Imaging Neuroscience
an open access journal

Research Article

Auditory dyadic interactions through the "eye" of the social brain: How visual is the posterior STS interaction region?

Julia Landsiedel, Kami Koldewyn

Department of Psychology, School of Human and Behavioural Sciences, Bangor University, Bangor, United Kingdom

Corresponding Author: Kami Koldewyn (k.koldewyn@bangor.ac.uk)

ABSTRACT

Human interactions contain potent social cues that meet not only the eye but also the ear. Although research has identified a region in the posterior superior temporal sulcus as being particularly sensitive to visually presented social interactions (SI-pSTS), its response to auditory interactions has not been tested. Here, we used fMRI to explore brain response to auditory interactions, with a focus on temporal regions known to be important in auditory processing and social interaction perception. In Experiment 1, monolingual participants listened to two-speaker conversations (intact or sentence-scrambled) and one-speaker narrations in both a known and an unknown language. Speaker number and conversational coherence were explored in separately localised regions-of-interest (ROI). In Experiment 2, bilingual participants were scanned to explore the role of language comprehension. Combining univariate and multivariate analyses, we found initial evidence for a heteromodal response to social interactions in SI-pSTS. Specifically, right SI-pSTS preferred auditory interactions over control stimuli and represented information about both speaker number and interactive coherence. Bilateral temporal voice areas (TVA) showed a similar, but less specific, profile. Exploratory analyses identified another auditory-interaction sensitive area in anterior STS. Indeed, direct comparison suggests modality specific tuning, with SI-pSTS preferring visual information while aSTS prefers auditory information. Altogether, these results suggest that right SI-pSTS is a heteromodal region that represents information about social interactions in both visual and auditory domains. Future work is needed to clarify the roles of TVA and aSTS in auditory interaction perception and further probe right SI-pSTS interaction-selectivity using non-semantic prosodic cues.

Keywords: superior temporal sulcus, social interaction, heteromodal, neuroimaging, conversation, narration, MVPA

1. INTRODUCTION

Every day, humans need to navigate social encounters which contain potent social cues that meet not only the eye (e.g., gestures and facial expressions) but also the ear (e.g., intonation and vocalisation timing). As such, gauging information from both visual and auditory social interactions is critical to support adaptive behaviour in a complex social world. Indeed, human sensitivity to interactions is such that the presence of an interaction facilitates processing speed (Papeo et al., 2019; Vestner et al., 2019), recognition accuracy (Papeo & Abassi, 2019; Papeo et al., 2017), and working memory efficiency (Ding et al., 2017; Vestner et al., 2019). Correspondingly, akin to evidence for face- (Kanwisher & Yovel, 2006), body- (Downing et al., 2001) or voice-selective (Belin et al., 2000, 2002) brain areas, neuroimaging studies have identified interaction-sensitive regions within the bilateral lateral occipito-temporal cortex. Specifically, the posterior superior temporal sulcus social interaction region (SI-pSTS) has been found to play a key role in the representation of *visually* perceived dynamic social interactions. Across a range of stimuli that vary in the strength of

Received: 16 May 2023 Accepted: 17 May 2023 Available Online: 19 July 2023

relevant social cues (e.g., point-light displays, animated shapes, videos of dyads), the SI-pSTS responds about twice as strongly to interacting dyads compared to two independently acting individuals (Isik et al., 2017; Walbrin & Koldewyn, 2019; Walbrin et al., 2018) and shows this selectivity even in naturalistic videos (Landsiedel et al., 2022; Masson & Isik, 2021). Further, multivariate decoding analyses have found that the right SI-pSTS not only discriminates between interactors and non-interactors, but also appears to represent the type and emotional content of interactions (e.g., competing/cooperating; Isik et al., 2017; Walbrin & Koldewyn, 2019; Walbrin et al., 2018). On the other hand, extrastriate body area (EBA) has been implicated in the processing of the relational properties; i.e., the facing direction, between two bodies, which could be classed as "prototypical" visual interactions (Papeo, 2020), for both static (Abassi & Papeo, 2020, 2021) and dynamic (Bellot et al., 2021) stimuli. Altogether, this suggests a special role of *visually* presented interactive cues within the social brain. However, although interactions in the world are conveyed through auditory as well as visual means, auditory cues to interaction have not received much attention thus far. The current study sought to address this.

Undeniably, perceiving interactions between others is not only a visual but also an *auditory* perceptual experience. For instance, a person might overhear two people behind them conversing with each other or listen to a radio interview. Even without visual information, a great deal can be derived about interactions based on not only *what* is heard, i.e., the semantic content, but also *how* that content is conveyed, i.e., the tone of interaction due to variations in prosody (e.g., changes in intonation or volume, use of pauses, etc.). Indeed, conversational characteristics can affect how interactors are perceived and which characteristics are attributed to them during conversations. For instance, when listening to a two-speaker conversation which culminates in one person asking for something, the listener's ratings of the respondent's willingness to agree to this request decreased as the gap between the request and their affirmative response increased (Roberts & Francis, 2013; Roberts et al., 2006). Henetz (2017) used similar procedures and found that the perception of a conversation's awkwardness as well as interlocutors' rapport and desire to interact in the future depended on inter-turn silences. This emphasises that cues derived whilst listening to interactions are no less informative than cues gathered from visually observing interactions. In fact, one could suggest that most visual cues to interaction have auditory counterpoints that convey nearly identical social information, including things like the identity of interactants (Awwad Shiekh Hasan et al., 2016; Stevenage et al., 2012), conversational turn-taking (Cañigueral & Hamilton, 2019; Pijper & Sanderman, 1994), as well as cues to the interactants' emotions (de Gelder et al., 2015; Demenescu et al., 2014; Schirmer & Adolphs, 2017), intentions (Enrici et al., 2011; Hellbernd & Sammler, 2016), and social traits (Ponsot et al., 2018; Todorov et al., 2015).

Despite the richness of information contained in auditory interactions, the neural underpinnings of social interaction perception have been investigated almost exclusively in the visual domain. Studies probing neural representation of purely auditory interactions are next to non-existent. The closest proxy are studies investigating the auditory motion of two people walking, which convey some sense of togetherness or interactiveness. Bidet-Caulet et al. (2005) asked participants to listen to footsteps of two people walking, one on their left side (left ear) and one on their right side (right ear). Subsequently, one of the walkers would cross the auditory scene; thus, their footstep sounds would move towards the same side as the other walker's, which required the participants' response. Compared to a simple noise detection task (requiring auditory attention), bilateral posterior superior temporal sulcus (pSTS) increased activation in the footstep condition. However, their study did not probe auditory interaction perception, per se, which would have required contrasting the auditory motion of one person vs two. Work by Saarela and Hari (2008) tested this directly and found no differences in brain activation in pSTS, or indeed any other brain region, when comparing footsteps of two walkers vs one. While these auditory motion studies do not provide evidence for a region that is selectively engaged by auditory interactions, to the best of our knowledge, no study has specifically investigated the perception of auditory interactions using actual conversational compared to non-conversational speech, or indeed probed whether regions characterised by its sensitivity to visual interactions might also be driven by auditory interactions.

In spite of the lack of interaction-specific studies, investigations focussed on the processing of other social stimuli (e.g., faces and voices) support the notion of heteromodal (i.e., responding to both visual and auditory stimuli) processing in the broader STS region (Deen et al., 2020; Watson et al., 2014), which is in line with reports of a significant overlap between face- and voice-sensitive voxels in parts of the pSTS (Deen et al., 2015). Furthermore, several studies have proposed the STS as an area

of audio-visual integration of both emotional and neutral facial and vocal expressions (Kreifelts et al., 2009; Robins et al., 2009; Watson et al., 2014; Wright et al., 2003), reflected by enhanced pSTS activation in response to both modalities compared to unimodal stimulus presentation. Given its proximity to auditory cortical areas, and nearby regions demonstrated to be integrative and/or heteromodal, the SI-pSTS seems an obvious candidate to investigate in the context of auditory interactions. This is in contrast to EBA, which may also be involved in interaction processing (Abassi & Papeo, 2020, 2021; Bellot et al., 2021), but which is considered to be a strictly visual region and not responsive to auditory information (Beer et al., 2013).

Across two experiments, we addressed the hypothesis that the SI-pSTS region might play a crucial role in the processing of not only visual but also auditory interactions using speech stimuli with three levels of interactiveness: interactions (conversations between two people) and their scrambled counterparts, as well as non-interactions (stories narrated by one person) in two languages. Importantly, scrambling recombined complete utterances taken from different interactions. Thus, speech was comprehensible at the sentence level but sentences were not semantically related to each other. In Experiment 1, participants were monolingual, whereas in Experiment 2, participants were bilingual. We used functional localisers (Fedorenko et al., 2010) to define bilateral regions of interest (ROIs): the visual SI-pSTS region, voice-selective temporal voice areas (TVA), and temporal parietal junction (TPJ) and tested their responses to auditory interactions across both univariate and multivariate pattern analyses. We hypothesised that if visual SI-pSTS was, in fact, heteromodal, it would show greater response to conversation stimuli involving two speakers compared to one-speaker narrations (regardless of semantic comprehension in monolinguals in Experiment 1). Beyond this broad test of auditory interaction sensitivity, we also expected SI-pSTS to be sensitive to the difference between conversations and scrambled conversations (which for monolingual participants deteriorated interactive cues of conversation coherence in their native language, and conversational flow/prosody in the unknown language). TVA (Agus et al., 2017; Belin et al., 2000, 2002) was included as an auditory control region that we expected to respond to all conditions, though given the dearth of information on auditory interaction processing, we did not have strong expectations regarding its interaction sensitivity. Our reasoning was that if SI-pSTS did not show heteromodal characteristics and sensitivity to auditory interactions could be "found" anywhere in the brain, such sensitivity might emerge in an area tuned to voices, like TVA. Finally, TPJ was included as a "social" control region that is spatially very near SI-pSTS but that we did not expect to be driven by either auditory stimuli in general, or manipulations of interactiveness specifically (Walbrin & Koldewyn, 2019; Walbrin et al., 2020). Additionally, whole-brain analyses were conducted to explore the wider brain networks implicated in auditory interaction processing.

2. METHODS

2.1. Experiment 1

2.1.1. Participants

Twenty-four right-handed participants were recruited to take part in this study. Handedness was confirmed using the Edinburgh Handedness Inventory (EHI; Oldfield, 1971). All participants had normal or corrected to normal vision, were native English speakers, and had no German language skills. After data exclusion (see 2.1.2), one participant was removed from the analyses (final sample of N = 23; mean age = 22.35, SD = 3.04; 7 males). All participants gave informed consent, were debriefed at the end of the study, and received monetary renumeration for their time. The protocol was approved by the School of Psychology's ethics committee at Bangor University and was pre-registered on AsPredicted.org (ID23865) on 23/05/2019.

2.1.2. Design & procedure

To investigate auditory interaction perception with and without language comprehension, the main experimental task consisted of a 2 × 3 repeated-measures fMRI event-related design. Across two *languages* (English and German), auditory interactiveness was manipulated using three *conditions*: non-interactive one-speaker narrations, interactive two-speaker conversations, as well as an intermediate condition using scrambled conversations (see 2.1.3 for details). This condition still contained interactive cues (two speakers taking turns); however, conversational content was not coherent. German stimuli were used to explore interactive effects independent of stimulus comprehension.

Each run contained 36 trials, including 24 task trials (4 per condition) and 12 catch trials (2 per condition). The order of conditions was pseudo-randomised using custom MATLAB code to optimise the efficiency of the design. The inter-stimulus interval was jittered (mean jitter 1.5 seconds, range = 0–3 seconds). Participants com-

pleted seven runs (28 task trials and 14 catch trials per condition across runs, thus 252 trials overall), each between 5.6–5.8 minutes in length. Due to the variability in stimulus length (range: 6–11 seconds), each run contained a specific set of stimuli. The order of the runs was counter-balanced across participants.

