

Research Articles | Behavioral/Cognitive

## Distinct portions of superior temporal sulcus combine auditory representations with different visual streams

<https://doi.org/10.1523/JNEUROSCI.1188-24.2025>

Received: 17 June 2024

Revised: 29 July 2025

Accepted: 16 September 2025

Copyright © 2025 the authors

---

*This Early Release article has been peer reviewed and accepted, but has not been through the composition and copyediting processes. The final version may differ slightly in style or formatting and will contain links to any extended data.*

**Alerts:** Sign up at [www.jneurosci.org/alerts](http://www.jneurosci.org/alerts) to receive customized email alerts when the fully formatted version of this article is published.

**Distinct portions of superior temporal sulcus combine auditory representations with different visual streams**

Abbreviated title: Audio-visual combination in the STS

Gabriel Fajardo<sup>1,2\*</sup>, Mengting Fang<sup>3\*</sup>, and Stefano Anzellotti<sup>1</sup>

<sup>1</sup> Boston College, Department of Psychology and Neuroscience, Chestnut Hill, MA, 02467

<sup>2</sup> Columbia University, Department of Psychology, New York, NY, 10027

<sup>3</sup> 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](mailto:stefano.anzellotti@bc.edu)

Number of pages: 20

Number of figures: 3

Number of tables: 3

Number of words Abstract: 98

Number of words Introduction: 621

## Number of words Discussion: 1251

30

31 Conflicts of interest:

32 The authors state no conflicts of interest.

33

34 Acknowledgements:

35

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

39

40

41 **Abstract**

42 In humans, the superior temporal sulcus (STS) combines auditory and visual  
43 information. However, the extent to which it relies on visual information from the ventral  
44 or dorsal stream remains uncertain. To address this, we analyzed open-source  
45 functional magnetic resonance imaging data collected from 15 participants (6 females  
46 and 9 males) as they watched a movie. We used artificial neural networks to investigate  
47 the relationship between multivariate response patterns in auditory cortex, the two  
48 visual streams, and the rest of the brain, finding that distinct portions of the STS  
49 combine information from the two visual streams with auditory information.

50

51 **Significance Statement**

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

62

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

## 65 Introduction

66 The human brain is adept at integrating visual and auditory information in order to  
67 create a coherent perception of the external world. Audio-visual integration contributes  
68 to sound localization (Zwiers et al., 2003), and plays a key role for emotion recognition  
69 (Piwek et al., 2015) as well as speech perception (Gentilucci and Cattaneo, 2005).  
70 Several phenomena demonstrate that the integration of visual and auditory cues shapes  
71 perceptual experience. In the McGurk effect, simultaneous presentation of a phoneme  
72 with a mismatched face video results in a distorted perception of the phoneme (McGurk  
73 and MacDonald, 1976). Similarly, presentation of mismatched auditory and visual  
74 stimuli can alter emotion recognition (Fagel, 2006), even when participants are explicitly  
75 instructed to focus only on one stimulus modality and ignore the other (Collignon et al.,  
76 2008), suggesting that audio-visual integration is automatic.

77

78 Audio-visual integration requires combining auditory information represented in the  
79 superior temporal gyrus with visual information encoded in occipitotemporal areas.  
80 Therefore, identifying brain regions that combine auditory and visual information is key  
81 for understanding the neural bases of audio-visual integration. Previous work found that  
82 the presentation of congruent audio-visual stimuli leads to supra-additive responses in  
83 the superior temporal sulcus (STS) compared to unimodal visual and auditory stimuli,  
84 whereas incongruent audio-visual stimuli leads to sub-additive responses (Calvert et al.,  
85 2000). In addition, participants' susceptibility to the McGurk effect correlates with the  
86 strength of STS responses (Nath and Beauchamp, 2012). Furthermore, response

87 patterns in the STS encode information about emotions and identity that generalizes  
88 across visual and auditory modalities (Peelen et al., 2010; Anzellotti and Caramazza,  
89 2017). These studies indicate that the STS plays a pivotal role in combining auditory  
90 and visual information.

91

92 However, little is known about the precise visual representations that are involved.

93 Visual

94 information is processed by multiple streams: a ventral and a dorsal stream  
95 (Ungerleider, 1982). The ventral stream originates in ventral area V3 (V3v) and area V4,  
96 and the dorsal stream in dorsal area V3 (V3d) and area V5 (Felleman and Van Essen,  
97 1987) (Fig. 1a). Area V5 is associated with motion perception, featuring a large number  
98 of direction-selective neurons (Born and Bradley, 2005). By contrast, many neurons in  
99 V4 show sensitivity to color (Schein and Desimone, 1990). Correspondingly, a large  
100 number of neurons in the dorsal part of V3 respond to motion, and a large number of  
101 neurons in the ventral portion of V3 are tuned for color processing (Felleman and Van  
102 Essen, 1987). The existence of these different visual streams prompts questions about  
103 their relative contributions to the combination of visual and auditory information.

