Distinct brain regions combine auditory representations with different visual streams

Abbreviated title: Audio-visual combination in the STS

Gabriel Fajardo[1*], Mengting Fang[2*], and Stefano Anzellotti[1]

[1] Boston College, Department of Psychology and Neuroscience, Chestnut Hill, MA, 02467
[2] University of Pennsylvania, Department of Psychology, Philadelphia, PA 19104

[*]: These authors contributed equally

To whom correspondence should be addressed: Stefano Anzellotti, E-mail: stefano.anzellotti@bc.edu

Number of pages: 20

Number of figures: 3

Number of tables: 3

Number of words Abstract: 70

Number of words Introduction: 623

Number of words Discussion: 953

Conflicts of interest:
The authors state no conflicts of interest.

Acknowledgements:

We thank Wei Qiu for technical support. We also thank the *StudyForrest* researchers for sharing their data. This work was supported by a startup grant from Boston College and by NSF grant 19438672 to Stefano Anzellotti.

Abstract

The superior temporal sulcus (STS) combines auditory and visual information. However, the extent to which it relies on visual information from the ventral or dorsal stream remains uncertain. We used artificial neural networks to investigate the relationship between response patterns in auditory cortex, the two visual streams, and the rest of the brain, finding that distinct portions of the STS combine information from the two visual streams with auditory information.

Significance Statement

In humans, the STS combines auditory and visual inputs. However, visual information is processed along a ventral and a dorsal stream, and the extent to which these streams contribute to the combination of audio-visual information is poorly understood. Is auditory information combined with visual information from both streams in a single centralized hub? Or do separate regions combine auditory information with ventral visual regions on one hand, and with dorsal visual regions on the other? To address this question, we employed a multivariate connectivity method based on artificial neural networks. Our findings reveal that information from the two visual streams is combined with auditory information in distinct portions of STS, offering new insights into the neural architecture underlying multisensory perception.

Keywords: audio-visual, multivariate statistical dependence, neural networks, superior temporal sulcus

Introduction

The human brain is adept at integrating visual and auditory information in order to create a coherent perception of the external world. Audio-visual integration contributes to sound localization (Zwiers et al., 2003), and plays a key role for emotion recognition (Piwek et al., 2015) as well as speech perception (Gentilucci and Cattaneo, 2005). Several phenomena demonstrate that the integration of visual and auditory cues shapes perceptual experience. In the McGurk effect, simultaneous presentation of a phoneme with a mismatched face video results in a distorted perception of the phoneme (McGurk and MacDonald, 1976). Similarly, presentation of mismatched auditory and visual stimuli can alter emotion recognition (Fagel, 2006), even when participants are explicitly instructed to focus only on one stimulus modality and ignore the other (Collignon et al., 2008), suggesting that audio-visual integration is automatic.

Audio-visual integration requires combining auditory information represented in the superior temporal gyrus with visual information encoded in occipitotemporal areas. Therefore, identifying brain regions that combine auditory and visual information is key for understanding the neural bases of audio-visual integration. Previous work found that the presentation of congruent audio-visual stimuli leads to supra-additive responses in the superior temporal sulcus (STS) compared to unimodal visual and auditory stimuli, whereas incongruent audio-visual stimuli leads to sub-additive responses (Calvert et al., 2000). In addition, participants' susceptibility to the McGurk effect correlates with the strength of STS responses (Nath and Beauchamp, 2012). Furthermore, response patterns in the STS encode information about emotions and identity that generalizes across visual and auditory modalities (Peelen et al., 2010; Anzellotti and Caramazza, 2017). These studies indicate that the STS plays a pivotal role in combining auditory and visual information.

However, little is known about the precise visual representations that are involved. Visual information is processed by multiple streams: a ventral and a dorsal stream (Ungerleider and Mishkin, 1982). The ventral stream originates in ventral area V3 (V3v) and area V4, and the dorsal stream in dorsal area V3 (V3d) and area V5 (Felleman and Van Essen, 1987) (Fig. 1a). Area V5 is associated with motion perception, featuring a large number of direction-selective neurons (Born and Bradley, 2005). By contrast, many neurons in V4 show sensitivity to color (Schein and Desimone, 1990). Correspondingly, a large number of neurons in the dorsal part of V3 respond to motion, and a large number of neurons in the ventral portion of V3 are tuned for color processing (Felleman and Van Essen, 1987). The existence of these different visual streams prompts questions about their relative contributions to the combination of visual and auditory information.

Auditory information could be combined with visual information from both streams, or with visual information from only one of the streams. If it is combined with visual information from

both streams, auditory information could be combined with information from both visual streams in a single hub, or distinct regions could combine auditory information with each visual stream separately. To investigate this, we used artificial neural networks to model the relationship between patterns of response in auditory brain regions, in the initial segments of the ventral and dorsal visual streams, and in the rest of the brain (Fig. 1b), following a strategy that has been recently adopted to investigate the combination of information from multiple category-selective regions (Fang et al., 2023). Functional magnetic resonance imaging (fMRI) data collected while participants viewed rich audio-visual stimuli (Hanke et al., 2016) were analyzed with multivariate pattern dependence networks (MVPN) (Anzellotti et al., 2017; Fang et al., 2022). Searching for brain regions where responses are better predicted using a combination of auditory responses and responses in different visual streams than using auditory or visual responses in isolation revealed two distinct portions of STS that combine information between auditory regions and the two visual streams.

