Supplementary materials

S1. Behavioural results: Catch trial detection during the main auditory task

In Experiment 1, on average, participants correctly detected 10.08 catch trials per included run (~84%). Participants correctly detected significantly more (t(22) =7.72, p < .001) catch trials per run for recordings in English (M=5.29, SD = 0.29) compared to German (M=4.80, SD = 0.33).

Due to technical issues with the response device, behavioural data is missing completely for four participants, and for one out of seven runs for another participant (remaining data included) in Experiment 2. On average, participants correctly detected 10.12 catch trials per included run (~84.5%). Participants correctly detected significantly more (t(7) =3.20, p = .02) catch trials per run for recordings in English (M=5.22, SD = 0.29) compared to German (M=4.91, SD = 0.27).

S2. Details of audio recordings and audio processing

All audio was recorded directly into Logic Pro X (Release 10.4.2, Apple Inc.) audio processing software using one Shure SM58 and one Shure SM57 microphone plugged into a Zoom H6n Audio Interface. During conversations, actors' voices were recorded simultaneously using the two microphones.

The separate audio channels were normalised and processed using the Logic Pro X Channel Equaliser. To remove unwanted background noise, the audio was first band-pass filtered (filtering signal below 61Hz and above 12000Hz) using the 'Spoken Word Vocal EQ' pre-set, and then filtered using two instances of RX iZotope RX 7 Voice De-Noise. Next, the audio was compressed using the Logic Pro X Compressor. To remove any colour added by the noise filters or the compressor, the audio was equalised again using the 'Spoken Word Vocal EQ'. Finally, the audio was sliced into the individual vocal samples using the Logic Pro X editing suit and rendered at a sample rate of 44100Hz and at 16-bit resolution. The final set of stimuli were root mean square amplitude-normalized and a custom Sensimetrics earphones EQ filter was applied.

S3. Stimulus selection process.

A total of 108 scripted conversations were recorded with a male and a female speaker pair for each language. Each conversation was recorded twice so that both speakers took the role of speaker A and speaker B. Two samples were taken for each recording to select the best quality one afterwards, resulting in 432 recordings per language & speaker pair. These recordings were first assessed for initial quality by two native English speakers regarding how natural they sounded. As a result of this "first-pass" selection, 69 English conversations (=138 recordings for speaker A & B total) were selected to be rated for naturalness, valence, interactiveness, and mental imagery on a 5-pt Likert scale (4 raters each rated half the set of conversations). Based on the highest ratings for naturalness and interactiveness, the final set of 26 English conversations was selected, scrambled conversations were generated, and then the final stimuli were rated again (see S4.3). German conversations were reduced to a set of 73 after a "first-pass" for initial quality by one native German speaker and one speaker with a good working knowledge of German. The final set of 26 German conversations was chosen to be not translated equivalents of the English conversations to minimise the occurrence of cognate words.

A total of 36 scripted narrations were recorded with two male and two female speakers for each language, resulting in 288 recordings total. The final set of 26 English narrations was based on ratings for naturalness (see S4.3) after excluding scripts that were deemed to potentially evoke theory-of-mind processes. Due to the smaller number of recorded narrations, German narrations included 16 stimuli that were translated English narrations.

Examples as well as the scripts of the final set of stimuli can be found as an Excel file on https://osf.io/4xedj/.

S4. Final stimulus set characteristics

S4.1 Duration, word, and letter count. Per condition, English and German stimulus sets were matched for mean stimulus duration (Conversations: t(27) = 0.57, p = .57, scrambled conversations: t(27) = -0.07, p = .95, narrations t(27) = 0.76, p = .45). To achieve this, English and German stimulus sets were well matched on the level of letters per stimulus (t(25) = -0.66, p = .52), whilst they differed on the mean number of words per stimulus (t(25) = 6.02, p < .001), due to word length differences between the two languages. Thus, German stimuli contained fewer words with more letters to match English stimuli with more words but fewer letters.