104

105 Auditory information could be combined with visual information from both streams, or  
106 with visual information from only one of the streams. If it is combined with visual  
107 information from both streams, auditory information could be combined with information  
108 from both visual streams in a single hub, or distinct regions could combine auditory

109 information with each visual stream separately. To investigate this, we used artificial  
110 neural networks to model the relationship between patterns of response in auditory  
111 brain regions, in the initial segments of the ventral and dorsal visual streams, and in the  
112 rest of the brain (Fig. 1b), following a strategy that has been recently adopted to  
113 investigate the combination of information from multiple category-selective regions  
114 (Fang et al., 2023). Functional magnetic resonance imaging (fMRI) data collected while  
115 participants viewed rich audio-visual stimuli (Hanke et al., 2016) were analyzed with  
116 multivariate pattern dependence networks (MVPN) (Anzellotti et al., 2017; Fang et al.,  
117 2022). Searching for brain regions where responses are better predicted using a  
118 combination of auditory responses and responses in different visual streams than using  
119 auditory or visual responses in isolation revealed two distinct portions of STS that  
120 combine information between auditory regions and the two visual streams.

121

## 122 **Materials and Methods**

### 123 **Experimental Design and Statistical Analyses**

#### 124 ***Experimental paradigm***

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

131

132 ***Data acquisition parameters***

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

143

144 ***Preprocessing***

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

153 components extracted from the union of cerebrospinal fluid and white matter (Behzadi  
154 et al., 2007). The raw data of one subject could not be preprocessed with the fMRIPrep  
155 pipeline. The remaining 14 subjects' data were used for the rest of the study.

156

157 ***ROI definition***

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

166

167 Since the anatomical location of auditory brain regions is more variable across subjects  
168 than visual brain regions (Rademacher et al., 2001), auditory ROIs were defined  
169 individually for each subject by identifying voxels where responses are parametrically  
170 modulated by the loudness of auditory stimuli. To this end, standard univariate GLM  
171 analyses were conducted using FSL FEAT (<https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/FEAT>)  
172 (Woolrich et al., 2001), with root mean square (RMS) levels as the predictor. The 80  
173 voxels with the highest t-scores were selected individually for each subject (example of  
174 a subject's auditory ROI mask in Fig. 1a, bottom panel). To ensure that the remaining

175 analyses are independent from the ROI selection, we used only data from the first fMRI  
176 run for auditory ROI selection, and this run was not used in the remaining analyses  
177 (which were therefore conducted on the remaining seven runs). There were no  
178 overlapping voxels between the ROIs.

179

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

184

185

186

187 **MVPN: Multivariate Pattern Dependence Network**

188 Recent research has taken advantage of the flexibility and computational power of  
189 artificial neural networks (ANNs) in order to analyze brain connectivity (Fang et al.,  
190 2022; Fang et al., 2023). The multivariate pattern dependence network (MVPN) method  
191 – an extension of MVPD (Anzellotti et al., 2017) – utilizes the power of ANNs to analyze  
192 the multivariate relationships between neural response patterns. It is important to note  
193 that MVPN measures the statistical relationship between response patterns in different  
194 regions, but it can not detect the direction of information flow. We implemented MVPN in  
195 PyTorch, and the neural networks were trained on Tesla V100 graphics processing

196 units (GPUs). In this study, we used 5-layer dense neural networks with 100 nodes per  
197 hidden layer. This architecture was selected based on prior work (Fang et al., 2022),  
198 which systematically compared different network architectures and found the 5-layer  
199 dense network to yield the highest overall predictive accuracy when using two different  
200 seed regions (FFA and PPA) to predict responses across the rest of the brain. The  
201 DNNs were optimized using stochastic gradient descent (SGD) with a mean squared  
202 error (MSE) loss function, a learning rate of 0.001 and a momentum of 0.9. The models  
203 were trained for 5000 epochs. We used a batch size of 32, and batch normalization was  
204 applied to each layer's inputs. The ANNs were given as input the multivariate response  
205 patterns in one or more sets of brain regions (Fig. 1): auditory regions, ventral visual  
206 regions (V3v and V4), dorsal visual regions (V3d and V5), and all pairwise  
207 combinations. ANNs were trained to predict the patterns of responses in all gray matter  
208 voxels.

209  
210 More precisely, the MVPN method works as follows. Consider an fMRI experiment with  
211  $m$  experimental runs. We label the multivariate time courses in a predictor region as  
212  $X_1, \dots, X_m$ . Each matrix  $X_i$  is of size  $n_X \times T_i$ , where  $n_X$  is the total number of voxels in the  
213 predictor region, and  $T_i$  is the number of timepoints in the  $i^{th}$  experimental run.  
214 Similarly, let  $Y_1, \dots, Y_m$  be the multivariate timecourses in the target region, where  $Y_i$  is an  
215  $n_Y \times T_i$  matrix,  $n_Y$  is the total number of voxels in the target region, and  $T_i$  is the number  
216 of timepoints in the  $i^{th}$  experimental run.

217

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

220 
$$Y_{train} = f(X_{train}) + E_{train},$$

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

224 
$$D_{\setminus i} = \{(X_1, Y_1), \dots, (X_{i-1}, Y_{i-1}), (X_{i+1}, Y_{i+1}), \dots, (X_m, Y_m)\},$$

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

226  
 227 We used the proportion of variance explained between the predictor region and all other  
 228 voxels in the gray matter mask in order to measure multivariate statistical dependence.  
 229 For each target region voxel  $j$ , the variance explained  $varExpl_i(j)$  was calculated as

230 
$$varExpl_i(j) = \max \left\{ 0, 1 - \frac{\text{var}(Y_i(j) - f_j(X_i))}{\text{var}(Y_i(j))} \right\},$$

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

234 **Combined-minus-max whole-brain analysis**

235 In order to identify brain regions that depend on the combination of auditory and visual  
 236 response patterns, we analyzed the *StudyForrest* dataset with a novel approach we

237 introduced in a recent study (Fang et al., 2023): the “combined-minus-max” approach,  
238 described in the following paragraphs. Since run 1 was used to functionally localize  
239 auditory regions (see the “ROI definition” section), to prevent circularity in the analysis,  
240 we used experimental runs 2 through 8 for the combined-minus-max analysis (a total of  
241 7 runs).

