked differing responses in subregions of the posterior superior temporal sulcus (pSTS). The anterior pSTS preferred trials in which participants fixated the mouth of the talker while the posterior pSTS preferred fixations on the eye of the talker. A second fMRI experiment demonstrated that anterior pSTS responded more strongly to auditory and audiovisual speech than posterior pSTS eye-preferring regions. These results provide evidence for functional specialization within the pSTS under more realistic viewing and stimulus conditions than in previous neuroimaging studies.

Keywords

audiovisual; face; multisensory; speech; eye tracking; fMRI

Introduction

Conversing with another human face-to-face exposes us to an abundance of sensory input. In the auditory modality, the voice of the talker conveys speech content. In the visual modality, the movements of the talker's mouth also convey speech content (since different mouth movements produce different speech sounds), while the talker's eyes carry other types of information, such as the spatial location of the talker's gaze. A growing body of evidence suggests that the human posterior superior temporal sulcus (pSTS) contains distinct regions specialized for processing these multiple information sources.

Address for correspondence: Michael S Beauchamp, Baylor College of Medicine, Department of Neurosurgery, Core of Advanced Magnetic Resonance Imaging, 1 Baylor Plaza, Houston, Texas, 77030, USA, Michael.Beauchamp@bcm.edu.

The authors declare no competing financial interests.

Wernicke first observed that damage to pSTS and nearby regions of lateral temporal cortex impairs speech perception. Functional neuroimaging has increased our understanding of the neural computations performed by this important piece of cortex. In the auditory domain, pSTS contains regions that are highly selective for the human voice (Belin et al., 2000) as well as particular speech sounds (Chang et al., 2010). Functional subdivisions of the pSTS also exist in the visual domain. Pelphrey and colleagues (2005) presented silent videos of a computer-generated face, either opening and closing its mouth or moving its eyes. BOLD fMRI activations within the pSTS to mouth movements were located more anteriorly while activations to eye movements were located more posteriorly. This direct comparison of mouth and eye movements was consistent with studies finding more posterior pSTS activity for eye-gaze observations compared to scrambled images (Hoffman and Haxby, 2000); that anterior pSTS activity is observed for mouth movements contrasted against still mouths (Calvert and Campbell, 2003); and that visual mouth movements related to speech production activate regions of the STS that are more anterior than those activated by non-speech mouth movements, such as yawns (Bernstein et al., 2011). Zhu and Beauchamp (2017) replicated the findings of Pelphrey and colleagues using silent videos of real human faces making mouth or eye movements. Mouth-preferring regions compared to eye preferring regions were located more anterior in the pSTS and responded strongly to unisensory auditory stimuli, especially speech.

A better understanding of the relationship between auditory speech and visual face processing in the STS requires presenting both unisensory and multisensory stimuli. However, most previous studies presented only unisensory auditory or visual stimuli (Belin et al., 2000; Bernstein et al., 2011; Chang et al., 2010; Pelphrey et al., 2005b; Zhu and Beauchamp, 2017). The unisensory visual stimuli in these studies consisted of silent videos of faces making mouth or eye movements in isolation, while in real world conditions, humans are confronted with audiovisual talking faces making both eye and mouth movements.

Under natural viewing conditions, humans fixate either the mouth or the eyes of the talker for time intervals that can extend to a second or more (Gurler et al., 2015). A recent fMRI study capitalized on the existence of these extended fixations to search for the neural correlates of fixating the eyes of a dynamic talking face (Jiang et al., 2016). The authors describe the existence of an “eye contact network” that includes portions of the pSTS, the temporo-parietal junction and other brain areas.

In the present study, we searched for brain areas that were more active when participants fixated the mouth of the talker. Our hypothesis was that this “mouth contact network” should include regions important for speech perception, especially areas responsive to visual mouth movements and auditory speech in anterior portions of the pSTS, and that these regions should demonstrate multisensory integration. To test this hypothesis, we performed two independent fMRI experiments. In the first, participants freely viewed dynamic talking faces while their eye movements were monitored in order to identify mouth and eye-selective regions. In the second, participants viewed blocks of auditory, visual, and audiovisual speech in order to determine functional specialization and multisensory integration of mouth and eye-selective regions.

Methods

34 healthy right-handed participants (16 females, mean age 26.5, range 18 – 45) with normal or corrected to normal vision and normal hearing provided written informed consent under an experimental protocol approved by the Committee for the Protection of Human Subjects of the Baylor College of Medicine, Houston, TX. 29 of 34 participants were native English speakers (2 German speakers, 3 Mandarin speakers).

Each participant was scanned using a 3T Siemens Trio MRI scanner equipped with a 32-channel head coil at Baylor College of Medicine's Core for Advanced MRI (CAMRI). During a single scanning session, participants performed two different fMRI experiments. The two experiments were analyzed independently to eliminate bias. Stimuli were presented using Matlab (The Mathworks, Inc., Natick, MA, USA) and the Psychophysics Toolbox (Brainard, 1997; Pelli, 1997) and viewed on an MR compatible screen (BOLDscreen32, Cambridge Research Systems, Rochester, UK) placed behind the bore of the scanner. Auditory speech was presented using high-fidelity MR compatible headphones (Sensimetrics, Malden, MA, USA). Behavioral responses were collected using a fiber-optic button response pad (Current Designs, Haverford, PA, USA) and eye movements were recorded during scanning using the Eye Link 1000 (SR Research Ltd., Ottawa, Ontario, Canada) in MR-compatible mode with a sampling rate of 500 Hz.

In the first fMRI experiment, eye-tracking was performed in the MR scanner while participants viewed audiovisual movies presented in an event-related design. Each 2-second movie consisted of a talker saying a single syllable. Each participant viewed 240 movies, equally distributed across six different types: three syllables (AbaVba, AgaVga, AbaVga) ×two talkers (one male and one female). Following each movie, participants identified the presented syllable with a button press.

In the second fMRI experiment, participants viewed long blocks (20 seconds) of auditory, visual or auditory-visual speech, with a single female talker reading Aesop's fables (Nath and Beauchamp, 2012). The eye image from the eye tracker was monitored to ensure the participant's alertness but no eye tracking was performed and there was no task.

Eye tracking data analysis

Figure 1A shows frames from a stimulus movie. To simulate natural viewing conditions, each face movie was preceded with a gray screen containing a fixation crosshairs presented in a random location distant from the spatial position where the face would later appear (Gurler et al., 2015). After the face appeared, the crosshairs disappeared, and participants were free to fixate anywhere. Preparatory mouth movements began ~400 ms after stimulus onset, voice onset occurred at ~800 ms, voice offset occurred at ~1100 ms, and articulatory mouth movements were complete by ~1400 ms. Only the speech relevant fixations between 400 ms to 1400 ms after stimulus onset were included in the analysis. If there were multiple fixations during this interval, their locations were averaged, weighted by fixation duration. Trials in which more than 40% of the eye position samples were invalid were discarded (approximately 6% of total trials). This analysis produced a two-dimensional spatial heat map of fixation locations (Figure 1B). To further reduce the data dimensionality, each

talker's face was divided into a mouth ROI, consisting of the lower half of the video frame, and an eye ROI, consisting of the upper half of the video frame (Figure 1C). For each trial, the percentage of fixation time spent within each ROI was calculated.

There were six different movie types (three syllables × two talkers). For each participant, we calculated the percent of mouth-looking time for each of the six stimuli. Eye movement behavior was consistent across stimuli, as demonstrated by high correlation across stimuli within each participant: mean $r = 0.95$, SD = 0.01, range: 0.93 – 0.97. Since there were no significant differences between stimuli and talkers, the different stimuli were grouped for further analysis.

Eye tracking calibration and quality control

Before each fMRI run, the eye tracker was calibrated by having the participant fixate each of a nine-point array of reference points presented on the display screen. This process was repeated until an acceptable calibration was obtained. Eye tracker drift was sometimes observed over the course of each run. This drift was corrected *post hoc* by using the fixation crosshairs presented at the center of the screen during non-stimulus epochs. The difference between the measured eye position during fixation epochs and the center of the screen (location of the crosshairs) was applied to correct the eye tracking data in the preceding stimulus epoch.

To measure the effectiveness of this drift correction, we searched for systematic changes in eye movement behavior over time, focusing on average gaze position per trial on the y-axis as this could bias our measures of mouth or eye looking. For each participant, we plotted the y-location of the mean fixation position against time for the entire session and measured the slope of the line (negative slope would indicate systematic shifts to lower fixations, positive slope would indicate systematic upper field shifts). The mean slope across participants was near zero (m = −0.041, SD: 0.21; *t*-test against zero: $t_{(33)} = -1.19$, $p = 0.24$). Another possible systematic drift could manifest as more central or more peripheral fixation location. To test for this possibility, we plotted the distance of the fixation location from the center of the display against time for the entire session. The slope of this line was also near zero (m = 0.011, SD: 0.12; *t*-test against zero: $t_{(33)} = 0.61$, $p = 0.55$).

Construction of fMRI regressors

fMRI analysis was conducted using a generalized linear model that was different for each participant as it was constructed from that participant's eye tracking data. The first fMRI regressor contained all trials in which the participant was more likely to fixate the mouth of the talker; the second regressor contained all trials in which the participant was more likely to fixate the eyes of the talker; the third regressor contained all trials which could not be classified because the eye tracking data was unreliable (mean of 6% of trials).

To ensure that the fMRI analysis was robust, we wished to include equal numbers of trials in the mouth and eye regressors. This was accomplished by classifying trials as mouth trials or eye trials using the median fixation-time across trials for each participant, so that within each participant, exactly half of the trials were mouth trials (with greater than the median amount of mouth-fixation) and the remaining trials were eye trials (with less than the median

amount of mouth-fixation). On average, the median trial contained 74% mouth fixation time and 26% eye fixation time (averaged across participants; SD: 29%; range: 1 % – 99 %). We also tested a classification strategy using the same classification measure for each participant: trials with greater than 50% of trial time spent on mouth fixations were classified as mouth trials and trials with greater than 50% of trial time spent on eye fixations were classified as eye trials. With this alternative strategy, five participants had few or none mouth or eye trials, requiring those participants to be excluded from the fMRI analysis. In order to include as much fMRI data as possible in the analysis, we used the individualized median fixation-time trial criterion for the primary analysis.

