OXFORD

Cerebral Cortex, 2022, 1–20
https://doi.org/10.1093/cercor/bhac095
Original Article

Neural decoding of emotional prosody in voice-sensitive auditory cortex predicts social communication abilities in children

Simon Leipold[1,†,*], Daniel A. Abrams[1,†], Shelby Karraker[1], Vinod Menon[1,2,3,*]

[1]Department of Psychiatry and Behavioral Sciences, Stanford University, Stanford, CA, USA,
[2]Department of Neurology and Neurological Sciences, Stanford University, Stanford, CA, USA,
[3]Stanford Neurosciences Institute, Stanford University, Stanford, CA, USA

*Corresponding authors: Department of Psychiatry and Behavioral Sciences, Stanford Cognitive and Systems Neuroscience Laboratory, 401 Quarry Rd. Stanford, CA 94305, USA. Email: leipold@stanford.edu; menon@stanford.edu

†These authors contributed equally.

Abstract

During social interactions, speakers signal information about their emotional state through their voice, which is known as emotional prosody. Little is known regarding the precise brain systems underlying emotional prosody decoding in children and whether accurate neural decoding of these vocal cues is linked to social skills. Here, we address critical gaps in the developmental literature by investigating neural representations of prosody and their links to behavior in children. Multivariate pattern analysis revealed that representations in the bilateral middle and posterior superior temporal sulcus (STS) divisions of voice-sensitive auditory cortex decode emotional prosody information in children. Crucially, emotional prosody decoding in middle STS was correlated with standardized measures of social communication abilities; more accurate decoding of prosody stimuli in the STS was predictive of greater social communication abilities in children. Moreover, social communication abilities were specifically related to decoding sadness, highlighting the importance of tuning in to negative emotional vocal cues for strengthening social responsiveness and functioning. Findings bridge an important theoretical gap by showing that the ability of the voice-sensitive cortex to detect emotional cues in speech is predictive of a child's social skills, including the ability to relate and interact with others.

Key words: development; emotion recognition; speech; superior temporal sulcus; voice.

Introduction

The human voice is a critical social stimulus in a child's environment. The voice not only conveys semantic information ("what") through speech, but it also provides information about the identity ("who") and the emotional state ("how") of the speaker (Belin et al. 2004), which is known as emotional prosody (Schirmer and Kotz 2006; Wildgruber et al. 2006; Brück, Kreifelts, and Wildgruber 2011; Grandjean 2021). Decoding these different pieces of information from the vocal signal is critical for navigating the social world. Understanding how a communication partner is feeling is crucial for providing empathy and support and is particularly important for building and maintaining interpersonal connections. While the human voice serves as a conduit for conveying emotional information in communication (Pell and Kotz 2021), little is known regarding the neurobiological mechanisms underlying emotional prosody perception in children and their links to broader measures of social function.

Vocal-emotional information is conveyed by a speaker's intonation, emphasis, rhythm, and speech rate (Hammerschmidt and Jürgens 2007), and these vocal gestures translate into an array of acoustical cues embedded in ongoing speech (Banse and Scherer 1996). For example, when a speaker is sad, vocal pitch and speech rate are reduced relative to neutral speech; however, when a speaker is happy, vocal pitch and speech rate typically increase compared to neutral speech. During the early stages of child development, young listeners begin to map these distinct acoustical features, which include changes in vocal pitch, timing, and timbre, on to speakers' emotional states (Flom and Bahrick 2007; Blasi et al. 2011). Following extensive experience and learning, this vocal-emotional mapping yields an efficient auditory mechanism for rapidly ascertaining the emotional state of a communication partner (for a review, see Morningstar et al. 2018).

An emerging literature has shown that, similar to other cognitive skills such as reading and memory, typically developing children reveal a wide range of abilities with regard to social skills: While some children effortlessly interact in social settings and form social bonds, others have more difficulty with these tasks (Constantino and Todd 2003; Posserud et al. 2006). Consistent with these findings, previous

Received: September 10, 2021. **Revised:** February 11, 2022. **Accepted:** February 12, 2022

studies investigating emotional prosody processing have revealed substantial interindividual variability in children for recognizing and categorizing emotions from voices (Nowicki and Duke 1994). Thus, both decoding of emotional prosody information and broader social communication skills are continuously distributed even among the general population of children without neurodevelopmental disorders associated with social communication and interaction deficits (Constantino and Todd 2003). Importantly, theoretical models have posited a link between individuals' ability to decode emotions of other individuals and the quality of their social interactions (Keltner and Haidt 1999; Van Kleef 2009). These theoretical links have been substantiated by behavioral evidence showing that children's emotional prosody recognition abilities are associated with their social abilities (Chronaki et al. 2015; Neves et al. 2021). While processing vocal emotions represents a crucial aspect of successful social interactions, it is unknown whether heterogeneity in neural decoding of emotional prosody information explains variance in social communication abilities in children.

Despite the importance of decoding vocal-emotional information for meaningful social interactions (Pell and Kotz 2021), little is known regarding the brain systems and neural representations underlying emotional prosody perception in typically developing children and whether these representations are related to social function. The vast majority of research in this area has been conducted in adult participants, and results from these studies have yielded inconsistent results. Supplementary Table 1 summarizes experimental and analytic approaches used in this literature as well as the brain regions identified in these studies in response to emotional prosody. Results from several functional magnetic resonance imaging (fMRI) studies support a crucial role for superior temporal cortex by showing that hearing a range of emotional prosody stimuli elicits increased activation in supratemporal plane (STP) as well as anterior superior temporal sulcus (aSTS), middle superior temporal sulcus (mSTS), and posterior superior temporal sulcus (pSTS) (Mitchell et al. 2003; Grandjean et al. 2005; Wildgruber et al. 2005; Ethofer et al. 2006, 2012; Johnstone et al. 2006; Beaucousin et al. 2007; Brück, Kreifelts, Kaza, et al. 2011; Goerlich-Dobre et al. 2014; Ceravolo et al. 2016). Additional support for the role of superior temporal cortex in emotional prosody perception includes an fMRI study which showed that multivariate patterns of fMRI activity within STP and superior temporal sulcus (STS) discriminate between different categories of vocal-emotional information; however, multivariate pattern analysis (MVPA) in this study was restricted to superior temporal regions and did not examine brain systems beyond temporal cortex (Ethofer et al. 2009). Apart from superior temporal cortex, other studies of emotional prosody processing have highlighted additional cortical and subcortical brain regions, including lateral prefrontal and orbitofrontal cortex (Buchanan et al. 2000; Kotz et al. 2003; Sander et al. 2005; Frühholz et al. 2012), the insula (Bach et al. 2008; Seydell-Greenwald et al. 2020; Giordano et al. 2021), and the amygdala (Frühholz and Grandjean 2013; Frühholz et al. 2015). Results from previous studies have shown increased activation in this extended collection of brain regions during the processing of emotional prosody stimuli; however, these findings beyond temporal cortex have been inconsistent between studies. Two studies that examined brain systems underlying emotional prosody processing in children identified an extensive brain network encompassing superior temporal, prefrontal, occipital, basal ganglia, and cerebellar brain regions; however, these studies used a silent baseline for their analyses and it therefore is unclear whether these effects specifically reflect emotional prosody processing or auditory processing more generally (Morningstar et al. 2019, 2020). Moreover, while a previous study of adults revealed above chance decoding of emotional voices in temporal and prefrontal regions using searchlight MVPA across the whole brain (Kotz et al. 2013), studies including children (Morningstar et al. 2019, 2020) have not applied an MVPA approach to decoding emotional prosody. Importantly, it is unknown how the brain systems serving emotional prosody processing relate to broader measures of social function.

Here, we build on prior studies and extend our knowledge regarding the precise brain systems underlying emotional prosody processing by examining this question in children with a specific focus on identifying links between prosody processing and broader measures of social function. We used event-related fMRI to measure neural responses in typically developing children aged 7–12 years while they listened to emotional prosody and neutral speech (Fig. 1). Our study had 3 major goals. Our first goal was to assess emotional prosody processing within auditory cortex, encompassing regions of both STP and STS. A limitation of previous studies has been a lack of anatomical specificity in auditory cortex with regard to emotional prosody processing (Ethofer et al. 2009; Grossmann et al. 2010; Morningstar et al. 2019, 2020; Zhang et al. 2019), and here, we sought to identify specific subregions of auditory cortex that are sensitive to these vocal features. Moreover, none of the studies in children used MVPA techniques to examine whether these auditory regions can reliably discriminate between emotional and neutral prosody stimuli. MVPA exploits multivariate information present in the functional imaging data and has high statistical sensitivity to detect differential activation patterns associated with emotional prosody and neutral speech (Kriegeskorte and Bandettini 2007; Kotz et al. 2013; Haynes 2015; Kragel and LaBar 2016). Furthermore, with cross-validation, MVPA provides a highly robust, neurobiologically plausible measure of dissociations in activation patterns induced by specific stimuli (Kriegeskorte and Douglas 2019). We therefore used MVPA to decode emotional prosody versus neutral speech, and distinct vocal emotions (sad

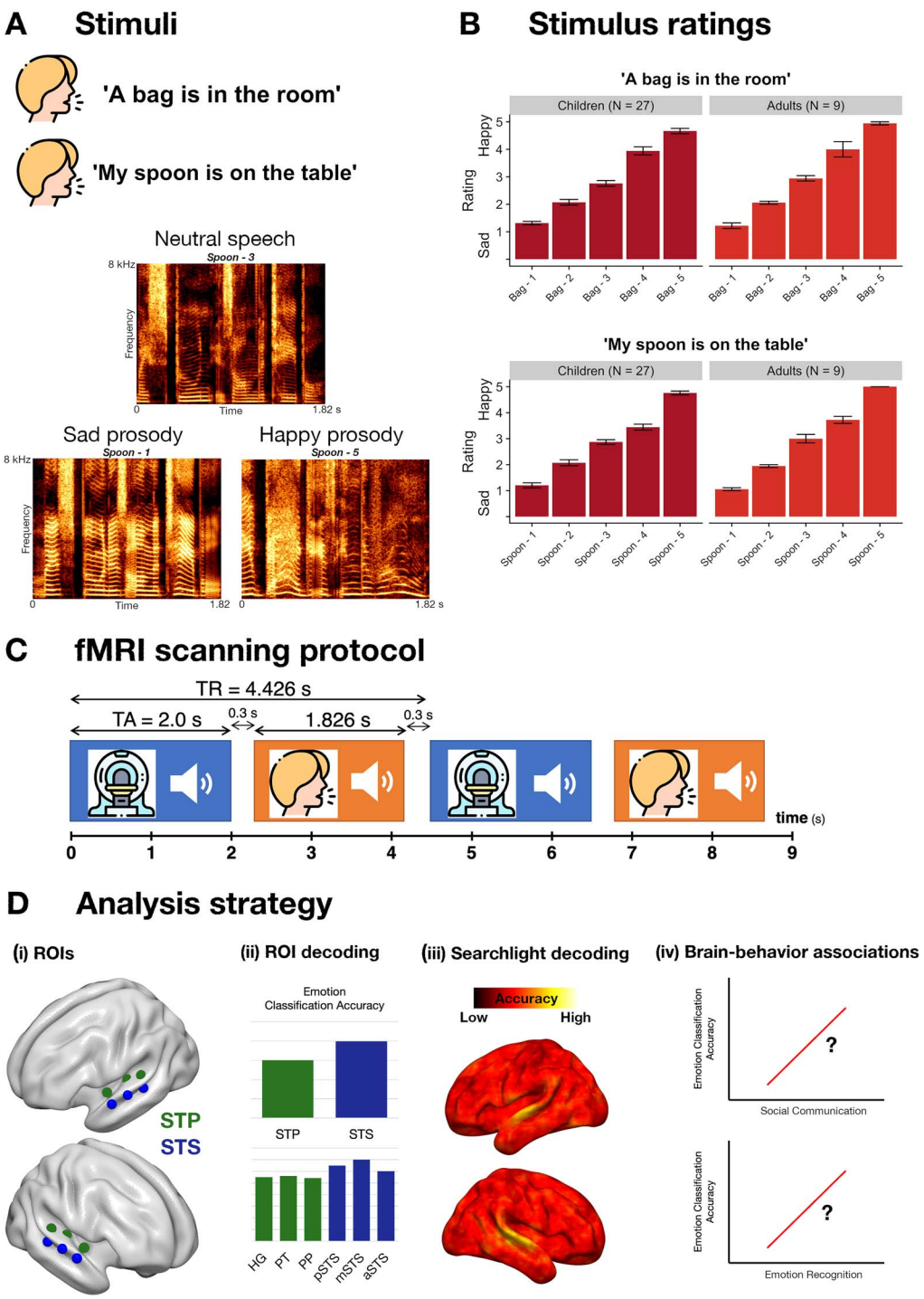

Fig. 1. fMRI stimuli, scanning protocol, and analysis strategy. A) fMRI stimuli consisted of acoustic sentences spoken in emotional and neutral prosody. Spectrograms of sentence #2, "my spoon is on the table," spoken in neutral speech (upper panel), in sad prosody (lower-left panel) and in happy prosody (lower-right panel). B) Stimuli were selected based on results from a behavioral experiment conducted in an independent cohort of 27 school-age, typically developing children and 9 adults, who provided ratings on a 5-point scale ("how sad or happy is this voice?"). C) A sparse sampling fMRI scanning protocol with a repetition time (TR) larger than the acquisition time (TA) was used to present acoustic stimuli during silent intervals between volume acquisitions to eliminate the effects of scanner noise on auditory perception. D) Schematic of the analyses employed in the study. (i) Definition of ROIs within auditory cortex. (ii) ROI-based multivariate pattern classification of neutral and emotional prosody was employed to compare decoding accuracy between auditory cortical subdivisions of the STP and STS. (iii) A whole-brain multivariate pattern classification method was used to examine whether brain regions beyond auditory cortex accurately discriminate emotional and neutral prosody stimuli. (iv) Associations between children's neural decoding of emotional prosody, social communication skills, and emotion recognition accuracy.

and happy), in regions of interest (ROIs) within both STP and STS. The second goal of the study was to further assess the brain systems underlying emotional prosody processing by examining the contributions of brain regions beyond auditory cortex for the decoding of these vocal cues using whole-brain searchlight MVPA in children. To facilitate comparison with prior work, we additionally analyzed neural response levels to emotional prosody in auditory ROIs and across the whole brain.