Table S1. Total and mean stimulus duration, as well as mean letter and word count per stimulus by experimental condition for the final stimulus set

		Total duration in s	Mean stimulus duration in s	Letter count per stimulus	Word count per stimulus
Conversation	English	211.55	7.56 (0.17)	107.35 (2.09)	26.65 (0.36)
	German	207.46	7.41 (0.12)	107.77 (1.51)	24.77 (0.36)
Scrambled conversation	English	211.77	7.56 (0.16)	106.35 (1.75)	26.42 (0.38)
	German	206.98	7.39 (0.13)	107.12 (1.60)	24.50 (0.33)
Narration	English	247.05	8.82 (0.14)	106.23 (1.62)	26.08 (0.39)
	German	247.41	8.84 (0.15)	108.04 (1.73)	24.65 (0.32)

S4.2 Mean pitch (F0). A Language × Condition repeated-measures ANOVA revealed that mean F0 was greater for English compared to German stimuli (F(1,27) = 6.29, p = 0.02, η_p^2 = 0.19). Additionally, the mean F0 of the two-speaker conditions (conversations, scrambled conversations) was greater than the mean F0 of narration stimuli (main effect of condition: F(2,54) = 43.92, p < 0.001, η_p^2 = 0.62). There was no significant interaction effect, F(2,54) = 2.42, p = 0.10, η_p^2 = 0.08.

Table S2. Mean pitch (F0 in Hz) and standard deviation of stimulus set by condition and gender

Language	Speaker gender	Conversation	Scrambled conversation	Narration
English	Male	159.51 (7.54)	155.66 (10.07)	134.18 (19.25)
	Female	217.14 (16.36)	215.64 (16.69)	191.81 (17.26)
	Across gender	188.33 (31.90)	185.65 (33.40)	163.00 (34.40)
German	Male	121.72 (7.77)	118.76 (11.80)	106.69 (12.26)
	Female	222.10 (7.29)	226.24 (8.72)	207.65 (7.22)
	Across gender	171.91 (51.64)	172.50 (55.60)	157.17 (52.35)

S4.3 Stimulus ratings. The final set of English narrations was rated by an independent sample of 8 raters (mean age = 30.8, SD = 5.08, 1 male) on perceived naturalness, valence, and the extent to which the narration evoked a mental image on a 5-point Likert scale (see Table S3). The final set of English conversations and scrambled conversations was rated by an independent sample of 27 raters (mean age = 22.3, SD = 5.1, 6 males) on perceived naturalness, valence, interactiveness (sounding like an interaction), closeness (relationship) of agents, as well as whether and to what extent the conversation evoked a mental image on a 5-point Likert scale.

Table S3. Mean and SD of English stimuli conditions

	Natural	Valence	% evoked mental image	Mental image strength	Interaction	Closeness (relationship) of agents
Conversations	3.94 (0.61)	3.38 (0.35)	74%	3.59 (0.74)	4.28 (0.53)	3.24 (0.37)
Scrambled conversations	2.02 (0.80)	2.79 (0.45)	25%	2.62 (0.77)	1.94 (0.60)	2.12 (0.66)
Narrations	4.10 (0.42)	3.47 (0.29)	75%	3.43 (0.66)	-	-

Independent sample t-tests revealed that ratings did not differ between conversations and narrations; paired sample t-tests revealed that conversations were rated higher on all dimensions compared to scrambled conversations (see Table S4).

Table S4. Rating data statistics

	Independent-sample t-test English conversations vs English narrations			Paired sample t-test English conversations vs English scrambled conversations				
	Naturalness	Valence	Mental Image Strength	Naturalness	Valence	Interactiveness	Closeness	Mental Image Strength
t	-0.71	-0.64	0.65	11.96	7.57	16.73	8.38	5.32
df	33	33	33	26	26	26	26	20
p	0.48	0.52	0.52	<.001	<.001	<.001	<.001	<.001

S5. TPJ as control ROI within the 'social brain'

Method. Participants completed an additional localiser task (Jacoby et al., 2016) to identify the temporo-parietal junction (TPJ), which served as a control region within the 'social brain'. TPJ was chosen based on previous findings that showed TPJ not to be sensitive to visual interactions across several stimulus types and paradigms (Isik et al., 2017; Masson & Isik, 2021; Walbrin et al., 2018; Walbrin & Koldewyn, 2019; Walbrin et al., 2020). In addition, the TPJ is very nearby the pSTS interaction region but not thought to be part of the voice-processing network. Thus, it was predicted that TPJ would not show sensitivity to the experimental conditions of the main auditory task. To localise the TPJ bilaterally, participants watched the short (5:49 minutes) Pixar animation short film 'Partly Cloudy' (2009) which has been found to reliably evoke responses in the mentalizing selective TPJ. The film scenes were coded by event type (mentalizing, pain, social, and control; Jacoby et al.,

2016) and the contrast mentalizing vs. pain was used to localise TPJ. The same procedures as described in the main text were used to define subject-specific TPJ ROIs bilaterally.