Participants were instructed to listen attentively, and an orthogonal catch-trial detection task was used to maintain (and confirm) participants' attention throughout. Catch trials consisted of recordings that were manipulated such that a single word was repeated, e.g., "Do you *need need* anything from the supermarket?". The occurrence of these repeated catch words was balanced for all conditions such that the repeated word occurred equally often in the first, second, third, or fourth sentence of a stimulus. Participants had to accurately detect seven out of 12 catch trials per run for that run to be included in subsequent analyses (see Supplementary S1 for the behavioural results). Based on this criterium, two participants had one, and another participant three run(s) removed from the analyses. Only participants with five or more runs after data exclusion were included in the final sample.

The task was presented in Psychtoolbox 3.0.14 (Brainard, 1997; Kleiner et al., 2007) using MATLAB 2018.a (The MathWorks Inc.) running on a Linux Ubuntu 16.04 distribution stimulus computer. Sensimetrics (model S15) MR-safe in-ear earphones were used for stimulus presentation. Stimulus volume was adjusted to a comfortable level individually for each participant.

2.1.3. Stimuli

Scripted conversations and narrations were developed in English, and subsequently translated into German by two native speakers. All stimuli were recorded specifically for this study (see Supplementary S2 for recording details) by native English or native German speakers. A large set of stimuli was recorded, from which the final stimulus set was selected (see Supplementary S3 and S4.1 for details).

Narrations were recorded for each speaker separately. Narration content was loosely based on and inspired by children's books. Care was taken that narrations remained descriptive rather than invoking mentalising processes.

Conversations were recorded in pairs (two same-gender pairs per language condition, one male, one female) to capture "true" interactions. They consisted of a short exchange (usually four sentences in total) between the two speakers taking turns (Agent A - Agent B - Agent A - Agent B). All conversations were recorded twice so that both speakers played both agent roles. Conversations varied in content, e.g., asking a friend about their exam or ordering food in a restaurant.

Scrambled conversations were created from the original conversation scripts by randomly combining individual speaker turns from different conversations into a new combination, which still consisted of two speakers taking turns, but where the turns were taken from different conversations and, thus, were unrelated in semantic meaning. Importantly, this process could result in a speaker turn being spoken by a different speaker in the scrambled compared to the original conversations whilst the script, i.e., the semantic content remained the same. For this, the original conversations (from each gender pair) were cut up into their individual speaker turns (two per agent) using Audacity (The Audacity Team) audio software. Using custom MATLAB code, turns from four random original conversations were selected and re-combined such that each turn was no longer in its original position of the conversation. For example, the opening turn of a conversation could only appear in the 2nd, 3rd, or 4th position in a scrambled conversation. This method was chosen to disrupt the natural conversational flow not only through mixing up content but also through disrupting prosodic and intonational cues. Scrambling generated equal numbers of stimuli, with either speaker taking the role of agent A or agent B.

All stimuli were analysed in Praat software (http://www.praat.org) to assess mean pitch (fundamental frequency, F0) for each condition (see Supplementary S4.2). Furthermore, the final set of English stimuli was rated on perceived naturalness, valence, and imaginability/mental imagery, and for the two-speaker conditions also on interactiveness and perceived closeness between the two speakers. In brief, conversation and narrations were closely matched on all non-interactive dimensions. In contrast, conversations and scrambled conversations significantly differed across all scales (see Supplementary S4.3 for rating data and statistics).

2.1.4. Localiser tasks

To define independent regions of interest (ROI), participants also completed a set of established localiser tasks. The *interaction localiser* (Isik et al., 2017; Walbrin et al., 2018) was used to localise the SI-pSTS region sensitive to visual social interactions. Across three runs, participants watched videos of two point-light agents in three conditions (interaction, scrambled interactions, and non-interaction/ independent actions). Each run contained two 16-second blocks per condition and three

16-second rest blocks (total run time 144 seconds). Bilateral SI-pSTS was localised using the contrast interactions > non-interactions. The *voice localiser* (Pernet et al., 2015) was used to localise the temporal voice areas (TVAs) along the anterior-posterior axis of the STS. The TVAs show sensitivity to human vocal (speech and non-speech) sounds (Agus et al., 2017; Belin et al., 2000, 2002; Pernet et al., 2015), and seem particularly involved in processing of paralinguistic information such as gender (Charest et al., 2013), identity (Latinus et al., 2013), or emotion (Ethofer et al., 2009, 2012). Therefore, we do not expect to find interaction sensitivity in the TVAs. Participants completed one run, listening to human vocal sounds (speech, e.g., words, or syllables; and non-speech sounds, e.g., laughs or sighs); and non-vocal sounds (natural sounds like waves or animals, and man-made object sounds like cars or alarms) across twenty 8-second blocks respectively. These condition blocks were interspersed with twenty 10-second blocks of silence (total run time 10.3-minutes). Bilateral TVA was localised using the contrast human vocal sounds > non-vocal sounds. Finally, a third localiser, described by Jacoby et al. (2016), was used to define temporo-parietal junction as a control regions within the "social brain" (see Supplementary S5 for details).

2.1.5. MRI parameters, pre-processing, & GLM estimation

Data were collected at the Bangor Imaging Centre using a Philips Achieva 3-T scanner using a 32-channel head coil (Philips, Eindhoven, the Netherlands). A T2*-weighted gradient-echo single-shot EPI pulse sequence (with SoftTone noise reduction, TR = 2000 ms, TE = 30 ms) was used for all tasks (with slightly different parameters depending on task for flip angle, FOV, number of slices, and slice order, see Table 1).

Structural images were obtained with the following parameters: T1-weighted image acquisition using a gradient echo, multi-shot turbo field echo pulse sequence, with a five echo average; TR = 12 ms, average TE = 3.4 ms, in 1.7 ms steps, total acquisition time = 136 seconds, flip angle = 8°, FOV = 240 × 240, acquisition matrix = 240 × 224 (reconstruction matrix = 240); 128 contiguous axial slices, acquired voxel size (mm) = 1.0 × 1.07 × 2.0 (reconstructed voxel size = 1 mm^3).

Pre-processing and general linear model (GLM) estimation were performed using SPM 12 (fil.ion.ucl.ac.uk/spm/software/spm12) in MATLAB 2018.a (The MathWorks Inc.). Pre-processing included slice-timing (event-related main auditory task only), realignment and re-slicing, co-registration to anatomical image, segmentation, normalisation (normalised MNI space with 2 mm isotropic voxels), and smoothing. All SPM12 default pre-processing parameters were used, except for the use of an initial 3 mm FWHM Gaussian smoothing kernel. This smoothing kernel is recommended when using ArtRepair toolbox (v5b; Mazaika et al., 2005). Specifically, ArtRepair was used to detect and repair noisy volumes (volumes that contained more than 1.3% variation in global intensity or 0.5 mm/TR scan-to-scan-motion). Thirteen subjects required repairs in at least one run. Prior to first-level modelling, data were smoothed again using a 5 mm FWHM kernel. Subsequently for each run, event or block durations and onsets for each experimental condition were modelled using a boxcar reference vector and convolved with a canonical hemodynamic response function (without time or dispersion derivatives) using a 128-second high-pass filter and autoregressive AR(1) model. Head motion was modelled using six nuisance regressors (translation and rotation). Additionally, for the main auditory task, catch trials were modelled as a regressor of no interest. Rest periods were modelled implicitly.

2.1.6. Whole-brain group analyses

For localiser tasks, respective contrasts were modelled at the group level using one-sample t-tests. For the main auditory task, the Multivariate and Repeated Measures (MRM) toolbox (McFarquhar et al., 2016) was used. Each participant's baseline contrast images of the six experi-

Table 1. Scanning parameters

	Dummy scans	Number of slices	Scan order	Flip angle	FOV	Voxel sizes acquired/ reconstructed	Matrix
Interaction localiser	4	36, Gap 0 mm	Ascending (foot to head)	83°	240 × 240 × 112	3 × 3.08 × 3.5 mm/3 mm	Acquisition 80 × 78 Reconstruction 80
Audio task Voice localiser	5	35, Gap 0 mm	Odd-even (1 3 2 4)	77°	240 × 240 × 105	3 × 3.08 × 3 mm/3 mm	

mental conditions were entered into a 2 (Language) × 3 (Condition) repeated-measures ANOVA. Within this model, *F*-contrasts were computed for the main effects and the interaction effect using Wilks' Lambda as multivariate test statistic. All reported F-contrasts were thresholded using an initial cluster-forming threshold of $p < .001$ uncorrected, followed by permutation tests (5000 iterations) to provide cluster-level FWE correction for multiple comparisons of $p_{FWE} < .05$.

2.1.7. ROI creation & percent signal change (PSC) analyses

Functional ROIs were defined following a "group-constrained" ROI definition approach (for details see Julian et al., 2012). This approach reduces subjectivity in how ROI locations are selected and ensures that ROI selection is not influenced by the data that will be extracted and analysed from a given ROI. To start, group-level T/F-maps were used to identify MNI coordinates of bilateral ROIs (see Supplementary S6). These coordinates formed the centres of initial *8 mm bounding spheres*. Subject-specific search spaces were then defined by running a group-level analysis to determine a peak coordinate for activation that was used to localise this search space using a leave-one-subject-out (LOSO) approach, i.e., the group contained all subjects except the "current" subject whose search space was being defined. The final *subject-specific search space* was defined based on the intersection of the original 8 mm bounding sphere with the group-level T-map of the LOSO scheme. To create the final *subject-specific ROI* (see Fig. S1), the top 100 *contiguous* voxels (highest T-values) within the *subject-specific search* were selected for each participant individually (Walbrin et al., 2020). Thus, while all ROIs were 100 voxels in size, they differed across participants in their exact placement. Additionally, a leave-one-run-out (LORO) scheme was implemented in this step (defining an ROI on all but one run, extracting PSC from the remaining one in an iterative n-fold partition scheme) in cases where ROI definition and extraction were based on the same task, i.e., when testing the response of the SI-pSTS itself to the interaction localiser. This procedure could not be applied when testing the response of the TVA to the voice localiser conditions, however, as the localiser consisted of only one run. PSC data were extracted from ROIs using the Marsbar toolbox (Brett et al., 2002).

For the main auditory task, PSC were analysed in a 2 (Language) × 3 (Condition) repeated-measures ANOVAs for each ROI respectively. Greenhouse-Geisser correction for violation of assumptions of sphericity was applied where necessary. Multiple comparison correction was implemented based on the number of ROIs tested in a given contrast; i.e., multiple tests for each contrast were considered as a "family" of statistical tests that should be corrected across. Given our four regions of interest (bilateral SI-pSTS and TVA), this resulted in a corrected *p*-value of $p < .0125$ (.05/4) for both contrasts used to test main and interaction effects. For the interaction localiser, differences between conditions were analysed using paired-sample t-tests, with particular focus on two contrasts of interest: interactions vs non-interactions and scrambled interactions respectively. For the auditory task, multiple comparison correction was applied based on the number of ROIs tested in a given contrast (corrected *p*-value of $p < .0125$). Furthermore, for selected ROIs, auditory- and visual-*interaction selectivity* was calculated as the t-value of differences in activation between conversations vs narrations across languages for the main experimental task, and interactions vs non-interactions for the interaction localiser, and compared using paired-sample t-tests.

For all paired t-test comparisons, effect sizes represent Cohen's *d* for repeated measures (d_{rm}), which represents the mean difference standardised by the standard deviation of the difference scores corrected for the correlation between the measurements (Lakens, 2013).

2.1.8. Multivariate pattern analyses (MVPA)

Pattern decoding analysis using an iterative n-folds partition scheme of LORO was implemented using the CoSMoMVPA toolbox (Oosterhof et al., 2016) with a focus on four contrasts of interest: conversations vs narrations and conversations vs scrambled conversations for each language. For each subject, a binary linear support vector machine (SVM) classifier was trained on a given ROI's voxel patterns for the conditions of interest (i.e., beta values in a subject's respective top 100-voxels ROI, see section 2.1.7, averaged across all trials per condition per run) in all but one run of data—with the "left-out" run of data used to independently test classification performance. This resulted in as many folds of cross-validation as a subject had valid task runs (usually seven). Prior to classification, voxel patterns were normalised (demeaned) for each run separately. Following cross-validation, classification accuracy was averaged across all n-folds iterations before being entered into group-level analysis. For each contrast of interest, average classification accuracy was tested against chance (50%) using one-tailed one-sample t-tests.