The third goal of our study was to investigate whether decoding of emotional prosody is related to social communication and emotion recognition abilities in typically developing children. This is a crucial question for establishing a link between children's sensitivity to emotional prosody cues in everyday speech and their proficiency in relating to their peers and establishing social bonds. Importantly, no previous studies have examined the relationship between neural decoding of emotional prosody and broader measures of social function in children. We therefore related neural decoding accuracy for emotional prosody stimuli to established measures of social communication skills using the Social Responsiveness Scale-2 (SRS-2; Constantino and Gruber 2012) and to behavioral emotion recognition using the Diagnostic Analysis System of Nonverbal Accuracy 2 (DANVA2; Baum and Nowicki 1998).

We investigated differential links for neural decoding of sad and happy prosody with children's social communication skills. Although a previous study in children with autism has hinted at a differential impact of sad versus happy social cue detection for social functioning (Williams and Gray 2013), the robustness of this result and its generalization to the auditory domain is not clear (Trevisan and Birmingham 2016), and moreover, it is unknown whether such a differentiation might be observed in typically developing children.

Materials and methods

Participants

The Stanford University Institutional Review Board approved the study protocol. Parental consent and children's assent were obtained for all evaluation procedures, and participants were paid for their participation in the study. We recruited a total of 31 typically developing children from the San Francisco Bay Area in CA, USA, to participate in the study. Ten participants were excluded after data acquisition as they did not meet data quality criteria based on maximal movement during fMRI scanning (see below for more details). The final sample for data analysis included $n = 21$ children between 7 and 12 years of age. No participants were excluded during data analysis. Detailed demographic and neuropsychological characteristics are given in Table 1.

All children were required to have a full-scale intelligence quotient (IQ) >80 as measured by the Wechsler Abbreviated Scale of Intelligence (WASI). All of the participants were right-handed and had no history of neurological, psychiatric, or learning disorders and no personal or family (first degree) history of developmental cognitive disorders or heritable neuropsychiatric disorders. Mothers of the participants reported no evidence of significant difficulty during pregnancy, labor, delivery, or the immediate neonatal period, and no abnormal developmental milestones as determined by neurologic history and examination.

fMRI stimuli

The stimuli presented during fMRI scanning consisted of acoustic sentences spoken in emotional and neutral prosody as well as nonspeech environmental sounds. The emotional and neutral prosody stimuli were recorded in a recording studio by a professional actress who produced 108 vocal samples of 2 sentences, "a bag is in the room" (sentence #1) and "my spoon is on the table" (sentence #2), using sad, happy, and neutral emotions (Fig. 1A). These sentences were previously validated to be neutral with regard to their emotional content (Ben-David et al. 2011). For both the sad and happy prosody conditions, the actress varied the intensity of prosodic cues with the goal of yielding both low-intensity and high-intensity emotional cues. All vocal stimuli were recorded using a Shure PG27-USB condenser microphone connected to a MacBook Air laptop computer and were digitized at a sampling rate of 44.1 kHz and D/A converted with 16-bit resolution. A second class of stimuli included in the study was nonspeech environmental sounds. These sounds, which included brief recordings of laundry machines, dishwashers, and other household sounds, were taken from a professional sound effects library. All vocal and environmental sound stimuli were band-pass-filtered (0.08–10.5 kHz), downsampled to 22.05 kHz, edited to equate stimulus intensity, and adjusted to a duration of 1,826 ms with a linear fade to prevent click-like sounds to occur at the end the stimuli. The vocal stimuli were equalized to a duration of 1,826 ms because this represents the average duration of the vocal stimuli. We used Praat software (RRID: SCR_016564), which does not alter the speech pitch properties of the recordings, to normalize the duration of the vocal stimuli to 1,826 ms. This process inherently alters the speech rate (i.e. the number of syllables or words within a given time); however, this process did not significantly affect the quality of these stimuli. All vocal and environmental stimuli can be downloaded from the Open Science Framework (https://dx.doi.org/10.17605/OSF.IO/TYFXS).

Stimulus selection experiment

Emotional and neutral prosody stimuli for the fMRI experiment were selected based on results from a behavioral experiment conducted in an independent cohort of 27 school-age typically developing children (mean age ± standard deviation [SD]: 11.1 ± 1.2 years; sex: 10 female, 17 male) who did not participate in

Table 1. Participants' demographic and neuropsychological characteristics. Continuous measures are given as mean ± SD.

Characteristic	Study sample	Population
Number of participants	21	
Sex (female/male)	8/13	
Age	10.71 ± 1.38 years	
WASI: full-scale IQ	120.29 ± 11.68	100.00 ± 15
WASI: verbal IQ	119.71 ± 14.94	100.00 ± 15
WASI: performance IQ	117.14 ± 11.19	100.00 ± 15
WIAT-II: word reading	110.76 ± 10.81	100.00 ± 15
WIAT-II: reading comprehension	113.48 ± 10.16	100.00 ± 15
SRS-2: total standard *t*-score	47.19 ± 8.08	50 ± 10
SRS-2: social communication standard *t*-score	46.14 ± 7.19	50 ± 10
DANVA2: receptive tests z-scored accuracy[a]	0.32 ± 0.43	0 ± 1

[a]z-scored emotion recognition accuracy averaged across 4 subtests: adult facial expressions, adult paralanguage, child facial expressions, and child paralanguage. Abbreviations: DANVA2 = Diagnostic Analysis System of Nonverbal Accuracy 2 (Baum and Nowicki 1998); SRS-2 = Social Responsiveness Scale-2 (Constantino and Gruber 2012); WIAT-II = Wechsler Individual Achievement Test, Second Edition.

the fMRI study as well as 9 adults. The 108 vocal samples initially produced by the professional actress were reduced to 24 samples for presentation during the stimulus selection experiment. Participants were seated in a quiet room in front of a laptop computer, and headphones were placed over their ears. Consistent with established methods for developing emotional prosody stimuli (Mazefsky and Oswald 2007; Ingersoll 2010; Nowicki 2010), participants were asked to rate the valence of candidate sentence stimuli on a 5-point scale ("how sad or happy is this voice?"). Each candidate vocal stimulus was presented twice to each participant, and the order of stimulus presentation was randomized for each participant. Stimuli that were consistently rated "1" and "5" by the child and adult participants were identified as the high-intensity "sad" and "happy" stimuli for the fMRI experiment; stimuli rated "2" and "4" were identified as the low-intensity "sad" and "happy" stimuli; and the stimulus consistently rated "3" was identified as the "neutral" control stimulus. Dependent samples *t*-tests comparing the mean ratings for the final stimuli for both sentences confirmed statistical differences between all happy, sad, and neutral stimuli (Fig. 1B).

fMRI task

Acoustic stimuli were presented in 10 separate fMRI runs, each lasting for ~3.5 min. One run consisted of 39 trials of acoustic sentence stimuli spoken in sad (high and low intensities), happy (high and low intensities), and neutral prosody as well as environmental sounds and catch trials. Stimuli were pseudorandomly presented within each run. Stimulus presentation order was kept constant across participants. Before each run, child participants were instructed to play the "kitty cat game" during the fMRI scanning. While lying down in the scanner, children were first shown a brief video of a cat and were told that the goal of the cat game was to listen to a variety of sounds and to push a button on a button box only when they heard kitty cat meows (catch trials). The function of the catch trials was to keep the children alert and engaged during stimulus presentation. During each run, we presented 6 sentence exemplars per stimulus condition (neutral, high-intensity sad, low-intensity sad, high-intensity happy, and low-intensity happy), 6 environmental sounds, and 3 catch trials. At the end of each run, the children were shown another engaging video of a cat. Across the 10 runs, a total of 30 repetitions of each sentence and prosody intensity combination were presented to each participant, including 30 repetitions of the sentence "a bag is in the room" in high-intensity sad, low-intensity sad, neutral, low-intensity happy, and high-intensity happy as well as 30 repetitions of the sentence "my spoon is on the table" in high-intensity sad, low-intensity sad, neutral, low-intensity happy, and high-intensity happy.

Speech stimuli were presented to participants in the scanner using E-Prime v2.0 (RRID:SCR_009567). Participants wore custom-built headphones designed to reduce the background scanner noise to ~70 adjusted dB (dBA) (Abrams et al. 2011; Abrams et al. 2013). Headphone sound levels were calibrated before each data collection session, and all stimuli were presented at a sound level of 75 dBA. Participants were scanned using a fast event-related design. Acoustic stimuli were presented during silent intervals between volume acquisitions to eliminate the effects of scanner noise on auditory perception (see below for details on the implementation).

Imaging data acquisition

Imaging data were acquired in a single session at the Richard M. Lucas Center for Imaging at Stanford University on a GE Signa 3.0 T magnetic resonance imaging (MRI) scanner using a custom-built 8-channel head coil. Participants were instructed to stay as still as possible during scanning, and head movement was further minimized by placing memory-foam pillows around the head. Reduction of blurring and reduction of signal loss arising from field inhomogeneities were accomplished by the use of an automated high-order shimming method before data acquisition. Whole-brain functional images were acquired using a T2*-weighted gradient-echo spiral in-out pulse sequence

(Glover and Law 2001) with the following parameters: repetition time (TR) = 4,426 ms, echo time = 30 ms, flip angle = 80°, slice acquisition order = ascending, number of axial slices = 31, slice thickness = 4 mm, spacing = 0.5 mm, field of view = 220 mm, matrix size = 64 × 64, voxel size = 3.44 × 3.44 × 4 mm^3, and total number of volumes = 43. The TR of 4,426 ms comprised the stimulus duration of 1,826 ms, a 300-ms silent interval buffering the beginning and end of each stimulus presentation (600 ms total of silent buffers) to avoid backward and forward masking effects, and the 2,000-ms acquisition time (TA) for a single volume. A linear shim correction was applied separately for each slice during reconstruction using a magnetic field map acquired automatically by the pulse sequence at the beginning of the scanning session. To assist preprocessing of the functional images, we additionally acquired an anatomical image using a T1-weighted sequence.

fMRI preprocessing

Functional images collected in each of the 10 runs were subjected to preprocessing procedures using SPM12 (version 7219; RRID:SCR_007037) in MATLAB R2019a (RRID:SCR_001622). The first 5 volumes were not analyzed to allow for signal equilibrium. Preprocessing included the following steps: (i) realignment using 6 motion parameters (3 translations and 3 rotations) to mitigate effects of participant motion; (ii) slice timing correction; (iii) coregistration of the mean functional image to the individual anatomical image; (iv) segmentation and bias-field correction of the individual anatomical image and estimation of the deformation field to map the image to the T1-weighted MNI152 template; (v) normalization of the functional images to Montreal Neurological Institute (MNI) space using the deformation field estimated in the previous step; (vi) interpolation to an isotropic voxel size of 2.0 mm; and (vii) smoothing of the functional images with an 6-mm full-width at half-maximum (FWHM) 3D Gaussian kernel. The quality of spatial normalization was manually inspected.

Scanner movement criteria for inclusion in statistical analyses

For inclusion in the fMRI analysis, we required that each functional run have a maximum volume-to-volume movement of <6 mm and that no more than 15% of volumes per run had movement exceeding 0.5 voxels (1.72 mm) or spikes in global signal exceeding 5%. Moreover, we required that all individual participant data included in the analysis consist of at least 7 functional runs that met our inclusion criteria (Abrams et al. 2019). Children who had fewer than 7 functional runs that met our inclusion criteria were excluded from the data analysis. All 21 participants included in the analysis had at least 7 functional runs that met our scanner movement criteria. Fourteen of the participants had 10 runs of data that met these movement criteria, 5 participants had 9 runs of data that met movement criteria, 1 participant had 8 runs of data, and 1 participant had 7 runs that met criteria.