Results. One-sample t-test against zero confirmed that the main auditory task did not drive activation in TPJ above baseline, suggesting these stimuli did not drive activity in the TPJ. Bilaterally, English conversations resulted in PSC that were not significantly different than baseline (all ts(22) > -1.17, all ps > .26), whereas marginally significant (German conversations, all ts(22) > -2.07, all ps > .05) or significant (all other conditions, all ts(22) < -2.11, all ps < .05) TPJ deactivation was observed otherwise. ANOVA revealed no significant main effects of Language (F(1,22) = 0.11, p = .74, η_p^2 < 0.01) or Condition (F(1,22) = 2.04, p = .14, η_p^2 = 0.09), nor a significant interaction (F(1.48,32.53) = 1.78, p = .18, η_p^2 = 0.08) in right TPJ. For left TPJ, there were no main effects of Language (F(1,22) = 0.15, p = .71, η_p^2 < 0.01) or Condition (F(1.40,30.77) = 2.56, p = .11, η_p^2 = 0.10), however, there was a significant interaction (F(1.38,30.31) = 5.46, p = .02, η_p^2 = 0.20). Explorative (uncorrected) post-hoc t-tests revealed that this was driven by a significant effect of language for narrations only: English narrations deactivated the region more than German narrations (t(22) = -2.47 , p = 0.02). Furthermore, English conversations deactivated the region significantly less than narrations (t(22) = 2.20 , p = 0.04).

Table S5. Bilateral TPJ PSC means (SE)

Language	Condition	Left TPJ	Right TPJ
English	Conversation	-0.07 (0.08)	-0.07 (0.06)
	Scrambled conversation	-0.15 (0.05)	-0.13 (0.05)
	Narration	-0.26 (0.07)	-0.18 (0.05)
German	Conversation	-013 (0.06)	-0.10 (0.05)
	Scrambled conversation	-0.16 (0.06)	-0.13 (0.05)
	Narration	-0.14 (0.06)	-0.11 (0.05)

S6. Additional ROI information

Table S6. MNI-space coordinates (based on group-level localiser task contrasts) used as centre for sphere creation identified at an uncorrected threshold of p < .01

Region	Defining task	Defining contrast	Left	Right
TPJ	Mentalizing localiser	Mentalizing > Pain	[-42 -68 40]	[48 -62 36]
SI-pSTS	Interaction localiser	Interaction > Non-Interaction	[-48 -60 12]	[52 -48 12]
TVA	Voice localiser	Vocal > Non-vocal sounds	[-60 -24 0]	[60 -24 0]
aSTS	Main audio task	Language × Condition interaction (F-contrast)	[-50 0 -24]	[52 2 -24]

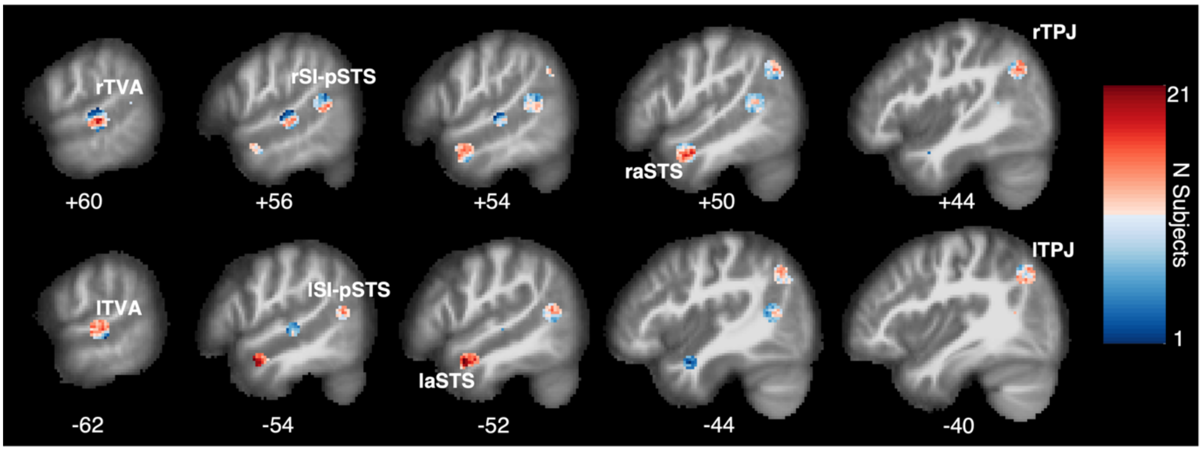

Figure S1. Sagittal view heatmap of subject-specific aSTS, TVA, SI-pSTS, and TPJ ROIs' overlap for right (top row) and left (bottom row) hemisphere. x-coordinate in MNI space shown below each slice. Figure created using bspmview toolbox (DOI: 10.5281/zenodo.168074).