To correct for multiple comparisons across four ROIs, we used a so-called "singleton" neighbourhood, where each ROI to be corrected for was treated as one feature. This means that each ROI was only a neighbour of itself. This neighbourhood was then tested using Monte-Carlo based clustering statistics. Here, we used Threshold-Free-Cluster-Enhancement (TFCE; Smith & Nichols, 2009) as a clustering statistic with 10.000 iterations of Monte Carlo simulations (cf. cosmo_montecarlo_cluster_stat.m). Although traditionally, TFCE is used to test cluster survival based on iteratively testing the spatial clustering at different height thresholds to determine how much local support each feature (voxel) has (using a neighbourhood composed of many features), the same principle can also be applied to correct for multiple comparison across ROIs when conceptualising each ROI as a cluster (cf. cosmo_singleton_neighborhood.m). Each iteration of Monte Carlo simulations generated null data based on the sample's classification accuracies in each ROI using a sign-based permutation approach (also implemented in FieldTrip; see Maris and Oostenveld, 2007). Significance was determined based on the comparison of "clustering" in the null data across all iterations compared to "clustering" observed in the original data. This method yields conservative estimates of significance. We report the resulting one-tailed Z-scores and *p*-values where Z-scores greater than 1.65 are indicative of significant above chance classification in each ROI after correction for multiple comparisons. As this method of multiple comparison correction is somewhat opaque, we additionally ran standard one-sample t-tests whose *p*-values can then be evaluated against a Bonferroni-corrected *p*-value threshold. Corrected *p*-values were determined as above for the PSC analyses, using the number of ROIs tested in a given contrast (corrected *p*-value of $p < .0125$ based on four ROIs per contrast 0.05/4).

2.2. Experiment 2

2.2.1. Participants

Twelve German native speakers took part in this study (mean age = 22.50, SD = 2.28, 3 males). All participants were right-handed as confirmed by the EHI, and had normal or corrected to normal vision. The sample's English language skills were at a minimum of B2 (upper intermediate, Common European Framework of Reference for Languages), which is the minimum level required by universities for international first-year students. Participants gave informed consent and were debriefed and paid at the end of the study.

2.2.2. Design & procedure

All procedures were the same as in Experiment 1, although participants only completed the main auditory experimental task as well as the pSTS interaction localiser. Due to technical difficulties, the TVA voice localiser could not be acquired for this sample.

2.2.3. MRI parameters & pre-processing

Compared to Experiment 1, images were acquired on a different scanner (Philips Igenia Elition X 3T scanner) with a 32-channel head coil (Philips, Eindhoven, the Netherlands) at the Bangor Imaging Centre. Acquisition parameters for functional runs were the same as in Experiment 1. The structural sequence was slightly different: for each participant, a high-resolution anatomical T1-weighted image acquired using a gradient echo, multi-shot turbo field echo pulse sequence, with a five-echo average; TR = 18 ms, average TE = 9.8 ms, in 3.2 ms steps, total acquisition time = 338 seconds, flip angle = 8°, FOV = 224 × 224, acquisition matrix = 224 × 220 (reconstruction matrix = 240); 175 contiguous slices, acquired voxel size (mm) = 1.0 × 1.0 × 2.0 (reconstructed voxel size = 1 mm^3).

Pre-processing and GLM estimation were performed using the same pipeline as in Experiment 1. Due to low levels of head-motion in this sample, ArtRepair was not used to repair noisy volumes. Due to human error, the first functional run of the main auditory task had to be discarded for two participants.

2.2.4. ROI creation & PSC analyses

For the SI-pSTS, the same group constrained ROI definition process was used as in Experiment 1, resulting in subject-specific ROIs consisting of the top 100 contiguous voxels in each hemisphere separately. However, due to the small sample size in this study, the initial 8 mm constraining sphere used the same group-level MNI coordinates as in Experiment 1, rather than using coordinates of our underpowered sample. Due to the missing voice localiser scan, for bilateral TVA, ROIs consisted of 6 mm spheres, again using centre MNI coordinates from Experiment 1. This radius was chosen to select a sphere size that contained a comparable number of voxels as the SI-pSTS ROI (6 mm sphere = 123 voxels). As before, PSC data were extracted for the experimental conditions of the main auditory task using Marsbar toolbox for the respective ROIs and analysed using 2 × 3 repeated-measures ANOVAs.

3. RESULTS

3.1. Experiment 1

Results of the PSC results (Fig. 1 left panel) and MVPA analyses (Fig. 1 right panel) will be presented below separately for the visually defined SI-pSTS and auditorily defined TVA (refer to Table 2 for ANOVA statistics and Supplementary S7, Table S7 for condition means). Results for TPJ as an additional control region within the "social brain" can be found in Supplementary S6, as it consistently showed activation at or below baseline. Bspmview toolbox was used for whole-brain data visualisation (DOI: 10.5281/zenodo.168074, see also https://www.bobspunt.com/software/bspmview/).

3.1.1. How does the visually defined SI-pSTS respond to auditory interactions?

PSC analyses. In line with our predictions, there was a main effect of Condition, where both conversations

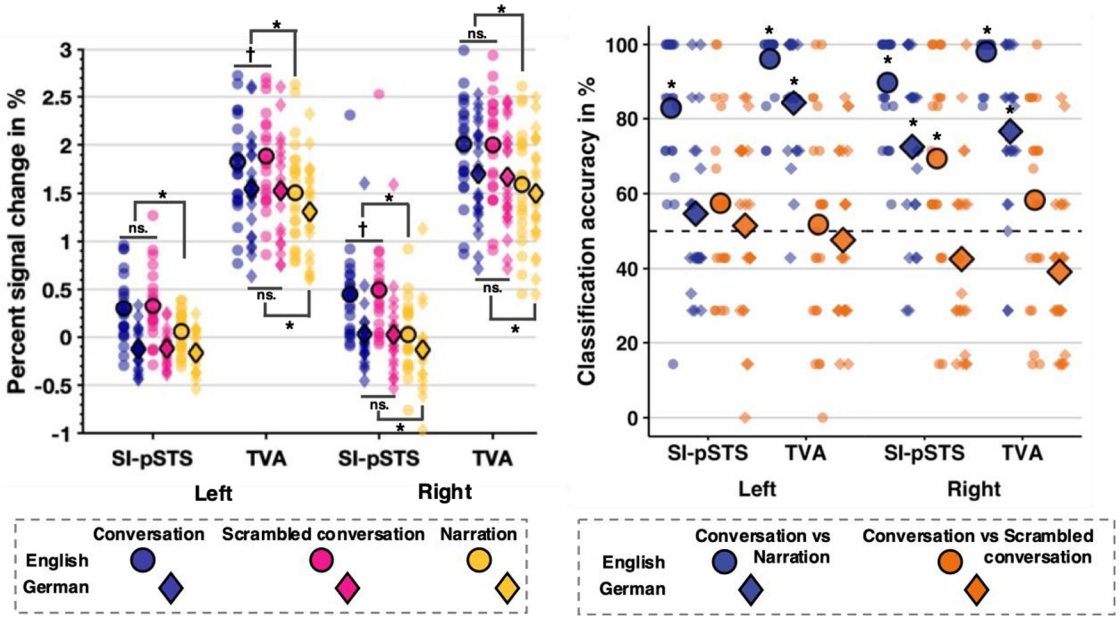

Fig. 1. Condition means (circle/diamond shape with bold edge) and data distribution dot plot for each ROI and hemisphere for Experiment 1. Left panel: Percent signal change data. Only significant post hoc t-test results are marked (*: $p < .0125$ corrected, †: $p < .05$ uncorrected). Effects for English conditions are shown above, and for German conditions below the condition mean. Right panel: SVM classification accuracy. Significant above chance classification accuracy is indicated using an asterisk (*: $p < .05$ TFCE-corrected). Chance level of 50% is represented using the dashed horizontal line.

Table 2. ANOVA results of PSC analyses for the main auditory task in Experiments 1 and 2

		Experiment 1				Experiment 2			
		SI-pSTS		TVA		SI-pSTS		TVA	
		L	R	L	R	L	R	L	R
Main effect of language	*F*	36.02	50.58	29.44	15.84	1.37	1.71	4.91	13.56
	p	*<.001*	*<.001*	*<.001*	*<.001*	.27	.22	*.049*	*.004*
	η_p^2	0.62	0.70	0.57	0.42	0.11	0.13	0.31	0.55
Main effect of condition	*F*	14.24	35.99	62.21	40.89	10.57	23.88	27.47	20.2
	P	*<.001*	*<.001*	*<.001*	*<.001*	*<.001*	*<.001*	*<.001*	*<.001*
	η_p^2	0.39	0.62	0.74	0.65	0.49	0.69	0.71	0.65
Interaction effect	*F*	12.18	24.71	9.86	25.42	0.81	3.5	0.10	0.25
	p	*<.001*	*<.001*	*<.001*	*<.001*	.41	*.048*	.91	.78
	η_p^2	0.36	0.53	0.31	0.54	0.07	0.24	0.01	0.02

Italicised *p*-values indicate significance ($p < .05$).

(rSI-pSTS: $t(22) = 6.17$, $p < .001$, $d_{rm} = 0.41$, lSI-pSTS: $t(22) = 3.82$, $p < .001$, $d_{rm} = 0.22$) and scrambled conversations (rSI-pSTS: $t(22) = 6.13$, $p < .001$, $d_{rm} = 0.44$, lSI-pSTS: $t(22) = 4.00$, $p < .001$, $d_{rm} = 0.20$) evoked greater responses than narrations in bilateral SI-pSTS. However, PSC did not differ for intact vs scrambled conversations. Thus, this effect was driven entirely by a difference between hearing two speakers vs hearing only one. Unexpectedly, there was also a large main effect of language, where responses in bilateral SI-pSTS were greater for English compared to German stimuli. Indeed, while PSC in bilateral SI-pSTS was significantly above baseline for English conversations and English scrambled conversations (all $t(22) \geq 4.06$, all ps $< .001$), for lSI-pSTS, German conditions led to a significant decrease in activation (all ts$(22) < -2.95$, all ps $< .007$) and response in rSI-pSTS was not significantly different than baseline. Both effects suggest that the *comprehensibility* of heard interactions is important to response within the SI-pSTS. These main effects were qualified by a significant Language × Condition interaction. Whilst responses in lSI-pSTS were only greater for the comprehensible English two-speaker conversation compared to single-speaker narrations ($t(22) = 4.14$, $p < .001$, $d_{rm} = 0.77$), rSI-pSTS showed this pattern independent of comprehensibility (English: $t(22) = 6.28$, $p < .001$, $d_{rm} = 0.90$; German: $t(22) = 4.76$, $p < .001$, $d_{rm} = 0.41$). Additionally, rSI-pSTS response to English scrambled conversations was slightly greater than conversations, albeit at an uncorrected level only ($t(22) = -2.16$, $p = .04$, $d_{rm} = -0.09$), suggesting weak sensitivity to the *coherence* of comprehensible interactions.

MVPA analyses. Classification analyses in the rSI-pSTS revealed that the SVM classifier could discriminate between voxel patterns representing conversations and narrations in both languages (English: M = 0.90, SE = 0.03, Z = 3.72, $p_{TFCE} < .001$, $p_{T\text{-}test} < .001$; German: M = 0.72, SE = 0.05, Z = 3.24, $p_{TFCE} < .001$, $p_{T\text{-}test} < .001$), in line with the PSC results. Crucially, strengthening the PSC results, the classifier could also decode voxel patterns of English conversations vs scrambled conversations (M = 0.69, SE = 0.06, Z = 2.49, $p_{TFCE} = .006$, $p_{T\text{-}test} = .002$) with above chance accuracy. This suggests that rSI-pSTS voxel patterns code for interaction information based on the number of speakers and interaction coherence when two speakers are present. Classification analyses in lSI-pSTS revealed above chance discrimination between English conversations vs narrations only (M = 0.83, SE = 0.04, Z = 3.72, $p_{TFCE} < .001$, $p_{T\text{-}test} < .001$), suggesting no strong auditory interaction sensitivity in the left hemisphere. German conversations vs German scrambled conversations were not decodable above chance in either region.