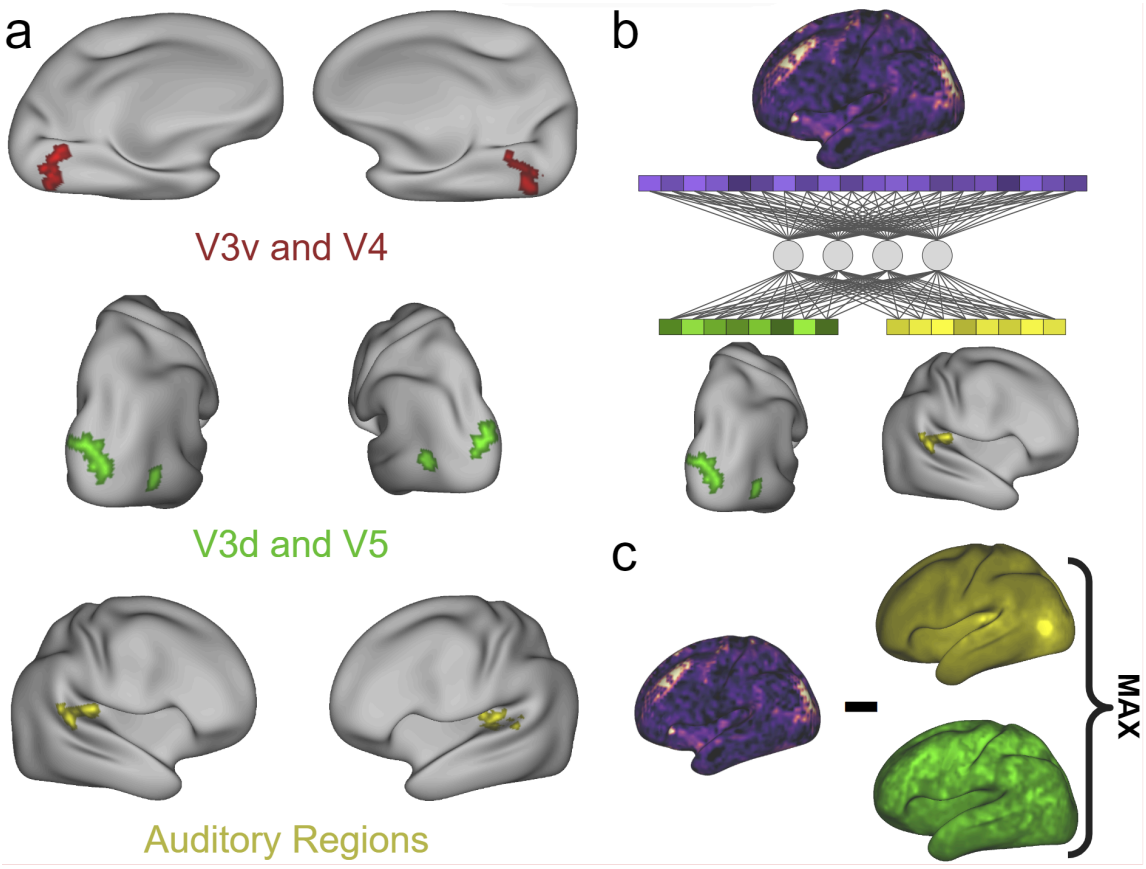

Fig. 1. a. Visual and auditory regions of interest (ROIs). **b.** Responses in a combination of visual (e.g., early dorsal visual stream; Fig. 1a, middle panel) and auditory regions were used to predict responses in the rest of the brain using MVPN. **c.** In order to identify brain regions that combine responses from auditory and visual regions, we identified voxels where predictions generated using the combined patterns from auditory regions and one set of visual regions jointly (as

shown in Fig. 1b) are significantly more accurate than predictions generated using only auditory regions or only that set of visual regions.

Materials and Methods

Experimental Design and Statistical Analyses

Experimental paradigm

The blood-oxygen-level-dependent (BOLD) functional magnetic resonance imaging (fMRI) data was obtained from the *StudyForrest* dataset (https://www.studyforrest.org) (Sengupta et al., 2016; Hanke et al., 2016). FMRI data was acquired while participants watched the movie 'Forrest Gump'. The movie was divided into 8 segments, each of which was approximately 15 minutes long. These segments were presented to subjects in chronological order in 8 separate scanner runs.

Data acquisition parameters

Fifteen right-handed subjects (6 females, 21-39 age range, mean = 29.4 years old), whose native language was German, were scanned in a 3T Philips Achieva dStream MRI scanner equipped with a 32 channel head coil. Functional MRI data was acquired with a T2*-weighted echo-planar imaging sequence (gradient-echo, 2s repetition time (TR), 30ms echo time, 90° flip angle, 1943 Hz/Px bandwidth, parallel acquisition with sensitivity encoding (SENSE) reduction factor). Scans captured 35 axial slices in ascending order, with 80 × 80 voxels (measuring 3.0 × 3.0 mm) of in-plane resolution, within a 240 mm field-of-view, utilizing an anterior-to-posterior phase encoding direction with a 10% gap between slices. The dataset also consists of root mean squared (RMS) annotations, which measure the loudness of the film.

Preprocessing

Data was first preprocessed using fMRIPrep (https://fmriprep.readthedocs.io/en/latest/index.html) (Esteban et al., 2019), a robust pipeline for preprocessing a wide range of fMRI data. Anatomical MRI images were skull-stripped using ANTs (http://stnava.github.io/ANTs/) (Avants et al., 2009), and FSL FAST was used for tissue segmentation. Functional MRI images were corrected for head movement using FSL MCFLIRT (https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/MCFLIRT) (Greve and Fischl, 2009), and were then coregistered with anatomical scans using FSL FLIRT (Jenkinson et al., 2002). Data was denoised with CompCor using 5 principal components extracted from the union of cerebrospinal fluid and

white matter (Behzadi et al., 2007). The raw data of one subject could not be preprocessed with the fMRIPrep pipeline. The remaining 14 subjects' data were used for the rest of the study.