Statistical analyses

The statistical analyses had 3 aims. First, we used ROI-based MVPA within distinct voice-sensitive subregions of auditory cortex to assess dissociable contributions of auditory cortical regions to children's emotional prosody decoding. Second, we used whole-brain searchlight MVPA analysis to explore prosody decoding beyond auditory cortex. Third, we examined relationships between children's neural prosody decoding and standardized measures of social communication and emotion recognition abilities.

Voxel-wise analysis of fMRI activation

The goal of the voxel-wise analysis of fMRI activation was to identify brain regions that showed differential activity levels in response to emotional prosody stimuli, neutral speech, and environmental sounds. For each participant, we modeled the voxel-wise blood oxygen level–dependent (BOLD) signal time series using a general linear model (GLM) implemented with SPM12. The first-level design matrix included, for each run separately, regressors modeling the speech stimulus conditions (neutral speech, high-intensity sad prosody, low-intensity sad prosody, high-intensity happy prosody, and low-intensity happy prosody; separate regressors for sentence #1 and sentence #2) and a regressor modeling the catch trials. Environmental sounds were not modeled and served as the baseline condition. Regressors were built as a boxcar function convolved with the canonical hemodynamic response function and the temporal derivative to account for voxel-wise latency differences in hemodynamic response. The 6 motion parameters estimated during preprocessing were included as nuisance regressors. Low-frequency drift was removed using a high-pass filter (0.5 cycles/min), and serial correlations were accounted for by modeling the voxel-wise BOLD signal time series as a first-degree autoregressive process. We generated a single contrast image per participant for the following contrasts: (neutral speech > environmental sounds), (sad prosody > neutral speech), and (happy prosody > neutral speech). To increase the signal-to-noise ratio of the contrast images, we combined the high-intensity and low-intensity variants for each emotion (i.e. a total of 120 stimulus repetitions for both sad and happy prosody conditions).

A second-level analysis used a one-sample *t*-test on the contrasts of interest (neutral speech > environmental sounds), (sad prosody > neutral speech), and (happy prosody > neutral speech). Statistically significant clusters of activation were obtained using a cluster-defining threshold of $P < 0.005$ and a spatial extent of 70 voxels, controlling the family-wise error (FWE) rate at $\alpha \leq 0.05$ across the whole brain as determined by using Monte Carlo simulations implemented in a custom MATLAB

script (Ward 2000). The unthresholded t-maps of the contrasts (neutral speech > environmental sounds), (sad prosody > neutral speech), and (happy prosody > neutral speech) are available on NeuroVault (RRID:SCR_003806; https://neurovault.org/collections/HRQEJGAZ/).

Multivariate decoding of emotional prosody in auditory cortex

An ROI-based multivariate pattern classification method was used to examine whether auditory cortical areas discriminate emotional prosody and neutral speech stimuli. Classification analysis was performed for each participant using PyMVPA (version 2.6.1; RRID:SCR_006099) in Python 2.7.18 (RRID:SCR_008394) on the Stanford University "Sherlock" cluster. Activation patterns extracted from run-wise contrast images generated by SPM served as inputs to the classification analysis. Specifically, voxel-wise fMRI time series were first modeled using a GLM with a design matrix that was identical to that previously described (see Voxel-wise analysis of fMRI activation above), including regressors modeling the speech stimulus conditions and catch trials, and motion parameters as nuisance regressors. Individual participants' contrast images were generated for the following contrasts: (sad prosody > environmental sounds), (happy prosody > environmental sounds), and (neutral speech > environmental sounds). To increase the signal-to-noise ratio of the contrast images, we combined the high-intensity and low-intensity variants for each emotion. To facilitate run-based cross-validation within each participant, we generated separate contrast images per run, resulting in 7–10 images per contrast and participant, depending on how many runs were included for a particular participant.

A linear support vector machine (SVM; C = 1) was used to identify brain regions that discriminated emotional prosody from neutral speech. Classification accuracy was estimated using a leave-one-run-out crossvalidation: Activation patterns within a particular ROI were assigned to k independent vectors, where k represents the number of runs and each vector contains the contrast beta values for each voxel in the ROI measured in response to a stimulus category for a single run. The SVM was trained on the data of $k - 1$ runs and was tested on the data of the remaining run. This procedure was repeated for k times with data from each of the runs used exactly once for testing. The average classification accuracy was used to evaluate the classifier's performance. Contrast images from different runs are thought to be independent; thus, the leave-one-run-out procedure provides an unbiased assessment of crossvalidated classifier performance (Etzel et al. 2011).

ROI-based multivariate pattern classification was used for 2 related analyses: First, decoding accuracy was compared between auditory cortical subdivisions of the STP, which are thought to encompass more primary regions of auditory cortex (Hickok and Poeppel 2007; Rauschecker and Scott 2009) and regions of the STS, a more ventral aspect of auditory cortex associated with human voice processing (Belin et al. 2000) (see subsection entitled Anatomical ROIs for details on these regions). An STP ROI was constructed by combining ROIs for Heschl's gyrus, planum temporale, and planum polare, while the STS ROI was constructed by combining the pSTS, mSTS, and aSTS ROIs. Separate STP and STS ROIs were constructed for left and right hemisphere auditory cortices. In the second analysis, decoding performance for each of the 12 ROIs were compared separately.

Classification accuracy values were subjected to second-level analyses in R (version 3.6.3; RRID:SCR_001905). First, decoding between STP and STS of both hemispheres were compared using a repeated-measures 2 × 2 ANOVA with two within-participant factors; ROI and hemisphere ($\alpha = 0.05$). Follow-up comparisons were performed using paired-samples t-tests ($\alpha = 0.05$, false discovery rate [FDR]-adjusted across hemispheres). Further, decoding accuracy within each ROI was examined to confirm that decoding was above chance (50%) using a one-sample t-test against 0.5 ($\alpha = 0.05$). Note that a t-test against chance level assesses if above chance decoding is present in our sample. We provide valid population inference below through the use of permutation-based prevalence inference (Allefeld et al. 2016). Second, decoding of each of the 12 bilateral auditory cortical ROIs were separately assessed using a repeated-measures 6 × 2 ANOVA with 2 within-participant factors: ROI and hemisphere ($\alpha = 0.05$). Follow-up comparisons were performed using paired-samples t-tests ($\alpha = 0.05$, FDR-adjusted within each hemisphere). For follow-up comparisons, we used Heschl's gyrus as a reference ROI to which all of the other ROIs were compared. We reasoned that decoding within Heschl's gyrus might primarily reflect differences between emotional prosody and neutral speech in low-level acoustical features, such as fundamental frequency and timbral cues, whereas we were primarily interested in decoding of emotional cues beyond low-level acoustical features (Ethofer et al. 2009). We also checked for each ROI if decoding was above chance level (50% accuracy) using a one-sample t-test against 0.5 ($\alpha = 0.05$). Finally, separate analyses were conducted for the contrasts (sad prosody versus neutral speech) and (happy prosody versus neutral speech) to identify similarities and differences related to the valence of the emotional prosody. We report effect sizes of ANOVA effects in terms of generalized eta-squared (η^2_G) and effect sizes for t-tests as Cohen's d.

To examine decoding of multiple vocal emotions in auditory cortical regions, a multi-emotion decoding analysis was performed for the contrasts (sad prosody versus happy prosody) and (sad prosody versus happy prosody versus neutral speech). The 2-class classification for (sad prosody versus happy prosody) was identical to the previously described contrasts, which differentiated between emotional prosody and neutral speech. For (sad prosody versus happy prosody versus neutral speech), activation patterns extracted from run-wise contrast images generated by SPM served as inputs to a multiclass classification analysis as implemented in PyMVPA using a

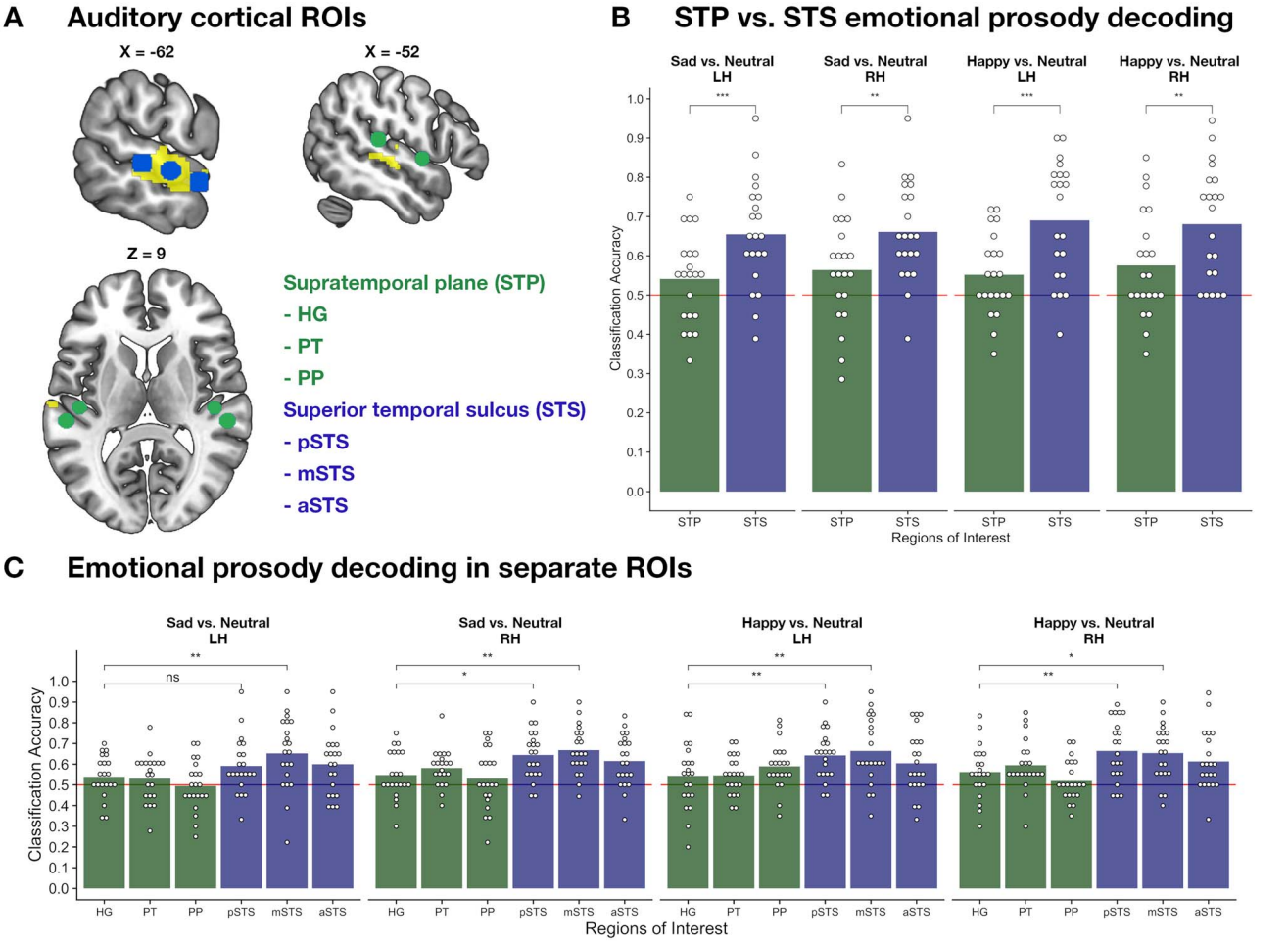

Fig. 2. ROI-based emotional prosody decoding. A) Auditory cortical brain regions included in the ROI-based emotional prosody decoding analysis. ROIs located on the STP (HG, PT, and PP) are colored in green, while ROIs located within STS (pSTS, mSTS, and aSTS) are colored in blue. B) Classification accuracies for bilateral STP and STS ROIs for the (sad prosody versus neutral speech) and (happy prosody versus neutral speech) contrasts. Results show consistently greater emotional prosody decoding in the STS compared to the STP across contrasts and hemispheres. C) Classification accuracies for all bilateral auditory cortical regions. Post hoc comparisons showed that classification accuracy within mSTS is consistently greater than those measured in HG, which served as a reference region for this analysis. Classification accuracy within pSTS was also greater than HG for 3 of the stimulus contrasts. Abbreviations: * = $P < 0.05$; ** = $P < 0.01$; *** = $P < 0.001$; HG = Heschl's gyrus; LH = left hemisphere; PP = planum polare; PT = planum temporale; RH = right hemisphere.

linear SVM in a "one-against-one" manner. All aspects of the multiclass classification analysis, including ROIs and cross-validation procedures, were identical to the 2-class classification described previously, with the exception of the classifier and chance level (33% accuracy).

Finally, we examined decoding across all bilateral voice-sensitive superior temporal cortex voxels using an ROI encompassing all statistically significant voxels of the (neutral speech > environmental sounds) GLM contrast (Ethofer et al. 2009). Apart from the ROI, all further aspects were identical to the 2-class classifications described above (see Supplementary Results).