3.1.2. How does the TVA, a region generally sensitive to voices, respond to auditory interactions?

PSC analyses. In bilateral TVA, while there was also a significant main effect of language, PSC was significantly greater than baseline for all conditions, regardless of language (all $t(22) \geq 12.62$, all ps $< .001$). Thus, bilateral TVA was clearly driven by voice stimuli regardless of comprehensibility. However, bilaterally, PSC was greater in response to English compared to German stimuli. As in the SI-pSTS, there was also a main effect of Condition bilaterally. PSC was smaller for narrations compared to both conversations (rTVA: $t(22) = 7.10$, $p < .001$, $d_{rm} = 0.54$, lTVA: $t(22) = 7.28$, $p < .001$, $d_{rm} = 0.54$) and scrambled conversations (rTVA: $t(22) = 6.09$, $p < .001$, $d_{rm} = 0.55$, lTVA: $t(22) = 9.58$, $p < .001$, $d_{rm} = 0.51$) but no difference was found between the latter two. Thus, the number of voices, hearing one or two speakers, clearly modulated TVA activation. These main effects were qualified by a significant Language × Condition interaction. Bilaterally, TVA responses were greater for conversations compared to narrations for both English (rTVA: $t(22) = 7.57$, $p < .001$, $d_{rm} = 0.74$, lTVA: $t(22) = 5.66$, $p < .001$, $d_{rm} = 0.55$) and German stimuli (rTVA: $t(22) = 5.18$, $p < .001$, $d_{rm} = 0.36$, lTVA: $t(22) = 8.98$, $p < .001$, $d_{rm} = 0.44$). Surprisingly, although at an uncorrected level, lTVA responded less to English conversations compared to scrambled conversations ($t(22) = -2.62$, $p = .02$, $d_{rm} = -0.09$), indicating potential sensitivity to the *coherence* of comprehensible interactions.

MVPA analyses. Classification analyses revealed that voxel patterns representing conversations and narrations could be decoded for each language respectively in right (English: M = 0.98, SE = 0.01, Z = 3.72, $p_{TFCE} < .001$, $p_{T\text{-}test} < .001$; German: M = 0.77, SE = 0.04, Z = 3.54, $p_{TFCE} < .001$, $p_{T\text{-}test} < .001$) and left (English: M = 0.96, SE = 0.02, Z = 3.72, $p_{TFCE} < .001$, $p_{T\text{-}test} < .001$; German: M = 0.84, SE = 0.03, Z = 3.72, $p_{TFCE} < .001$, $p_{T\text{-}test} < .001$) TVA (consistent with the PSC results). Importantly, bilaterally, discrimination of conversations vs scrambled conversations based on TVA voxel patterns was not successful for either language.

3.1.3. What is the evidence for heteromodal social interaction processing in SI-pSTS (and TVA)?

To examine heteromodal processing in response to social interactions, we examined and, if appropriate, compared

responses to visual interactions with responses to auditory interaction within our ROIs. Analyses confirmed sensitivity to visual interactions in bilateral SI-pSTS, which showed greater responses to interactions compared with both non-interactions and scrambled interactions (all *ts*(22) > 5.84, all *ps* < .001; see Fig. 2 left panel and Supplementary S7, Table S8). In contrast, bilateral TVA responded at or below baseline to the interaction localiser conditions (all *t*(23) < -2.00, *p* < .06). Therefore, the subsequent comparison of interaction-selectivity across modalities focussed on SI-pSTS only. Analyses revealed that visual interaction selectivity (rSI-pSTS: M = 0.62, SE = 0.08, lSI-pSTS: M = 0.53, SE = 0.06) in bilateral SI-pSTS was significantly greater (rSI-STS: $t(22) = 5.81$, $p < .001$, $d_{rm} = 0.34$, lSI-pSTS: $t(22) = 6.60$, $p < .001$, $d_{rm} = 0.26$) than auditory interaction selectivity (rSI-pSTS: M = 0.29, SE = 0.05, lSI-pSTS: M = 0.14, SE = 0.04). Altogether, this suggests heteromodal processing of social interaction in SI-pSTS, though with clear preference for visual stimuli.

3.1.4. Does general voice sensitivity explain responses to auditory interactions in SI-pSTS?

General responsiveness to voice stimuli (see also Supplementary S7, Table S9) was examined by extracting PSC from the voice localiser in SI-pSTS (see Fig. 2, right panel). Although vocal sounds activated rSI-pSTS above baseline ($t(23) = 2.16$, $p = .04$), the region was not strongly driven by human voices. In fact, compared to comprehensible English conversations (M = 0.45, SE = 0.10), rSI-pSTS responded about 50% less to vocal stimuli (M = 0.25, SE = 0.12). Overall, this analysis suggests that SI-pSTS responses in the main experimental task were *not* due to a general sensitivity to voices.

3.1.5. Does whole-brain data reveal an additional region sensitive to auditory interactions?

To explore whole-brain auditory interaction sensitivity, we followed a data-driven approach. Rather than focussing on the main effect of condition, the ROI PSC indicated an unexpected but robust Language × Condition interaction effect for our key ROIs. Therefore, the whole-brain interaction effect contrast was used to identify potential candidate regions that may be sensitive to auditory interactions. As it is evident from both Figure 3 and Table 3, brain activity was modulated by our factors within large clusters in bilateral STS, including substantial portions of the sulcus along much of its anterior-posterior axis. Other activations included prefrontal clusters in bilateral inferior frontal gyrus, right middle frontal gyrus, left superior medial gyrus, right anterior cingulate cortex, left precuneus, as well as the right cerebellum (Crus 2). Bilaterally, the global peak of *F*-values fell within the anterior portion of the STS clusters. In an exploratory post hoc analysis, coordinates close to

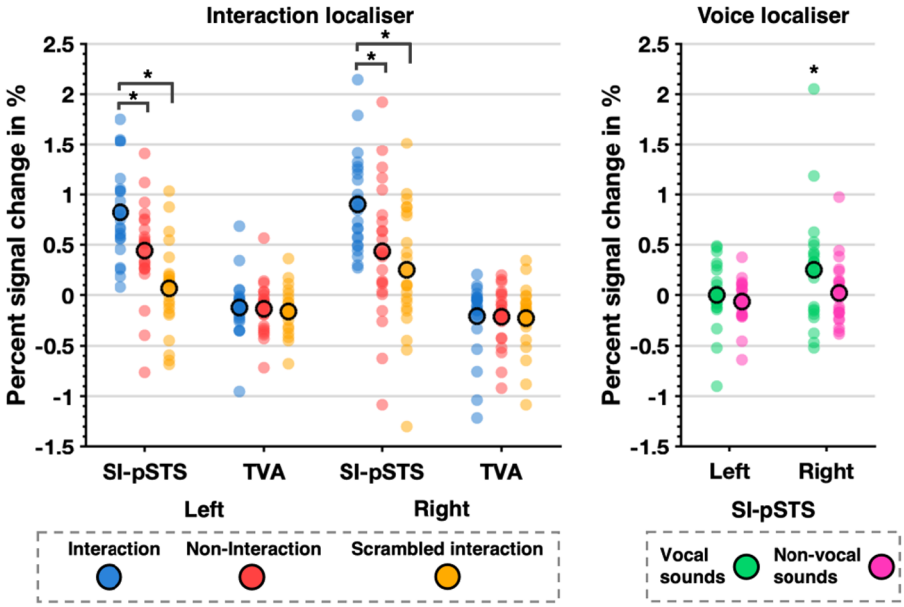

Fig. 2. Illustration of PSC condition means (circle with bold edge) and data distribution (scatter plot) for localiser data of Experiment 1. Left panel: Interaction localiser data for SI-pSTS and TVA. Significant condition differences are marked (*$p < .0125$ corrected). Right panel: Voice localiser data for SI-pSTS only. Above baseline responses are marked (*$p < .05$).

this peak were used to define bilateral anterior STS (aSTS) ROIs for PSC extraction using an iterative LORO process (see 2.1.7 above).

As in both SI-pSTS and TVA regions, PSC analyses revealed a significant main effect of condition in bilateral aSTS, indicating sensitivity to auditory interactions (see Table 4 and Fig. 4 left panel, as well as Supplementary S7, Table S10 for condition means). PSC was smaller for narrations (compared to both conversations (raSTS: $t(22) = 9.69$, $p < .001$, $d_{rm} = 0.22$, laSTS: $t(22) = 7.29$, $p < .001$, $d_{rm} = 0.15$) and scrambled conversations (raSTS: $t(22) = 8.60$, $p < .001$, $d_{rm} = 0.22$, laSTS: $t(22) = 8.85$, $p < .001$, $d_{rm} = 0.13$). Interestingly, right but not left aSTS responded more to conversations compared to scrambled conversations ($t(22) = 3.08$, $p = .005$, $d_{rm} = 0.24$), indicating that the region was not merely driven by the difference of hearing two speakers vs hearing one. Furthermore, there was a main effect of language. Responses were greater for English compared to German stimuli. Indeed, PSC in bilateral aSTS was significantly above baseline for English conversations and English scrambled conversations, and for left aSTS also for English narrations (all $ts(22) \geq 6.06$, all $ps < .001$). Additionally, German narrations led to a significant decrease in activation (raSTS: $t(22) = -3.50$, $p = .002$, laSTS: $t(22) = -1.97$, $p = .06$), whilst for all other conditions, PSC was at baseline. Thus, aSTS showed a similar effect of comprehensibility as SI-pSTS. Finally, these main effects were qualified by a significant Language × Condition interaction. For bilateral aSTS, PSC was greater for conversations compared to narrations for both English and German stimuli (all $ts(22) > 3.03$, $p \leq .006$). Finally, right aSTS showed a significantly greater response to conversations compared to scrambled conversations ($t(22) = 2.78$,

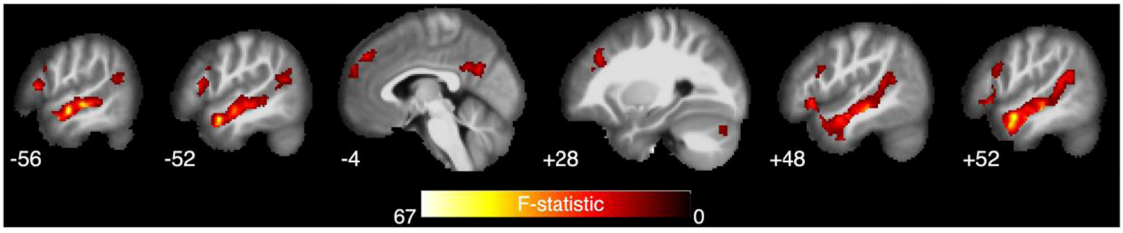

Fig. 3. Sagittal view of whole-brain group analysis Language × Condition interaction F-contrast. Slices in MNI space with x-coordinate shown next to each slice.

Table 3. Significant clusters for whole-brain language × condition interaction F-contrast, cluster-corrected $p_{FWE} < .05$

Side	Cluster	Region label	Cluster size	*F*-value	x	y	z
R	1	Medial Temporal Pole	2552	56.54	54	6	-28
	2	ACC	102	27.89	16	22	26
	3	Cerebellum (Crus 2)	161	26.62	20	-82	-36
	4	Middle Frontal Gyrus	250	21.84	28	36	36
L	1	Middle Temporal Gyrus	1512	67.98	-54	4	-22
	2	Inferior Frontal Gyrus (p. Triangularis)	147	29.40	-56	20	10
	3	Precuneus	457	24.88	-2	-62	26
	4	Superior Medial Gyrus	268	23.30	-4	46	40
	5	Posterior Middle Temporal Gyrus	463	19.611	-52	-62	18
	6	Inferior Frontal Gyrus (p. Triangularis/Opercularis)	121	17.844	-36	18	24

All x, y, and z coordinates in MNI space.