Stimulus Ordering and MRI acquisition

For the four runs of the first fMRI experiment (event-related with eye-tracking) each run contained eighty 2-second trials: twenty events of each stimulus type (AbaVba, AgaVga, AbaVga) and 20 Null events (fixation only). The events were ordered in an optimal rapid event-related design specified by optseq2 (Dale, 1999; https://surfer.nmr.mgh.harvard.edu/optseq).

For the two runs of the second fMRI experiment (block design, no eye tracking), each run contained nine blocks, each consisting of twenty seconds of stimulation followed by ten seconds of fixation baseline, consisting of three blocks each of auditory, visual and audiovisual stimulation in optimal pseudo-random order (total duration of each run, 4.5 minutes). The stimuli were Aesop's fables read aloud by a female talker. In auditory blocks, only a fixation crosshair was visible; in visual blocks, no sound was presented.

In each participant's scanning session, we collected two T1-weighted MP-RAGE anatomical MRI scans and six runs of functional imaging (4 runs of first experiment, 2 runs of second experiment) using a 64-channel head coil. Each functional run began with 5 TRs of dummy scans to reach equilibrium magnetization.

19 participants were scanned using a continuous multislice echo planar imaging sequences (Setsompop et al., 2012). For these participants, the parameters were identical for the runs containing the first and second fMRI experiments: TR = 1500 ms, TE = 30 ms, flip angle = 72°, in-plane resolution of 2 × 2 mm, 69 2 mm axial slices, multiband factor: 3, GRAPPA factor: 2.

15 participants were scanned with a slightly different pulse sequence which permitted a clustered acquisition (sparse sampling) in which there were periods of silence in which the stimulus was presented (Moeller et al., 2010). The sequences parameters were TE = 30 ms, flip angle = 90°, in-plane resolution of 2 × 2 mm, 69 2 mm axial slices, multiband factor = 2. For the first fMRI experiment, a TR of 4000 ms was used, consisting of 2 seconds of EPI acquisition followed by 2 seconds of no acquisition in which the stimulus was presented. For the second fMRI experiment (block design, no eye tracking) the same parameters were used, except with a TR of 2000 ms (continuous instead of sparse sampling). Results were similar for both sequences, so they were combined for the analyses reported in the paper.

fMRI analysis

fMRI analysis was conducted using the standard AFNI processing stream (Cox, 1996) consisting of slice-timing correction, motion correction with local Pearson correlation (Saad et al., 2009) and fitting each voxel's time series with a generalized linear model that included baseline drift and the six motion parameters (roll, pitch, yaw; linear movement into x-, y-, z-directions) as regressors of no interest. For the first fMRI experiment, the experimental regressors consisted of the three eye-tracking regressors (trials in which the participant looked mainly at the mouth, trials in which the participant looked mainly at the eyes, invalid eye-tracking trials). For the second fMRI experiment, the regressors consisted of regressors coding for auditory, visual and audiovisual blocks.

ROI construction

For every participant, we created a cortical surface model from two repetitions of a T1-weighted image with FreeSurfer (Dale et al., 1999; Fischl et al., 1999) and manipulated it with SUMA (Argall et al., 2006). An anatomical pSTS mask was created by combining the Freesurfer-defined superior temporal gyrus, superior temporal sulcus, and middle temporal gyrus (Destrieux et al., 2010) followed by selection of the posterior half of this ROI using a cutoff placed at the individual midpoint of the full anterior-to-posterior extent of the ROI. The average cutoff location was y = −23 ± 0.4 mm (left hemisphere) and y = −24 ± 0.4 mm (right hemisphere); co-ordinates in MNI standard space (N27). This anatomical ROI was combined with a functional criterion. Voxels that showed a significant overall omnibus *F*-test ($F > 5$, $q < 0.0001$, false discovery rate corrected) and a significant effect of the mouth *vs.* eye contrast ($q < 0.05$) were included in the analysis.

fMRI Group Maps

Group analysis was performed on the cortical surface. A spherical version of each participant's cortical surface model was aligned to the Freesurfer template (Dale et al., 1999; Fischl et al., 1999). Functional data was mapped to each participant's template-aligned surface and smoothed with a 5 mm kernel. The AFNI program 3dttest++ was used to compare the responses to mouth and eye preferring trials within each participant (paired *t*-test) at each node of the standard surface. Anatomical labels were obtained from the most recent FreeSurfer atlas (Destrieux et al., 2010) and the average dataset was visualized on the N27 atlas surface. The group analysis revealed significant activation in occipital lobe, so individual subject maps of occipital activation were created using the following atlas labels from (Destrieux et al., 2010): collateral, parieto-occipital, cuneus, calcarine, occipito-temporal medial, occipital superior transversal, occipital superior, occipital middle, occipital middle lunatus, occipital pole, occipital inferior.

Cross-participant analysis using Mixed Models

The average BOLD fMRI response in each ROI in each hemisphere was calculated and entered into linear mixed-effects models created with the *lme4* package in R (Bates et al. 2015). We compared two models, both with BOLD percent signal change as the dependent variable. In the first model we applied the fixed factors of ROI (mouth, eye) hemisphere (left, right) and stimulus (auditory, visual, audiovisual). The second model was identical,

except that it excluded the hemisphere factor. For both models, there were two random factors: participant and participant by stimulus interaction. The models were compared using the Bayesian Information Criterion (BIC). The BIC of the model without the hemisphere factor was significantly smaller (358.8 *vs.* 408.2) indicating a better representation of this data, and only the results of this model are reported in the manuscript. For each statistical test, the degrees of freedom, t value, and p value were calculated according to the Satterthwaite approximation using the *lmerTest* package (Kuznetsova et al., 2015) and ANOVA-like tests (Type II Wald chi square test resulting in χ^2 and p values) were calculated using the *Anova* function of the *car* package.

Results

Two independent fMRI experiments were performed. In the first experiment, participants freely viewed dynamic talking faces while their eye movements were monitored. *Post hoc* trial sorting was used to identify brain areas that responded more when participants fixated either the mouth or eyes of the talker. In the second experiment, participants viewed blocks of auditory, visual, and audiovisual speech in order to determine the functional properties of mouth and eye-selective regions.

Eye movements

In the first experiment, eye movements were recorded from 34 participants in the MR scanner as they viewed 8160 trials of brief audiovisual movies of talking faces. Fixations to the talker's face accounted for 91% of the total fixation time during the time window in which auditory speech and speech-related mouth movements occurred (from 400 ms to 1400 ms after stimulus onset, Figure 1A). Figure 1B shows the grand mean of face-looking behavior during this time window across all trials. Most fixations were located in the central region of the talker's face, especially the mouth and eyes. As shown in Figure 1C, we measured the amount of time participants spent fixating the lower half of the face (containing the mouth) and the upper half of the face (containing the eyes). Within individual trials, participants tended to fixate either the mouth or the eyes of the talker, but not both. This bimodal distribution led us to classify each trial as either a mouth-looking trial or an eye-looking trial. Figure 1D illustrates the average fixation pattern for both types of trials.

BOLD fMRI data sorted by eye movements

Next, we examined the BOLD fMRI data, using the eye movement recordings to classify each trial as either a mouth-viewing or eye-viewing trial. Note that the two trials types contained physically identical stimuli, but participants viewed them with different patterns of eye movements. Our initial analysis focused on an *a priori* anatomical region of interest (ROI), the posterior STS. First, an anatomically-defined pSTS ROI was created in the left and right hemispheres for each of the 34 participants (68 total hemispheres). Next, voxels within the pSTS that responded more strongly to either mouth or eye trials were mapped. Within each hemisphere, mouth-preferring voxels were grouped into a pSTS mouth ROI and eye-preferring voxels were grouped into a pSTS eye ROI. Most hemispheres contained both mouth and eye-preferring voxels in the pSTS (51/68).

As shown in Figure 2A and 2C for four example participants, mouth-preferring voxels were located more anteriorly in the pSTS, while eye-preferring voxels were located more posteriorly. To quantify this effect, we calculated the center of mass of the activation for the mouth and eye ROIs. The average Euclidean distance between the centers of mass of the mouth and eye ROIs (in 25 participants where both kinds of ROIs were present) was 17 ± 8 mm (SD) in standard space in the left hemisphere (*t*-test against zero: $t_{(24)} = 10.16$, $p = 5.6 \times 10^{-10}$) and 16 ± 8 mm in the right hemisphere (*t*-test against zero: $t_{(25)} = 9.70$, $p = 5.5 \times 10^{-10}$).

The primary driver of this effect was a more anterior location for mouth-preferring voxels (results for all cardinal axes shown in Table 1). There was no significant difference in volume of activation between the mouth and eye ROIs in either the left hemisphere (average volume: 313 ± 88 mm^3 for eye and 283 ± 93 mm^3 for mouth, unpaired *t*-test, $t_{(57)} = 0.32$, $p = 0.749$) or the right hemisphere (315 ± 85 mm^3 for mouth and 254 ± 82 mm^3 for eye, $t_{(58)} = 0.69$, $p = 0.490$).

To verify that the mouth and eye ROI responses were characteristic of those typically observed with BOLD fMRI, we examined the average hemodynamic response in both ROIs averaged across participants to single trials of audiovisual speech (Figures 2B and 2D). Both ROIs showed the characteristic BOLD response pattern, with a peak equal to a 0.5% deviation from the mean intensity at 4 to 6 seconds following stimulus onset, followed by a return to baseline and a post-undershoot below baseline. As the voxels were assigned to ROIs based on their preference for mouth or eye trials, we did not compare the response amplitudes to the two trial types to avoid bias.

Functional properties of mouth and eye preferring regions

Using the ROIs created from the fMRI and eye tracking data collected in the first experiment, we examined BOLD fMRI responses from the second experiment in which participants viewed blocks of unisensory and multisensory speech (this analysis was unbiased, because the first and second fMRI experiments were completely independent).