ROI definition

Two sets of visual regions were identified by creating anatomical masks using Probabilistic Maps of Visual Topography in Human Cortex (Wang et al., 2015). This atlas provides probabilistic maps in MNI space of the likelihood that a voxel is a part of a certain brain region. The early ventral stream ROI was created by choosing the 80 voxels with the highest probability to be in the ventral parts of V3 (V3v) and V4 (Fig. 1a, top panel), and the early dorsal stream ROI was created by choosing the 80 voxels with the highest probability to be in the dorsal parts of V3 (V3d) and V5 (Fig. 1a, middle panel).

Since the anatomical location of auditory brain regions is more variable across subjects than visual brain regions (Rademacher et al., 2001), auditory ROIs were defined individually for each subject by identifying voxels where responses are parametrically modulated by the loudness of auditory stimuli. To this end, standard univariate GLM analyses were conducted using FSL FEAT (https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/FEAT) (Woolrich et al., 2001), with root mean square (RMS) levels as the predictor. The 80 voxels with the highest t-scores were selected individually for each subject (example of a subject's auditory ROI mask in Fig. 1a, bottom panel). To ensure that the remaining analyses are independent from the ROI selection, we used only data from the first fMRI run for auditory ROI selection, and this run was not used in the remaining analyses (which were therefore conducted on the remaining seven runs). There were no overlapping voxels between the ROIs.

Additionally, a group-average gray matter mask was created using the gray matter probability maps that were generated during preprocessing. This gray matter mask had a total of 53,539 voxels, and was used as the target of prediction in the multivariate pattern dependence analyses, explained in the following section.

MVPN: Multivariate Pattern Dependence Network

Recent research has taken advantage of the flexibility and computational power of artificial neural networks (ANNs) in order to analyze brain connectivity (Fang et al., 2022; Fang et al., 2023). The multivariate pattern dependence network (MVPN) method – an extension of MVPD (Anzellotti et al., 2017) – utilizes the power of ANNs to analyze the multivariate relationships between neural response patterns. It is important to note that MVPN measures the statistical relationship between response patterns in different regions, but it can not detect the direction of information flow. We implemented MVPN in PyTorch, and ANNs were trained on Tesla V100 graphics processing units (GPUs). In this study, we used 5-layer dense neural networks, as this

architecture produced the best predictive accuracy in prior work (Fang et al., 2022). The ANNs were given as input the multivariate response patterns in one or more sets of brain regions (Fig. 1): auditory regions, ventral visual regions (V3v and V4), dorsal visual regions (V3d and V5), and all pairwise combinations. ANNs were trained to predict the patterns of responses in all gray matter voxels.

More precisely, the MVPN method works as follows. Consider an fMRI experiment with m experimental runs. We label the multivariate time courses in a predictor region as $X_1, \ldots, X_m$. Each matrix X_i is of size $n_X \times T_i$, where n_X is the total number of voxels in the predictor region, and T_i is the number of timepoints in the i^{th} experimental run. Similarly, let $Y_1, \ldots, Y_m$ be the multivariate timecourses in the target region, where Y_i is an $n_Y \times T_i$ matrix, n_Y is the total number of voxels in the target region, and T_i is the number of timepoints in the i^{th} experimental run.

The neural networks were trained with a leave-one-run-out procedure to learn a function f such that

$$Y_{train} = f\left(X_{train}\right) + E_{train},$$

where X_{train} and Y_{train} are data in the predictor region and data in the target region, respectively, during training. E_{train} is the error term. Formally, for the i^{th} experimental run, data in the rest of the runs made up the training set $D_{\setminus i}$, where

$$D_{\setminus i} = \left\{\left(X_1, Y_1\right), \ldots, \left(X_{i-1}, Y_{i-1}\right), \left(X_{i+1}, Y_{i+1}\right), \ldots, \left(X_m, Y_m\right)\right\},$$

while the dataset $D_i = \left\{\left(X_i, Y_i\right)\right\}$ is the left out run i testing set.

We used the proportion of variance explained between the predictor region and all other voxels in the gray matter mask in order to measure multivariate statistical dependence. For each target region voxel j, the variance explained $varExpl_i(j)$ was calculated as

$$varExpl_i(j) = max\left\{0,\ 1 - \frac{var\left(Y_i(j) - f_j\left(X_i\right)\right)}{var\left(Y_i(j)\right)}\right\},$$

where X_i is the time course in the predictor region for the i^{th} run, and $f_j\left(X_i\right)$ is the MVPN prediction for the j^{th} voxel. The values $varExpl_i(j)$ obtained for the different runs $i = 1, \ldots, m$ were averaged, thus yielding $\overline{\text{varExpl}}(j)$.

Combined-minus-max whole-brain analysis

In order to identify brain regions that depend on the combination of auditory and visual response patterns, we analyzed the *StudyForrest* dataset with a novel approach we introduced in a recent study (Fang et al., 2023): the "combined-minus-max" approach, described in the following paragraphs. Since run 1 was used to functionally localize auditory regions (see the "ROI definition" section), to prevent circularity in the analysis, we used experimental runs 2 through 8 for the combined-minus-max analysis (a total of 7 runs).

In the combined-minus-max approach, first, we used MVPN to calculate the variance explained in each gray matter voxel using individual ROIs as predictors (early dorsal stream, early ventral stream, auditory stream). Then, we used pairs of these ROIs as joint inputs of the MVPN model in order to predict the neural responses of each gray matter voxel (Fig. 1b). We tested all pairs of the three streams: (1) posterior dorsal stream and auditory stream, (2) posterior ventral stream and auditory stream, and (3) posterior ventral stream and posterior dorsal stream.