Table 4. ANOVA results of global aSTS peak ROI PSC analyses for main auditory task in Experiment 1

		Main effect of language			Main effect of condition			Interaction effect		
	Side	*F*	*p*	η_p^2	*F*	*p*	η_p^2	*F*	*p*	η_p^2
aSTS	L	161.20	<.001	0.88	55.84	<.001	0.72	54.21	<.001	0.71
	R	74.56	<.001	0.77	78.07	<.001	0.78	75.66	<.001	0.78

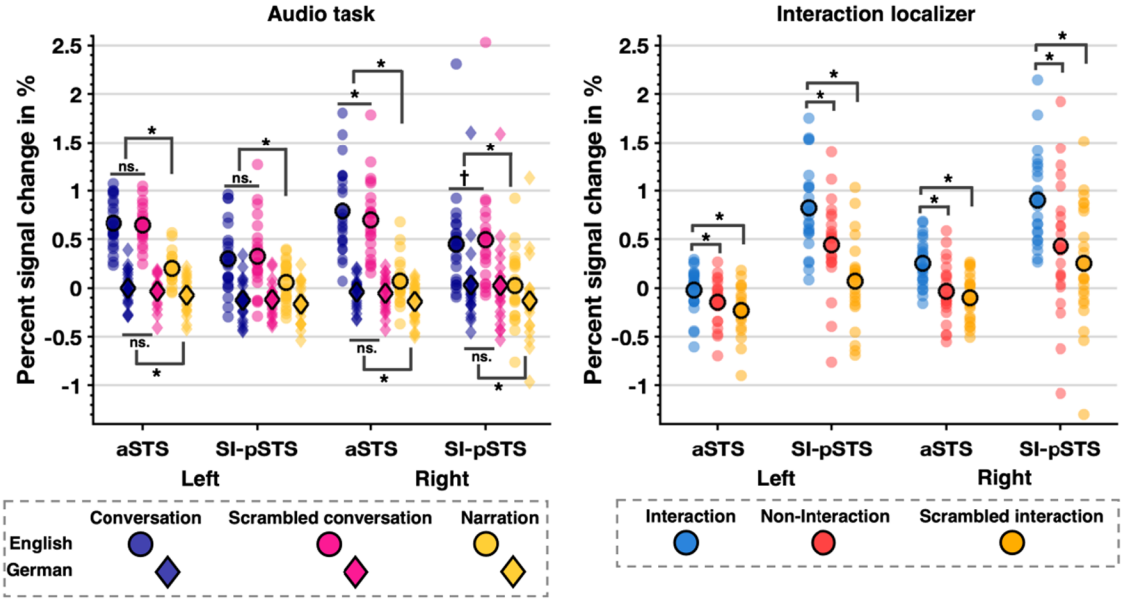

Fig. 4. Percent signal change data displaying condition means (circle with bold edge) and data distribution for each aSTS and SI-pSTS ROI by hemisphere for audio task (left panel) and interaction localiser (right panel). Please note that pSTS data for the audio task is the same as in Figure 1 and for the interaction localiser is the same as Figure 2. Significant post hoc t-test results are marked by an asterisk (*: $p < .0125$ corrected, †: $p < .05$ uncorrected).

$p = .01$) for English stimuli only, suggesting sensitivity to the *coherence* of comprehensible interactions.

3.1.6. What is the evidence for heteromodal social interaction processing in aSTS?

Further exploratory analyses were conducted to investigate whether a region sensitive to auditory social interactions identified using our main auditory experimental task would also be responsive to visual social interactions. As such, this analysis was a reversal of our main experimental hypothesis, i.e., how does the *auditorily* defined aSTS region respond to *visual* interactions, to explore whether social interactions are processed cross-modally within the social brain. Thus, bilateral aSTS was used to extract PSC from the interaction localiser.

Visual interactions only activated right aSTS significantly above baseline ($t(22) = 5.03$, $p < .001$), whereas non-interactions ($t(22) = -0.61$, $p = .55$) and scrambled interactions ($t(22) = -1.96$, $p = .06$) were at or marginally below baseline. For left aSTS, all conditions were at (interactions, $t(22) = -0.44$, $p = .66$) or significantly below (non-interactions $t(22) = -2.95$, $p < .01$; scrambled interactions, $t(22) = -4.56$, $p < .001$) baseline. Paired-sample t-tests comparing interactions with non-interactions as well as interactions with scrambled interactions found significantly greater responses to interactions bilaterally for aSTS (all $ts(22) > 3.71$, all $ps \leq .001$; see Fig. 4 right panel and Supplementary S7, Table S11).

Finally, a comparison of interaction-selectivity in the auditory vs visual domain revealed the reverse pattern to SI-pSTS (see 3.1.3), greater auditory interaction selectivity (raSTS: M = 0.83, SE = 0.09, laSTS: M = 0.54, SE = 0.07) compared to visual interaction selectivity (raSTS: M = 0.29, SE = 0.05, laSTS: M = 0.12, SE = 0.07) (raSTS: $t(22) = -7.07$, $p < .001$, $d_{rm} = 0.36$, laSTS: $t(22) = -5.94$, $p < .001$, $d_{rm} = 0.29$).

3.1.7. Summary

The main aim of this experiment was to investigate whether the interaction-sensitive SI-pSTS region is not only responsive to visual interactions but also responsive to auditory interactions. Both univariate and multivariate ROI analyses suggest that bilateral SI-pSTS displays interaction sensitivity to a broad contrast of *two speakers vs one speaker.* Univariate results also lend tentative support that right SI-pSTS exhibits interaction sensitivity *beyond* the number of speakers. This notion was corroborated more strongly using decoding analyses. Specifically, right SI-pSTS was the only region which could decode conversations vs scrambled conversations, indicating that it also represents information about the meaningfulness of an auditory interaction. Unexpectedly,

there were strong effects of language: bilateral TVA responded above baseline across both languages, whereas SI-pSTS was not driven by German stimuli. In contrast to our predictions, it seems likely that language comprehension was an important factor in some of our results. However, right SI-pSTS could discriminate between German conversations and narrations, suggesting that comprehension is not a pre-requisite when processing interactions at the level of speaker number. Finally, bilateral TVA also exhibited sensitivity to interaction based on the number of speakers, and unexpectedly, like SI-pSTS, left TVA also exhibited weak interaction sensitivity *beyond* the number of speakers, but only in univariate analyses.

Furthermore, this experiment explored (1) whether there was another brain region particularly sensitive to auditory interactions, and (2) whether visual and/or auditory interaction-sensitive regions may exhibit a heteromodal response profile. We used whole-brain group response to "find" a region in bilateral aSTS and explored its auditory and visual interaction sensitivity. Interestingly, right aSTS displayed a response profile characterised by greater sensitivity to auditory than visual interactions, whereas the right SI-pSTS showed the opposite pattern of greater sensitivity to visual compared to auditory interactions. Importantly, both regions showed sensitivity to interactive content *across* modality.

3.2. Experiment 2

This experiment was conducted as a small-scale follow-up study to address the unexpectedly strong language effects observed in Experiment 1. Here, participants were fluent in English but not German; thus, language comprehension might have driven PSC responses. For instance, SI-pSTS bilaterally was either at or below baseline for German conditions. This might be the result of listening to recordings in a language one does not comprehend in the context of a language you understand very well within the same run. A similar native vs unknown language comprehension effect has been found in prior work (Cotosck et al., 2021) using a target word detection task whilst listening to stories. On the other hand, language-specific acoustic differences (Mennen et al., 2012, see also Supplementary S4.2) might have driven some of these differences, particularly in the TVA. Experiment 2 set out to address this question with particular focus on the SI-pSTS by using the same stimulus set but testing German-English bilingual participants.

Please refer to Table 2 for ANOVA statistics, Supplementary S7, Table S7 for condition means, and Figure 5 for an illustration of the PSC results of Experiment 2.

3.2.1. Does SI-pSTS respond differently across languages when both are understood?

When both languages were comprehensible to participants, there was no main effect of language. Responses in bilateral SI-pSTS were similar for English and German stimuli. Indeed, PSC in bilateral SI-pSTS was significantly above baseline for both English and German conversations and scrambled conversations (all $t(11) \geq 2.36$, all ps $< .04$). Thus, regardless of language, only narrations *did not activate* the SI-pSTS. Replicating Experiment 1, there *was* a main effect of condition, driven by a difference between hearing two speakers vs hearing only one. PSC was smaller for narrations compared to both conversations (rSI-pSTS: $t(11) = 4.41$, $p < .001$, $d_{rm} = 1.04$, lSI-pSTS: $t(11) = 3.66$, $p < .001$, $d_{rm} = 0.25$) and scrambled conversations (rSI-pSTS: $t(11) = 6.35$, $p < .001$, $d_{rm} = 0.80$, lSI-pSTS: $t(11) = 3.17$, $p < .001$, $d_{rm} = 0.27$) but not different between conversations and scrambled conversations. These main effects were qualified by a significant Language × Condition interaction in the right SI-pSTS only. PSC was greater for conversations compared to narrations for both English ($t(11) = 5.23$, $p < .001$, $d_{rm} = 0.41$) and German stimuli ($t(11) = 3.31$, $p = .007$, $d_{rm} = 0.26$). In contrast, response to *German* ($t(11) = -2.38$, $p = .04$, $d_{rm} = 0.18$) but not English ($t(11) = -0.59$, $p = .57$) scrambled conversations was significantly greater than conversations, albeit at an uncorrected level only. This suggests some sensitivity to the *coherence* of interactions in the participants' *native* language.

3.2.2. How does the TVA response compare in this case?

As in Experiment 1, all conditions strongly activated bilateral TVA above baseline (all $t(11) \geq 7.69$, all ps $< .001$). Bilaterally, although marginally in the left hemisphere, a significant effect of language remained even when participants understood both languages. Similarly, PSC was greater for English compared to German conditions. Further replicating Experiment 1, there was also a main effect of condition. PSC was smaller for narrations compared to both conversations (rTVA: $t(11) = 4.70$, $p < .001$, $d_{rm} = 0.57$, lTVA: $t(11) = 4.82$, $p < .001$, $d_{rm} = 0.52$) and scrambled conversations (rTVA: $t(11) = 4.53$, $p < .001$, $d_{rm} = -0.18$, lTVA: $t(11) = 5.89$, $p < .001$, $d_{rm} = 0.56$). Finally,

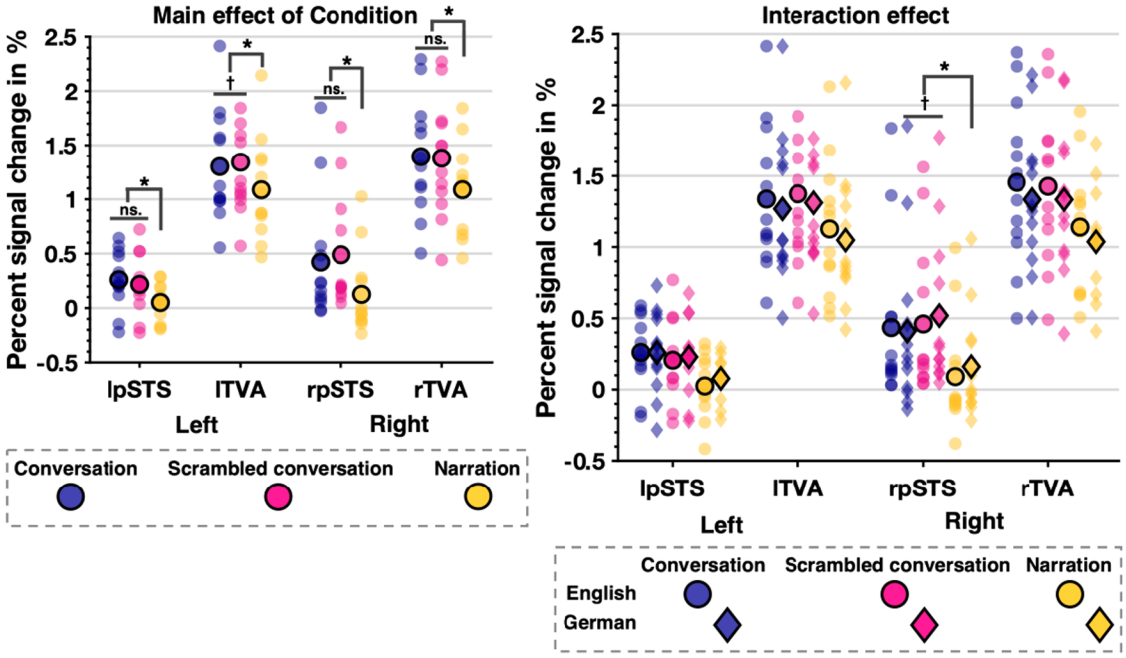

Fig. 5. Experiment 2 PSC data illustrating the main effect of Condition (*: $p < .017$ corrected) and Language × Condition interaction effect (*: $p < .0125$ corrected, †: $p < .05$ uncorrected).

sustaining the unexpected finding from Experiment 1 of lTVA sensitivity to not only number of speakers but also coherence of conversation, responses to scrambled conversations were slightly but significantly greater than for intact conversations in left TVA only, albeit at an uncorrected level ($t(11) = -2.61$, $p < .03$, $d_{rm} = 0.56$). No significant interaction effect emerged.