As shown in Figure 3A, responses to blocks of auditory speech and audiovisual speech in the second fMRI experiment were greater in the mouth ROI than in the eye ROI. This observation was quantified using linear mixed-effects models (complete results in Table 2). There were significant main effects for ROI (mouth *vs.* eye; $\chi^2_{(1)} = 34.26$, $p = 4.8 \times 10^{-9}$) and stimulus (A, V, AV speech; $\chi^2_{(2)} = 60.67$, $p = < 2.2 \times 10^{-16}$) as well as a significant interaction between ROI and stimulus ($\chi^2_{(2)} = 9.94$, $p = 0.007$). These effects were driven by larger responses in the mouth ROI to auditory speech (0.66% for mouth ROI vs 0.34% for eye ROI, $\chi^2_{(1)} = 18.53$, $p = 1.7 \times 10^{-5}$) and audiovisual speech (1.00% vs. 0.69%, $\chi^2_{(1)} = 12.00$, $p = 5.3 \times 10^{-4}$) but similar responses in the mouth and eye ROIs to visual speech (0.30% vs 0.24%, $\chi^2_{(1)} = 1.46$, $p = 0.23$). To assess the reliability of these effects, we plotted each participant separately (Figure 3B). The majority of hemispheres showed greater responses to auditory and audiovisual speech in the mouth ROI than in the eye ROI (39/51 for auditory; 38/51 for audiovisual) but not for visual speech (31/51).

Alternative classification methods

Trials were classified using a criterion that differed for each participant based on that participant's eye looking behavior. This had the advantage of evenly balancing the number of eye and mouth trials in the fMRI analysis for each participant, but could be criticized as comparing incommensurate items, since a trial classified as an eye trial in one participant might be considered a mouth trial in another participant. Therefore, we considered an alternative classification method in which the identical criterion was used for each participant: trials with greater than 50% of trial time spent on mouth fixations were classified as mouth trials and trials with greater than 50% of trial time spent on eye fixations were classified as eye trials. This method excluded 5/34 participants with few (or no) trials of one type or the other, but for the remaining participants, the results were consistent with the main analysis. Table 3 shows the complete statistical results of the linear mixed effects model. As for the median split criterion, using the 50/50 split criterion, the mouth ROI responded significantly more than the eye ROI to both auditory (0.69% *vs.* 0.38%, $p = 9.5 \times 10^{-5}$) and audiovisual (1.04% *vs.* 0.71%, $p = 9.2 \times 10^{-5}$) speech.

Whole brain analysis

Our initial analysis examined brain regions within the pSTS that preferred trials on which the eyes or mouth of the talker were fixated. This analysis was extended to the whole brain by normalizing each participant's cortical surface model to a surface template, allowing statistical comparisons at each node in standard space ($p < 0.05$, FDR corrected). As shown in Figure 4A and Table 4, the group average dataset revealed a number of brain regions that preferred eye or mouth trials.

The largest of these regions was in occipital lobe. To further study these occipital responses, we parcellated the occipital lobe in each individual participant. As shown in Figure 4B for eight sample hemispheres, mouth-preferring responses in occipital lobe were restricted to the occipital pole, extending slightly onto the lateral surface, while eye-preferring responses were more widespread and covered much of calcarine cortex and the medial face of the hemisphere.

To determine if mouth and eye-preferring occipital lobe regions showed responses similar to those observed in the pSTS, we calculated the response to the blocks of speech presented in the second fMRI experiment. Unlike in the pSTS, mouth and eye-preferring regions in visual cortex did not differ significantly in their response to auditory speech (−0.08% in mouth regions *vs.* 0.12% in eye regions; $\chi^2_{(1)} = 1.23$, $p = 0.27$), visual speech (1.24% *vs.* 1.11%; $\chi^2_{(1)} = 2.32$, $p = 0.13$) or audiovisual speech (0.89% *vs.* 0.84%; $\chi^2_{(1,97)} = 0.66$, $p = 0.42$). A linear mixed-effects model showed a significant main effect of stimulus (levels: auditory, visual, audiovisual speech; $\chi^2_{(2)} = 659.47$, $p = 2.0 \times 10^{-16}$), driven by a greater response to visual and audiovisual speech, but no significant effect of ROI (levels: mouth, eye ROI; $\chi^2_{(1)} = 3.49$, $p = 0.06$) or interaction ($\chi^2_{(2)} = 1.03$, $p = 0.60$).

The whole-brain group analysis revealed a number of other brain regions outside of visual cortex and pSTS that preferred mouth or eye trials. Since the total volume of these regions was less than occipital regions, we grouped them all together (excluding the pSTS and the

visual cortex) and calculated responses to blocks of speech presented in the second fMRI experiment. Mouth and eye-preferring regions did not differ in their response to auditory speech (0.15% in mouth regions *vs.* 0.22% in eye regions; $\chi^2_{(1)} = 2.61$, $p = 0.11$), visual speech (0.35% *vs.* 0.44%; $\chi^2_{(1)} = 2.75$, $p = 0.10$) or audiovisual speech (0.80% *vs.* 0.74%; $\chi^2_{(1)} = 1.48$, $p = 0.22$). A linear mixed-effects model showed a significant main effect of stimulus (levels: auditory, visual, audiovisual speech; $\chi^2_{(2)} = 115.97$, $p = 2.0 \times 10^{-16}$), driven by a greater response audiovisual speech, but no significant effect of ROI (levels: mouth, eye ROI; $\chi^2_{(1)} = 1.75$, $p = 0.19$) or interaction ($\chi^2_{(2)} = 2.06$, $p = 0.13$).

Behavioral data and correlations between variables

For the behavioral data, participants identified which of the three syllables was presented in each trial with a manual button press. Participants identified the congruent syllables (AbaVba and AgaVga) with high accuracy (mean 95%, range 75% − 100%, SD 7%) resulting in too few incorrect trials for meaningful analysis. However, there was significant variability in the responses to the incongruent McGurk syllable (AbaVga). Participants reported the McGurk fusion percept “da” on 60% of the trials and the auditory component of the syllable (“ba”) on the remaining 40% of trials. We tested whether there was a relationship between participants’ perceptual reports to McGurk syllables and their eye movements. For instance, trials on which participants fixated the mouth might be expected to result in more frequent perception of the McGurk effect. However, participants reported the illusion at similar rates during mouth and eye trials (64% *vs.* 60%, $t_{(33)} = 1.56$, $p = 0.15$).

Next, we searched for a relationship between McGurk perception and brain responses in the pSTS. In the first analysis, we generated ROIs using data collected in the first, event-related experiment in which eye movements were recorded. The ROIs consisted of pSTS voxels that responded to either mouth or eye trials; voxels that that responded more to mouth trials; and voxels that responded more to eye trials. The three types of ROIs were generated separately for the left and the right hemispheres, resulting in a total of six ROIs. To avoid bias, the BOLD response in each of the six ROIs was measured in the completely independent second experiment, in which participants listened to blocks of auditory, visual or audiovisual stories, resulting in a total of 18 fMRI measures for each participant (six ROIs generated from the first experiment times three conditions in the second experiment). Each of the 18 measures was then correlated with McGurk susceptibility across participants. Although there was significant variability in both McGurk susceptibility (ranging from 0% in some participants to 100% in others) and the fMRI measures (ranging from 0.1% to 1.5%), none of the 18 brain-behavior correlations were significant, even without correction for multiple comparisons (Table 5).

In the second analysis, the analysis was reversed: ROIs were generated using data collected in the second fMRI experiment, and the responses within the ROIs measured in the conditions of the first fMRI experiment.

The ROIs consisted of pSTS voxels that responded to either blocks of audiovisual stories (AV); or to both unisensory auditory and visual stories (A∩V). The two types of ROIs were generated separately for the left and the right hemispheres, resulting in a total of four ROIs. To avoid bias, the BOLD response in each of the ROIs was measured in the conditions of the

first experiment, consisting of all trials; trials on which the mouth was fixated more often; and trials on which the eyes were fixated more often. The 12 fMRI measures for each participant (four times three) were then correlated across participants with each participant's McGurk susceptibility. None of the 12 brain-behavior correlations were significant, even without correction for multiple comparisons (Table 6).

Discussion

Using simultaneous BOLD fMRI and infrared eye-tracking, we identified a region in anterior pSTS that responded more strongly to trials in which participants primarily fixated the mouth of the talker and a region in posterior pSTS that preferred trials in which participants primarily fixated the talker's eyes. In a second fMRI experiment, we showed that the anterior pSTS region responded more to auditory and audiovisual speech than the posterior pSTS region. These results are consistent with a model in which anterior pSTS serves as a locus for audiovisual speech perception, integrating visual information from mouth movements and auditory information from heard speech in the service of accurate speech perception.

Functional specialization within the STS

Functional neuroimaging has significantly advanced our understanding of face processing in the human STS. fMRI was used to demonstrate responses to static faces in the STS (Kanwisher et al., 1997) as well as STS responses to both mouth and eye movements (Puce et al., 1998) and a greater response to mouth movements than scrambled images (Puce et al., 2003), to radial movements patterns (Thompson et al., 2007) or still images of mouths (Calvert and Campbell, 2003). Hoffman and Haxby (2000) reported greater activity for eye observations compared to scrambled images while (Calder et al., 2002) demonstrated greater response for direct than averted gaze. Other studies (Pelphrey et al., 2005; Zhu and Beauchamp, 2017) directly compared eye and mouth movements and found more anterior pSTS preference for moving mouths and more posterior pSTS activity for moving eyes (Bernstein et al., 2011).

The present study replicated and extends the results of Zhu & Beauchamp (2017) and Pelphrey et al. (2005) by showing by showing functional specialization within the pSTS for viewing the eyes or mouth of a face. Critically, in the present study functional specialization was demonstrated using comparisons between brain responses to physically identical stimuli sorted by eye movements rather than between conditions containing silent videos in which only the eyes or only the mouth of the face moved, a scenario unlikely to be encountered in natural vision.

While we were most interested in the anterior pSTS mouth-preferring region, our study complements the recent study of Jiang et al. (2016) who focused on brain regions, including posterior pSTS, that were more active when participant fixated the eyes of talking faces. Like Jiang et al. (2016), we observed a preference for eye observations in the occipital lobe and adjacent inferior ventral and parieto-occipital cortex.

These results agree with evidence for shared cortical mechanisms for language processing and the perception of moving faces (Deen et al., 2015). We show that the overlapping activity pattern in pSTS observed by Deen and colleagues (2015) for the contrasts of [moving faces *vs.* objects] and [voice stimuli *vs.* environmental sounds] can be partially explained by an STS region that prefers both mouth movements and human voices.