242

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

250

251 If a voxel only encodes information from one of the streams, using the responses from  
252 multiple streams as predictors should not improve the variance explained. On the  
253 contrary, if the responses in the voxel are better predicted by a neural network including  
254 multiple streams combined than by a single stream, we can conclude that the voxel  
255 combines information from multiple streams. Therefore, we searched for voxels that  
256 combine information from multiple streams by computing an index given by the  
257 difference between the proportion of variance explained by a model using two streams  
258 jointly (the “combined” model), and the proportion of variance explained by a model

259 using the best predicting stream among the two (the “max” model). This procedure is  
260 illustrated in Fig. 1c.

261

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

266 
$$\Delta\text{varExpl}(j) = \text{varExpl}_{\text{pair}}(j) - \text{varExpl}_{\text{max}}(j)$$

267 This  $\Delta\text{varExpl}(j)$  gives us a multi-stream dependence (MSD) index for each voxel, that  
268 allowed us to identify candidate brain regions that jointly combine information from  
269 different streams. We calculated the statistical significance of  $\Delta\text{varExpl}$  values across  
270 subjects using statistical non-parametric mapping, utilizing the SnPM extension for SPM  
271 (<http://nisox.org/Software/SnPM13/>) (Nichols and Holmes, 2002).

272

### 273 **Control analysis**

274 When using the combined-minus-max approach, there is still the possibility that the  
275 better predictive accuracy of the combined model might be due to the larger number of  
276 voxels in the combined analysis. To control for this possibility, we conducted a control  
277 analysis using voxels from the primary motor cortex (M1) as predictors (see Fang et al.,  
278 2023 as an example of an analogous approach). In this analysis, we randomly selected  
279 three non-overlapping groups of 80 voxels in M1 (this number was chosen to match the

280 number of voxels selected from the three streams: the posterior ventral, posterior  
281 dorsal, and auditory). We then used the responses from the three groups of M1 voxels  
282 to run a control analysis following the same procedure as the combined-minus-max  
283 analysis, and we computed the statistical significance of  $\Delta varExpl$  for each voxel in  
284 gray matter across subjects. Any regions showing statistical significance in this control  
285 analysis ( $p < 0.05$ , FWE-corrected with SnPM) were due to the larger number of voxels in  
286 the combined model, not multi-stream information combination. Therefore, they were  
287 excluded from the multi-stream dependence (MSD) analysis described above.

288

### 289 **Face-selective ROI analysis**

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

296

297 We used the first run in the category localizer to identify three face-selective ROIs: the  
298 occipital face area (OFA), the fusiform face area (FFA), and the face-selective posterior  
299 superior temporal sulcus (STS). Data were modeled with a standard GLM using FSL  
300 FEAT (Woolrich et al., 2001). Each seed ROI was defined as a sphere with a 9mm  
301 radius centered in the peak for the contrast faces > bodies, artifacts, scenes, scrambled

302 images. Data from both the left and the right hemisphere were combined for each ROI,  
303 and the 80 voxels that showed the highest z-value for the contrast were selected.  
304 Visualizations of these ROIs can be found in Fig. 3a. We then analyzed the variance  
305 explained measures for each voxel in these face-selective ROIs across our three  
306 pairings (posterior dorsal stream and auditory stream, posterior ventral stream and  
307 auditory stream, and posterior dorsal stream and posterior ventral stream).

308

### 309 **Code/Software Accessibility**

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

313

## 314 **Results**

### 315 **STS combines information from auditory regions with information from different 316 visual streams.**

317 To identify brain regions that jointly encoded information from different streams, we  
318 calculated the multi-stream dependence (MSD) index for each voxel. This index was  
319 computed as the difference between the proportion of variance explained by the  
320 combined model and that of the max model (see Materials and Methods section for a  
321 detailed explanation of the “combined-minus-max” approach). Group-level analyses  
322 were used to identify voxels with MSD indices significantly greater than zero. These

323 voxels were considered as candidate multi-stream dependence brain regions. Clusters  
324 with peaks having  $p < 0.05$  (FWE corrected) were included.

325

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

333

334 The control analysis showed significant effects in the sensorimotor cortex (peak MNI  
335 coordinates = [0, -21, 64], [33, -42, 67], [-39, -18, 41]), premotor cortex (peak MNI  
336 coordinates = [-57, -9, 44], [57, 12, 31]), the bilateral intraparietal sulcus (peak MNI  
337 coordinates = [30, -69, 54], [-24, -72, 50]), and the angular gyrus (peak MNI coordinates  
338 = [-45, -69, 37]). Importantly, the control analyses did not show significant effects in  
339 ventral and lateral occipitotemporal regions. Therefore, significant findings in these  
340 regions in the main analysis could not be explained just by a difference between the  
341 number of predictor voxels in the combined analysis and the max analysis. Voxels that  
342 yielded significant effects in the control analysis ( $p < 0.05$ , FWE-corrected) were  
343 excluded before calculating the MSD indices in the main analysis.