3.2.3. Summary

Experiment 2 set out to test whether language comprehension may have driven some of the effects seen in Experiment 1. Testing bilingual participants revealed that when participants comprehended both languages, language effects disappeared in bilateral SI-pSTS whilst condition effects remained. In line with the results from Experiment 1, SI-pSTS responded more strongly to conversations compared to narrations. Crucially, right SI-pSTS was sensitive to the difference between German conversations and German scrambled conversations. This replicates Experiment 1 which found this difference for English stimuli. Taken together, these findings suggest that right SI-pSTS is sensitive to meaningful auditory interactions, at least in participants' native language. For bilateral TVA, language effects remained relatively stable with the participants' non-native language resulting in greater activation. This might point to more effortful processing of the participants' second language (Hasegawa et al., 2002). Like Experiment 1, left TVA showed a greater response to scrambled compared to intact conversations. Thus, across both experiments, left TVA was less responsive to meaningful conversations.

4. DISCUSSION

Whilst everyday social interactions provide a rich multi-sensory experience, neuroimaging studies of social interaction perception have predominantly focussed on the abundance of visual cues they provide. Conversely, not much is known about auditory interaction perception in the social brain. Combining univariate and multivariate analyses, we confirmed our key prediction that visual SI-pSTS exhibits heteromodal processing of social interactions. In contrast, although voice-selective TVA shows an unexpectedly similar response profile to auditory interactions, it is clearly a unimodal region. Specifically, both bilateral SI-pSTS and TVA were sensitive to interactive information in a broad contrast between two-speaker conversations and one-speaker narrations, in line with similar recent work focussed on language processing (Olson et al., 2023). More importantly, right SI-pSTS and left TVA showed some weak sensitivity to auditory interactions when coherence of comprehensible (native language) conversations was manipulated. However,

multivariate decoding analyses only corroborated this finding for right SI-pSTS, suggesting that the information represented in SI-pSTS voxel patterns is qualitatively different from that represented in TVA.

These findings are in line with previous results that put the broader pSTS region at the heart of heteromodal or even multimodal integrative processing of social information (Kreifelts et al., 2009; Lahnakoski et al., 2012; Robins et al., 2009; Watson et al., 2014; Wright et al., 2003). Indeed, regions along the pSTS show tuning to a variety of both visual and auditory social stimuli (Deen et al, 2015) and the pSTS is widely referred to as the "hub" of the social brain because of its involvement across varied social tasks (e.g., Yang et al., 2015). However, much of the prior literature has investigated heteromodal processing in the context of social signals from individuals, making this study's focus on the perception of social interactions relatively unique. Importantly, our data make clear that auditory interaction sensitivity in right SI-pSTS reflects more than tuning to voice stimuli in general. Indeed, the SI-pSTS region shows negligible sensitivity to vocal sounds in response to the voice localiser (see Fig. 2). At the same time, responses in SI-pSTS to visual and auditory interactive stimuli were not fully equivalent. Right SI-pSTS interaction selectivity for visual stimuli was about 50% greater than for auditory stimuli. It could be that the nature of the interaction-region localiser might, in part, account for this. Essentially, we tested how SI-pSTS voxels sensitive to interaction information conveyed by human body- and biological motion cues responded to interaction cues conveyed by voice. Body and voice cues, however, are less strongly associated with each other compared to face and voice cues. Had we used stimuli that relied on facial cues of interaction in our localiser, we might have found a greater degree of correspondence between visual and auditory SI-pSTS response profiles. Indeed, heteromodal responses in the broader STS region to voices have previously been shown in conjunction with face stimuli (Deen et al., 2015, 2020; Watson et al., 2014), though not in the context of social interactions. As such, our approach is a strong test of whether SI-pSTS shows sensitivity to interactive information across modalities.

Nonetheless, the response profile of SI-pSTS to auditory interactions was more nuanced and less definitive than originally predicted. As a broad test of auditory interaction sensitivity, we expected that the mere presence of two speakers taking conversational turns would drive SI-pSTS activation regardless of language comprehension. However, testing monolinguals (Experiment 1) and bilinguals (Experiment 2) revealed that comprehension mattered. The SI-pSTS was only driven by the two-speaker conditions in monolingual participants' native language, whereas this language effect was abolished in bilingual speakers. Nevertheless, in monolingual English speakers, MVPA analysis of voxel patterns of SI-pSTS revealed that the two-speaker conditions could be discriminated from narrations even in the German condition. While this is perhaps not surprising, it does suggest that language comprehension is not a pre-requisite for representation of information; i.e., number of speakers, that clearly distinguishes auditory interactions from non-interactions within the SI-pSTS. When this distinction is less obvious however, interactive cues might well be derived through language comprehension. Indeed, monolingual participants listening to intact and scrambled conversations presented in their native language would be able to differentiate them based on detecting conversational coherence and presence of overall gist, whereas without comprehension, they would have to rely on subtle prosodic cues. We found that *right* SI-pSTS only distinguished between the two-speaker conditions when participants could access their meaning. Thus, language comprehension clearly mattered when extracting cues to conversational coherence. Notably, SI-pSTS lies in proximity to a bilateral brain network (including STG, STS, MTG, and left IFG, see Bookheimer, 2002; Mar, 2011; Vigneau et al., 2011; Walenski, 2019; Yang, 2014) implicated in higher-level discourse comprehension processes such as evaluation of global coherence, pragmatic interpretations, and text integration at the gist-level. However, right SI-pSTS's overlap with this network is unclear. Although domain-general language processes designed to detect coherence could contribute to the response difference between intact and scrambled interactions, they cannot explain the drop in response to coherent narrations. Instead, right SI-pSTS might receive and integrate coherence or gist information from nearby language regions as a cue to evaluate interactiveness. Furthermore, although we found no support that SI-pSTS was sensitive to conversational flow in an unknown language, strong between-language effects may have overshadowed the potential to detect more subtle effects of prosody. Future studies investigating the role of SI-pSTS independent of language comprehension are needed to firmly establish its role when cues to interaction are harder to extract; e.g., using low-pass filtered muffled stimuli containing only prosodic but no lexico-semantic cues to interaction.

Unexpectedly, voice-selective TVA, especially in the left hemisphere, exhibited a similar response profile to right

SI-pSTS. Firstly, a greater response to and decoding of conversations vs narrations across languages, and secondly, a slightly greater response to scrambled conversations compared to conversations. Importantly, however, classification analyses in left TVA could not discriminate scrambled from intact conversations. Thus, it is unclear how distinctly (left) TVA responses could be attributed to pure interaction sensitivity. Indeed, it is possible that response difference between the one-speaker narration condition and the two-speaker conditions (conversations, scrambled) in both TVA and SI-pSTS could partially be driven by these regions adapting to vocal quality or speaker identity in the narration condition. However, pSTS responsivity to voices is thought to reflect higher-level social process because individuals with lesions in pSTS are still able to discriminate between and recognise individual voices (Jiahui et al, 2017). In addition, we think it unlikely that adaptation can fully explain our effects in the SI-pSTS because we do not see strong differences between these conditions when participants do not understand what is being said, making simple adaptation effects unlikely. However, as TVA is known to be involved in the spectro-temporal analysis of human vocal sounds (and speech) (Agus et al., 2017; Belin et al., 2000, 2002), adaptation to vocal quality or identity may partially explain our effects in this region. Similarly, previous research has found a significant positive association between mean F0 of speech and TVA activation (Wiethoff et al., 2008). More generally, TVA is part of the STS/STG engaged in phonological language processing (Vigneau et al., 2006, 2011). Notably, due to the complexity and diversity of the stimuli used in this study, acoustic features of the stimulus set could not be as tightly controlled as we might have liked, which may drive some between-condition differences in TVA activation. Specifically, F0 was greater for conversations compared to narrations, and greater for English compared to German stimuli (see Supplementary S4.2). Thus, greater TVA activation to conversations compared to narrations, and English compared to German observed across both experiments could be at least partially explained by their corresponding differences in F0. However, as intact and scrambled conversations were matched on F0, these differences cannot explain higher left TVA response to scrambled conversations. Importantly, although scrambled and intact conversations contained identical sentences, they were not exact phonological equivalents as our scrambling process allowed a change in speaker. Thus, differential responses in left TVA might reflect sensitivity to speaker-dependent variations in phonation between conditions. Importantly, the above explanations would not apply to responses in right SI-pSTS, which is not known to be involved in phonological processes (Vigneau et al., 2006, 2011) or modulated by F0 (Wiethoff et al., 2008). In addition, between-language differences were no longer present in the SI-pSTS when participants understood both languages (Experiment 2). Altogether, our findings might motivate future work into auditory interaction perception in the brain to corroborate evidence on the unique role of right SI-pSTS using more tightly controlled stimuli that could additionally clarify the role of TVA.

Finally, we confirmed our prediction that TPJ, a social cognition region selectively engaged by mentalising processes (Saxe et al., 2009; Van Overwalle & Baetens, 2009), is neither driven by our stimuli nor sensitive to differences between conditions (see Supplementary S5 for details). TPJ can be engaged by auditory stimuli when participants are engaged in a mentalising task (Kandylaki et al., 2015; Saxe et al., 2009). However, while some prior work has suggested that TPJ is involved in processing social interactions (e.g., Canessa et al., 2012; Centelles et al., 2011), it is likely that the region is involved only when mentalising is required (Masson & Isik, 2021; Walbrin et al., 2018), which was not the case in our task. Indeed, we took care to select stimuli that did not imply nor require mentalising. Similarly, our whole-brain analysis suggests that occipital and temporal areas outside the STS, including EBA, are not involved in auditory interaction perception. This makes sense, as "early" social perception regions like EBA are not usually considered to be heteromodal or, indeed, particularly responsive to auditory stimuli (Beer et al., 2013). Instead, our exploratory whole-brain analysis identified an area in the right anterior STS, close to the temporal poles, which showed sensitivity to interactive information both through greater activation to two speakers than to one and through a higher response to comprehensible conversations compared to scrambled conversations (Experiment 1). While this region's responsiveness to auditory interactions (and perhaps to interactive information in general) needs to be replicated (though see Olson et al, 2023), this is a particularly intriguing finding in light of prior work that highlights the dorsolateral anterior temporal lobe (ATL) as a region that may be involved in social semantics (Arioli et al, 2021; Lin et al., 2018; Zhang et al., 2021; but see also Balgova et al., 2022; Binney et al., 2016; Binney & Ramsey, 2020 for a broader perspective regarding the role of the ATL in social cognition). Furthermore, in the context of speech comprehension, it has been suggested that the meaning of speech is processed in bilateral anterior temporal cortex, including aSTS (e.g., Mitchell et al., 2003; Scott et al., 2000, 2009, for a review see Price, 2012). Thus, right aSTS might be

particularly involved in the semantic analysis of auditory interactions, given that its activation was not driven by comprehensible narrative stimuli.

5. CONCLUSION

Our results present initial evidence that SI-pSTS, initially defined visually, is also sensitive to interactive cues presented in the auditory domain. In other words, this region is characterised by a heteromodal response profile that appears to be particularly sensitive to social interactions. Future research is needed to both replicate these novel findings and look beyond the number of speakers and interaction coherence to investigate whether SI-pSTS codes for other auditory cues involved in understanding social interactions, for instance, subtle prosodic cues or interactional turn duration. In addition, our results may motivate future work to determine how SI-pSTS integrates multimodal audio-visual social interaction information to inform our understanding of highly naturalistic everyday life social interactions. Finally, this initial work also prompts further research into the role of aSTS regions in social interaction perception more broadly and in conversation/language-based interactions specifically.

DATA AND CODE AVAILABILITY

Stimulus examples as well as subject-level summary data that support the findings of this study are openly available at the project's OSF page at https://osf.io/4xedj/.