An anterior-to-posterior organization of mouth and eye preferring regions is consistent with previous descriptions. In the present study, the center-of-mass of mouth-preferring activity in the pSTS was located 6 mm more anterior than eye-preferring activity (y = −43 *vs.* −49 in the left hemisphere, −40 *vs.* −46 on the right; see Table 1), similar to the values reported by Zhu & Beauchamp (2017) (6.5 mm; y = −46 *vs.* −52 in left hemisphere and y = −43 *vs.* −50, right hemisphere). Pelphrey et al. (2005) reported that right-hemisphere mouth-preferring activity was 21 mm anterior to eye-preferring activity (y = −37 *vs.* −58) while Jiang et al. (2016) reported 24 mm, y = −33 *vs.* −57). The larger anterior-posterior differences in these studies could be attributable to their use of spatial-smoothing filters on the fMRI data. The present study, like Zhu & Beauchamp (2017), used no spatial smoothing and within-subject comparisons.

In a relevant study, Bernstein and colleagues (Bernstein et al., 2011) presented six different types of visual-only face movements: videos of visual speech (eleven different syllables); videos of non-speech facial gestures (eleven different types, including yawns, smirks, chews, and kisses); point-light displays of visual speech; point-light displays of non-speech facial gestures; and scrambled control stimuli for the video and the point-light displays. Bernstein and colleagues found that visual speech (in both video and point-light form) activated regions of the STS that were more anterior than regions responding to non-speech facial gestures. This raises important questions about the fine-scale parcellation of the STS that will require additional studies to address. Posterior regions of the STS are active during viewing of video or point-light displays of whole body movements (Beauchamp et al., 2003) as well as during viewing of eye movements (Pelphrey et al., 2005; Zhu and Beauchamp, 2017) and non-speech facial gestures (Bernstein et al., 2011). It is unknown if posterior STS responses to whole body movements, facial gestures, and eye movements are anatomically segregated into separate subregions within the pSTS. If they are not, it suggests a two-compartment model of the STS (anterior: speech *vs.* posterior: all other biological motion).

Why did participants fixate the mouth of the talker on some trials and the eye of the talker on other trials?

The resolution of the human visual system varies greatly depending on the location in the visual field, from 1 arc minute (minimum angle of resolution) at the center of gaze to 20 arc minutes at 20 degrees in the periphery. Foveating visual targets therefore greatly increases the amount of visual information available about the target, at the cost of decreasing the information available about non-foveated targets. Observers who fixate the eyes of a talking face will therefore have less information about the talker's mouth and *vice versa*. This decrease in visual information was tested using the McGurk Effect by Paré and colleagues (2003). Participants perceived the Effect (illustrating the influence of information about the talker's mouth) on 76% of trials when foveating the mouth but perceived the effect on only

56% of trials when the mouth was at 20 degrees in the periphery, and on 12% of trials when the mouth was at 60 degrees in the periphery. Fixating the mouth of the talker provides participants with the greatest access to mouth information, while fixating the eyes provides less information about the talker's mouth movements (but presumably more information about the talker's gaze location). Our neuroimaging data suggest that these results may be related to enhanced responses in mouth-preferring regions of the STS during fixation of the talker's mouth.

Of course, some mouth information can be extracted even if the mouth of the talker is in the visual periphery. Temporal cues provided by visual mouth movements contribute to enhanced auditory speech perception, and this enhancement can be produced by visual cues that contain a temporally aligned visual stimulus, even if it is not a mouth (Munhall et al., 2004; Tjan et al., 2014).

Early visual cortex is thought to play an important role in visual tasks requiring high spatial acuity. Detecting eye gaze or emotional content in eyes may require higher acuity than detection of mouth movements, providing an explanation for the greater activation of early visual cortex during eye trials compared with mouth trials.

We observed a bimodal distribution of eye movement patterns. When viewing repeated presentations of identical audiovisual speech stimuli, on some trials participants primarily fixated the mouth of the talker, while on other trials, participants primarily fixated the eyes of the talker. This distribution was likely influenced by several factors, including the behavioral task. During presentation of noisy audiovisual speech, participants preferentially fixate the mouth to extract visual speech information (Buchan et al., 2008; Vatikiotis-Bateson et al., 1998). In the present study, while no noise was added to the auditory speech, the noise of the MR scanner may have contributed to a predominance of trials in which participants fixated the mouth *vs.* fixating the eyes. A second contributor to the distribution of eye and mouth looking trials is the intrinsic preference of observers to fixate different parts of the face (Gurler et al., 2015; Peterson and Eckstein, 2013, 2012). Some observers prefer to fixate the eyes of a face, while others prefer to fixate the mouth, a preference that is stable over time scales of at least 18 months (Mehoudar et al., 2014).

While both the behavioral task and the intrinsic preference of participants are likely important contributors to the distribution of mouth and eye trials, our finding of STS eye and mouth voxels cannot be explained simply by the use of an auditory stimulus and task, as previous studies have demonstrated mouth or eye preferring voxels in the STS using completely silent stimuli (Hoffman and Haxby, 2000; Pelphrey et al., 2005; Zhu and Beauchamp, 2017).

Multisensory Enhancement in pSTS

Many studies have shown multisensory enhancement in the pSTS, with a greater response to audiovisual than unisensory auditory or visual stimuli (Beauchamp et al., 2004a, 2004b, van Atteveldt et al., 2004, 2007; Wright et al., 2003). Our results replicate these findings and demonstrate a similar degree of enhancement in mouth and eye ROIs, suggesting that

multisensory integration may be a general property of pSTS and not restricted to specific functionally-specialized zones.

Correlation of pSTS activity and McGurk susceptibility

Previous studies have reported a relationship between activity in the left pSTS and perception of the McGurk Effect: participants with a higher BOLD response were more likely to perceive the illusory fusion percept (Nath et al., 2011; Nath and Beauchamp, 2012). Using a variety of different analysis strategies to measure activity in the pSTS, we calculated 30 different brain-behavior correlations, but none of them were significant (Tables 5 and 6).

Experimental differences may explain our failure to replicate (Nath et al., 2011; Nath and Beauchamp, 2012). Different stimuli were used in the current study, and there is a large influence of stimulus on McGurk susceptibility (Magnotti and Beauchamp, 2015; Mallick et al., 2015); even ostensibly similar face stimuli can evoke very different fMRI responses (Westfall et al., 2016). In the current study, perceptual data were collected within the MRI scanner (as opposed to outside the scanner as in the earlier studies). Added auditory noise increases McGurk susceptibility (Fixmer and Hawkins, 1998). The high levels of auditory noise in the bore of the MR scanner may have increased estimates of McGurk susceptibility and decreased the brain-behavior correlation.

A second possible explanation is that the earlier studies (Nath et al., 2011; Nath and Beauchamp, 2012) were underpowered. Underpowered studies may inflate observed correlations between behavioral measures and BOLD signal changes, reducing replicability (Cremers et al., 2017; Yarkoni, 2009). Given the large interindividual variability in the McGurk effect and other multisensory phenomena, almost all published estimates of group differences in multisensory integration are inflated (Magnotti and Beauchamp, 2018). Since the present study used a much larger sample size than the two earlier studies (34 participants in the present study *vs.* 14 participants in Nath and Beauchamp, 2012 and 17 in Nath et al., 2011) it may provide a more accurate assessment of the true effect size.

Clinical relevance and conclusions

These findings have implications for clinical disorders in which impaired social cognition and face processing is apparent, especially autism spectrum disorder. Abnormalities in eye movements during face viewing have been frequently reported in ASD, both avoidance of eyes in faces expressing emotions (Kliemann et al., 2010; Klin et al., 2002; Neumann et al., 2006; Spezio et al., 2007) and avoidance of the mouth region of a talking face (Grossman et al., 2015; Irwin and Brancazio, 2014). We speculate that these abnormal eye movements might be linked to abnormalities in the two important subdivisions of the pSTS, mouth-preferring regions in anterior pSTS and eye-preferring regions in posterior pSTS.

Conclusions

Our results provide new insights into the functional specialization of the human pSTS. The results are relevant to natural conditions because these regions were identified based on natural viewing behaviors during free viewing of talking faces. The posterior mouth-preferring subregion of pSTS has a number of qualities that suggest a critical role in

audiovisual speech perception. First, it lies anterior to the eye-preferring pSTS subregion and is adjacent to auditory cortex, an ideal anatomical location for integrating auditory and visual information. Second, it shows strong responses to both auditory speech information (unisensory auditory) and visual speech information in the form of mouth movements (unisensory visual) as would be expected for a region that integrates auditory and visual speech.

Acknowledgements

This work was supported by the National Institutes of Health (R01NS065395 to M.S.B) and the Deutsche Forschungsgemeinschaft (RE 3693/1–1 to J.R.). We acknowledge the Core for Advanced MRI at Baylor College of Medicine.