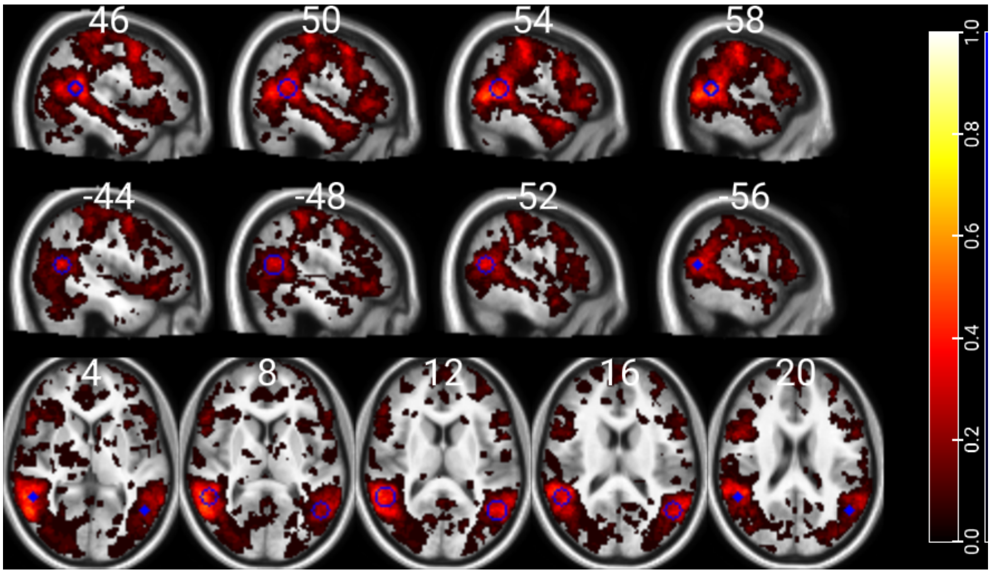

Figure S2. Map showing the proportion of individual participants showing activation for the interaction localizer, illustrated by overlaying all participants' first level T-maps for the contrast interaction > non-interaction, thresholded at p < .01, uncorrected. Activation overlap is measured from 0 (0%) to 1 (100). The blue lines indicate the edges of the original bilateral 8mm bounding spheres.

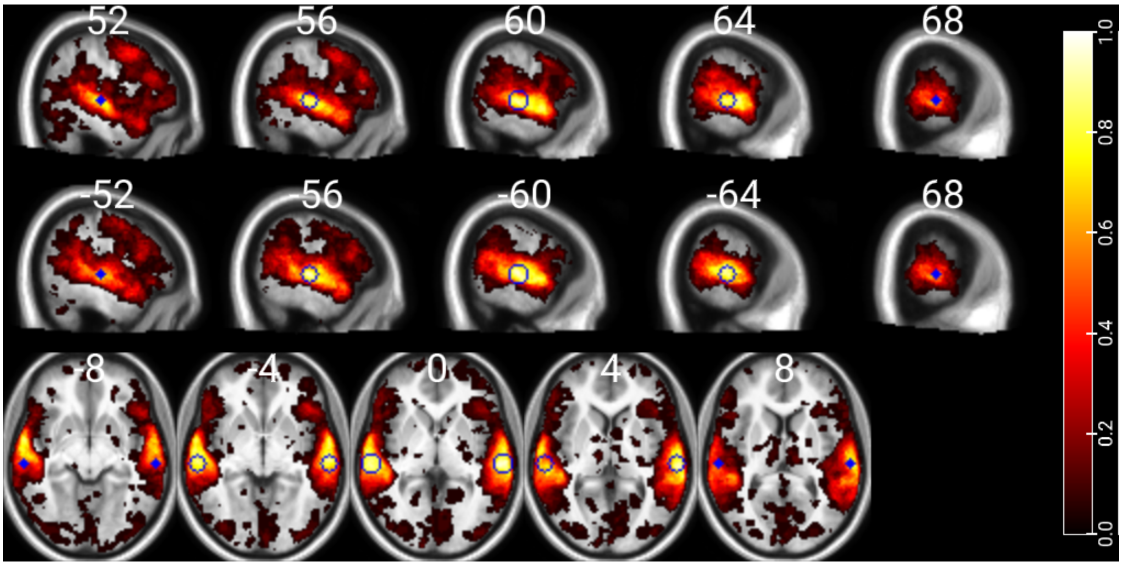

Figure S3. Map showing the proportion of individual participants showing activation for the voice localizer, illustrated by overlaying all participants' first level T-maps for the contrast vocal > non-vocal sounds, thresholded at p < .01, uncorrected. Activation overlap is measured from 0 (0%) to 1 (100). The blue lines indicate the edges of the original bilateral 8mm bounding spheres.