344

345 Combining response patterns from auditory regions and the early dorsal stream  
346 revealed significant effects in the bilateral STS (peak MNI coordinates = [-66, -42, 4],  
347 [45, -57, 18]) and within the posterior cingulate cortex (peak MNI coordinates = [15, -27,  
348 41]) (Table 1;  $p < 0.05$ , FWE corrected). Combining responses from auditory regions and  
349 the early ventral stream also revealed effects in the right STS (Table 2;  $p < 0.05$ , FWE  
350 corrected), but in a more posterior portion (peak MNI coordinates = [48, -57, 8]), at the  
351 boundary with the occipital lobe (Fig. 2a).

352

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

357

358

359

### 360 **Robustness of the results across different data splits**

361 In order to further evaluate the robustness of the results, we defined a broad bilateral  
362 STS region of interest via the “Superior Temporal Gyrus” map from WFU Pick Atlas. We  
363 then extracted the patterns of the combined-minus-max effects across voxels as  
364 vectors. For each split of the data, this procedure yielded a vector for the  
365 auditory+dynamic combined-minus-max results and another vector for the

366 auditory+static combined-minus-max results. The robustness of the patterns across the  
367 two splits of the data was assessed by computing the Pearson correlation between the  
368 vectors for the two halves. The correlation for vectors from the same analysis (for  
369 example, between the first and second halves of the auditory+dynamic analysis) was  
370 compared to the correlation for vectors from different analyses (for example, between  
371 the first half of the auditory+dynamic analysis and the second half of the auditory+static  
372 analysis), following a procedure inspired by prior work (Haxby et al. 2001). If the results  
373 are robust across different splits of the data, we expected to observe higher correlations  
374 between the patterns for the same analysis across the splits compared to the patterns  
375 for two different analyses. The results were in line with the prediction: correlations  
376 between the vectors for the same analysis were higher than correlations for vectors for  
377 different analyses across splits (Figure 2d).

378

### 379 **Quantifying distinct spatial distributions of auditory+ventral and auditory+dorsal** 380 **effects**

381 Using the STS ROI introduced in the previous section, for each subject individually, we  
382 retrieved the 50 voxels with the highest  $\Delta varExpl$  across both models (auditory+dorsal  
383 combined-minus-max and auditory+ventral combined-minus-max). We then computed  
384 the Pearson correlation between the  $\Delta varExpl$  values of both models across these  
385 voxels, using a strategy inspired by previous work (Peelen et al 2006). Since  
386 correlations range from -1 to 1 they violate the normality assumption, correlation values  
387 were Fisher transformed and submitted to a two-tailed t-test across subjects to probe for  
388 spatial correlations between the two effects (auditory+dorsal and auditory+ventral),

389 Figure 2c). This revealed a significantly negative correlation ( $t(13)=-2.16$ ,  $p<0.05$ ),  
390 suggesting that combination effects for auditory information with different visual streams  
391 involve spatially distinct neural substrates. To expand this investigation to other regions,  
392 we ran this analysis again at the whole brain level, this time using the 100 voxels with  
393 the highest  $\Delta varExpl$  across both models. The Fisher-z transformed t-test also revealed  
394 a significantly negative interaction ( $t(13) = -4.54$ ,  $p<0.001$ ). These findings indicate that  
395 dorsal and ventral areas do indeed contribute to spatially distinct effects of combination  
396 with auditory cortex.

397

398 **Ventral temporal cortex combines information from different visual streams.**

399 These results raise the question of whether and where information from early dorsal  
400 (V3d and V5) and ventral (V3v and V4) visual regions is combined. We adopted the  
401 same strategy to test this, searching for voxels that are better predicted by both visual  
402 streams jointly than by either stream in isolation. This analysis identified regions in the  
403 calcarine sulcus (V1 and V2) that are located upstream of V3, V4, and V5, and in  
404 regions in ventral occipitotemporal cortex, that are located downstream (peak MNI  
405 coordinates = [21, -102, 1]) (Table 3;  $p<0.05$ , FWE corrected, Fig. 2b). Notably, no  
406 effects for the combination of the two visual streams were observed in the STS. This is  
407 consistent with the finding that the combination of auditory information with different  
408 visual streams involves distinct cortical regions: if it happened in a single STS  
409 subregion, we would also expect to observe effects in that subregion for combining both  
410 visual streams.

411 **Combination of visual and auditory information outside the STS**

412 Our results also suggest the involvement of brain regions outside of the STS in  
413 combining audio-visual information. The combined-minus-max analysis for the  
414 combination of auditory and the early dorsal visual stream responses also identified  
415 brain regions in the anterior temporal lobe (ATL; peak MNI coordinates = [-54, -6, -15]),  
416 the primary somatosensory cortex (S1; peak MNI coordinates = [3, -42, 61]), the  
417 supramarginal gyrus (peak MNI coordinates = [-54, -6, -15]), and the retrosplenial cortex  
418 (peak MNI coordinates = [30, -54, 4]).

419

420 The combined-minus-max analysis of auditory and early ventral visual stream  
421 responses revealed brain regions in the intraparietal sulcus (IPS; peak MNI coordinates  
422 = [-39, -51, 57]), retrosplenial cortex (peak MNI coordinates = [6, -42, 4]), caudate  
423 nucleus (peak MNI coordinates = [15, -9, 24]), and the lingual gyrus (peak MNI  
424 coordinates = [-27, -57, 4]).

425

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

432

433 Inspecting the combined-minus-max maps of individual participants in search of other  
434 regions that might show these effects, and that might not appear in the second-level  
435 analyses due to greater topographic variability across individuals, did not reveal other  
436 clear candidate regions. This does not rule out that additional regions might be identified  
437 in the future using more powerful data acquisition and analysis methods.