If a voxel only encodes information from one of the streams, using the responses from multiple streams as predictors should not improve the variance explained. On the contrary, if the responses in the voxel are better predicted by a neural network including multiple streams combined than by a single stream, we can conclude that the voxel combines information from multiple streams. Therefore, we searched for voxels that combine information from multiple streams by computing an index given by the difference between the proportion of variance explained by a model using two streams jointly (the "combined" model), and the proportion of variance explained by a model using the best predicting stream among the two (the "max" model). This procedure is illustrated in Fig. 1c.

Formally, for each voxel j, we can compute the variance explained by MVPN using as input responses from pairs of ROIs, $varExpl_{pair}(j)$, and the variance explained using as input responses from the best-predicting individual ROIs, $varExpl_{max}(j)$. For each voxel j, the difference in variance explained is then calculated as

$$\Delta\, varExpl(j) = varExpl_{pair}(j) - varExpl_{max}(j) .$$

This $\Delta\, varExpl(j)$ gives us a multi-stream dependence (MSD) index for each voxel, that allowed us to identify candidate brain regions that jointly combine information from different streams. We calculated the statistical significance of $\Delta\, varExpl$ values across subjects using statistical non-parametric mapping, utilizing the SnPM extension for SPM (http://nisox.org/Software/SnPM13/) (Nichols and Holmes, 2001).

Control analysis

When using the combined-minus-max approach, there is still the possibility that the better predictive accuracy of the combined model might be due to the larger number of voxels in the combined analysis. To control for this possibility, we conducted a control analysis using voxels from the primary motor cortex (M1) as predictors (see Fang et al., 2023 as an example of an analogous approach). In this analysis, we randomly selected three non-overlapping groups of 80 voxels in M1 (this number was chosen to match the number of voxels selected from the three streams: the posterior ventral, posterior dorsal, and auditory). We then used the responses from the three groups of M1 voxels to run a control analysis following the same procedure as the combined-minus-max analysis, and we computed the statistical significance of $\Delta\, varExpl$ for each voxel in gray matter across subjects. Any regions showing statistical significance in this control analysis ($p<0.05$, FWE-corrected with SnPM) were due to the larger number of voxels in the combined model, not multi-stream information combination. Therefore, they were excluded from the multi-stream dependence (MSD) analysis described above.

Face-selective ROI analysis

Face perception requires the combination of both static and dynamic information (Dobs et al., 2014). In addition, some face-selective regions have been found to represent identity during the perception of both visual and auditory stimuli (Anzellotti and Caramazza, 2017). Therefore, we applied the combined-minus-max approach to investigate the multi-stream dependence effect in face-selective regions (Kanwisher et al., 2002; Yovel, 2016).

We used the first run in the category localizer to identify three face-selective ROIs: the occipital face area (OFA), the fusiform face area (FFA), and the face-selective posterior superior temporal sulcus (STS). Data were modeled with a standard GLM using FSL FEAT (Woolrich et al., 2001). Each seed ROI was defined as a sphere with a 9mm radius centered in the peak for the contrast faces > bodies, artifacts, scenes, scrambled images. Data from both the left and the right hemisphere were combined for each ROI, and the 80 voxels that showed the highest z-value for the contrast were selected. Visualizations of these ROIs can be found in Fig. 3a. We then analyzed the variance explained measures for each voxel in these face-selective ROIs across our three pairings (posterior dorsal stream and auditory stream, posterior ventral stream and auditory stream, and posterior dorsal stream and posterior ventral stream).

Code/Software Accessibility

The code to implement the analysis can be obtained at https://github.com/sccnlab/PyMVPD. A description of the code can be found in Fang et al. (2022).

Results

STS combines information from auditory regions with information from different visual streams.

To identify brain regions that jointly encoded information from different streams, we calculated the multi-stream dependence (MSD) index for each voxel. This index was computed as the difference between the proportion of variance explained by the combined model and that of the max model (see Materials and Methods section for a detailed explanation of the "combined-minus-max" approach). Group-level analyses were used to identify voxels with MSD indices significantly greater than zero. These voxels were considered as candidate multi-stream dependence brain regions. Clusters with peaks having $p<0.05$ (FWE corrected) were included.

To ensure that the combined model's predictive accuracy was not merely due to the larger number of voxels used in comparison to the max analysis, we conducted a control analysis. In the control analysis, we used three non-overlapping groups of 80 voxels from the primary motor cortex (M1) as predictors, matching the number of voxels used from the auditory cortex and two visual streams in the main analyses. We then ran the combined-minus-max analysis with these M1 voxel groups and obtained statistical significance for each gray matter voxel across subjects.

The control analysis showed significant effects in the sensorimotor cortex (peak MNI coordinates = [0, -21, 64], [33,-42,67], [-39, -18, 41]), premotor cortex (peak MNI coordinates = [-57, -9, 44], [57, 12, 31]), the bilateral intraparietal sulcus (peak MNI coordinates = [30, -69, 54], [-24, -72, 50]), and the angular gyrus (peak MNI coordinates = [-45, -69, 37]). Importantly, the control analyses did not show significant effects in ventral and lateral occipitotemporal regions. Therefore, significant findings in these regions in the main analysis could not be explained just by a difference between the number of predictor voxels in the combined analysis and the max analysis. Voxels that yielded significant effects in the control analysis ($p<0.05$, FWE-corrected) were excluded before calculating the MSD indices in the main analysis.

Combining response patterns from auditory regions and the early dorsal stream revealed significant effects in the bilateral STS (peak MNI coordinates = [-66, -42, 4], [45, -57, 18]) and within the posterior cingulate cortex (peak MNI coordinates = [15, -27, 41]) ($p<0.05$, FWE corrected). Combining responses from auditory regions and the early ventral stream also revealed effects in the right STS ($p<0.05$, FWE corrected), but in a more posterior portion (peak MNI coordinates = [48, -57, 8]), at the boundary with the occipital lobe (Fig. 2a).

These findings indicate that auditory information is not combined with information from both visual streams within one single STS hub. Instead, distinct portions of STS combine information from auditory regions and information from ventral and dorsal visual regions, respectively.