Anatomical ROIs

To examine decoding within specific subregions of auditory cortex, ROIs encompassing bilateral superior temporal auditory areas were constructed. These regions were defined as 3 ROIs along the anterior–posterior axis of auditory cortex in both the STP and the STS, which is consistent with recent work highlighting dissociation in the functional architecture between these subregions of auditory cortex (Abrams et al. 2020). The first group of ROIs consisted of auditory areas located along the STP, including Heschl's gyrus, planum temporale, and planum polare (see Fig. 1D and Fig. 2A). These ROIs were anatomically defined based on probabilistic maps included in the Harvard-Oxford cortical atlas in FSL (RRID:SCR_002823). The center coordinate for Heschl's gyrus was placed within the medial aspect of the structure to capture the putative location of primary auditory cortex (Moerel et al. 2014). The second group of ROIs consisted of areas along the anterior–posterior extent of STS; pSTS, mSTS, and aSTS. The pSTS, mSTS, and aSTS ROIs were equidistantly placed along the full anterior–posterior extent of the cluster of statistically significant activation derived from the (neutral speech > environmental sounds) contrast in both hemispheres to capture functionally defined voice-sensitive cortex (see Fig. 2A; the significant cluster is given in yellow). The center coordinates for bilateral mSTS ROIs are

Table 2. Auditory cortex ROI coordinates for multivariate decoding and activation analyses. Coordinates are given in MNI space. Homotopic ROIs in bilateral auditory cortex were placed equidistant from the midsagittal plane on the *x*-axis.

ROI	Subdivision of auditory cortex	X	Y	Z
HG	STP	−46/46	−21	6
PT	STP	−55/55	−30	12
PP	STP	−50/50	0	−1
pSTS	STS	−60/60	−30	2
mSTS	STS	−60/60	−15	−2
aSTS	STS	−60/60	0	−8

Abbreviations: HG = Heschl's gyrus; PP = planum polare; PT = planum temporale.

proximal to peak coordinates reported by seminal papers on the localization of voice-sensitive STS (Belin et al. 2000; Pernet et al. 2015). All ROIs were constructed as nonoverlapping spheres (radius = 6 mm) in MNI space centered on the coordinates listed in Table 2.

Whole-brain searchlight MVPA

A whole-brain multivariate pattern classification method was used to examine whether brain regions beyond auditory cortex accurately discriminate emotional and neutral prosody stimuli. Therefore, a whole-brain searchlight analysis as implemented in PyMVPA was performed. This analysis used the same contrast images described for the 2- and multiclass analyses described previously (see above for details on contrast image generation). For each participant, we built whole-brain, voxel-wise maps of classification accuracies reflecting how well emotional and neutral prosody stimuli could be discriminated based on the activation pattern of a particular voxel and its surrounding voxels. Specifically, a sphere (radius = 6 mm; 123 voxels) was moved across all brain voxels. In every sphere, multivariate pattern classification was performed using identical procedures as described above for the ROI-based classification (linear SVM; leave-one-run-out cross-validation). The average cross-validated classification accuracy was recorded for the center voxel of the sphere, resulting in a whole-brain accuracy map per participant for each stimulus contrast of interest, including (sad prosody versus neutral speech) and (happy prosody versus neutral speech).

Searchlight maps were subjected to second-level analysis using permutation-based prevalence inference using the minimum statistic (Allefeld et al. 2016). This approach provides population inference regarding the proportion of participants in the population, e.g. 50+% (i.e. the majority), exhibiting above chance classification at a particular voxel in the brain. Note that a simple voxel-wise *t*-test against chance level cannot provide valid population inference because the "true" single-participant accuracy can never be below chance level (Allefeld et al. 2016). To perform prevalence inference, 100 searchlight maps were constructed for each contrast and participant using permuted class labels. Class labels were permuted within each run, and importantly, the permutation was fixed across all center voxels of a map to preserve spatial dependencies (Stelzer et al. 2013). All properties of the searchlight analysis with permuted class labels were identical to the analysis with the unpermuted labels. Subsequently, we smoothed both the searchlight maps obtained using unpermuted and permuted labels with a 6-mm FWHM Gaussian kernel. Smoothing reduces residual anatomical misalignment, and voxel-wise second-level inference critically relies on anatomical alignment between participants. Finally, searchlight maps were inputted to the prevalence inference algorithm by Allefeld et al. (2016), as implemented in MATLAB (https://github.com/allefeld/prevalence-permutation). The algorithm outputs voxel-wise FWE-corrected *P* values (Nichols and Holmes 2002), which were converted to z-scores. This procedure resulted in separate statistical maps for each contrast and identifies each voxel in the brain in which at least half of the population showed above chance decoding (chance level = 50% accuracy) of (sad prosody versus neutral speech) and (happy prosody versus neutral speech) contrasts. Clusters of statistically significant voxels ($z > 1.65$) were extracted using AtlasReader (Notter et al. 2019) in Python. We report the spatial extent (in mm^3) and the minimum FWE-corrected *P* value ($p_{FWE\text{-}min}$) of statistically significant clusters consisting of ≥5 voxels.

To examine decoding of multiple vocal emotions across the whole brain, a multi-emotion decoding analysis was performed for the contrasts (sad prosody versus happy prosody) and (sad prosody versus happy prosody versus neutral speech). The 2-class classification for (sad prosody versus happy prosody) was identical to the previously described contrasts differentiating between emotional prosody and neutral speech. For (sad prosody versus happy prosody versus neutral speech), activation patterns extracted from run-wise contrast images generated by SPM served as inputs to a multiclass classification analysis as implemented in PyMVPA using a linear SVM in a "one-against-one" manner. All aspects of the multiclass classification analysis, including cross-validation procedures and prevalence inference, were identical to the 2-class classification described previously, with the exception of the classifier and chance level (33% accuracy). The unthresholded classification accuracy maps of the 4 contrasts, (sad prosody versus neutral speech), (happy prosody versus neutral speech), (sad prosody versus happy prosody), and (sad prosody versus happy prosody versus neutral

speech), averaged across participants, are available on NeuroVault (RRID:SCR_003806; https://neurovault.org/collections/HRQEJGAZ/).

Restricted anatomical search space

Permutation-based prevalence inference provides conservative voxel-wise inference. Thus, in addition to a whole-brain analysis, we also performed an analysis restricting the search space to voxels covering regions that have been implicated in emotional prosody perception (for an overview, see Frühholz and Ceravolo 2018), including the following bilateral parcels of the Harvard-Oxford atlas: Heschl's gyrus, planum temporale, planum polare, anterior and posterior divisions of superior temporal gyrus (STG) and middle temporal gyrus (Grandjean et al. 2005; Ethofer et al. 2006, 2009, 2012), pars triangularis and pars opercularis of inferior frontal gyrus (IFG) (Frühholz et al. 2012), and the amygdala (Frühholz and Grandjean 2013) (see Supplementary Results).

ROI-based signal-level analysis

Group mean signal levels in response to emotional prosody stimuli were computed for auditory cortical brain regions. Signal level within each auditory cortical ROI (see Table 2) was calculated by extracting the *t*-value from individual participants' contrast *t*-maps for the (sad prosody > neutral speech) and (happy prosody > neutral speech) comparisons. The mean *t*-value within each ROI was computed for both contrasts in all participants (see Supplementary Results).

Brain–behavior associations

The third aim of the study was to investigate associations between children's neural decoding of emotional prosody, social communication abilities, and emotion recognition accuracy. Social communication abilities were measured with the social communication subscale of the SRS-2 (Constantino and Gruber 2012) and emotion recognition was assessed using the DANVA2 (Nowicki and Duke 1994; Baum and Nowicki 1998).

Association between neural decoding of emotional prosody and social communication

The SRS-2 is a validated and widely used parent-report rating scale that measures social behavior in children (Constantino and Gruber 2012). The social communication subscale has 22 items and assesses reciprocal communication in social situations. For the analyses using the SRS-2 social communication subscale, the sign of the standardized *t*-scores outputted by the scale was flipped to obtain a score of parent-reported social communication abilities. Note that the original standardized *t*-scores are reported in Table 1 to facilitate an assessment of overall social functioning in our sample. For the statistical analysis, the relationships between neural decoding of emotional prosody and social communication abilities (SRS-2 *t*-scores) were computed using linear mixed effects models and follow-up Pearson correlations. This brain–behavior analysis was focused on brain regions which showed the highest decoding performance from the ROI-based prosody decoding analysis. First, to examine whether an association of neural decoding and social communication abilities was present across emotion categories or was primarily present for one emotion category (i.e. sad or happy), we performed linear mixed effects modeling using the R packages "lme4" (RRID:SCR_015654) and "lmerTest" (RRID:SCR_015656). The mixed effects models included neural decoding as the dependent variable; SRS-2 scores, emotion category, and the interaction between SRS-2 and emotion category as regressors; and a random intercept for each participant. To examine whether emotions of different valence showed distinct relationships with social communication ability, the relationship of decoding of (sad prosody versus neutral speech) and (happy prosody versus neutral speech) with SRS-2 scores were separately examined using Pearson's correlation coefficient *r*. Given that auditory cortical regions revealed similar levels of classification accuracy in both hemispheres, mixed effects models and correlations were computed using the average classification accuracy across hemispheres. The significance level for all brain–behavior analyses was set to $\alpha = 0.05$, FDR-adjusted for multiple ROIs.

Association between neural decoding of emotional prosody and behavioral emotion recognition

DANVA2 measures children's ability to recognize and express nonverbal emotional information. Given that the brain measures in our study probe receptive aspects of prosody, the brain–behavior analyses focused on the 4 receptive (rather than expressive) subtests of the DANVA2 that measure recognition of emotions in (i) adult faces (subtest "adult facial expressions"), (ii) adult voices (subtest "adult paralanguage"), (iii) child faces (subtest "child facial expressions"), and (iv) child voices (subtest "child paralanguage"). These subtests have been validated for internal consistency, test–retest reliability, and construct validity (Nowicki and Duke 1994; Baum and Nowicki 1998). The adult and child facial expressions subtests consist of 24 photographs of adults and 24 photographs of children, respectively, showing happy, sad, angry, and fearful faces. The adult paralanguage subtest consists of 24 samples of a sentence spoken by adults in happy, sad, angry, and fearful prosody, while the child paralanguage subtest consists of 32 samples of a sentence spoken by children in happy, sad, angry, and fearful prosody (Nowicki 2010). A combined emotion recognition accuracy z-score was computed. Specifically, each child's raw score for each subtest, representing the total number of emotion classification errors made, was transformed to a z-score using age-specific norms provided in the DANVA2 manual (Nowicki 2010). The z-scores were then averaged across the 4 subtests, including adult facial expressions, adult paralanguage,

child facial expressions, and child paralanguage, for each participant. Finally, because the z-scores reflected recognition errors, the sign of the scores was flipped to obtain scores representing emotion recognition accuracy and therefore scores of recognition abilities rather than impairments, relative to a norm population. Our approach is consistent with empirical evidence showing that emotion recognition ability on the behavioral level is best conceptualized as a broad, mostly unitary ability (Schlegel et al. 2012; Connolly et al. 2020). This conceptualization is consistent with data from our sample in which the scores from facial and paralanguage subtests were highly correlated ($r = 0.54$, $P = 0.01$).

For the statistical analysis, the relationships between neural decoding of emotional prosody and behavioral emotion recognition (DANVA2 z-scores) were computed using linear mixed effects models. These brain–behavior analyses were focused on brain regions which showed the highest decoding performance from the ROI-based prosody decoding analysis. To examine whether an association of neural decoding and emotion recognition abilities was present across emotions or primarily present for one emotion category (i.e. sad or happy), we performed linear mixed effects modeling in R. The mixed effects model included neural decoding as the dependent variable; DANVA2 scores, emotion category, and the interaction between DANVA2 and emotion category as regressors; and a random intercept for each participant. Given that auditory cortical regions revealed similar levels of classification accuracy in both hemispheres, mixed effects models were computed using the average classification accuracy across hemispheres. The significance level for all brain–behavior analyses was set to $\alpha = 0.05$, FDR-adjusted for multiple ROIs.

Cross-validation of brain–behavior associations

To confirm the robustness of the statistically significant brain–behavior correlations, we estimated $r_{(\text{observed, predicted})}$, a cross-validated measure of how well the independent variable (e.g. SRS-2 *t*-scores) predicts the dependent variable (e.g. decoding of sad prosody versus neutral speech), using a repeated 4-fold cross-validation (100 iterations). Data points were randomly assigned to 4-folds, with the constraint that the mean values of both the independent and dependent variable did not differ across folds; this constraint was implemented by repeating the random assignment as necessary until there was no evidence for differences between folds ($P > 0.50$) according to a one-way ANOVA. Then, a linear model was built using 3 folds, leaving out the fourth, and this model was used to predict the data in the omitted fold. This procedure was repeated 4 times leaving out each fold once. Finally, the statistical significance of the model was assessed using a permutation test. The empirical null distribution of $r_{(\text{observed, predicted})}$ was estimated by generating 1,000 surrogate datasets under the null hypothesis of no association between the independent variable and the dependent variable (Cohen et al. 2010).