438

439 Overlap between the auditory+ventral and auditory+dorsal effects was observed in  
440 posterior cingulate and in pulvinar in some individual participants, but these effects were  
441 variable across participants - further work will be needed to establish whether these  
442 regions combine auditory information with both ventral and dorsal representations. To  
443 probe for three-way combination effects across auditory, ventral and dorsal regions we  
444 conducted a combined-minus-max analysis of the three regions combined minus the  
445 maximum across each of the 3 different pairs (auditory+dynamic combined,  
446 auditory+static combined, dynamic+static combined). Statistical non-parametric  
447 analysis did not reveal anything past the threshold (FWE corrected  $p < .05$ ); future work  
448 with more sensitive methods or greater statistical power might reveal some effects.

449

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

452 Considering the importance of combining facial information with auditory information for  
453 the recognition of speech and emotions (Piwek et al., 2015; Gentilucci and Cattaneo,  
454 2005), we studied the combination of auditory and visual representations from different

455 streams within functionally localized face-selective regions (Fig. 3a). In the face-  
456 selective STS, the effect of combining auditory and dorsal responses was significantly  
457 greater than that of combining auditory and ventral responses ( $t(13)=3.82$ ,  $p<0.05$ ) and  
458 than that of combining ventral and dorsal responses ( $t(13)=4.55$ ,  $p<0.01$ , Fig. 3b, top  
459 panel). This finding could be due to the type of visual information encoded in V3d and  
460 V5: previous work has shown that these regions respond to motion (Felleman and Van  
461 Essen, 1987; Born and Bradley, 2005). Combining information about visual motion with  
462 auditory information might support audio-visual integration during speech perception  
463 and emotion recognition.

464

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

471

## 472 Discussion

473 Audio-visual integration is a fundamental process that allows for the unified perception  
474 of everyday experiences. Given that distinct visual streams encode different kinds of  
475 representations, this study sought to uncover what visual representations are combined  
476 with auditory information when engaging in audio-visual integration, and what brain

477 regions support the combination of responses from auditory regions and the different  
478 visual streams. The results demonstrate that both ventral and dorsal visual information  
479 is combined with auditory information, but that distinct portions of posterior STS  
480 combine auditory information with visual information encoded in the two streams. The  
481 topography of combined-minus-max effects observed in the STS could be related to the  
482 types of features encoded in dorsal and ventral visual regions. Importantly, however,  
483 these results are only possible in the presence of audio-visual combination effects. If  
484 posterior STS encoded visual features that are well predicted by dorsal visual regions in  
485 isolation, and anterior STS encoded visual features that are well predicted by ventral  
486 visual regions in isolation, subtracting the max in the combined-minus-max analysis  
487 would remove these effects.

488

489 What are the specific factors that drive the observed topography of STS effects remains  
490 an open question. Meta analyses suggest that different portions of posterior STS play  
491 different functional roles, including audio-visual integration, biological motion perception,  
492 theory of mind, and face processing (Hein and Knight, 2008). Meta-analyses, however,  
493 make it difficult to assess the degree of overlap between areas engaged in different  
494 functions: since different functions are probed in different participants, variability in  
495 response locations due to different functions is confounded with variability arising from  
496 individual differences. More recently, the investigation of multiple stimulus types within  
497 the same participants led to a more precise characterization of the distinct portions of  
498 the STS responsible for processing language, theory of mind, faces, voices, and  
499 biological motion (Deen et al., 2015). Relevant to the present results, Deen et al (2015)

500 analyzed posterior-to-anterior changes in functional specialization in posterior STS,  
501 observing greater responses for Theory of Mind tasks in more posterior portions,  
502 followed by biological motion, and ultimately by greater responses to faces and voices  
503 in anterior portions. The posterior-to-anterior organization observed in the present  
504 study, therefore, could indicate that different visual inputs are combined with auditory  
505 representations to serve the needs of distinct functional subsystems that occupy  
506 adjacent areas within STS. In order to study the relationship between the topography of  
507 the effects we identified in the present work and other functional subdivisions of STS, it  
508 will be necessary to perform both sets of analyses within the same group of participants.

509

510 Previous research on ventral stream representations suggests a possible functional role  
511 for the more posterior of the two STS hubs identified in this study. Effects for the  
512 combination of auditory information and the ventral visual stream were observed in a  
513 more posterior portion of the STS, and previous research has implicated the ventral  
514 visual stream in the recognition of the identity of objects (Ungerleider, 1982). Posterior  
515 portions of the STS that combine information from ventral visual regions and auditory  
516 regions might contribute to encoding the typical sounds produced by different kinds of  
517 objects, associating dogs with barking, cars with vrooming, and so on. By contrast,  
518 more anterior portions might encode the way different movements are associated with  
519 sounds - even when the identity or category of an object is held constant. For example,  
520 in face perception, the relationship between lip movements and phonemes is known to  
521 involve audiovisual integration mechanisms that lead to phenomena such as the  
522 McGurk effect (McGurk and MacDonald 1976). In many other instances, sounds are

523 produced by the dynamic interactions between multiple objects. Experiments with  
524 tailored designs, that include distinct conditions that separate between these different  
525 kinds of audio-visual information, will be needed to test this hypothesis. As an  
526 alternative hypothesis, the organization of the combination of auditory and visual  
527 information into two distinct portions of posterior STS might not be due to their  
528 engagement in supporting different functions, but to unique computational requirements  
529 of integrating auditory representations with different kinds of visual representations.