Ventral temporal cortex combines information from different visual streams.

These results raise the question of whether and where information from early dorsal (V3d and V5) and ventral (V3v and V4) visual regions is combined. We adopted the same strategy to test this, searching for voxels that are better predicted by both visual streams jointly than by either stream in isolation. This analysis identified regions in the calcarine sulcus (V1 and V2) that are located upstream of V3, V4, and V5, and in regions in ventral occipitotemporal cortex, that are located downstream (peak MNI coordinates = [21, -102, 1]) ($p<0.05$, FWE corrected, Fig. 2b). Notably, no effects for the combination of the two visual streams were observed in the STS. This is consistent with the finding that the combination of auditory information with different visual streams involves distinct cortical regions: if it happened in a single STS subregion, we would also expect to observe effects in that subregion for combining both visual streams.

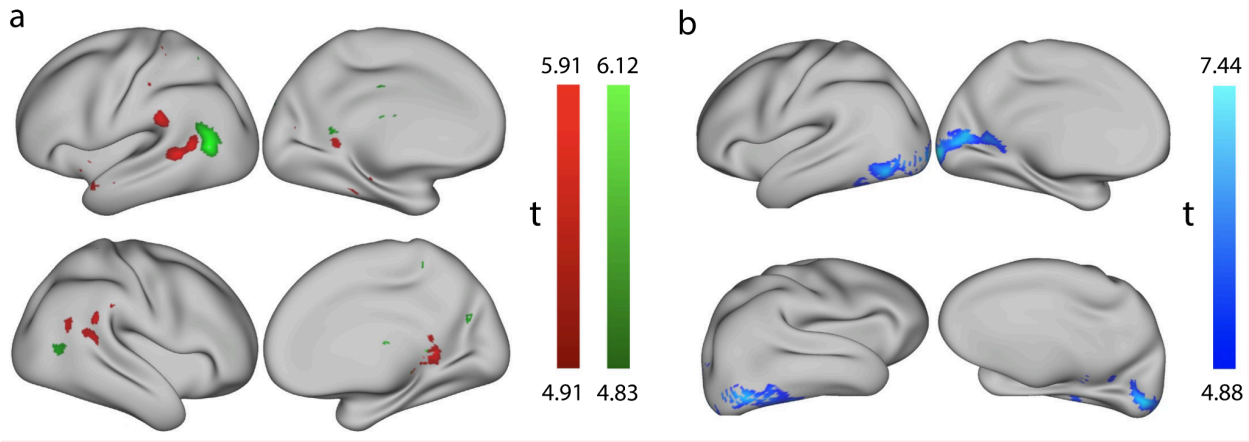

Fig. 2. a. Voxels showing significant effects ($p < 0.05$, FWE corrected) for the combination of auditory responses with responses in V3d and V5 (red), and auditory responses with responses in V3v and V4 (green). **b.** Voxels showing significant effects for the combination of responses in V3v and V4 with responses in V3d and V5 (blue).

Combination of visual and auditory information outside the STS

Our results also suggest the involvement of brain regions outside of the STS in combining audio-visual information. The combined-minus-max analysis for the combination of auditory and the early dorsal visual stream responses also identified brain regions in the anterior temporal lobe (ATL; peak MNI coordinates = [-54, -6, -15]), the primary somatosensory cortex (S1; peak MNI

coordinates = [3, -42, 61]), the supramarginal gyrus (peak MNI coordinates = [-54, -6, -15]), and the retrosplenial cortex (peak MNI coordinates = [30, -54, 4]).

Number of Voxels	t-values	MNI Coordinates	Label
126	6.64	(-66,-42,4)	STS
40	6.36	(45,-57, 18)	STS
66	5.9	(51,-45,14)	STS
8	5.77	(57,-9, -9)	STS
24	6.6	(-54, -6, -15)	ATL
24	6.44	(3,-42, 61)	S1
18	5.81	(24,-36, 70)	S1
11	5.76	(-3,-42, 57)	S1
44	6.35	(-54, -42, 31)	Supramarginal Gyrus
153	6.28	(30,-54, 4)	Retrosplenial Cortex
23	6.22	(15, -27, 41)	Posterior Cingulate
11	6.22	(18,-15,24)	Caudate Nucleus
26	6.18	(-15, -33, 41)	Middle Cingulate
6	6.04	(-51, -57, -32)	Cerebellum
8	5.83	(0,-24, 54)	M1

Table 1: Regions combining responses between auditory regions and V3d and V5 showing significant t-values ($p<0.01$, FWE-corrected) computed from the combined-max analysis.

The combined-minus-max analysis of auditory and early ventral visual stream responses revealed brain regions in the intraparietal sulcus (IPS; peak MNI coordinates = [-39, -51, 57]), retrosplenial cortex (peak MNI coordinates = [6, -42, 4]), caudate nucleus (peak MNI coordinates = [15, -9, 24]), and the lingual gyrus (peak MNI coordinates = [-27, -57, 4]).

Number of Voxels	t-values	MNI Coordinates	Label
36	6.44	(-39, -51, 57)	IPS
116	6.31	(-48, -72, 11)	STS, Occipitotemporal
29	6.07	(6, -42, 4)	Retrosplenial Cortex
80	5.88	(48, -57, 8)	STS
10	5.88	(15, -9, 24)	Caudate Nucleus
37	5.78	(-27, -57, 4)	Lingual Gyrus

Table 2: Regions combining responses between auditory regions and V3v and V4 showing significant t-values ($p<0.01$, FWE-corrected) computed from the combined-max analysis.