References

Argall BD, Saad ZS, Beauchamp MS, 2006 Simplified intersubject averaging on the cortical surface using SUMA. Hum. Brain Mapp 27, 14–27. doi:10.1002/hbm.20158 [PubMed: 16035046]

Beauchamp MS, Argall BD, Bodurka J, Duyn JH, Martin A, 2004a Unraveling multisensory integration: patchy organization within human STS multisensory cortex. Nat. Neurosci 7, 1190–2. doi:10.1038/nn1333 [PubMed: 15475952]

Beauchamp MS, Lee KE, Argall BD, Martin A, 2004b Integration of auditory and visual information about objects in superior temporal sulcus. Neuron 41, 809–23. doi:10.1016/S0896-6273(04)00070-4 [PubMed: 15003179]

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. doi:10.1162/089892903770007380 [PubMed: 14614810]

Belin P, Zatorre RJ, Lafaille P, Ahad P, Pike B, 2000 Voice-selective areas in human auditory cortex. Nature 403, 309–12. doi:10.1038/35002078 [PubMed: 10659849]

Bernstein LE, Jiang J, Pantazis D, Lu Z-L, Joshi A, 2011 Visual phonetic processing localized using speech and nonspeech face gestures in video and point-light displays. Hum. Brain Mapp 32, 1660–76. doi:10.1002/hbm.21139 [PubMed: 20853377]

Brainard DH, 1997 The Psychophysics Toolbox. Spat. Vis 10, 433–6. [PubMed: 9176952]

Buchan JN, Paré M, Munhall KG, 2008 The effect of varying talker identity and listening conditions on gaze behavior during audiovisual speech perception. Brain Res. 1242, 162–71. doi:10.1016/j.brainres.2008.06.083 [PubMed: 18621032]

Calder AJ, Lawrence AD, Keane J, Scott SK, Owen AM, Christoffels I, Young AW, 2002 Reading the mind from eye gaze. Neuropsychologia 40, 1129–38. doi:10.1162/089892903321107828 [PubMed: 11931917]

Calvert GA, Campbell R, 2003 Reading speech from still and moving faces: the neural substrates of visible speech. J. Cogn. Neurosci 15, 57–70. doi:10.1162/089892903321107828 [PubMed: 12590843]

Chang EF, Rieger JW, Johnson K, Berger MS, Barbaro NM, Knight RT, 2010 Categorical speech representation in human superior temporal gyrus. Nat. Neurosci 13, 1428–1432. doi:10.1038/nn.2641 [PubMed: 20890293]

Cox RW, 1996 AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. Comput. Biomed. Res 29, 162–73. [PubMed: 8812068]

Cremers HR, Wager TD, Yarkoni T, 2017 The relation between statistical power and inference in fMRI. PLoS One 12, e0184923. doi:10.1371/journal.pone.0184923 [PubMed: 29155843]

Dale AM, 1999 Optimal experimental design for event-related fMRI. Hum. Brain Mapp 8, 109–14. [PubMed: 10524601]

Dale AM, Fischl B, Sereno MI, 1999 Cortical Surface-Based Analysis I. Segmentation and Surface Reconstruction. Neuroimage 9, 179–194. doi:10.1006/nimg.1998.0395 [PubMed: 9931268]

Deen B, Koldewyn K, Kanwisher N, Saxe R, 2015 Functional Organization of Social Perception and Cognition in the Superior Temporal Sulcus. Cereb. Cortex 25, 4596–609. doi:10.1093/cercor/bhv111 [PubMed: 26048954]

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. doi:10.1016/j.neuroimage.2010.06.010 [PubMed: 20547229]

Fischl B, Sereno MI, Dale AM, 1999 Cortical Surface-Based Analysis II: Inflation, Flattening, and a Surface-Based Coordinate System. Neuroimage 9, 195–207. doi:10.1006/nimg.1998.0396 [PubMed: 9931269]

Fixmer E, Hawkins S, 1998 The influence of quality of information on the McGurk effect. Proc. Int. Conf. Audit. Speech Process 27–32.

Grossman RB, Steinhart E, Mitchell T, McIlvane W, 2015 "Look who's talking!" Gaze Patterns for Implicit and Explicit Audio-Visual Speech Synchrony Detection in Children With High-Functioning Autism. Autism Res 8, 307–16. doi:10.1002/aur.1447 [PubMed: 25620208]

Gurler D, Doyle N, Walker E, Magnotti J, Beauchamp M, 2015 A link between individual differences in multisensory speech perception and eye movements. Atten. Percept. Psychophys 77, 1333–41. doi:10.3758/s13414-014-0821-1 [PubMed: 25810157]

Hoffman EA, Haxby J V, 2000 Distinct representations of eye gaze and identity in the distributed human neural system for face perception. Nat. Neurosci 3, 80–4. doi:10.1038/71152

Irwin JR, Brancazio L, 2014 Seeing to hear? Patterns of gaze to speaking faces in children with autism spectrum disorders. Front. Psychol 5, 397. doi:10.3389/fpsyg.2014.00397 [PubMed: 24847297]

Jiang J, Borowiak K, Tudge L, Otto C, von Kriegstein K, 2016 Neural mechanisms of eye contact when listening to another person talking. Soc. Cogn. Affect. Neurosci. nsw127. doi:10.1093/scan/nsw127

Kanwisher N, McDermott J, Chun MM, 1997 The fusiform face area: a module in human extrastriate cortex specialized for face perception. J. Neurosci 17, 4302–11. doi:10.3410/f.717989828.793472998 [PubMed: 9151747]

Kliemann D, Dziobek I, Hatri A, Steimke R, Heekeren HR, 2010 Atypical reflexive gaze patterns on emotional faces in autism spectrum disorders. J. Neurosci 30, 12281–7. doi:10.1523/JNEUROSCI.0688-10.2010 [PubMed: 20844124]

Klin A, Jones W, Schultz R, Volkmar F, Cohen D, 2002 Visual Fixation Patterns During Viewing of Naturalistic Social Situations as Predictors of Social Competence in Individuals With Autism. Arch. Gen. Psychiatry 59, 809. doi:10.1001/archpsyc.59.9.809 [PubMed: 12215080]

Kuznetsova A, Brockhoff P, Christensen R, 2015 Package 'lmerTest.' R Packag. version 2.

Magnotti JF, Beauchamp MS, 2018 Published estimates of group differences in multisensory integration are inflated. BioRxiv 1–19.

Magnotti JF, Beauchamp MS, 2015 The noisy encoding of disparity model of the McGurk effect. Psychon. Bull. Rev 22, 701–9. doi:10.3758/s13423-014-0722-2 [PubMed: 25245268]

Mallick DB, Magnotti JF, Beauchamp MS, 2015 Variability and stability in the McGurk effect: contributions of participants, stimuli, time, and response type. Psychon. Bull. Rev 22, 1299–307. doi:10.3758/s13423-015-0817-4 [PubMed: 25802068]

Mehoudar E, Arizpe J, Baker CI, Yovel G, 2014 Faces in the eye of the beholder: unique and stable eye scanning patterns of individual observers. J. Vis 14, 6. doi:10.1167/14.7.6

Moeller S, Yacoub E, Olman CA, Auerbach E, Strupp J, Harel N, Uğurbil K, 2010 Multiband multislice GE-EPI at 7 tesla, with 16-fold acceleration using partial parallel imaging with application to high spatial and temporal whole-brain fMRI. Magn. Reson. Med 63, 1144–53. doi:10.1002/mrm.22361 [PubMed: 20432285]

Munhall KG, Jones JA, Callan DE, Kuratate T, Vatikiotis-Bateson E, 2004 Visual prosody and speech intelligibility: head movement improves auditory speech perception. Psychol. Sci 15, 133–7. doi:10.1111/j.0963-7214.2004.01502010.x [PubMed: 14738521]

Nath AR, Beauchamp MS, 2012 A neural basis for interindividual differences in the McGurk effect, a multisensory speech illusion. Neuroimage 59, 781–7. doi:10.1016/j.neuroimage.2011.07.024 [PubMed: 21787869]

Nath AR, Fava EE, Beauchamp MS, 2011 Neural correlates of interindividual differences in children's audiovisual speech perception. J. Neurosci 31, 13963–71. doi:10.1523/JNEUROSCI.2605-11.2011 [PubMed: 21957257]

Neumann D, Spezio ML, Piven J, Adolphs R, 2006 Looking you in the mouth: abnormal gaze in autism resulting from impaired top-down modulation of visual attention. Soc. Cogn. Affect. Neurosci 1, 194–202. doi:10.1093/scan/nsl030 [PubMed: 18985106]

Paré M, Richler RC, ten Hove M, Munhall KG, 2003 Gaze behavior in audiovisual speech perception: the influence of ocular fixations on the McGurk effect. Percept. Psychophys 65, 553–567. doi: 10.3758/BF03194582 [PubMed: 12812278]

Pelli DG, 1997 The VideoToolbox software for visual psychophysics: transforming numbers into movies. Spat. Vis 10, 437–42. [PubMed: 9176953]

Pelphrey KA, Morris JP, Michelich CR, Allison T, McCarthy G, 2005a Functional anatomy of biological motion perception in posterior temporal cortex: An fMRI study of eye, mouth and hand movements. Cereb. Cortex 15, 1866–1876. doi:10.1093/cercor/bhi064 [PubMed: 15746001]

Pelphrey KA, Morris JP, Michelich CR, Allison T, McCarthy G, 2005b Functional anatomy of biological motion perception in posterior temporal cortex: an FMRI study of eye, mouth and hand movements. Cereb. Cortex 15, 1866–76. doi:10.1093/cercor/bhi064 [PubMed: 15746001]

Peterson MF, Eckstein MP, 2013 Individual differences in eye movements during face identification reflect observer-specific optimal points of fixation. Psychol. Sci 24, 1216–25. doi: 10.1177/0956797612471684 [PubMed: 23740552]

Peterson MF, Eckstein MP, 2012 Looking just below the eyes is optimal across face recognition tasks. Proc. Natl. Acad. Sci. U. S. A 109, E3314–23. doi:10.1073/pnas.1214269109 [PubMed: 23150543]

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. [PubMed: 9482803]

Puce A, Syngeniotis A, Thompson JC, Abbott DF, Wheaton KJ, Castiello U, 2003 The human temporal lobe integrates facial form and motion: evidence from fMRI and ERP studies. Neuroimage 19, 861–9. doi:10.1016/S1053-8119(03)00189-7 [PubMed: 12880814]

Saad ZS, Glen DR, Chen G, Beauchamp MS, Desai R, Cox RW, 2009 A new method for improving functional-to-structural MRI alignment using local Pearson correlation. Neuroimage 44, 839–848. doi:10.1016/j.neuroimage.2008.09.037 [PubMed: 18976717]

Setsompop K, Gagoski BA, Polimeni JR, Witzel T, Wedeen VJ, Wald LL, 2012 Blipped-controlled aliasing in parallel imaging for simultaneous multislice echo planar imaging with reduced g-factor penalty. Magn. Reson. Med 67, 1210–24. doi:10.1002/mrm.23097 [PubMed: 21858868]

Spezio ML, Adolphs R, Hurley RSE, Piven J, 2007 Abnormal use of facial information in high-functioning autism. J. Autism Dev. Disord 37, 929–39. doi:10.1007/s10803-006-0232-9 [PubMed: 17006775]

Thompson JC, Hardee JE, 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. doi:10.1016/j.neuroimage.2007.05.058 [PubMed: 17616403]