530

531 Focusing on face-selective regions of interest, we found that the combination of audio-  
532 visual information in the face-selective STS relies disproportionately on visual  
533 information encoded in dorsal visual regions. This is consistent with the observation that  
534 effects for the combination of auditory information with visual information from dorsal  
535 regions were located in more anterior portions of posterior STS in our whole-brain  
536 analyses, and with the previous studies indicating that face responses also peak in  
537 more anterior portions of posterior STS (Deen et al. 2015). The latter finding could be  
538 due to the type of visual information encoded in V3d and V5: previous work has shown  
539 that these regions contain neurons that respond to motion (Felleman and Van Essen,  
540 1987; Born and Bradley, 2005). Combining information about visual motion with auditory  
541 information might support audio-visual integration during speech perception. It will be  
542 interesting to test whether the effects for the combination of auditory information and  
543 dorsal visual representations reported here are localized to the same voxels showing an  
544 association with individual differences in susceptibility to the McGurk effect reported in  
545 previous work (Nath and Beauchamp, 2012).

546

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

556

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

562

563 Methodologically it is worth noting that the results obtained from the MVPN combined-  
564 minus-max analyses only establish correlational relationships. To establish causality  
565 between the joint responses from the auditory and different visual streams in MSD sites,  
566 future research could employ techniques that infer causality, such as transcranial  
567 magnetic stimulation-fMRI (TMS-fMRI). Further, our method shows that two regions

568 jointly contribute to predict responses in a third region (i.e., statistical dependence), but  
569 we can not determine precisely whether and how this information is integrated into a  
570 multi-modal representation. In addition, we used a 5-layer dense neural network to  
571 model multivariate pattern dependence across all ROI sets tested in this study.  
572 However, it is possible that the optimal model architecture for capturing brain  
573 interactions may differ depending on the specific set of predictor regions. Future work  
574 using different neural network architectures may potentially uncover additional effects.  
575 Despite these limitations, the results reveal a novel aspect of the large-scale topography  
576 of STS, and provide insights into the neural architecture that supports our unified  
577 perception of the world.

578

579 The present work provides evidence for distinct portions of the multi-sensory posterior  
580 STS: a more posterior portion characterized by the combination of auditory and ventral  
581 representations, and a more anterior portion characterized by the combination of  
582 auditory and dorsal representations. Clarifying the functional and causal contributions of  
583 these subdivisions of STS to behavior will require additional work, including importantly  
584 studies with causal methodologies.

585 **References**

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

588

- 589 Anzellotti, S., Caramazza, A., & Saxe, R. (2017). Multivariate pattern dependence.
- 590 PLoS computational biology, 13(11), e1005799.
- 591
- 592 Avants, B. B., Tustison, N., & Song, G. (2009). Advanced normalization tools (ANTS).
- 593 Insight j, 2(365), 1-35.
- 594
- 595 Behzadi, Y., Restom, K., Liau, J., & Liu, T. T. (2007). A component based noise
- 596 correction method (CompCor) for BOLD and perfusion based fMRI. Neuroimage, 37(1),
- 597 90-101.
- 598
- 599 Born, R. T., & Bradley, D. C. (2005). Structure and function of visual area MT. Annu.
- 600 Rev. Neurosci., 28, 157-189.
- 601
- 602 Calvert, G. A., Campbell, R., & Brammer, M. J. (2000). Evidence from functional
- 603 magnetic resonance imaging of crossmodal binding in the human heteromodal cortex.
- 604 Current biology, 10(11), 649-657.
- 605
- 606 Calvert, G. A., Hansen, P. C., Iversen, S. D., & Brammer, M. J. (2001). Detection of
- 607 audio-visual integration sites in humans by application of electrophysiological criteria to
- 608 the BOLD effect. Neuroimage, 14(2), 427-438.

609

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

613

614 Collignon, O., Girard, S., Gosselin, F., Roy, S., Saint-Amour, D., Lassonde, M., &  
615 Lepore, F. (2008). Audio-visual integration of emotion expression. *Brain research*, 1242,  
616 126-135.

617

618 Deen, B., Koldewyn, K., Kanwisher, N., & Saxe, R. (2015). Functional organization of  
619 social perception and cognition in the superior temporal sulcus. *Cerebral cortex*, 25(11),  
620 4596-4609.

621

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

625

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

629

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

632

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

635

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

639

640 Felleman, D. J., & Van Essen, D. C. (1987). Receptive field properties of neurons in  
641 area V3 of macaque monkey extrastriate cortex. *Journal of neurophysiology*, 57(4), 889-  
642 920.

643

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

646

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

649

- 650 Hanke, M., Adelhöfer, N., Kottke, D., Iacobella, V., Sengupta, A., Kaule, F. R., ... &  
651 Stadler, J. (2016). A studyforrest extension, simultaneous fMRI and eye gaze  
652 recordings during prolonged natural stimulation. *Scientific data*, 3(1), 1-15.
- 653
- 654 Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001).  
655 Distributed and overlapping representations of faces and objects in ventral temporal  
656 cortex. *Science*, 293(5539), 2425-2430.
- 657
- 658 Hein, G., & Knight, R. T. (2008). Superior temporal sulcus—it's my area: or is it?.  
659 *Journal of cognitive neuroscience*, 20(12), 2125-2136.
- 660
- 661 Jenkinson, M., Bannister, P., Brady, M., & Smith, S. (2002). Improved optimization for  
662 the robust and accurate linear registration and motion correction of brain images.  
663 *Neuroimage*, 17(2), 825-841.
- 664
- 665 Kanwisher, N., McDermott, J., & Chun, M. M. (2002). The fusiform face area: a module  
666 in human extrastriate cortex specialized for face perception.
- 667
- 668 Lewis, J. W., Beauchamp, M. S., & DeYoe, E. A. (2000). A comparison of visual and  
669 auditory motion processing in human cerebral cortex. *Cerebral Cortex*, 10(9), 873-888.