AUTHOR CONTRIBUTIONS

Julia Landsiedel: Conceptualisation, Methodology, Investigation, Formal analysis, Project administration, Visualisation, Writing—Original Draft, and Writing—Review & Editing; Kami Koldewyn: Conceptualisation, Funding acquisition, Methodology, Investigation, Supervision, Project administration, Writing—Original Draft, and Writing—Review & Editing.

FUNDING

This work has received funding from the European Research Council under the European Union's Horizon 2020 research and innovation programme (ERC-2016-STG-716974: Becoming Social).

DECLARATION OF COMPETING INTEREST

None.

ACKNOWLEDGEMENTS

The authors are grateful to Corinne Voigt-Hill for her help with data collection as well as to members of the Social Neuroscience and Cognition group and Bangor Imaging Group at Bangor University for general feedback, helpful discussion, and suggestions throughout the research process.

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available with the online version here: https://doi.org/10.1162/imag_a_00003.

REFERENCES

Abassi, E., & Papeo, L. (2020). The representation of two-body shapes in the human visual cortex. *J Neurosci*, *40*(4), 852–863. https://doi.org/10.1523/JNEUROSCI.1378-19.2019

Abassi, E., & Papeo, L. (2021). A new behavioral and neural marker of social vision. *bioRxiv*, 2021.2006.2009.447707. https://doi.org/10.1101/2021.06.09.447707

Agus, T. R., Paquette, S., Suied, C., Pressnitzer, D., & Belin, P. (2017). Voice selectivity in the temporal voice area despite matched low-level acoustic cues. *Sci Rep*, *7*(1), 11526. https://doi.org/10.1038/s41598-017-11684-1

Arioli, M., Gianelli, C., & Canessa, N. (2021). Neural representation of social concepts: A coordinate-based meta-analysis of fMRI studies. *Brain Imaging Behav*, *15*(4), 1912–1921. https://doi.org/10.1007/s11682-020-00384-6

Awwad Shiekh Hasan, B., Valdes-Sosa, M., Gross, J., & Belin, P. (2016). "Hearing faces and seeing voices": Amodal coding of person identity in the human brain. *Sci Rep*, *6*, 37494. https://doi.org/10.1038/srep37494

Balgova, E., Diveica, V., Walbrin, J., & Binney, R. J. (2022). The role of the ventrolateral anterior temporal lobes in social cognition. *Hum Brain Mapp*, *43*(15), 4589–4608. https://doi.org/10.1002/hbm.25976

Beer, A., Plank, T., Meyer, G., & Greenlee, M. (2013). Combined diffusion-weighted and functional magnetic resonance imaging reveals a temporal-occipital network involved in auditory-visual object processing. *Front Integr Neurosci*, *7*, 5. https://doi.org/10.3389/fnint.2013.00005

Belin, P., Zatorre, R. J., & Ahad, P. (2002). Human temporal-lobe response to vocal sounds. *Brain Res Cogn Brain Res*, *13*(1), 17–26. https://doi.org/10.1016/s0926-6410(01)00084-2

Belin, P., Zatorre, R. J., Lafaille, P., Ahad, P., & Pike, B. (2000). Voice-selective areas in human auditory cortex. *Nature*, *403*(6767), 309–312. https://doi.org/10.1038/35002078

Bellot, E., Abassi, E., & Papeo, L. (2021). Moving toward versus away from another: How body motion direction changes the representation of bodies and actions in the visual cortex. *Cereb Cortex*, *31*(5), 2670–2685. https://doi.org/10.1093/cercor/bhaa382

Bidet-Caulet, A., Voisin, J., Bertrand, O., & Fonlupt, P. (2005). Listening to a walking human activates the temporal

biological motion area. *Neuroimage*, *28*(1), 132–139. https://doi.org/10.1016/j.neuroimage.2005.06.018
Binney, R. J., Hoffman, P., & Lambon Ralph, M. A. (2016). Mapping the multiple graded contributions of the anterior temporal lobe representational hub to abstract and social concepts: Evidence from distortion-corrected fMRI. *Cereb Cortex*, *26*(11), 4227–4241. https://doi.org/10.1093/cercor/bhw260
Binney, R. J., & Ramsey, R. (2020). Social semantics: The role of conceptual knowledge and cognitive control in a neurobiological model of the social brain. *Neurosci Biobehav Rev*, *112*, 28–38. https://doi.org/10.1016/j.neubiorev.2020.01.030
Bookheimer, S. (2002). Functional MRI of language: New approaches to understanding the cortical organization of semantic processing. *Annu Rev Neurosci*, *25*, 151–188. https://doi.org/10.1146/annurev.neuro.25.112701.142946
Brainard, D. H. (1997). The psychophysics toolbox. *Spat Vis*, *10*(4), 433–436. https://doi.org/10.1163/156856897X00357
Brett, M., Anton, J.-L., Valabregue, R., & Poline, J.-B. (2002). Region of interest analysis using an SPM toolbox 8th International Conference on Functional Mapping of the Human Brain, Sendai, Japan. https://matthew.dynevor.org/research/abstracts/marsbar/marsbar_abstract.pdf
Canessa, N., Alemanno, F., Riva, F., Zani, A., Proverbio, A. M., Mannara, N., Perani, D., & Cappa, S. F. (2012). The neural bases of social intention understanding: The role of interaction goals. *PLoS One*, *7*(7), e42347. https://doi.org/10.1371/journal.pone.0042347
Cañigueral, R., & Hamilton, A. F. d. C. (2019). The role of eye gaze during natural social interactions in typical and autistic people. *Front Psychol*, *10*. https://doi.org/10.3389/fpsyg.2019.00560
Centelles, L., Assaiante, C., Nazarian, B., Anton, J. L., & Schmitz, C. (2011). Recruitment of both the mirror and the mentalizing networks when observing social interactions depicted by point-lights: A neuroimaging study. *PLoS One*, *6*(1), e15749. https://doi.org/10.1371/journal.pone.0015749
Charest, I., Pernet, C., Latinus, M., Crabbe, F., & Belin, P. (2013). Cerebral processing of voice gender studied using a continuous carryover FMRI design. *Cereb Cortex*, *23*(4), 958–966. https://doi.org/10.1093/cercor/bhs090
Cotosck, K. R., Meltzer, J. A., Nucci, M. P., Lukasova, K., Mansur, L. L., & Amaro, E. (2021). Engagement of language and domain general networks during word monitoring in a native and unknown language. *Brain Sci*, *11*(8), 1063. https://doi.org/10.3390/brainsci11081063
de Gelder, B., de Borst, A. W., & Watson, R. (2015). The perception of emotion in body expressions. *Wiley Interdiscip Rev Cogn Sci*, *6*(2), 149–158. https://doi.org/10.1002/wcs.1335
Deen, B., Koldewyn, K., Kanwisher, N., & Saxe, R. (2015). Functional organization of social perception and cognition in the superior temporal sulcus. *Cereb Cortex*, *25*(11), 4596–4609. https://doi.org/10.1093/cercor/bhv111
Deen, B., Saxe, R., & Kanwisher, N. (2020). Processing communicative facial and vocal cues in the superior temporal sulcus. *Neuroimage*, *221*, 117191. https://doi.org/10.1016/j.neuroimage.2020.117191
Demenescu, L. R., Mathiak, K. A., & Mathiak, K. (2014). Age- and gender-related variations of emotion recognition in pseudowords and faces. *Exp Aging Res*, *40*(2), 187–207. https://doi.org/10.1080/0361073X.2014.882210
Ding, X., Gao, Z., & Shen, M. (2017). Two equals one: Two human actions during social interaction are grouped as one unit in working memory. *Psychol Sci*, *28*(9), 1311–1320. https://doi.org/10.1177/0956797617707318
Downing, P. E., Jiang, Y., Shuman, M., & Kanwisher, N. (2001). A cortical area selective for visual processing of the human body. *Science*, *293*(5539), 2470–2473. https://doi.org/10.1126/science.1063414
Enrici, I., Adenzato, M., Cappa, S., Bara, B. G., & Tettamanti, M. (2011). Intention processing in communication: A common brain network for language and gestures. *J Cogn Neurosci*, *23*(9), 2415–2431. https://doi.org/10.1162/jocn.2010.21594
Ethofer, T., Bretscher, J., Gschwind, M., Kreifelts, B., Wildgruber, D., & Vuilleumier, P. (2012). Emotional voice areas: Anatomic location, functional properties, and structural connections revealed by combined fMRI/DTI. *Cereb Cortex*, *22*(1), 191–200. https://doi.org/10.1093/cercor/bhr113
Ethofer, T., Van De Ville, D., Scherer, K., & Vuilleumier, P. (2009). Decoding of emotional information in voice-sensitive cortices. *Curr Biol*, *19*(12), 1028–1033. https://doi.org/10.1016/j.cub.2009.04.054
Fedorenko, E., Hsieh, P. J., Nieto-Castanon, A., Whitfield-Gabrieli, S., & Kanwisher, N. (2010). New method for fMRI investigations of language: Defining ROIs functionally in individual subjects. *J Neurophysiol*, *104*(2), 1177–1194. https://doi.org/10.1152/jn.00032.2010
Hasegawa, M., Carpenter, P. A., & Just, M. A. (2002). An fMRI study of bilingual sentence comprehension and workload. *Neuroimage*, *15*(3), 647–660. https://doi.org/10.1006/nimg.2001.1001
Hellbernd, N., & Sammler, D. (2016). Prosody conveys speaker's intentions: Acoustic cues for speech act perception. *J Mem Lang*, *88*, 70–86. https://doi.org/10.1016/j.jml.2016.01.001
Henetz, T. (2017). Don't hesitate! The length of inter-turn gaps influences observers' interactional attributions. Stanford University. http://purl.stanford.edu/zk615rk5483
Isik, L., Koldewyn, K., Beeler, D., & Kanwisher, N. (2017). Perceiving social interactions in the posterior superior temporal sulcus. *Proc Natl Acad Sci U S A*, *114*(43), E9145–E9152. https://doi.org/10.1073/pnas.1714471114
Jacoby, N., Bruneau, E., Koster-Hale, J., & Saxe, R. (2016). Localizing pain matrix and theory of mind networks with both verbal and non-verbal stimuli. *Neuroimage*, *126*, 39–48. https://doi.org/10.1016/j.neuroimage.2015.11.025
Jiahui, G., Garrido, L., Liu, R. R., Susilo, T., Barton, J. J. S., & Duchaine, B. (2017). Normal voice processing after posterior superior temporal sulcus lesion. *Neuropsychologia*, *105*, 215–222. https://doi.org/10.1016/j.neuropsychologia.2017.03.008
Julian, J. B., Fedorenko, E., Webster, J., & Kanwisher, N. (2012). An algorithmic method for functionally defining regions of interest in the ventral visual pathway. *Neuroimage*, *60*(4), 2357–2364. https://doi.org/10.1016/j.neuroimage.2012.02.055
Kandylaki, K. D., Nagels, A., Tune, S., Wiese, R., Bornkessel-Schlesewsky, I., & Kircher, T. (2015).