The combined-minus-max analysis for the posterior dorsal and posterior ventral visual stream response pairings identified a distinct set of brain regions compared to the previous two analyses. The largest cluster size was located in V1 (peak MNI coordinates = [21, -102, 1]) (Fig. 2b). Other brain regions included the bilateral parahippocampal place area (PPA; peak MNI coordinates = [-30, -48, -9], [30, -51, -9]) and the cerebellum (peak MNI coordinates = [-24, -78, -25]).

Number of Voxels	t-values	MNI Coordinates	Label
1365	7.63	(21,-102,1)	V1
48	6.01	(-30, -48, -9)	PPA
16	5.7	(30,-51,-9)	PPA
12	5.93	(-24, -78, -25)	Cerebellum

Table 3: Regions combining responses between V3v and V4, and V3d and V5, showing significant t-values ($p<0.01$, FWE-corrected) computed from the combined-max analysis.

Combination of information from auditory regions and different visual streams within face-selective ROIs

Considering the importance of combining facial information with auditory information for the recognition of speech and emotions (Piwek et al., 2015; Gentilucci and Cattaneo, 2005), we studied the combination of auditory and visual representations from different streams within functionally localized face-selective regions (Fig. 3a). In the face-selective STS, the effect of combining auditory and dorsal responses was significantly greater than that of combining auditory and ventral responses ($t(13)=3.82$, $p<0.05$) and than that of combining ventral and dorsal responses ($t(13)=4.55$, $p<0.01$, Fig. 3b, top panel). This finding could be due to the type of visual information encoded in V3d and V5: previous work has shown that these regions respond to motion (Felleman and Van Essen, 1987; Born and Bradley, 2005). Combining information about visual motion with auditory information might support audio-visual integration during speech perception and emotion recognition.

Unlike the face-selective STS, the fusiform face area (FFA) did not show significant differences between the pairwise combinations (Fig. 3b, middle panel). In the occipital face area (OFA), the effect of combining information from the two visual streams was significantly stronger than combining auditory and dorsal visual responses ($t(13)=5.11$, $p<0.01$) and than combining auditory and ventral visual responses ($t(13)=6.73$, $p<0.001$) (Fig. 3b, bottom panel).

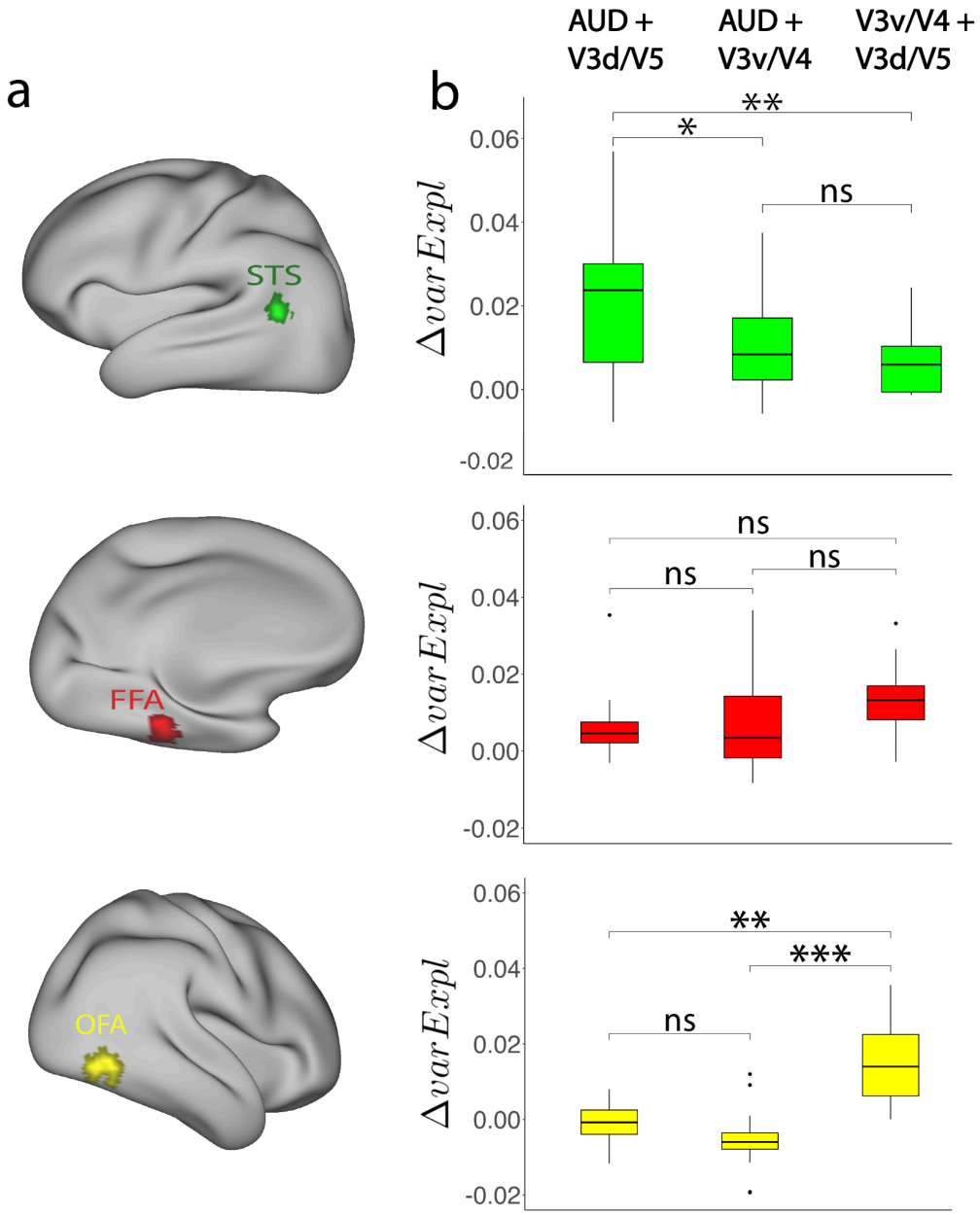

Fig. 3. a. Face-selective ROIs: STS, FFA, and OFA. **b.** Box plots depicting the difference in variance explained between the "combined" and "max" analyses across subjects in different face-selective ROI voxels. * signifies $p < 0.05$, ** signifies $p < 0.01$, and *** signifies $p < 0.001$. Significantly higher combined-minus-max effects were observed in the face-selective STS for the combination of the auditory and posterior dorsal stream than for the other pairings. No significant differences were observed in the FFA across the different pairings. Significantly higher combined-minus-max effects were observed in the OFA for the combination of the posterior dorsal and posterior ventral streams than for the other pairings.