Control analysis: association between neural activation and behavior

To examine whether the statistically significant associations that we detected between neural decoding and broader measures of social function could also be observed by using the mean activation of the ROIs instead of classification accuracy, we computed mixed effects models and correlations (*r*) between the signal levels extracted from the (sad prosody > neutral speech) and (happy prosody > neutral speech) contrasts (see ROI-based signal-level analysis) and the behavioral measures (see Supplementary Results).

Results

Multivariate decoding of emotional prosody in anatomically distinct subdivisions of auditory cortex

We assessed decoding of emotional prosody in children within distinct subregions of auditory cortex using ROI-based multivariate pattern classification. Consistent with previous work examining prosody decoding in auditory cortex (Ethofer et al. 2009), we restricted our decoding analysis to auditory regions extending from the STP, including primary auditory cortex, to more ventral regions of auditory cortex, including STS. Voice-sensitive areas were defined by first identifying brain regions that showed greater activity for the (neutral speech > environmental sounds) GLM contrast measured across all children. Similar to canonical findings of voice-sensitive cortex from previous studies (Belin et al. 2000; Pernet et al. 2015; Abrams et al. 2016), this contrast revealed significant clusters in bilateral superior temporal cortex encompassing anterior, middle, and posterior regions of STS (Fig. 2A, yellow).

The significant clusters from the (neutral speech > environmental sounds) GLM contrast were used as the basis for the definition of STS ROIs. To enable for a greater degree of anatomical specificity in the prosody decoding analysis, ROIs were constructed in anterior, middle, and posterior subregions of STS (Deen et al. 2015). Consistent with recent work highlighting dissociation in the functional architecture between STS and more dorsally located STP (Abrams et al. 2020), we further included ROIs of anatomically defined STP subregions which are thought to encompass more primary regions of auditory cortex, including Heschl's gyrus, planum polare, and planum temporale (Hickok and Poeppel 2007; Rauschecker and Scott 2009). Multivariate decoding of emotional prosody stimuli was then performed using these 6 bilateral ROIs.

Results from multivariate decoding analyses revealed a striking and consistent pattern of regional specificity within auditory cortex for discriminating emotional

prosody stimuli. We first examined whether decoding of emotional prosody differed between STP and STS subregions. This analysis was performed by first combining the voxels from the 3 STP ROIs in each hemisphere, then separately combining the three STS ROIs in each hemisphere, and comparing classification accuracies between these merged ROIs with a 2 × 2 (ROI × hemisphere) repeated-measures ANOVA. For both of the primary contrasts of interest, including (sad prosody versus neutral speech) and (happy prosody versus neutral speech), results revealed a main effect of ROI (sad vs. neutral: $F(1, 20) = 17.27$, $P < 0.001$, $\eta^2_G = 0.15$; happy vs. neutral: $F(1, 20) = 28.47$, $P < 0.001$, $\eta^2_G = 0.18$) but no effect of hemisphere or ROI × hemisphere interaction. Post hoc comparisons revealed that STS showed greater emotional prosody decoding than the STP in both hemispheres and for both investigated contrasts ($P < 0.009$, $q_{FDR} < 0.01$ for all comparisons). With the exception of left STP for the (sad prosody versus neutral speech) contrast, all other ROIs decoded above chance level (all $P < 0.05$; Fig. 2B).

In a second analysis, we further examined regional specificity within auditory cortex for decoding emotional prosody stimuli by comparing classification accuracies between all 6 bilateral superior temporal cortex ROIs. For both of the primary contrasts of interest, including (sad prosody versus neutral speech) and (happy prosody versus neutral speech), a 6 × 2 (ROI × hemisphere) repeated-measures ANOVA revealed a main effect of ROI (sad vs. neutral: $F(5, 100) = 9.36$, $P < 0.001$, Greenhouse–Geisser-corrected P value (p_{GG}) < 0.001, $\eta^2_G = 0.14$; happy vs. neutral: $F(5, 100) = 6.91$, $P < 0.001$, $p_{GG} < 0.001$, $\eta^2_G = 0.10$) but no effect of hemisphere or ROI × hemisphere interaction. For the follow-up comparisons within each hemisphere, we used Heschl's gyrus, which comprises primary auditory cortex, as a reference ROI. Results from the (sad prosody versus neutral speech) contrast revealed greater emotional prosody classification accuracies in left hemisphere mSTS ($P = 0.006$, $q_{FDR} = 0.03$, $d = 0.67$) as well as right hemisphere pSTS ($P = 0.01$, $q_{FDR} = 0.03$, $d = 0.61$) and mSTS ($P = 0.002$, $q_{FDR} = 0.01$, $d = 0.78$) compared to Heschl's gyrus in their respective hemisphere (Fig. 2C). The (happy prosody versus neutral speech) contrast similarly revealed greater classification accuracies in left hemisphere pSTS ($P = 0.006$, $q_{FDR} = 0.02$, $d = 0.67$) and mSTS ($P = 0.004$, $q_{FDR} = 0.02$, $d = 0.71$) and right hemisphere pSTS ($P = 0.006$, $q_{FDR} = 0.03$, $d = 0.66$), with a trend toward greater decoding accuracy in right hemisphere mSTS ($P = 0.028$, $q_{FDR} = 0.07$, $d = 0.52$), compared to Heschl's gyrus (Fig. 2C). By contrast, neither the planum polare nor planum temporale of the STP showed greater decoding accuracy compared to Heschl's gyrus. We next examined whether and which ROIs showed above chance classification accuracy for decoding emotional prosody stimuli, and results revealed that the 3 STS ROIs, including pSTS, mSTS, and aSTS, showed classification accuracies that were greater than chance across both stimulus contrasts and hemispheres ($P < 0.007$ for all comparisons). By contrast, not all STP ROIs consistently showed above chance classification accuracies (see Supplementary Table 2 for details). A multi-emotion analysis, which simultaneously examined classification of (sad prosody versus happy prosody versus neutral speech) in auditory cortical areas, revealed similar results. We did not detect above chance classification accuracies for (sad prosody versus happy prosody), which suggests that the 3-class classification was primarily driven by the contrasts between emotional prosody and neutral speech. To complement these regionally specific classification analyses, we next examined classification across all voice-sensitive superior temporal cortex voxels, similar to previous work (Ethofer et al. 2009). Classification results from this analysis were consistent with those obtained from the regionally specific analysis and show above chance classification accuracies for decoding emotional prosody compared to neutral stimuli, however, results did not identify above chance classification accuracies for (sad prosody versus happy prosody). Detailed results and a discussion of the multi-emotion decoding analyses are given in the Supplementary Material.

Multivariate decoding of emotional prosody across the whole brain

The second major goal of the data analysis was to examine whether brain areas outside of superior temporal cortex accurately decode emotional prosody stimuli. We therefore performed a whole-brain searchlight analysis with permutation-based prevalence inference to identify brain regions that showed above chance decoding of emotional versus neutral prosody. The spatial extent (in mm^3) and the minimum FWE-corrected P value ($p_{FWE\text{-}min}$) of statistically significant clusters consisting of ≥5 voxels are reported. Consistent with the results from ROI-based analysis of auditory cortex, results from the whole-brain analysis revealed that multiple regions of superior temporal cortex discriminate emotional prosody stimuli from neutral speech (see Supplementary Table 3 for MNI coordinates). For the (sad prosody versus neutral speech) contrast, the whole-brain searchlight identified a cluster in the right pSTS (112 mm^3, $p_{FWE\text{-}min} = 0.006$). No significant clusters outside of superior temporal cortex were identified for the (sad prosody versus neutral speech) contrast (Fig. 3A). Furthermore, for the (happy prosody versus neutral speech) contrast, the whole-brain analysis identified multiple clusters in STS, including right (576 mm^3, $p_{FWE\text{-}min} = 0.005$) and left mSTS (544 mm^3, $p_{FWE\text{-}min} = 0.005$) and right pSTS (80 mm^3, $p_{FWE\text{-}min} = 0.008$). Additional clusters were identified in the right planum temporale (240 mm^3, $p_{FWE\text{-}min} = 0.008$) and the left posterior superior temporal gyrus (pSTG; 224 mm^3, $p_{FWE\text{-}min} = 0.006$). Whole-brain analysis for the (happy prosody versus neutral speech) contrast additionally revealed a large cluster in the left central operculum (528 mm^3, $p_{FWE\text{-}min} = 0.005$) and several smaller clusters located in the left parahippocampus

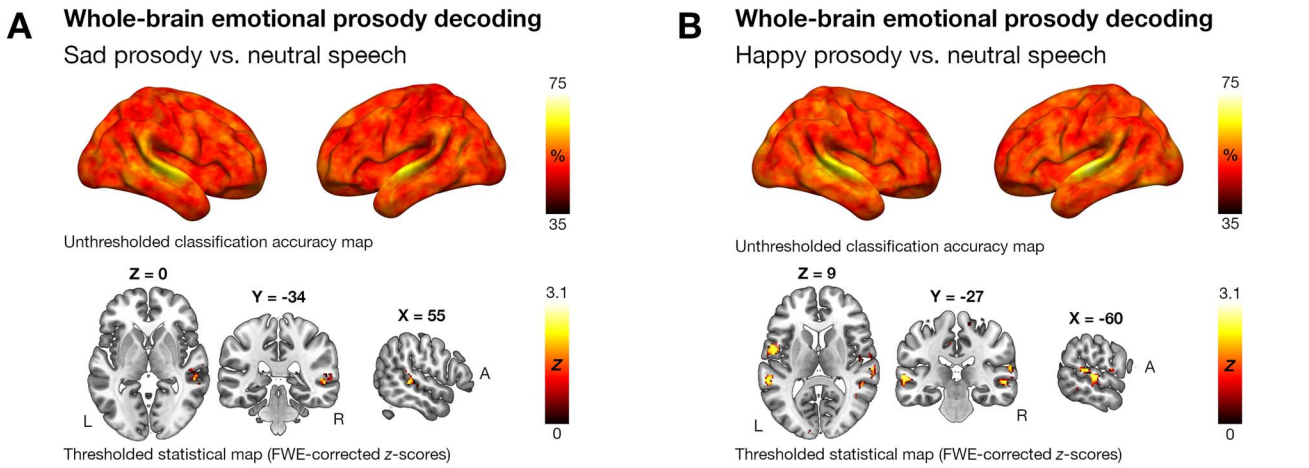

Fig. 3. Searchlight-based prosody decoding. Chance level for the 2-class emotional prosody decoding was at 50%. A) For the (sad prosody versus neutral speech) contrast, the whole-brain searchlight identified a significant cluster in the right pSTS. No significant clusters outside of superior temporal cortex were identified for the (sad prosody versus neutral speech) contrast. B) For the (happy prosody versus neutral speech) contrast, the whole-brain analysis identified multiple clusters in bilateral STS and STG. In addition, a large cluster in the left central operculum was identified. Abbreviations: FWE = family-wise error.

(80 mm^3, $p_{FWE\text{-}min} = 0.01$), the right supramarginal gyrus (48 mm^3, $p_{FWE\text{-}min} = 0.01$), and the left cuneus (40 mm^3, $p_{FWE\text{-}min} = 0.008$) (see Fig. 3B). Importantly, whole-brain searchlight results for both stimulus contrasts did not identify significant clusters in either IFG or the amygdala. Moreover, voxel-wise analyses restricting the search space to IFG and the amygdala in addition to superior temporal cortex also did not reveal significant decoding of emotional prosody stimuli in IFG and amygdala despite the less conservative FWE-correction (see Supplementary Table 3). A multi-emotion searchlight decoding analysis, which simultaneously examined classification of (sad prosody versus happy prosody versus neutral speech) across the whole brain, revealed similar results. We did not detect statistically significant above chance classification accuracies in any clusters across the brain for (sad prosody versus happy prosody), which suggests that the whole-brain 3-class classification was primarily driven by the contrasts between emotional prosody and neutral speech. Detailed results and a discussion of the multi-emotion whole-brain searchlight analyses are given in the Supplementary Material.

Association between neural decoding of emotional prosody and social communication ability

The next major goal of the data analysis was to examine whether neural measures of prosody decoding are related to children's social communication abilities. Given the extensive experimental and theoretical literatures implicating the STS as a key node for voice-related processes, and the finding of consistent decoding of emotional prosody in the STS using ROI-based analyses (Fig. 2) and whole-brain classification approaches (Fig. 3), the brain–behavior analysis focused on decoding in STS. Based on the results of the ROI-based decoding analyses, we specifically focused on neural decoding in mSTS and pSTS, the regions which showed the highest classification accuracies in differentiating emotional prosody from neutral speech. We performed a linear mixed effects analysis to examine associations between mSTS decoding and social communication abilities, measured using the SRS-2, across emotion categories (i.e. sad and happy). This analysis revealed a statistically significant main effect of social communication abilities on mSTS decoding ($t(27.05) = 2.85$, $P = 0.008$, $q_{FDR} = 0.02$) and an interaction between social communication abilities and emotion category ($t(21) = -3.32$, $P = 0.003$, $q_{FDR} = 0.006$). Further analyses revealed that the relationship between mSTS decoding and social communication abilities was significant for sad but not happy prosody: We found a positive association between decoding of (sad prosody versus neutral speech) in the mSTS and social communication abilities ($r = 0.56$, $P = 0.009$, $q_{FDR} = 0.02$). Specifically, children with greater emotional prosody decoding in the STS had greater social communication skills (Fig. 4). The 4-fold cross-validation confirmed the robustness of the positive relationship between mSTS decoding of (sad prosody versus neutral speech) and social communication abilities, $r_{(observed,\ predicted)} = 0.50$, $P = 0.002$. Correlation analysis with decoding of (happy prosody versus neutral speech) failed to uncover a significant relationship ($r = 0.10$, $P = 0.68$). No brain–behavior relationships involving social communication abilities were detected in the pSTS (main effect: $t(33.11) = 0.82$, $P = 0.42$).