Tjan BS, Chao E, Bernstein LE, 2014 A visual or tactile signal makes auditory speech detection more efficient by reducing uncertainty. Eur. J. Neurosci 39, 1323–31. doi:10.1111/ejn.12471 [PubMed: 24400652]

van Atteveldt N, Formisano E, Goebel R, Blomert L, 2004 Integration of letters and speech sounds in the human brain. Neuron 43, 271–82. doi:10.1016/j.neuron.2004.06.025 [PubMed: 15260962]

van Atteveldt NM, Formisano E, Blomert L, Goebel R, 2007 The effect of temporal asynchrony on the multisensory integration of letters and speech sounds. Cereb. Cortex 17, 962–74. doi:10.1093/cercor/bhl007 [PubMed: 16751298]

Vatikiotis-Bateson E, Eigsti IM, Yano S, Munhall KG, 1998 Eye movement of perceivers during audiovisual speech perception. Percept. Psychophys 60, 926–40. doi:10.3758/BF03211929 [PubMed: 9718953]

Westfall J, Nichols TE, Yarkoni T, 2016 Fixing the stimulus-as-fixed-effect fallacy in task fMRI. Wellcome open Res. 1, 23. doi:10.12688/wellcomeopenres.10298.2 [PubMed: 28503664]

Wright TM, Pelphrey KA, Allison T, McKeown MJ, McCarthy G, 2003 Polysensory interactions along lateral temporal regions evoked by audiovisual speech. Cereb. Cortex 13, 1034–43. [PubMed: 12967920]

Yarkoni T, 2009. Big Correlations in Little Studies: Inflated fMRI Correlations Reflect Low Statistical Power-Commentary on Vul et al. (2009). Perspect. Psychol. Sci 4, 294–8. doi:10.1111/j.1745-6924.2009.01127.x [PubMed: 26158966]

Zhu LL, Beauchamp MS, 2017 Mouth and Voice: A Relationship between Visual and Auditory Preference in the Human Superior Temporal Sulcus. J. Neurosci 37, 2697–2708. doi:10.1523/JNEUROSCI.2914-16.2017 [PubMed: 28179553]

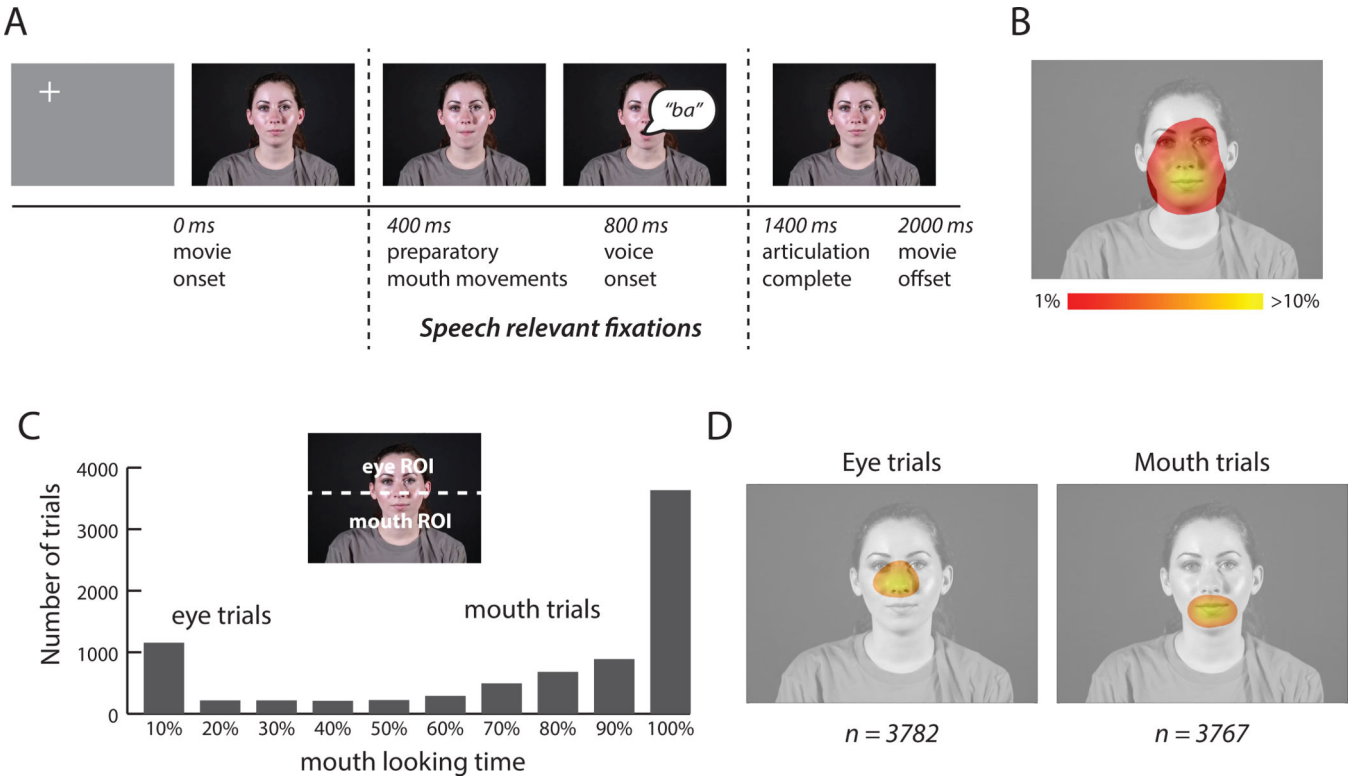

Figure 1:

Stimulus and eye movement analysis. **A.** Within each trial, participants viewed a 2-second duration audiovisual movie of a talker speaking a single syllable (still frames from single movie shown for illustration). Preparatory mouth movements began ~400 ms after stimulus onset, voice onset occurred at ~800 ms and articulation was complete by 1400 ms. Only the speech relevant fixations between 400 ms to 1400 ms were included in the analysis. **B.** Fixations for 34 participants viewing the audiovisual movies. Color scale indicates percent fixation time for each image location. **C.** Each movie was divided into an upper region, corresponding to the eye region of the face, and a lower region, corresponding to the mouth region of the face (dashed white line, not present in actual stimulus). For each trial, the percent of time fixating the eye and mouth regions of the face was calculated. The histogram shows the number of trials in each bin, with bins sorted by increasing amounts of time fixating the mouth. Within each participant, each trial was classified as an eye or a mouth trial, based on that participant's median fixation time. **D.** Average fixation locations across 34 participants for all eye and all mouth trials (*n* shows number of trials used for the fMRI analysis).

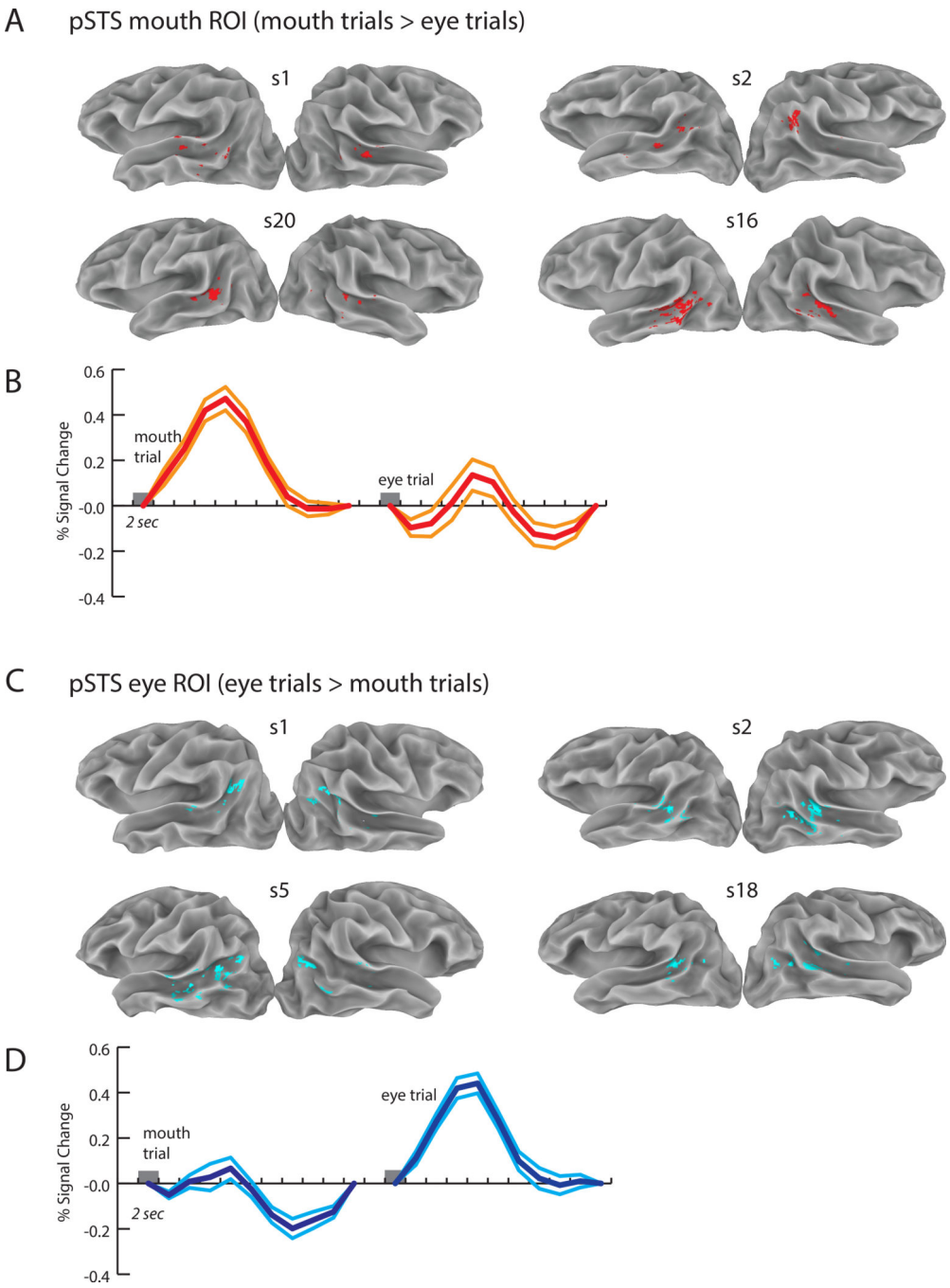

Figure 2:

Individual participant ROIs and mean BOLD fMRI responses. Eye movements were used to classify each trial as an eye trial (greater time spent fixating the eyes of the talker) or a mouth trial (greater time spent fixating the mouth of the talker). **A.** Eight individual hemispheres from four participants showing the pSTS mouth ROI, defined as regions of the anatomically-defined posterior STS (pSTS) that responded more strongly to mouth trials. Group map shown in Figure 4. **B.** Response across all 34 participants in the pSTS mouth ROI to mouth trials (left trace) and eye trials (right trace). Center trace shows the mean and

flanking lines show the SEM. Trial duration was 2 seconds (gray box). **C.** Location of the pSTS eye ROI, defined as regions of the pSTS responding more strongly to eye trials. **D.** Response across participants in the pSTS eye ROI for the two trial types.