670

671 McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*,  
672 264(5588), 746-748.

673

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

676

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

679

680 Peelen, M. V., Wiggett, A. J., & Downing, P. E. (2006). Patterns of fMRI activity  
681 dissociate overlapping functional brain areas that respond to biological motion. *Neuron*,  
682 49(6), 815-822.

683

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

686

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

689

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

693

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

697

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

700

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

702

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

705

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

708

709 Woolrich, M. W., Ripley, B. D., Brady, M., & Smith, S. M. (2001). Temporal  
710 autocorrelation in univariate linear modeling of fMRI data. *Neuroimage*, 14(6), 1370-  
711 1386.

712

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

715

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

718

719 **Legends**

720

721 **Fig. 1.** **a.** Visual and auditory regions of interest (ROIs). **b.** Responses in a combination  
722 of visual (e.g., early dorsal visual stream; Fig. 1a, middle panel) and auditory regions  
723 were used to predict responses in the rest of the brain using MVPN. **c.** In order to  
724 identify brain regions that combine responses from auditory and visual regions, we  
725 identified voxels where predictions generated using the combined patterns from auditory  
726 regions and one set of visual regions jointly (as shown in Fig. 1b) are significantly more  
727 accurate than predictions generated using only auditory regions or only that set of visual  
728 regions.

729

730 **Fig. 2.** **a.** Voxels showing significant effects ( $p < 0.05$ , FWE corrected) for the  
731 combination of auditory responses with responses in V3d and V5 (red), and auditory  
732 responses with responses in V3v and V4 (green). **b.** Voxels showing significant effects  
733 for the combination of responses in V3v and V4 with responses in V3d and V5 (blue). **c.**  
734 Fisher transformed Pearson correlation values between the auditory+dorsal and  
735 auditory+ventral combined-minus-max models, computed across the top 50 voxels in  
736 the STS (left) and the top 100 voxels across the whole brain (right) showing the greatest  
737 change in variance explained across both models. **d.** Pearson correlation values  
738 between combined-minus-max effect patterns from the auditory+dorsal and  
739 auditory+ventral models within an STS ROI. We computed these correlations across  
740 500 splits of the participants into two equal groups, comparing pattern similarity within

741 the same model across splits (e.g. AUD+dorsal and AUD+dorsal) to the similarity of  
742 patterns between different models across splits (e.g. AUD+dorsal in split 1 to  
743 AUD+ventral in split 2: “AD1 / AV2”).

744

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

754

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

758

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

762

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

766

767 **Table 1**

<b>Number of Voxels</b>	<b>t-values</b>	<b>MNI coordinates</b>	<b>Label</b>
126	6.64	(-66,-42,4)	STS
40	6.36	(45,-57,18)	STS
66	5.90	(51,-45,14)	STS
8	5.77	(57,-9,-9)	STS
24	6.60	(-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

768

769

770 **Table 2**

<b>Number of Voxels</b>	<b>t-values</b>	<b>MNI coordinates</b>	<b>Label</b>
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

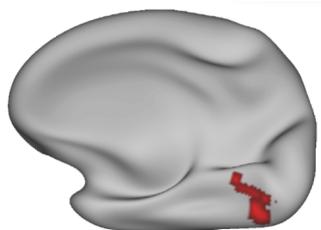
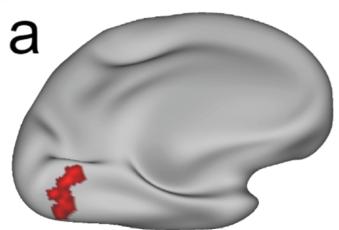
771

**Table 3**

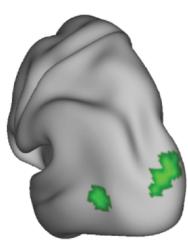
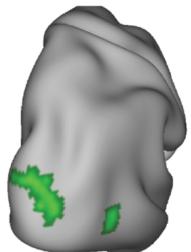
<b>Number of Voxels</b>	<b>t-values</b>	<b>MNI coordinates</b>	<b>Label</b>
1365	7.63	(21,-102,1)	V1
48	6.01	(-30,-48,-9)	PPA
16	5.70	(30,-51,-9)	PPA
12	5.93	(-24,-78,-25)	Cerebellum

772

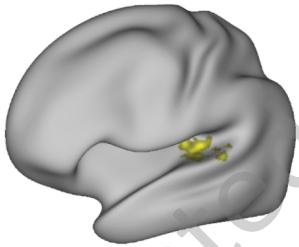
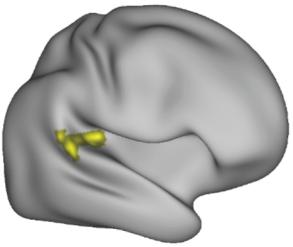
773