Association between neural decoding of emotional prosody and emotion recognition

We examined whether individual differences in neural measures of prosody decoding are related to behavioral emotion recognition accuracy measured with the DANVA2. We performed a linear mixed effects analysis to determine associations between mSTS decoding and

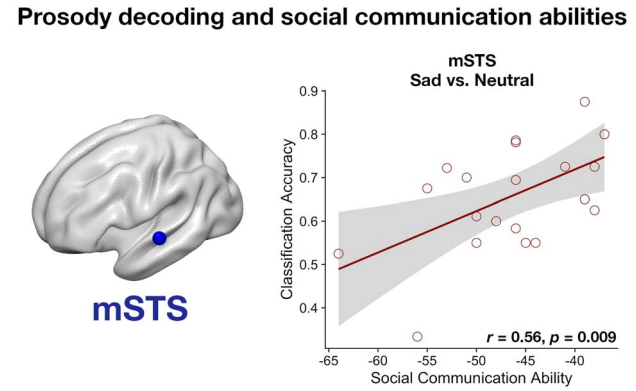

Fig. 4. Neural decoding of sad prosody predicts social communication abilities. Results revealed a significant positive association between decoding of the (sad prosody versus neutral speech) contrast in the mSTS and social communication abilities. Specifically, more accurate emotional prosody decoding in the mSTS was associated with greater social communication abilities in children.

DANVA2 emotion recognition accuracy across emotion categories (i.e. sad and happy). Results from this analysis showed a trend toward a main effect of DANVA2 emotion recognition accuracy on mSTS decoding ($t(30.80) = 1.93$, $P = 0.06$) and no DANVA2 × emotion category interaction $t(21) = 0.72$, $P = 0.48$). No significant brain–behavior relationships involving emotion recognition were detected in the pSTS (main effect: $t(35.45) = 1.17$, $P = 0.25$).

Discussion

Understanding the emotional state of a speaker is an essential skill for navigating the social world and empathizing with others, which is critical for forming and strengthening relationships (Pell and Kotz 2021). Little is known regarding the brain systems underlying emotional prosody processing in children and whether it is related to their social abilities. We examined these questions in school-age children and found that, similar to reports from studies of adult listeners, bilateral voice-sensitive regions of the STS decode emotional prosody information. Surprisingly, brain regions outside the superior temporal cortex failed to reliably decode this information. Crucially, decoding in mSTS was positively related to social communication abilities, and this relationship was specific to sad prosody; more accurate decoding of sad prosody stimuli in the mSTS was predictive of greater social communication abilities in children. Findings bridge an important theoretical gap by showing that the auditory system's ability to detect emotional cues produced by the voice is predictive of a child's social skills, including the ability to relate and interact with others.

Voice-sensitive STS shows greater decoding of emotional prosody information compared to more primary regions of auditory cortex

A major finding of the current study is that voice-sensitive regions of the STS decode emotional prosody information in children, a result that was consistently identified across ROIs, whole-brain classification analyses, and emotional prosody contrasts. Moreover, STS regions showed greater decoding of emotional prosody relative to auditory processing regions of the STP, including Heschl's gyrus, which contains primary auditory cortex, as well as planum polare and planum temporale. The STS has long been implicated as the primary voice sensitive region of the cerebral cortex based on the fact that this brain region consistently shows greater activation in response to human vocal sounds compared to nonvocal control sounds (Belin et al. 2000; Pernet et al. 2015). By contrast, regions of the STP typically show comparable activation profiles for vocal and nonvocal sounds, which is consistent with the hypothesis that these low-level regions of auditory cortex are responsible for processing spectro-temporal acoustical features irrespective of whether they contain vocal information (Hickok and Poeppel 2007; Okada et al. 2010; De Heer et al. 2017). Based on its sensitivity for the human voice, models of voice perception have consistently implicated the STS as a hub for distributing vocal information to other brain systems for subsequent emotional, reward, and cognitive processing (Belin et al. 2004; Young et al. 2020). Additional models have further implicated the STS in more subtle aspects of human voice processing, including the discrimination of emotional prosodic stimuli (Schirmer and Kotz 2006; Wildgruber et al. 2006; Brück, Kreifelts, and Wildgruber 2011; Grandjean 2021). For example, a previous study in adults showed that vocal emotions are decoded from auditory cortex (Ethofer et al. 2009). However, this study did not differentiate between contributions of auditory cortical regions of the STP and more ventral regions of auditory cortex within STS. This differentiation is critical given that these 2 subregions of auditory cortex differ in cytoarchitecture (Zachlod et al. 2020) and myelination (Glasser et al. 2016), have differential intrinsic functional (Abrams et al. 2020) and structural connectivity profiles (Turken and Dronkers 2011), and importantly, perform distinct computations as shown by voxel-wise computational modeling (Norman-Haignere and McDermott 2018). Here, we show that STS regions show greater accuracy for decoding emotional prosody information relative to regions of the STP, where rudimentary spectro-temporal acoustical features are likely to be decoded from the speech signal. Findings from the current study support the hypothesis that emotional prosody decoding in the STS reflects higher-order operations that form the basis for the recognition of emotions expressed through the speaker's voice. Specifically, we hypothesize that the STS categorizes vocal-emotional cues using a form of template matching by which patterns of amplitude, pitch, and duration features in a vocal stimulus are integrated over time and are subsequently matched to templates of vocal patterns which are associated with specific human emotions. Importantly, this model proposes distinct roles for different auditory cortical regions, with structures of the STP underlying

spectro-temporal processing of incoming speech signals and the STS integrating these signals over time and matching specific neural patterns of activity to distinct emotional categories.

Decoding of emotional prosody in prefrontal cortex and amygdala

While results from the current study showed that decoding of emotional prosody was largely restricted to auditory processing regions of superior temporal cortex, one exception was the finding of above chance decoding of happy prosody in the left central operculum region. This finding is intriguing given that this region in the ventral portion of motor cortex represents the larynx, the organ housing the vocal folds which are the sound source of vocalizations (Penfield and Boldrey 1937; Belyk and Brown 2017). Previous research has reported motor involvement in prosody and vocal emotion perception. For example, studies have provided causal evidence in adults (Banissy et al. 2010; Sammler et al. 2015) and correlative evidence in children (Correia et al. 2019), supporting a role for the motor system in the perception of prosodic and vocal-emotional cues. In particular, left ventral motor cortex activation has been reported during the perception of positive vocal emotions and has been interpreted as preparation for responsive gestures (e.g. laughter) in response to happy stimuli (Warren et al. 2006). While the passive listening design in our study is unable to identify a precise role for the motor cortex in prosody perception, findings from the current study add to the expanding empirical and theoretical literatures that highlight a key role for motor systems in voice and speech perception.

A notable pattern of findings from the emotional prosody literature involves studies reporting increased activation of the IFG (Buchanan et al. 2000; Frühholz et al. 2012) and amygdala (Bach et al. 2008; Frühholz and Grandjean 2013) in response to emotional prosody stimuli. Results from both ROI-based and whole-brain classification analyses in our study did not provide evidence for emotional prosody decoding in IFG. Theoretical accounts have suggested that the IFG is engaged when participants are required to explicitly evaluate and categorize emotional prosody stimuli (Schirmer and Kotz 2006). Therefore, a plausible explanation as to why the IFG did not decode emotional prosody stimuli in the current study is that participants passively attended to the speech stimuli and did not perform an emotion categorization task. On the other hand, the contribution of the amygdala to emotional prosody processing has remained elusive, with some studies showing differential activation for vocal emotional stimuli in the amygdala (Schirmer et al. 2008; Frühholz and Grandjean 2013) while other studies have not shown effects in this brain region (Ethofer et al. 2006; Warren et al. 2006; Kotz et al. 2013). Given that the amygdala has been implicated in both implicit and explicit processing of emotional prosody (Frühholz et al. 2012), it is unclear why many studies, including the current study, have failed to reveal differential activation or decoding of emotional prosody in the amygdala (Schirmer 2018). A plausible hypothesis is that amygdala activation only occurs for vocal emotions which signal a credible threat to the individual (Frühholz and Grandjean 2013). As our study did not include such stimuli, this hypothesis could be further explored in future studies.

Emotional prosody processing in different cohorts across the lifespan

Findings from the current study provide new information germane to the neurodevelopment of emotional prosody perception. From the first days of life, children are highly attuned to human vocal sounds (DeCasper and Fifer 1980), and much of an infant's acoustical input comes from her primary caregivers, who often use infant-directed speech ("motherese") which is characterized by exaggerated prosody (i.e. speech melody) to convey emotionality (Trainor et al. 2000). Thus, children are exposed to emotional prosody information in speech from a very young age. Behavioral studies have shown that the ability to detect and discriminate emotional prosody cues becomes evident within the first year of a child's life, at a time when these young listeners also show a preference for the sound structure of their native language. The ability to identify and discriminate emotional prosody cues continue to be refined throughout childhood (Chronaki et al. 2018), with children reaching adult-like levels of vocal emotion recognition by late adolescence (Grosbras et al. 2018; Morningstar et al. 2018; Amorim et al. 2019). Studies investigating the brain bases of emotional prosody perception have focused primarily on infants, and results have revealed that newborns (Zhang et al. 2019) and 7-month-old infants (Grossmann et al. 2010) show temporal lobe activation increases in response to emotional prosody. Importantly, these infant studies employed functional near-infrared spectroscopy which lacks the spatial resolution to both resolve differential contributions of low-level and extended auditory cortex areas to emotional prosody decoding and to examine contributions of cortical regions beyond auditory cortex. Despite middle and late childhood being a crucial period of development for refining sensitivity for emotional prosody, only 2 studies have examined the brain bases for prosody perception within this age range (Morningstar et al. 2019, 2020). These studies focused on age-related effects and showed that children aged 8–19 years reveal age-related increases in activation for emotional prosody stimuli in prefrontal cortical regions. Importantly, previous studies did not directly contrast emotional and neutral prosody stimuli (Morningstar et al. 2019, 2020). Therefore, a major gap in our understanding of the neurodevelopment of emotional prosody perception is that it is unknown whether children in this age range, who often show adult-like proficiency for identifying and discriminating emotional prosody stimuli, rely on similar or different

neural resources to decode these stimuli. Findings from the current study inform this neurodevelopmental literature by showing that children between 7 and 12 years of age reveal adult-like patterns of emotional prosody decoding in the STS and that regions outside of superior temporal cortex do not consistently decode these stimuli. Findings suggest that adult-like neural mechanisms underlying emotional prosody decoding in superior temporal cortex are present by middle childhood, even as these perceptual systems continue to be refined into late childhood and adolescence (Chronaki et al. 2012; Morningstar et al. 2018).

Emotional prosody decoding is related to social abilities

A major finding from the current study is that greater decoding of emotional prosody in STS is related to greater scores on standardized measures of social communication. When children interact with caregivers and peers, extracting and interpreting the prosodic cues embedded in the speech signal, which are indicative of the speaker's emotional state, is crucial for meaningful social interactions (Lemerise and Arsenio 2000; Pell and Kotz 2021). Moreover, an inability to understand the emotion a speaker is expressing creates barriers to effective social communication and may impact key aspects of relationship building, including the ability to empathize and show compassion to a communication partner. Importantly, typically developing children show a wide range of abilities with regard to both emotion recognition (Nowicki and Duke 1994) and social communication (Constantino and Todd 2003), and theoretical models (Keltner and Haidt 1999; Van Kleef 2009) have posited a link between a person's ability to decode emotions of other individuals and the quality of their social interactions (Chronaki et al. 2015; Neves et al. 2021). Findings from the current study linking neural prosody decoding in the STS and social communication abilities in children have important theoretical implications. Specifically, an assumption of previous studies is that behavioral and neural measures of prosody processing index a crucial aspect of social function: the ability to interpret prosodic cues during discourse as a means of improving communication and relating to the speaker. Importantly, previous neural studies of prosody processing have not examined relationships to social communication abilities and therefore have been unable to link experimental findings with skills that impact social function in everyday life. Findings from the current study advance our understanding of emotional prosody processing by showing that neural discrimination of these vocal cues predicts broader measures of social function which reflect the ability of individuals to make and sustain social connections. Results add to an emerging literature showing that sensitivity of auditory cortical areas to dissociable aspects of vocal information, including processing of a mother's voice compared to unfamiliar voices (Abrams et al. 2016, 2019), is linked to social communication abilities in children. Results highlight the importance of "tuning in" to vocal cues for strengthening social connections, which is crucial for children's well-being.