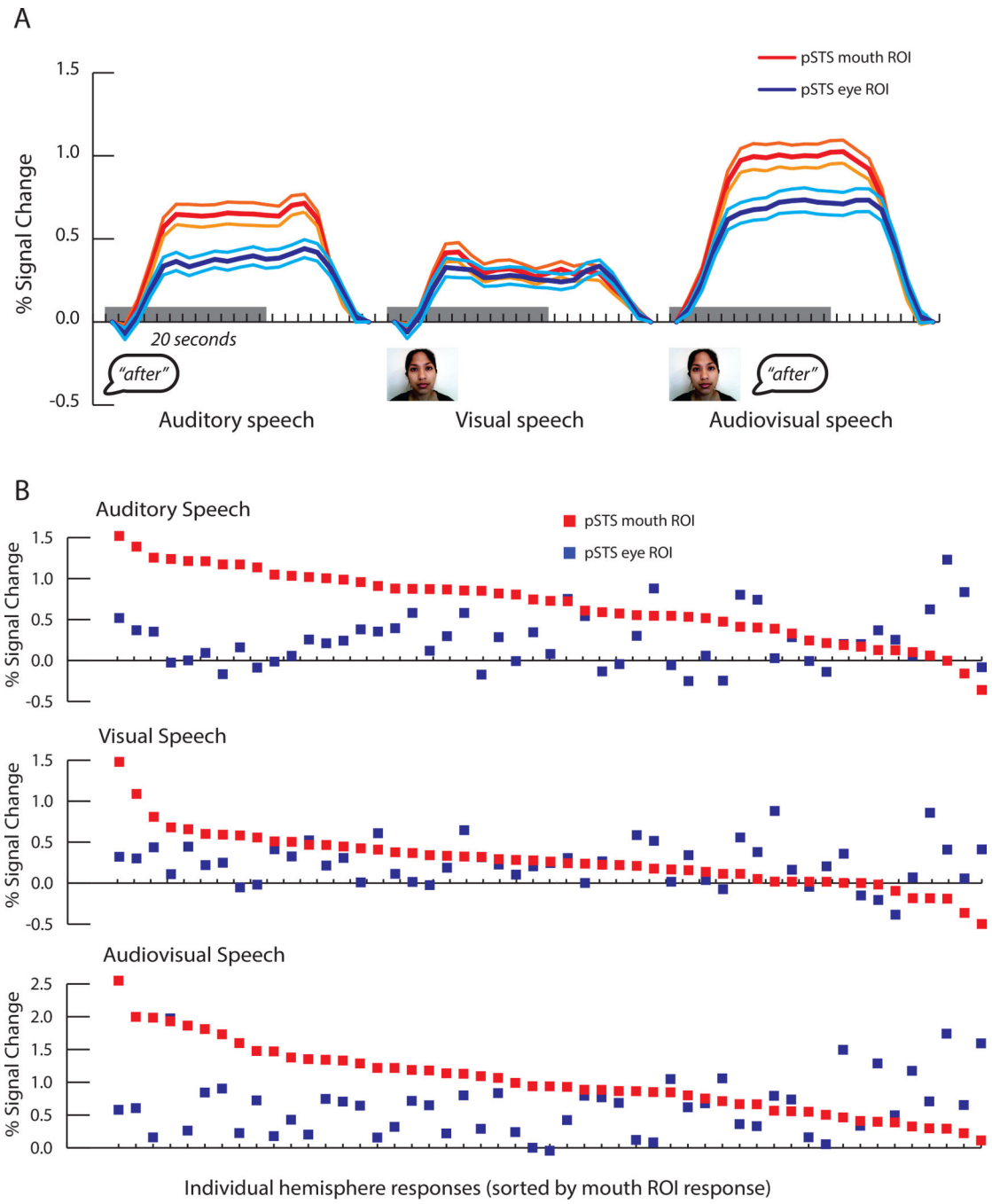

Figure 3:

Responses to blocks of speech in the second fMRI experiment. This data was analyzed independently from the eye tracking data used to define the ROIs in the first fMRI experiment. **A.** Time course of the average BOLD fMRI response to blocks of auditory, visual and audiovisual speech in the pSTS mouth (red traces) and eye ROIs (blue traces). Center lines show mean, flanking lines show SEM. Gray bar indicates 20-second stimulus duration. **B.** Response amplitudes within individual participants to blocks of auditory, visual and audiovisual speech. Each participant is represented by a vertical dot pair, where the red

dot represents the response in the mouth ROI and the blue dot represents the response in the eye ROI (participants sorted by decreasing amplitude of mouth ROI response).

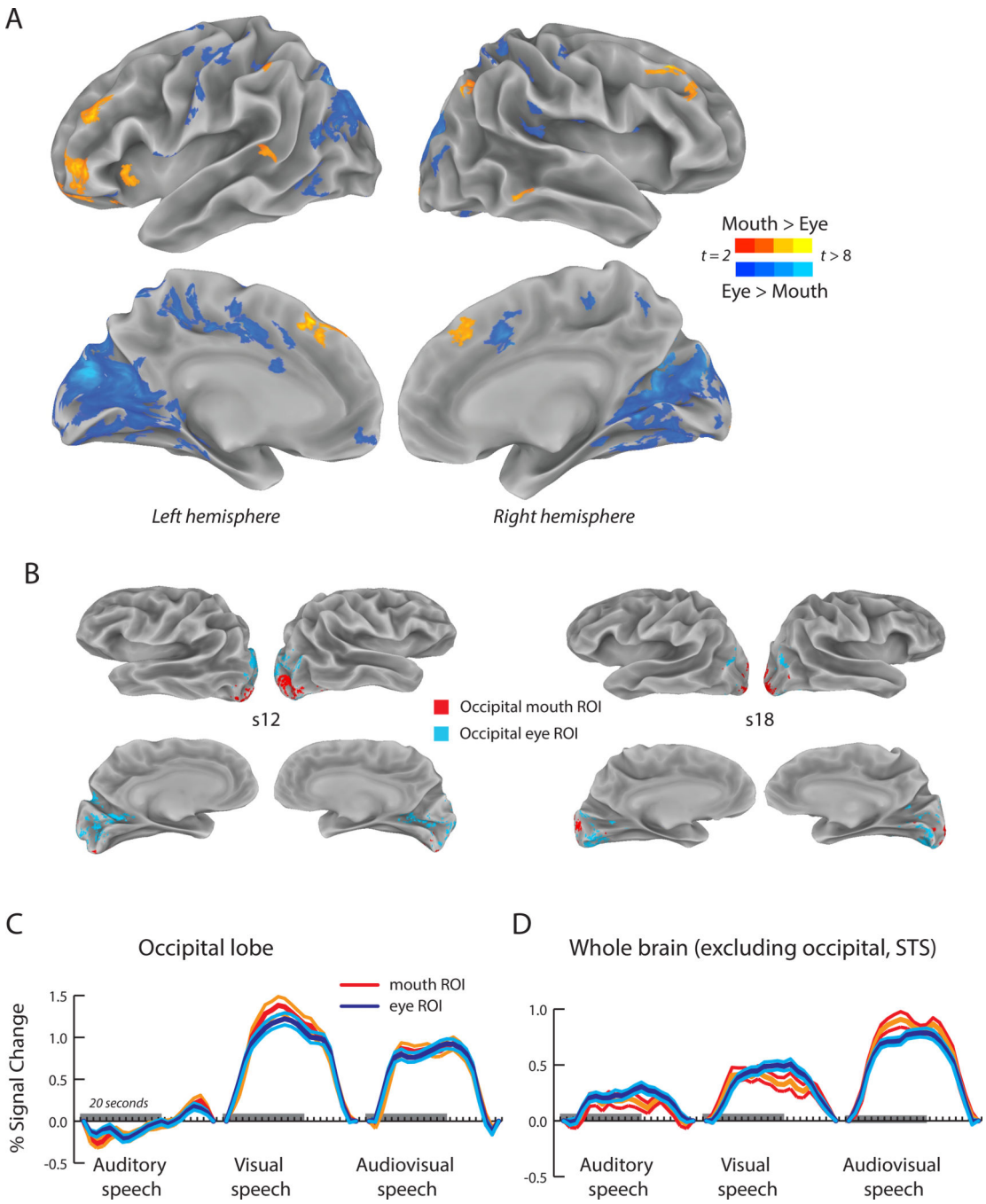

Figure 4:

A. Whole brain analysis using a cortical surface average across participants ($N = 34$) visualized on the standardized surface of the N27 atlas brain. Yellow color scale shows regions preferring trials in which participants fixated the mouth of the talker (yellow color scale), blue color scale shows regions preferring trials in which participants fixated the eyes of the talker. Only areas with a significant positive response to both mouth and eye trials ($q < 0.05$) and a surface area greater than 50 mm^2 are shown. **B.** The largest area of selectivity in the group map was in occipital lobe. Eight hemispheres from individual participants are

shown to illustrate interindividual variability. Colored regions show significant positive response to both mouth and eye trials and significant preference for either mouth trials (red) or eye trials (blue) in the first fMRI experiment. **C.** Responses of mouth-preferring and eye-preferring regions of occipital lobe were calculated in the second fMRI experiment in which blocks of auditory, visual and audiovisual speech were presented. Center lines show mean response, flanking lines show SEM. Gray bar indicates 20-second stimulus duration. **D.** Additional brain regions outside of occipital lobe and pSTS showed greater responses to mouth trials or eye trials. All mouth-preferring and eye-preferring regions (excluding occipital lobe and pSTS) in each participant were grouped, and the response in the second fMRI experiment calculated. Time course of the average response across participants to blocks of auditory, visual and audiovisual speech in the whole-brain mouth (red) and eye ROIs (blue).