S7. Additional data tables for Experiment 1 and Experiment 2

Table S7. SI-pSTS and TVA PSC means (SE) for all experimental conditions for both experiments

		Experiment 1				Experiment 2			
		SI-pSTS		TVA		SI-pSTS		TVA	
Language	Condition	L	R	L	R	L	R	L	R
English	Conversation	0.30 (0.07)	0.45 (0.10)	1.83 (0.13)	2.01 (0.12)	0.26 (0.07)	0.43 (0.17)	1.34 (0.15)	1.45 (0.17)
	Scrambled conversation	0.33 (0.08)	0.50 (0.11)	1.88 (0.12)	2.00 (0.12)	0.21 (0.08)	0.46 (0.16)	1.38 (0.16)	1.43 (0.16)
	Narration	0.06 (0.05)	0.03 (0.08)	1.50 (0.11)	1.59 (0.12)	0.02 (0.06)	0.09 (0.11)	1.13 (0.14)	1.14 (0.13)
German	Conversation	-0.12 (0.04)	0.03 (0.08)	1.55 (0.11)	1.70 (0.11)	0.26 (0.09)	0.41 (0.17)	1.27 (0.15)	1.33 (0.15)
	Scrambled conversation	-0.12 (0.04)	0.03 (0.09)	1.52 (0.10)	1.67 (0.11)	0.23 (0.08)	0.52 (0.16)	1.31 (0.15)	1.34 (0.15)
	Narration	-0.16 (0.04)	-0.13 (0.08)	1.31 (0.10)	1.50 (0.12)	0.08 (0.05)	0.16 (0.11)	1.05 (0.14)	1.04 (0.11)

Table S8. Bilateral PSC means (SE) for interaction localiser (Experiment 1 only)

	SI-pSTS		TVA		aSTS	
Condition	L	R	L	R	L	R
Interaction	0.82 (0.10)	0.90 (0.10)	-0.12 (0.06)	-0.21 (0.07)	-0.02 (0.05)	0.26 (0.05)
Non-Interaction	0.44 (0.10)	0.43 (0.14)	-0.13 (0.05)	-0.21 (0.06)	-0.14 (0.05)	-0.04 (0.06)
Scrambled interaction	0.07 (0.09)	0.25 (0.13)	-0.16 (0.05)	-0.23 (0.07)	-0.23 (0.05)	-0.10 (0.05)

Table S9. Bilateral PSC means (SE) for bilateral SI-pSTS responses to voice localiser (Experiment 1 only)

	SI-pSTS	
Condition	L	R
Vocal sounds	0.01 (0.07)	0.25 (0.12)
Non-vocal sounds	-0.06 (0.04)	0.02 (0.06)

Table S10. Bilateral aSTS PSC means (SE) for auditory task (Experiment 1 only)

Language	Condition	Left aSTS	Right aSTS
English	Conversation	0.67 (0.05)	0.79 (0.09)
	Scrambled conversation	0.65 (0.04)	0.70 (0.08)
	Narration	0.20 (0.03)	0.07 (0.05)
German	Conversation	< -0.01 (0.04)	-0.04 (0.03)
	Scrambled conversation	-0.03 (0.03)	-0.06 (0.04)
	Narration	-0.07 (0.04)	-0.14 (0.04)

Table S11. Results of paired sample t-tests for aSTS and SI-pSTS PSC comparison of interaction localiser conditions (Experiment 1 only)

Region	Side	Comparison	Mean difference	t(22)	p
SI-pSTS	R	Interaction vs Non-Interaction	0.47	5.84	<.001
		Interaction vs Scrambled interaction	0.65	7.08	<.001
	L	Interaction vs Non-Interaction	0.38	6.19	<.001
		Interaction vs Scrambled interaction	0.75	10.03	<.001
aSTS	R	Interaction vs Non-Interaction	0.29	5.41	<.001
		Interaction vs Scrambled interaction	0.35	7.48	<.001
	L	Interaction vs Non-Interaction	0.12	3.71	.001
		Interaction vs Scrambled interaction	0.21	6.93	<.001