Discussion

Audio-visual integration is a fundamental process that allows for the unified perception of everyday experiences. Given that distinct visual streams encode different kinds of representations, this study sought to uncover what visual representations are combined with auditory information when engaging in audio-visual integration, and what brain regions support the combination of responses from auditory regions and the different visual streams. The results demonstrate that both ventral and dorsal visual information is combined with auditory information, but that distinct portions of posterior STS combine auditory information with visual information encoded in the two streams.

What might drive this organization? Meta analyses suggest that different portions of posterior STS play different functional roles, including audio-visual integration, biological motion perception, theory of mind, and face processing (Hein and Knight, 2008). Meta-analyses, however, make it difficult to assess the degree of overlap between areas engaged in different functions: since different functions are probed in different participants, variability in response locations due to different functions is confounded with variability arising from individual differences. More recently, the investigation of multiple stimulus types within the same participants led to a more precise characterization of the distinct portions of the STS responsible for processing language, theory of mind, faces, voices, and biological motion (Deen et al., 2015). Relevant to the present results, Deen et al (2015) analyzed posterior-to-anterior changes in functional specialization in posterior STS, observing greater responses for Theory of Mind tasks in more posterior portions, followed by biological motion, and ultimately by greater responses to faces and voices in anterior portions. The posterior-to-anterior organization observed in the present study, therefore, could indicate that different visual inputs are combined with auditory representations to serve the needs of distinct functional subsystems that occupy adjacent areas within STS. In order to study the relationship between the topography of the effects we identified in the present work and other functional subdivisions of STS, it will be necessary to perform both sets of analyses within the same group of participants.

Previous research on ventral stream representations suggests a possible functional role for the more posterior of the two STS hubs identified in this study. Effects for the combination of auditory information and the ventral visual stream were observed in a more posterior portion of the STS, and previous research has implicated the ventral visual stream in the recognition of the identity of objects (Ungerleider and Mishkin, 1982). Posterior portions of the STS that combine information from ventral visual regions and auditory regions might contribute to encoding the typical sounds produced by different kinds of objects, associating dogs with barking, cars with vrooming, and so on. Additional research will be needed to test this hypothesis. As an alternative hypothesis, the organization of the combination of auditory and visual information into two distinct portions of posterior STS might not be due to their engagement in supporting different

functions, but to unique computational requirements of integrating auditory representations with different kinds of visual representations.

Focusing on face-selective regions of interest, we found that the combination of audio-visual information in the face-selective STS relies disproportionately on visual information encoded in dorsal visual regions. This is consistent with the observation that effects for the combination of auditory information with visual information from dorsal regions were located in more anterior portions of posterior STS in our whole-brain analyses, and with the previous studies indicating that face responses also peak in more anterior portions of posterior STS (Deen et al. 2015). The latter finding could be due to the type of visual information encoded in V3d and V5: previous work has shown that these regions contain neurons that respond to motion (Felleman and Van Essen, 1987; Born and Bradley, 2005). Combining information about visual motion with auditory information might support audio-visual integration during speech perception. It will be interesting to test whether the effects for the combination of auditory information and dorsal visual representations reported here are localized to the same voxels showing an association with individual differences in susceptibility to the McGurk effect reported in previous work (Nath and Beauchamp, 2012).

Finally, the combination of visual information from the two visual streams was observed in ventral occipitotemporal cortex, and ROI analyses showed that the extent of these effects includes the OFA. Classical work has proposed the importance of motion to identify and segment objects (Spelke, 1990), leading to recent computational models of motion-based segmentation (Chen et al., 2022). We hypothesize that the combination of information from the two visual streams within occipitotemporal cortex could support motion-based segmentation. Considering the anatomical location of the effects that are co-localized with the earliest stages of category-selectivity (e.g. OFA), we hypothesize that motion-based segmentation might provide the basis for category-selectivity.

Our findings also implicate brain regions beyond the STS. Regarding the candidate MSD sites that were statistically dependent on information from the auditory and posterior ventral streams, the intraparietal sulcus (IPS) was the region with the highest t-value. This region has been implicated in audio-visual integration in prior work (Lewis et al., 2000, Calvert et al., 2001).

Methodologically it is worth noting that the results obtained from the MVPN combined-minus-max analyses only establish correlational relationships. To establish causality between the joint responses from the auditory and different visual streams in MSD sites, future research could employ techniques that infer causality, such as transcranial magnetic stimulation-fMRI (TMS-fMRI). Further, our method shows that two regions jointly contribute to predict responses in a third region (i.e., statistical dependence), but we can not determine precisely whether and how this information is integrated into a multi-modal representation.

Despite these limitations, the results reveal a novel aspect of the large-scale topography of STS, and provide insights into the neural architecture that supports our unified perception of the world.

References

Anzellotti, S., & Caramazza, A. (2017). Multimodal representations of person identity individuated with fMRI. Cortex, 89, 85-97.

Anzellotti, S., Caramazza, A., & Saxe, R. (2017). Multivariate pattern dependence. PLoS computational biology, 13(11), e1005799.