Table 1.

ROI sizes and location

Average size and location of ROIs created from the within-subject contrast mouth vs. eye observations averaged across all participants (mean ± SEM). Talairach coordinates are given in mm for the x- (left/right position), y- (anterior/posterior) and z-axis (inferior/superior). The test values (*t, p*) indicate the statistical difference between position of mouth and eye ROIs for the respective axis using a paired t-test. Significantly different values are in bold.

		Talairach coordinates (mm) and statistical difference					
ROI	**Size (mm^3)**	**x**	***t, p***	**y**	***t, p***	**z**	***t, p***
Left pSTS							
Mouth	313 ± 88	**−56 ± 1**	2.20	**−43 ± 1**	2.60	9 ± 1	0.59
Eye	283 ± 93	**−52 ± 11**	0.038	**−49 ± 2**	0.016	11 ± 2	0.56
Right pSTS							
Mouth	315 ± 85	55 ± 1	1.10	**−40 ± 1**	2.51	**8 ± 1**	2.53
Eye	254 ± 82	53 ± 1	0.29	**−46 ± 2**	0.019	**11 ± 1**	0.018

Table 2.
Analyses of ROIs created from trials classified with individual median criterion

Participants viewed short trials of audiovisual speech. For each subject, all trials were sorted by the amount of time fixating the mouth region of the talker's face; trials with greater than the median amount of time spent fixating the mouth were classified as *mouth* trials and the remained were classified as *eye* trials. The contrast of brain activations between mouth *vs.* eye trials was used to define mouth and eye selective regions of the pSTS. The response of these mouth and eye ROIs to a separate experimental condition consisting of blocks of auditory-only (A), visual-only (V) and audiovisual (AV) speech were calculated for each participant and entered into a linear mixed-effects model. The linear mixed-effects model had fixed factors of ROI (mouth, eye) and stimulus (A, V, AV) with participant as a random factor and participant × stimulus as a random interaction. The first three rows show the mean response to A, V and AV conditions in mouth and eye ROIs and the contrast between mouth and eye ROIs using reduced linear mixed-effects (tested with chi square tests). The next rows show the parameter estimates for each factor in the model (the baseline condition was the response to A speech in the eye ROI). The final rows show the results of chi square tests of the models.

Mouth ROI *vs.* Eye ROI	***df***	χ^2	***p***	**Mouth ROI**	**Eyes ROI**
A Stim	1	18.53	1.7×10^{-5}	0.66%	0.34%
V Stim	1	1.46	0.23	0.30%	0.24%
AV Stim	1	12.00	5.3×10^{-4}	1.00%	0.69%

	Estimate	**Std. Error**	***df***	***t***	***p***
Fixed effects					
Baseline (A Stim)	0.330	0.059	88.6	5.56	2.9×10^{-7}
Mouth ROI	0.344	0.073	330.3	4.70	3.8×10^{-6}
AV Stim	0.344	0.072	316.0	4.81	2.3×10^{-6}
V Stim	−0.093	0.078	112.6	−1.18	0.240
Mouth ROI × AV Stim	0.002	0.102	322.7	0.02	0.983
Mouth ROI × V Stim	−0.281	0.103	326.7	−2.73	0.007
Random effects	Variance				
Participant	0.033				
Participant × AV Stim	0.000				
Participant × V Stim	0.035				

Main effects	***df***	χ^2	***p***
ROI	1	34.26	4.8×10^{-9}
Stim	2	60.67	$< 2.2 \times 10^{-16}$
Interaction			
ROI ×Stim	2	9.94	0.007

Table 3.
Analyses of ROIs created from trials classified with alternative method

Participants viewed short trials of audiovisual speech. For each subject, all trials were sorted by the amount of time fixating the mouth region of the talker's face; trials with greater than 50% of time spent fixating the mouth were classified as *mouth* trials and the remained were classified as *eye* trials. The contrast of brain activations between mouth *vs.* eye trials was used to define mouth and eye selective regions of the pSTS. The response of these mouth and eye ROIs to a separate experimental condition consisting of blocks of auditory-only (A), visual-only (V) and audiovisual (AV) speech were calculated for each participant and entered into a linear mixed-effects model. The linear mixed-effects model had fixed factors of ROI (mouth, eye) and stimulus (A, V, AV) with participant as a random factor and participant × stimulus as a random interaction. The first three rows show the mean response to A, V and AV conditions in mouth and eye ROIs and the contrast between mouth and eye ROIs using reduced linear mixed-effects (tested with chi square tests). The next rows show the parameter estimates for each factor in the model (the baseline condition was the response to A speech in the eye ROI). The final rows show the results of chi square tests of the models.

Mouth ROI *vs.* Eye ROI	***df***	χ^2	***p***	**Mouth ROI**	**Eye ROI**
A Stim	1	15.24	9.5×10^{-5}	0.69%	0.38%
V Stim	1	0.13	0.72	0.27%	0.25%
AV Stim	1	15.29	9.2×10^{-5}	1.04%	0.71%

	Estimate	**Std. Error**	***df***	***t***	***p***
Fixed effects					
Baseline (A)	0.387	0.056	228.5	6.91	4.7×10^{-11}
Mouth ROI	0.307	0.075	305.4	4.11	5.0×10^{-5}
AV Stim	0.325	0.073	290.6	4.42	1.4×10^{-5}
V Stim	−0.133	0.073	290.6	−1.81	0.071
Mouth ROI × AV Stim	0.027	0.104	290.6	0.26	0.797
Mouth ROI × V Stim	−0.283	0.104	290.6	−2.71	0.007
Random effects	Variance				
Participant	0.000				
Participant × AV Stim	0.029				
Participant × V Stim	0.004				

Main effects	***df***	χ^2	***p***
ROI	1	26.58	4.7×10^{-7}
Stimulus	2	137.83	$< 2.2 \times 10^{-16}$
Interaction			
ROI ×Stimulus	2	10.85	0.004

Table 4.

Whole-brain analysis

Size (measured as cortical surface area in mm^2) and location (in standard co-ordinates) of mouth-preferring and eye-preferring clusters identified during the whole-brain group analysis.

		Talairach coordinates (mm)		
Cluster	**Size (mm^2)**	**x**	**y**	**z**
Left hemisphere				
Mouth preferring				
Superior Parietal Gyrus	102	6	66	58
Eye preferring				
Middle Occipital Sulcus	2908	30	73	16
Calcarine Sulcus	859	24	66	7
Middle Anterior Cingulate Gyrus	132	4	−20	23
Right hemisphere				
Mouth preferring				
Superior Frontal Sulcus	155	−45	−19	40
Eye preferring				
Parieto Occipital Sulcus	3932	−16	64	19
Middle Anterior Cingulate Sulcus	293	−11	−3	37
Posterior Lateral Fissure	263	−40	33	13
Middle Occipital Sulcus	118	−38	72	10

Table 5.

Correlation between McGurk perception and BOLD signal change in ROIs from the first fMRI experiment

For each participant, six ROIs were generated using data from the first fMRI experiment (one table column per ROI). Audiovisual (AV) ROIs were created from all pSTS regions that showed a significant overall omnibus *F*-test ($F > 5$, $q < 0.0001$, false discovery rate corrected) and a significant positive response to mouth and eye trials ($q < 0.05$). The mouth ROI was restricted to pSTS regions that significantly preferred mouth trials. The eye ROI was restricted to pSTS regions that significantly preferred eye trials. The percent BOLD signal change in each ROI was calculated for the blocks of AV, A and V speech presented in the second fMRI experiment (one table row per condition). For each participant, the fraction of trials on which a McGurk fusion percept was reported in response to a McGurk stimulus was calculated. Across 34 participants, we correlated each ROI fMRI value with each participant's McGurk fraction to obtain *r* and *p* values. No correction for multiple comparisons was performed.

	AV ROI L	Mouth ROI L	Eye ROI L	AV ROI R	Mouth ROI R	Eye ROI R
AV	r = −0.23 p = 0.17	r = −0.32 p = 0.08	r = −0.12 p = 0.51	r = −0.01 p = 0.97	r = 0.10 p = 0.57	r = −0.03 p = 0.85
A	r = −0.15 p = 0.37	r = −0.15 p = 0.42	r = −0.09 p = 0.60	r = 0.08 p = 0.69	r = 0.08 p = 0.65	r = 0.10 p = 0.58
V	r = −0.11 p = 0.52	r = −0.35 p = 0.06	r = 0.27 p = 0.14	r = 0.07 p = 0.67	r = 0.16 p = 0.38	r = −0.10 p = 0.58

Table 6.

Correlation between McGurk perception and BOLD signal change in ROIs from the second fMRI experiment

For each participant, four ROIs were generated using data from the second fMRI experiment (one table column per ROI). AV ROIs were created from pSTS regions that showed a significant positive response to blocks of AV speech ($q < 0.01$). A∩V ROIs were created from pSTS regions that showed a significant positive response to both unisensory A and unisensory V stimulation (both $q < 0.05$). The percent BOLD signal change in each ROI was calculated for all trials, mouth-viewing trials, and eye-viewing trials in the first fMRI experiment (one table row per trial type). For each participant, the fraction of trials on which a McGurk fusion percept was reported in response to a McGurk stimulus was calculated. Across 34 participants, we correlated each ROI fMRI value with each participant's McGurk fraction to obtain r and p values. No correction for multiple comparisons was performed.

	AV ROI L	A∩V ROI L	AV ROI R	A∩V ROI R
All trials	r = 0.08 p = 0.67	r = 0.05 p = 0.76	r = 0.11 p = 0.52	r = 0.12 p = 0.51
Mouth trials	r = 0.03 p = 0.86	r = 0.02 p = 0.91	r = 0.07 p = 0.69	r = 0.07 p = 0.70
Eye trials	r = 0.12 p = 0.50	r = 0.09 p = 0.60	r = 0.16 p = 0.37	r = 0.17 p = 0.33