Results from brain–behavior analyses revealed anatomical specificity within the STS in which effects were restricted to the mSTS relative to more posterior aspects of the STS. This result suggests a privileged role for the mSTS within extended auditory cortex for the extraction and interpretation of prosodic cues that are crucial for children's everyday social functioning. Empirical evidence regarding the functional anatomy within STS suggests that mSTS is especially tuned for vocal sounds in general, and for speech prosody in particular, when compared to nonvocal sounds (Belin et al. 2000; Liebenthal et al. 2014; Pernet et al. 2015; Deen et al. 2020). We confirmed this using automated meta-analysis on the term "vocal" using Neurosynth, which found that mSTS was consistently and most strongly activated by vocal sounds across previous studies (see Supplementary Results). Moreover, while pSTS is also involved in speech and voice processing (Pernet et al. 2015), the functional anatomy of this structure is much more diverse compared to mSTS and includes functions as diverse as perception of facial expressions and their integration with voices (Watson et al. 2014), expression of emotional states, nonsocial audiovisual integration, biological motion perception, and theory of mind (Hein and Knight 2008; Deen et al. 2015). Results from the current study provide new evidence for regional specificity within the STS and suggest that the mSTS has a pronounced role for decoding a range of vocal signals relative to pSTS, which has stronger associations with multisensory integrative processes.

Finally, the link between children's social communication abilities and neural processing of emotional speech stimuli was specific to decoding of sad prosody and was not present for decoding of happy prosody. These results suggest that decoding of sad prosody is particularly relevant for social functioning in children. A link between decoding sad emotional information and broader measures of social skills has been previously reported in individuals with autism. These studies showed that individuals with autism were specifically impaired in behaviorally decoding sad emotions from visual stimuli and decoding of sad emotions correlated with social function in affected children (Williams and Gray 2013), adolescents (Wallace et al. 2011), and adults (Boraston et al. 2007). Importantly, all of these studies showed specificity for sad stimuli and the same relationships were not found for happy stimuli. We add to this literature by showing a relationship between decoding sad emotions in voices and social functioning in typically developing children. Together, these findings suggest that understanding when a communication partner is feeling sad might be crucial for building and maintaining interpersonal connections through the provision of empathy and support.

Task-based neuroimaging of auditory social information processing in school-age children

Examining the functional brain bases of vocal and social–emotional information processing in children represents a crucial approach for understanding why some children excel, while others struggle, at decoding this information during communication. Nevertheless, surprisingly little task-based fMRI research has been performed in children in the context of auditory social information processing. Importantly, data collection for task-based neuroimaging studies in relatively young children represents a significant challenge compared to studies in adult populations and resting-state or structural MRI in children (Yerys et al. 2009; Yuan et al. 2009). Importantly, resting-state and structural MRI studies cannot address specific research questions related to the neural decoding of emotional prosody and its associations with broader social function in children. Crucially, an additional consideration is that the robustness and replicability of findings in task-based neuroimaging is not only dependent on the number of participants but also on the amount of individual participant-level data (Baker et al. 2021). A recent report demonstrated that sample sizes comparable to the number of participants in this study yield replicable results with only 4 runs of fMRI data with a similar number of trials per run (Nee 2019). In comparison, we required that each child participant had at least 7 runs with 39 trials per run that met our rigorous scanner movement inclusion criteria. Nevertheless, future studies with larger samples, and multiple runs as used here, are needed to ensure the replicability of the findings reported here.

Conclusion

In conclusion, we have identified brain systems instantiated in the STS that decode emotional prosody in typically developing children, and decoding is linked to social communication skills in these children. Findings suggest that decoding emotional information from vocal cues is a crucial component of social function in children and that "tuning in" to these vocal cues facilitates the formation and strengthening of children's social bonds. Our findings provide a neurobiological template for investigating emotional prosody decoding in children with clinical disorders, such as autism, who show insensitivity to emotional prosody and voices more generally.

Acknowledgments

We thank all the children and their parents who participated in our study and the staff at the Stanford Lucas Center for Imaging for assistance with data collection. We thank Dawlat El-Said and Carlo de los Angeles for assistance with data analysis and K. D'Arcey for help with stimulus production.

Supplementary material

Supplementary material is available at *Cerebral Cortex Journal* online.

Funding

This work was supported by National Institutes of Health grants (K01MH102428 to D.A.A., R21DC011095 to V.M. and D.A.A., R21DC017950 and R21DC017950-S1 to D.A.A. and V.M., and R01MH084164 and R01MH121069 to V.M.); the Brain and Behavior Research Foundation (NARSAD Young Investigator Grant to D.A.A.); the Singer Foundation; the Simons Foundation/SFARI (308939 to V.M.); and the Swiss National Science Foundation (P2ZHP1_187704 to S.L.).

Conflict of interest statement: None declared.

References

Abrams DA, Bhatara A, Ryali S, Balaban E, Levitin DJ, Menon V. Decoding temporal structure in music and speech relies on shared brain resources but elicits different fine-scale spatial patterns. *Cereb Cortex*. 2011:21:1507–1518.

Abrams DA, Ryali S, Chen T, Balaban E, Levitin DJ, Menon V. Multivariate activation and connectivity patterns discriminate speech intelligibility in Wernicke's, Broca's, and Geschwind's areas. *Cereb Cortex*. 2013:23:1703–1714.

Abrams DA, Chen T, Odriozola P, Cheng KM, Baker AE, Padmanabhan A, Ryali S, Kochalka J, Feinstein C, Menon V. Neural circuits underlying mother's voice perception predict social communication abilities in children. *Proc Natl Acad Sci U S A*. 2016:113:6295–6300.

Abrams DA, Padmanabhan A, Chen T, Odriozola P, Baker AE, Kochalka J, Phillips JM, Menon V. Impaired voice processing in reward and salience circuits predicts social communication in children with autism. *elife*. 2019:8:e39906.

Abrams DA, Kochalka J, Bhide S, Ryali S, Menon V. Intrinsic functional architecture of the human speech processing network. *Cortex*. 2020:129:41–56.

Allefeld C, Görgen K, Haynes J-D. Valid population inference for information-based imaging: from the second-level t-test to prevalence inference. *NeuroImage*. 2016:141:378–392.

Amorim M, Anikin A, Mendes AJ, Lima CF, Kotz SA, Pinheiro AP. Changes in vocal emotion recognition across the life span. *Emotion*. 2019:21:315–325.

Bach DR, Grandjean D, Sander D, Herdener M, Strik WK, Seifritz E. The effect of appraisal level on processing of emotional prosody in meaningless speech. *NeuroImage*. 2008:42:919–927.

Baker DH, Vilidaite G, Lygo FA, Smith AK, Flack TR, Gouws AD, Andrews TJ. Power contours: optimising sample size and precision in experimental psychology and human neuroscience. *Psychol Methods*. 2021:26:295–314.

Banissy MJ, Sauter DA, Ward J, Warren JE, Walsh V, Scott SK. Suppressing sensorimotor activity modulates the discrimination of auditory emotions but not speaker identity. *J Neurosci*. 2010:30:13552–13557.

Banse R, Scherer KR. Acoustic profiles in vocal emotion expression. *J Pers Soc Psychol*. 1996:70:614–636.

Baum KM, Nowicki S. Perception of emotion: measuring decoding accuracy of adult prosodic cues varying in intensity. *J Nonverbal Behav*. 1998:22:89–107.

Beaucousin V, Lacheret A, Turbelin M-R, Morel M, Mazoyer B, Tzourio-Mazoyer N. FMRI study of emotional speech comprehension. *Cereb Cortex*. 2007:17:339–352.

Belin P, Zatorre RJ, Lafaille P, Ahad P, Pike B. Voice-selective areas in human auditory cortex. *Nature*. 2000:403:309–312.

Belin P, Fecteau S, Bédard C. Thinking the voice: neural correlates of voice perception. *Trends Cogn Sci*. 2004:8:129–135.

Belyk M, Brown S. The origins of the vocal brain in humans. *Neurosci Biobehav Rev*. 2017:77:177–193.

Ben-David BM, van PHHM L, Leszcz T. A resource of validated affective and neutral sentences to assess identification of emotion in spoken language after a brain injury. *Brain Inj*. 2011:25:206–220.

Blasi A, Mercure E, Lloyd-Fox S, Thomson A, Brammer M, Sauter D, Deeley Q, Barker GJ, Renvall V, Deoni S, et al. Early specialization for voice and emotion processing in the infant brain. *Curr Biol*. 2011:21:1220–1224.

Boraston Z, Blakemore S-J, Chilvers R, Skuse D. Impaired sadness recognition is linked to social interaction deficit in autism. *Neuropsychologia*. 2007:45:1501–1510.

Brück C, Kreifelts B, Kaza E, Lotze M, Wildgruber D. Impact of personality on the cerebral processing of emotional prosody. *NeuroImage*. 2011:58:259–268.

Brück C, Kreifelts B, Wildgruber D. Emotional voices in context: a neurobiological model of multimodal affective information processing. *Phys Life Rev*. 2011:8:383–403.

Buchanan TW, Lutz K, Mirzazade S, Specht K, Shah NJ, Zilles K, Jäncke L. Recognition of emotional prosody and verbal components of spoken language: an fMRI study. *Cogn Brain Res*. 2000:9:227–238.

Ceravolo L, Frühholz S, Grandjean D. Proximal vocal threat recruits the right voice-sensitive auditory cortex. *Soc Cogn Affect Neurosci*. 2016:11:793–802.

Chronaki G, Broyd S, Garner M, Hadwin JA, Thompson MJJ, Sonuga-Barke EJS. Isolating N400 as neural marker of vocal anger processing in 6–11-year old children. *Dev Cogn Neurosci*. 2012:2:268–276.

Chronaki G, Garner M, Hadwin JA, Thompson MJJ, Chin CY, Sonuga-Barke EJS. Emotion-recognition abilities and behavior problem dimensions in preschoolers: evidence for a specific role for childhood hyperactivity. *Child Neuropsychol*. 2015:21:25–40.

Chronaki G, Wigelsworth M, Pell MD, Kotz SA. The development of cross-cultural recognition of vocal emotion during childhood and adolescence. *Sci Rep*. 2018:8:8659.

Cohen JR, Asarnow RF, Sabb FW, Bilder RM, Bookheimer SY, Knowlton BJ, Poldrack RA. Decoding developmental differences and individual variability in response inhibition through predictive analyses across individuals. *Front Hum Neurosci*. 2010:4.

Connolly HL, Lefevre CE, Young AW, Lewis GJ. Emotion recognition ability: evidence for a supramodal factor and its links to social cognition. *Cognition*. 2020:197:104166.

Constantino JN, Gruber CP. *Social responsiveness scale: SRS-2*. Los Angeles, CA: Western Psychological Services; 2012

Constantino JN, Todd RD. Autistic traits in the general population: a twin study. *Arch Gen Psychiatry*. 2003:60:524–530.

Correia AI, Branco P, Martins M, Reis AM, Martins N, Castro SL, Lima CF. Resting-state connectivity reveals a role for sensorimotor systems in vocal emotional processing in children. *NeuroImage*. 2019:201:116052.

De Heer WA, Huth AG, Griffiths TL, Gallant JL, Theunissen FE. The hierarchical cortical organization of human speech processing. *J Neurosci*. 2017:37:6539–6557.

DeCasper AJ, Fifer WP. Of human bonding: newborns prefer their mothers' voices. *Science*. 1980:208:1174–1176.

Deen B, Koldewyn K, Kanwisher N, Saxe R. Functional organization of social perception and cognition in the superior temporal sulcus. *Cereb Cortex*. 2015:25:4596–4609.

Deen B, Saxe R, Kanwisher N. Processing communicative facial and vocal cues in the superior temporal sulcus. *NeuroImage*. 2020:221:117191.

Ethofer T, Anders S, Wiethoff S, Erb M, Herbert C, Saur R, Grodd W, Wildgruber D. Effects of prosodic emotional intensity on activation of associative auditory cortex. *Neuroreport*. 2006:17:249–253.

Ethofer T, Van De Ville D, Scherer K, Vuilleumier P. Decoding of emotional information in voice-sensitive cortices. *Curr Biol*. 2009:19:1028–1033.

Ethofer T, Bretscher J, Gschwind M, Kreifelts B, Wildgruber D, Vuilleumier P. Emotional voice areas: anatomic location, functional properties, and structural connections revealed by combined fMRI/DTI. *Cereb Cortex*. 2012:22:191–200.

Etzel JA, Valchev N, Keysers C. The impact of certain methodological choices on multivariate analysis of fMRI data with support vector machines. *NeuroImage*. 2011:54:1159–1167.

Flom R, Bahrick LE. The development of infant discrimination of affect in multimodal and unimodal stimulation: the role of intersensory redundancy. *Dev Psychol*. 2007:43:238.