Avants, B. B., Tustison, N., & Song, G. (2009). Advanced normalization tools (ANTS). Insight j, 2(365), 1-35.

Behzadi, Y., Restom, K., Liau, J., & Liu, T. T. (2007). A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. Neuroimage, 37(1), 90-101.

Born, R. T., & Bradley, D. C. (2005). Structure and function of visual area MT. Annu. Rev. Neurosci., 28, 157-189.

Calvert, G. A., Campbell, R., & Brammer, M. J. (2000). Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. Current biology, 10(11), 649-657.

Calvert, G. A., Hansen, P. C., Iversen, S. D., & Brammer, M. J. (2001). Detection of audio-visual integration sites in humans by application of electrophysiological criteria to the BOLD effect. Neuroimage, 14(2), 427-438.

Chen, Y., Mancini, M., Zhu, X., & Akata, Z. (2022). Semi-supervised and unsupervised deep visual learning: A survey. IEEE transactions on pattern analysis and machine intelligence.

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

Deen, B., Koldewyn, K., Kanwisher, N., & Saxe, R. (2015). Functional organization of social perception and cognition in the superior temporal sulcus. Cerebral cortex, 25(11), 4596-4609.

Dobs, K., Bülthoff, I., Breidt, M., Vuong, Q. C., Curio, C., & Schultz, J. (2014). Quantifying human sensitivity to spatio-temporal information in dynamic faces. Vision Research, 100, 78-87.

Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., ... & Gorgolewski, K. J. (2019). fMRIPrep: a robust preprocessing pipeline for functional MRI. Nature methods, 16(1), 111-116.

Fagel, S. (2006, May). Emotional mcgurk effect. In Proceedings of the international conference on speech prosody (Vol. 1).

Fang, M., Poskanzer, C., & Anzellotti, S. (2022). Pymvpd: a toolbox for multivariate pattern dependence. Front Neuroinform 16: 835772.

Fang, M., Aglinskas, A., Li, Y., & Anzellotti, S. (2023). Angular gyrus responses show joint statistical dependence with brain regions selective for different categories. Journal of Neuroscience, 43(15), 2756-2766.

Felleman, D. J., & Van Essen, D. C. (1987). Receptive field properties of neurons in area V3 of macaque monkey extrastriate cortex. Journal of neurophysiology, 57(4), 889-920.

Gentilucci, M., & Cattaneo, L. (2005). Automatic audiovisual integration in speech perception. Experimental Brain Research, 167, 66-75.

Greve, D. N., & Fischl, B. (2009). Accurate and robust brain image alignment using boundary-based registration. Neuroimage, 48(1), 63-72.

Hanke, M., Adelhöfer, N., Kottke, D., Iacovella, V., Sengupta, A., Kaule, F. R., ... & Stadler, J. (2016). A studyforrest extension, simultaneous fMRI and eye gaze recordings during prolonged natural stimulation. Scientific data, 3(1), 1-15.

Hein, G., & Knight, R. T. (2008). Superior temporal sulcus—it's my area: or is it?. Journal of cognitive neuroscience, 20(12), 2125-2136.

Jenkinson, M., Bannister, P., Brady, M., & Smith, S. (2002). Improved optimization for the robust and accurate linear registration and motion correction of brain images. Neuroimage, 17(2), 825-841.

Kanwisher, N., McDermott, J., & Chun, M. M. (2002). The fusiform face area: a module in human extrastriate cortex specialized for face perception.

Lewis, J. W., Beauchamp, M. S., & DeYoe, E. A. (2000). A comparison of visual and auditory motion processing in human cerebral cortex. Cerebral Cortex, 10(9), 873-888.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. Nature, 264(5588), 746-748.

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

Nichols, T. E., & Holmes, A. P. (2002). Nonparametric permutation tests for functional neuroimaging: a primer with examples. Human brain mapping, 15(1), 1-25.

Peelen, M. V., Atkinson, A. P., & Vuilleumier, P. (2010). Supramodal representations of perceived emotions in the human brain. Journal of Neuroscience, 30(30), 10127-10134.

Piwek, L., Pollick, F., & Petrini, K. (2015). Audiovisual integration of emotional signals from others' social interactions. Frontiers in psychology, 6, 137846.

Rademacher, J., Morosan, P., Schormann, T., Schleicher, A., Werner, C., Freund, H. J., & Zilles, K. (2001). Probabilistic mapping and volume measurement of human primary auditory cortex. Neuroimage, 13(4), 669-683.

Sengupta, A., Kaule, F. R., Guntupalli, J. S., Hoffmann, M. B., Häusler, C., Stadler, J., & Hanke, M. (2016). A studyforrest extension, retinotopic mapping and localization of higher visual areas. Scientific data, 3(1), 1-14.

Schein, S. J., & Desimone, R. (1990). Spectral properties of V4 neurons in the macaque. Journal of Neuroscience, 10(10), 3369-3389.

Spelke, E. S. (1990). Principles of object perception. Cognitive science, 14(1), 29-56.

Ungerleider, L. G. (1982). Two cortical visual systems. Analysis of visual behavior, 549, chapter-18.

Wang, L., Mruczek, R. E., Arcaro, M. J., & Kastner, S. (2015). Probabilistic maps of visual topography in human cortex. Cerebral cortex, 25(10), 3911-3931.

Woolrich, M. W., Ripley, B. D., Brady, M., & Smith, S. M. (2001). Temporal autocorrelation in univariate linear modeling of FMRI data. Neuroimage, 14(6), 1370-1386.

Yovel, G. (2016). Neural and cognitive face-selective markers: An integrative review. Neuropsychologia, 83, 5-13.

Zwiers, M. P., Van Opstal, A. J., & Paige, G. D. (2003). Plasticity in human sound localization induced by compressed spatial vision. Nature neuroscience, 6(2), 175-181.