Processing of false belief passages during natural story comprehension: An fMRI study. *Hum Brain Mapp*, *36*(11), 4231–4246. https://doi.org/10.1002/hbm.22907
Kanwisher, N., & Yovel, G. (2006). The fusiform face area: A cortical region specialized for the perception of faces. *Philos Trans R Soc Lond B Biol Sci*, *361*(1476), 2109–2128. https://doi.org/10.1098/rstb.2006.1934
Kleiner, M., Brainard, D. H., & Pelli, D. (2007). What's new in Psychtoolbox-3? *Perception*, *36*, 1–6. https://doi.org/10.1177/03010066070360S101
Kreifelts, B., Ethofer, T., Shiozawa, T., Grodd, W., & Wildgruber, D. (2009). Cerebral representation of non-verbal emotional perception: fMRI reveals audiovisual integration area between voice- and face-sensitive regions in the superior temporal sulcus. *Neuropsychologia*, *47*(14), 3059–3066. https://doi.org/10.1016/j.neuropsychologia.2009.07.001
Lahnakoski, J. M., Glerean, E., Slmi, J., Jääskeläinen, L. P., Sams, M., Hari, R., & Nummemaa, L. (2012). Naturalistic fMRI mapping reveals superior temporal sulcus as a hub for the distributed brain network for social perception. *Front Hum Neurosci*, *13*(6), 233. https://doi.org/10.3389/fnhum.2012.00233
Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Front Psychol*, *4*, 863. https://doi.org/10.3389/fpsyg.2013.00863
Landsiedel, J., Daughters, K., Downing, P. E., & Koldewyn, K. (2022). The role of motion in the neural representation of social interactions in the posterior temporal cortex. *Neuroimage*, *262*, 119533. https://doi.org/10.1016/j.neuroimage.2022.119533
Latinus, M., McAleer, P., Bestelmeyer, P. E., & Belin, P. (2013). Norm-based coding of voice identity in human auditory cortex. *Curr Biol*, *23*(12), 1075–1080. https://doi.org/10.1016/j.cub.2013.04.055
Lin, N., Wang, X., Xu, Y., Wang, X., Hua, H., Zhao, Y., & Li, X. (2018). Fine subdivisions of the semantic network supporting social and sensory-motor semantic processing. *Cereb Cortex*, *28*(8), 2699–2710. https://doi.org/10.1093/cercor/bhx148
Mar, R. A. (2011). The Neural Bases of Social Cognition and Story Comprehension. *Annual Review of Psychology, 62*(1), 103–134. https://doi.org/10.1146/annurev-psych-120709-145406
Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *J Neurosci Methods*, *164*(1), 177–190. https://doi.org/10.1016/j.jneumeth.2007.03.024
Masson, H. L., & Isik, L. (2021). Functional selectivity for social interaction perception in the human superior temporal sulcus during natural viewing. *Neuroimage*, *245*, 118741. https://doi.org/10.1016/j.neuroimage.2021.118741
Mazaika, P., Whitfield, S., & Cooper, J. C. (2005). Detection and Repair of Transient Artifacts in fMRI Data. 11th Annual Meeting of the Organization for Human Brain Mapping, Toronto, Canada.
McFarquhar, M., McKie, S., Emsley, R., Suckling, J., Elliott, R., & Williams, S. (2016). Multivariate and repeated measures (MRM): A new toolbox for dependent and multimodal group-level neuroimaging data. *Neuroimage*, *132*, 373–389. https://doi.org/10.1016/j.neuroimage.2016.02.053
Mennen, I., Schaeffler, F., & Docherty, G. (2012). Cross-language differences in fundamental frequency range: A comparison of English and German. *J Acoust Soc Am*, *131*(3), 2249–2260. https://doi.org/10.1121/1.3681950
Mitchell, R. L. C., Elliott, R., Barry, M., Cruttenden, A., & Woodruff, P. W. R. (2003). The neural response to emotional prosody, as revealed by functional magnetic resonance imaging. *Neuropsychologia*, *41*(10), 1410–1421. https://doi.org/10.1016/s0028-3932(03)00017-4
Oldfield R. C. (1971). The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia, 9*(1), 97–113. https://doi.org/10.1016/0028-3932(71)90067-4
Olson, H., Chen, E., Lydic, K., & Saxe, R. (2023). Left-hemisphere cortical language regions respond equally to dialogue and monologue. *bioRxiv*. https://doi.org/10.1101/2023.01.30.526344
Oosterhof, N. N., Connolly, A. C., & Haxby, J. V. (2016). CoSMoMVPA: Multi-modal multivariate pattern analysis of neuroimaging data in Matlab/GNU Octave. *Front Neuroinform*, *10*, 27. https://doi.org/10.3389/fninf.2016.00027
Papeo, L. (2020). Twos in human visual perception. *Cortex*, *132*, 473–478. https://doi.org/10.1016/j.cortex.2020.06.005
Papeo, L., & Abassi, E. (2019). Seeing social events: The visual specialization for dyadic human-human interactions. *J Exp Psychol Hum Percept Perform*, *45*(7), 877–888. https://doi.org/10.1037/xhp0000646
Papeo, L., Goupil, N., & Soto-Faraco, S. (2019). Visual search for people among people. *Psychol Sci*, *30*(10), 1483–1496. https://doi.org/10.1177/0956797619867295
Papeo, L., Stein, T., & Soto-Faraco, S. (2017). The two-body inversion effect. *Psychol Sci*, *28*(3), 369–379. https://doi.org/10.1177/0956797616685769
Pernet, C. R., McAleer, P., Latinus, M., Gorgolewski, K. J., Charest, I., Bestelmeyer, P. E., Watson, R. H., Fleming, D., Crabbe, F., Valdes-Sosa, M., & Belin, P. (2015). The human voice areas: Spatial organization and inter-individual variability in temporal and extra-temporal cortices. *Neuroimage*, *119*, 164–174. https://doi.org/10.1016/j.neuroimage.2015.06.050
Pijper, J. R. d., & Sanderman, A. A. (1994). On the perceptual strength of prosodic boundaries and its relation to suprasegmental cues. *J Acoust Soc Am*, *96*(4), 2037–2047. https://doi.org/10.1121/1.410145
Ponsot, E., Burred, J. J., Belin, P., & Aucouturier, J. J. (2018). Cracking the social code of speech prosody using reverse correlation. *Proc Natl Acad Sci U S A*, *115*(15), 3972–3977. https://doi.org/10.1073/pnas.1716090115
Price, C. J. (2012). A review and synthesis of the first 20 years of PET and fMRI studies of heard speech, spoken language and reading. *Neuroimage*, *62*(2), 816–847. https://doi.org/10.1016/j.neuroimage.2012.04.062
Roberts, F., & Francis, A. L. (2013). Identifying a temporal threshold of tolerance for silent gaps after requests. *J Acoust Soc Am*, *133*(6), EL471–EL477. https://doi.org/10.1121/1.4802900
Roberts, F., Francis, A. L., & Morgan, M. (2006). The interaction of inter-turn silence with prosodic cues in listener perceptions of "trouble" in conversation. *Speech Commun*, *48*(9), 1079–1093. https://doi.org/https://doi.org/10.1016/j.specom.2006.02.001

Robins, D. L., Hunyadi, E., & Schultz, R. T. (2009). Superior temporal activation in response to dynamic audio-visual emotional cues. *Brain Cogn*, *69*(2), 269–278. https://doi.org/10.1016/j.bandc.2008.08.007

Saarela, M. V., & Hari, R. (2008). Listening to humans walking together activates the social brain circuitry. *Soc Neurosci*, *3*(3–4), 401–409. https://doi.org/10.1080/17470910801897633

Saxe, R. R., Whitfield-Gabrieli, S., Scholz, J., & Pelphrey, K. A. (2009). Brain regions for perceiving and reasoning about other people in school-aged children. *Child Dev*, *80*(4), 1197–1209. https://doi.org/10.1111/j.1467-8624.2009.01325.x

Schirmer, A., & Adolphs, R. (2017). Emotion perception from face, voice, and touch: Comparisons and convergence. *Trends Cogn Sci*, *21*(3), 216–228. https://doi.org/10.1016/j.tics.2017.01.001

Scott, S. K., Blank, C. C., Rosen, S., & Wise, R. J. (2000). Identification of a pathway for intelligible speech in the left temporal lobe. *Brain, 123 Pt 12*, 2400–2406. https://doi.org/10.1093/brain/123.12.2400

Scott, S. K., Rosen, S., Beaman, C. P., Davis, J. P., & Wise, R. J. (2009). The neural processing of masked speech: Evidence for different mechanisms in the left and right temporal lobes. *J Acoust Soc Am*, *125*(3), 1737–1743. https://doi.org/10.1121/1.3050255

Smith, S. M., & Nichols, T. E. (2009). Threshold-free cluster enhancement: Addressing problems of smoothing, threshold dependence and localisation in cluster inference. *Neuroimage*, *44*(1), 83–98. https://doi.org/10.1016/j.neuroimage.2008.03.061

Stevenage, S. V., Hugill, A., & Lewis, H. G. (2012). Integrating voice recognition into models of person perception. *J Cognit Psychol*, *24*, 409–419. https://doi.org/10.1080/20445911.2011.642859

Todorov, A., Olivola, C. Y., Dotsch, R., & Mende-Siedlecki, P. (2015). Social attributions from faces: Determinants, consequences, accuracy, and functional significance. *Annu Rev Psychol*, *66*, 519–545. https://doi.org/10.1146/annurev-psych-113011-143831

Van Overwalle, F., & Baetens, K. (2009). Understanding others' actions and goals by mirror and mentalizing systems: A meta-analysis. *Neuroimage*, *48*(3), 564–584. https://doi.org/10.1016/j.neuroimage.2009.06.009

Vestner, T., Tipper, S. P., Hartley, T., Over, H., & Rueschemeyer, S. A. (2019). Bound together: Social binding leads to faster processing, spatial distortion, and enhanced memory of interacting partners. *J Exp Psychol Gen*, *148*(7), 1251–1268. https://doi.org/10.1037/xge0000545

Vigneau, M., Beaucousin, V., Herve, P. Y., Duffau, H., Crivello, F., Houde, O., Mazoyer, B., & Tzourio-Mazoyer, N. (2006). Meta-analyzing left hemisphere language areas: Phonology, semantics, and sentence processing. *Neuroimage*, *30*(4), 1414–1432. https://doi.org/10.1016/j.neuroimage.2005.11.002

Vigneau, M., Beaucousin, V., Herve, P. Y., Jobard, G., Petit, L., Crivello, F., Mellet, E., Zago, L., Mazoyer, B., & Tzourio-Mazoyer, N. (2011). What is right-hemisphere contribution to phonological, lexico-semantic, and sentence processing? Insights from a meta-analysis. *Neuroimage*, *54*(1), 577–593. https://doi.org/10.1016/j.neuroimage.2010.07.036

Walbrin, J., Downing, P., & Koldewyn, K. (2018). Neural responses to visually observed social interactions. *Neuropsychologia*, *112*, 31–39. https://doi.org/10.1016/j.neuropsychologia.2018.02.023

Walbrin, J., & Koldewyn, K. (2019). Dyadic interaction processing in the posterior temporal cortex. *Neuroimage*, *198*, 296–302. https://doi.org/10.1016/j.neuroimage.2019.05.027

Walbrin, J., Mihai, I., Landsiedel, J., & Koldewyn, K. (2020). Developmental changes in visual responses to social interactions. *Dev Cogn Neurosci*, *42*, 100774. https://doi.org/10.1016/j.dcn.2020.100774

Walenski, M., Europa, E., Caplan, D., & Thompson, C. K. (2019). Neural networks for sentence comprehension and production: An ALE-based meta-analysis of neuroimaging studies. *Human Brain Mapping, 40*(8), 2275–2304. https://doi.org/10.1002/hbm.24523

Watson, R., Latinus, M., Charest, I., Crabbe, F., & Belin, P. (2014). People-selectivity, audiovisual integration and heteromodality in the superior temporal sulcus. *Cortex*, *50*, 125–136. https://doi.org/10.1016/j.cortex.2013.07.011

Wiethoff, S., Wildgruber, D., Kreifelts, B., Becker, H., Herbert, C., Grodd, W., & Ethofer, T. (2008). Cerebral processing of emotional prosody—influence of acoustic parameters and arousal. *Neuroimage*, *39*(2), 885–893. https://doi.org/10.1016/j.neuroimage.2007.09.028

Wright, T. M., Pelphrey, K. A., Allison, T., McKeown, M. J., & McCarthy, G. (2003). Polysensory interactions along lateral temporal regions evoked by audiovisual speech. *Cereb Cortex*, *13*(10), 1034–1043. https://doi.org/10.1093/cercor/13.10.1034

Yang, D. Y.-J., Rosenblau, G., Keifer, C., & Pelphrey, K. A. (2015). An integrative neural model of social perception, action observation, and theory of mind. *Neurosci Biobehav Rev*, *51*, 263–275. https://doi.org/10.1016/j.neubiorev.2015.01.020

Yang, J. (2014). The role of the right hemisphere in metaphor comprehension: A meta-analysis of functional magnetic resonance imaging studies. *Human Brain Mapping, 35*(1), 107–122. https://doi.org/10.1002/hbm.22160

Zhang, G., Xu, Y., Zhang, M., Wang, S., & Lin, N. (2021). The brain network in support of social semantic accumulation. *Soc Cogn Affect Neurosci*, *16*(4), 393–405. https://doi.org/10.1093/scan/nsab003