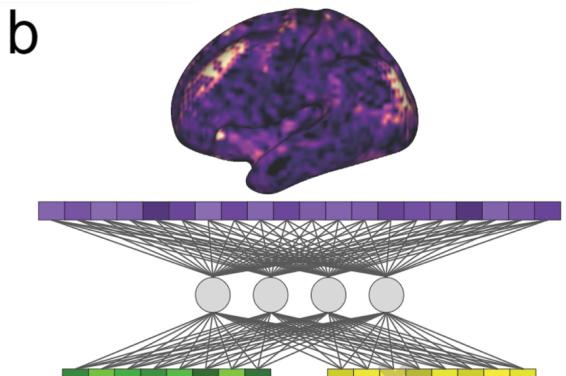
V3v and V4



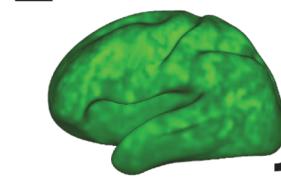
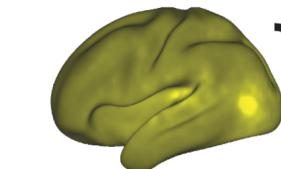
V3d and V5



Auditory Regions

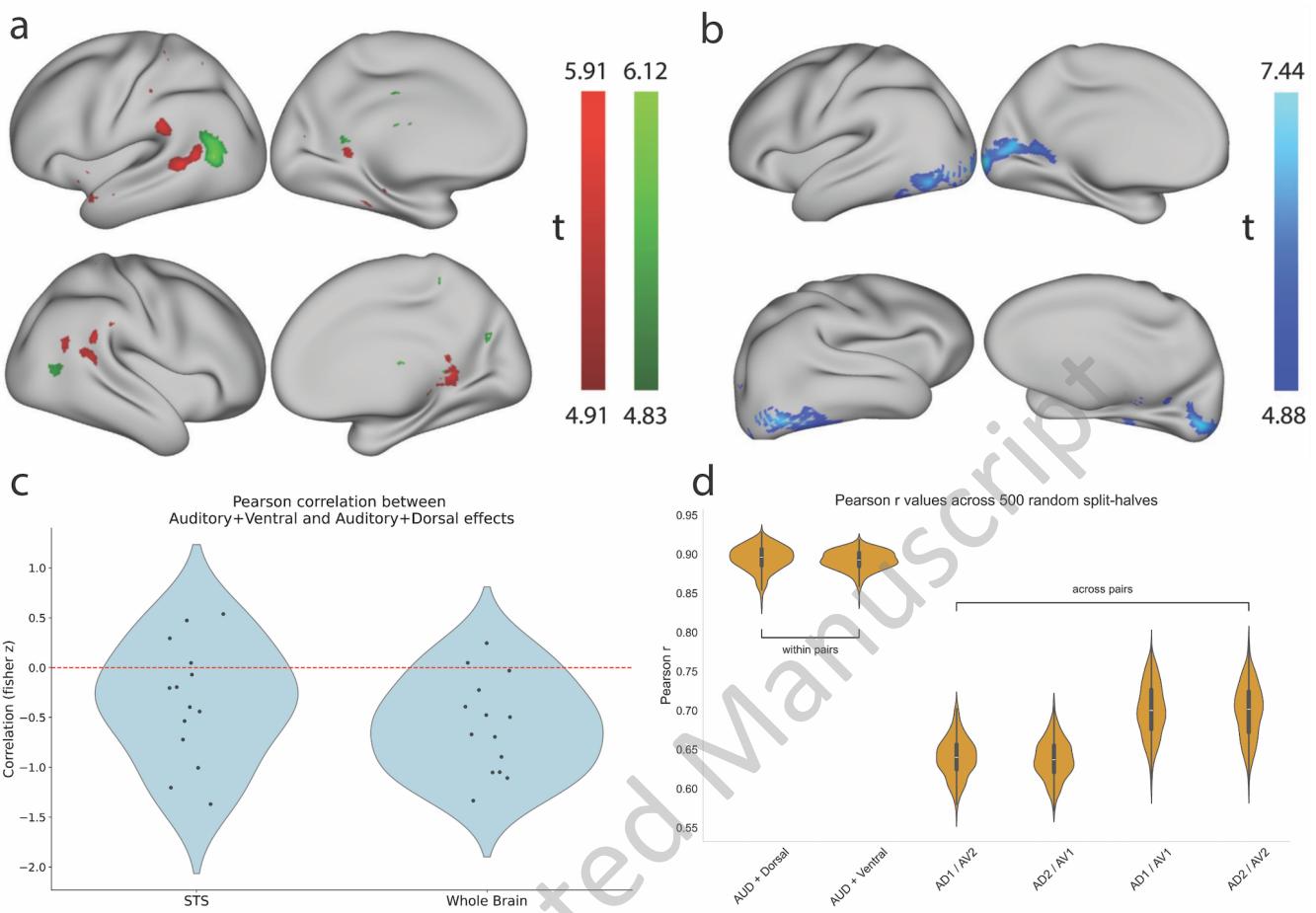


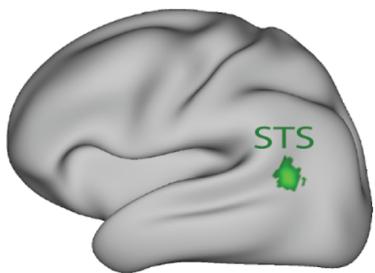
c



-

MAX



**a****b**

AUD + V3d/V5   AUD + V3v/V4   V3v/V4 + V3d/V5