Frühholz S, Ceravolo L. The neural network underlying the processing of affective vocalizations. In: Frühholz S, Belin P, editors. *The Oxford handbook of voice perception*. Oxford University Press; 2018

Frühholz S, Grandjean D. Amygdala subregions differentially respond and rapidly adapt to threatening voices. *Cortex*. 2013:49:1394–1403.

Frühholz S, Ceravolo L, Grandjean D. Specific brain networks during explicit and implicit decoding of emotional prosody. *Cereb Cortex*. 2012:22:1107–1117.

Frühholz S, Hofstetter C, Cristinzio C, Saj A, Seeck M, Vuilleumier P, Grandjean D. Asymmetrical effects of unilateral right or left amygdala damage on auditory cortical processing of vocal emotions. *Proc Natl Acad Sci*. 2015:112:1583–1588.

Giordano BL, Whiting C, Kriegeskorte N, Kotz SA, Gross J, Belin P. The representational dynamics of perceived voice emotions evolve from categories to dimensions. *Nat Hum Behav*. 2021:5:1203–1213.

Glasser MF, Coalson TS, Robinson EC, Hacker CD, Harwell J, Yacoub E, Ugurbil K, Andersson J, Beckmann CF, Jenkinson M, et al. A multimodal parcellation of human cerebral cortex. *Nature*. 2016:536:171–178.

Glover GH, Law CS. Spiral-in/out BOLD fMRI for increased SNR and reduced susceptibility artifacts. *Magn Reson Med*. 2001:46:515–522.

Goerlich-Dobre KS, Witteman J, Schiller NO, van Heuven VJP, Aleman A, Martens S. Blunted feelings: alexithymia is associated with a diminished neural response to speech prosody. *Soc Cogn Affect Neurosci*. 2014:9:1108–1117.

Grandjean D. Brain networks of emotional prosody processing. *Emot Rev*. 2021:13:34–43.

Grandjean D, Sander D, Pourtois G, Schwartz S, Seghier ML, Scherer KR, Vuilleumier P. The voices of wrath: brain responses to angry prosody in meaningless speech. *Nat Neurosci*. 2005:8:145–146.

Grosbras M-H, Ross PD, Belin P. Categorical emotion recognition from voice improves during childhood and adolescence. *Sci Rep*. 2018:8:14791.

Grossmann T, Oberecker R, Koch SP, Friederici AD. The developmental origins of voice processing in the human brain. *Neuron*. 2010:65:852–858.

Hammerschmidt K, Jürgens U. Acoustical correlates of affective prosody. *J Voice*. 2007:21:531–540.

Haynes J-D. A primer on pattern-based approaches to fMRI: principles, pitfalls, and perspectives. *Neuron*. 2015:87:257–270.

Hein G, Knight RT. Superior temporal sulcus—It's my area: or is it? *J Cogn Neurosci*. 2008:20:2125–2136.

Hickok G, Poeppel D. The cortical organization of speech processing. *Nat Rev Neurosci*. 2007:8:393–402.

Ingersoll B. Broader autism phenotype and nonverbal sensitivity: evidence for an association in the general population. *J Autism Dev Disord*. 2010:40:590–598.

Johnstone T, van Reekum CM, Oakes TR, Davidson RJ. The voice of emotion: an FMRI study of neural responses to angry and happy vocal expressions. *Soc Cogn Affect Neurosci*. 2006:1:242–249.

Keltner D, Haidt J. Social functions of emotions at four levels of analysis. *Cognit Emot*. 1999:13:505–521.

Kotz SA, Meyer M, Alter K, Besson M, von Cramon DY, Friederici AD. On the lateralization of emotional prosody: an event-related functional MR investigation. *Brain Lang*. Understanding Language. 2003:86:366–376.

Kotz SA, Kalberlah C, Bahlmann J, Friederici AD, Haynes J-D. Predicting vocal emotion expressions from the human brain. *Hum Brain Mapp*. 2013:34:1971–1981.

Kragel PA, LaBar KS. Decoding the nature of emotion in the brain. *Trends Cogn Sci*. 2016:20:444–455.

Kriegeskorte N, Bandettini PA. Analyzing for information, not activation, to exploit high-resolution fMRI. *NeuroImage*. 2007:38:649–662.

Kriegeskorte N, Douglas PK. Interpreting encoding and decoding models. *Curr Opin Neurobiol*. Machine Learning, Big Data, and Neuroscience. 2019:55:167–179.

Lemerise EA, Arsenio WF. An integrated model of emotion processes and cognition in social information processing. *Child Dev*. 2000:71:107–118.

Liebenthal E, Desai RH, Humphries C, Sabri M, Desai A. The functional organization of the left STS: a large scale meta-analysis of PET and fMRI studies of healthy adults. *Front Neurosci*. 2014:8.

Mazefsky CA, Oswald DP. Emotion perception in Asperger's syndrome and high-functioning autism: the importance of diagnostic criteria and cue intensity. *J Autism Dev Disord*. 2007:37:1086–1095.

Mitchell RLC, Elliott R, Barry M, Cruttenden A, Woodruff PWR. The neural response to emotional prosody, as revealed by functional magnetic resonance imaging. *Neuropsychologia*. 2003:41:1410–1421.

Moerel M, De Martino F, Formisano E. An anatomical and functional topography of human auditory cortical areas. *Front Neurosci*. 2014:8:225–225.

Morningstar M, Nelson EE, Dirks MA. Maturation of vocal emotion recognition: insights from the developmental and neuroimaging literature. *Neurosci Biobehav Rev*. 2018:90:221–230.

Morningstar M, Mattson WI, Venticinque J, Singer S, Selvaraj B, Hu HH, Nelson EE. Age-related differences in neural activation and functional connectivity during the processing of vocal prosody in adolescence. *Cogn Affect Behav Neurosci*. 2019:19:1418–1432.

Morningstar M, Mattson WI Jr, SS, Venticinque JS, Nelson EE. Children and adolescents' neural response to emotional faces and voices: age-related changes in common regions of activation. *Soc Neurosci*. 2020:15:613–629.

Nee DE. fMRI replicability depends upon sufficient individual-level data. *Commun Biol*. 2019:2:1–4.

Neves L, Martins M, Correia AI, Castro SL, Lima CF. Associations between vocal emotion recognition and socio-emotional adjustment in children. *R Soc Open Sci*. 2021:8:211412.

Nichols TE, Holmes AP. Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Hum Brain Mapp*. 2002:15:1–25.

Norman-Haignere SV, McDermott JH. Neural responses to natural and model-matched stimuli reveal distinct computations in primary and nonprimary auditory cortex. *PLoS Biol*. 2018:16:e2005127.

Notter MP, Gale D, Herholz P, Markello R, Notter-Bielser M-L, Whitaker K. AtlasReader: a Python package to generate coordinate tables, region labels, and informative figures from statistical MRI images. *J Open Source Softw*. 2019:4:1257.

Nowicki S. *Manual for the receptive tests of the diagnostic analysis of nonverbal accuracy 2 (DANVA2)*. Atlanta (GA): Emory University; 2010

Nowicki S, Duke MP. Individual differences in the nonverbal communication of affect: the diagnostic analysis of nonverbal accuracy scale. *J Nonverbal Behav*. 1994:18:9–35.

Okada K, Rong F, Venezia J, Matchin W, Hsieh I-H, Saberi K, Serences JT, Hickok G. Hierarchical organization of human auditory cortex: evidence from acoustic invariance in the response to intelligible speech. *Cereb Cortex*. 2010:20:2486–2495.

Pell MD, Kotz SA. Comment: the next frontier: prosody research gets interpersonal. *Emot Rev*. 2021:13:51–56.

Penfield W, Boldrey E. Somatic motor and sensory representation in the cerebral cortex of man as studied by electrical stimulation. *Brain*. 1937:60:389–443.

Pernet CR, McAleer P, Latinus M, Gorgolewski KJ, Charest I, Bestelmeyer PEG, Watson RH, Fleming D, Crabbe F, Valdes-Sosa M, et al. The human voice areas: spatial organization and inter-individual variability in temporal and extra-temporal cortices. *NeuroImage*. 2015:119:164–174.

Posserud M-B, Lundervold AJ, Gillberg C. Autistic features in a total population of 7–9-year-old children assessed by the ASSQ (Autism Spectrum Screening Questionnaire). *J Child Psychol Psychiatry*. 2006:47:167–175.

Rauschecker JP, Scott SK. Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. *Nat Neurosci*. 2009:12:718–724.

Sammler D, Grosbras M-H, Anwander A, Bestelmeyer PEG, Belin P. Dorsal and ventral pathways for prosody. *Curr Biol*. 2015:25:3079–3085.

Sander D, Grandjean D, Pourtois G, Schwartz S, Seghier ML, Scherer KR, Vuilleumier P. Emotion and attention interactions in social cognition: brain regions involved in processing anger prosody. *NeuroImage*. Special Section: Social Cognitive Neuroscience. 2005:28:848–858.

Schirmer A. Is the voice an auditory face? An ALE meta-analysis comparing vocal and facial emotion processing. *Soc Cogn Affect Neurosci*. 2018:13:1–13.

Schirmer A, Kotz SA. Beyond the right hemisphere: brain mechanisms mediating vocal emotional processing. *Trends Cogn Sci*. 2006:10:24–30.

Schirmer A, Escoffier N, Zysset S, Koester D, Striano T, Friederici AD. When vocal processing gets emotional: on the role of social orientation in relevance detection by the human amygdala. *NeuroImage*. 2008:40:1402–1410.

Schlegel K, Grandjean D, Scherer KR. Emotion recognition: unidimensional ability or a set of modality- and emotion-specific skills? *Personal Individ Differ*. 2012:53:16–21.

Seydell-Greenwald A, Chambers CE, Ferrara K, Newport EL. What you say versus how you say it: comparing sentence comprehension and emotional prosody processing using fMRI. *NeuroImage*. 2020:209:116509.

Stelzer J, Chen Y, Turner R. Statistical inference and multiple testing correction in classification-based multi-voxel pattern analysis (MVPA): random permutations and cluster size control. *NeuroImage*. 2013:65:69–82.

Trainor LJ, Austin CM, Desjardins RN. Is infant-directed speech prosody a result of the vocal expression of emotion? *Psychol Sci*. 2000:11:188–195.

Trevisan DA, Birmingham E. Are emotion recognition abilities related to everyday social functioning in ASD? A meta-analysis. *Res Autism Spectr Disord*. 2016:32:24–42.

Turken AU, Dronkers NF. The neural architecture of the language comprehension network: converging evidence from lesion and connectivity analyses. *Front Syst Neurosci*. 2011:5.

Van Kleef GA. How emotions regulate social life: the emotions as social information (EASI) model. *Curr Dir Psychol Sci*. 2009:18: 184–188.

Wallace GL, Case LK, Harms MB, Silvers JA, Kenworthy L, Martin A. Diminished sensitivity to sad facial expressions in high functioning autism spectrum disorders is associated with symptomatology and adaptive functioning. *J Autism Dev Disord*. 2011:41: 1475–1486.

Ward BD. *Simultaneous inference for fMRI data*; 2000.

Warren JE, Sauter DA, Eisner F, Wiland J, Dresner MA, Wise RJS, Rosen S, Scott SK. Positive emotions preferentially engage an auditory–motor "mirror" system. *J Neurosci*. 2006:26: 13067–13075.

Watson R, Latinus M, Charest I, Crabbe F, Belin P. People-selectivity, audiovisual integration and heteromodality in the superior temporal sulcus. *Cortex*. 2014:50:125–136.

Wildgruber D, Riecker A, Hertrich I, Erb M, Grodd W, Ethofer T, Ackermann H. Identification of emotional intonation evaluated by fMRI. *NeuroImage*. 2005:24:1233–1241.

Wildgruber D, Ackermann H, Kreifelts B, Ethofer T. Cerebral processing of linguistic and emotional prosody: fMRI studies. In: Anders S, Ende G, Junghofer M, Kissler J, Wildgruber D, editors. *Progress in brain research. Understanding emotions*. Elsevier; 2006. pp. 249–268.

Williams BT, Gray KM. The relationship between emotion recognition ability and social skills in young children with autism. *Autism*. 2013:17:762–768.

Yerys BE, Jankowski KF, Shook D, Rosenberger LR, Barnes KA, Berl MM, Ritzl EK, VanMeter J, Vaidya CJ, Gaillard WD. The fMRI success rate of children and adolescents: typical development, epilepsy, attention deficit/hyperactivity disorder, and autism spectrum disorders. *Hum Brain Mapp*. 2009:30:3426–3435.

Young AW, Frühholz S, Schweinberger SR. Face and voice perception: understanding commonalities and differences. *Trends Cogn Sci*. 2020:24:398–410.

Yuan W, Altaye M, Ret J, Schmithorst V, Byars AW, Plante E, Holland SK. Quantification of head motion in children during various fMRI language tasks. *Hum Brain Mapp*. 2009:30:1481–1489.

Zachlod D, Rüttgers B, Bludau S, Mohlberg H, Langner R, Zilles K, Amunts K. Four new cytoarchitectonic areas surrounding the primary and early auditory cortex in human brains. *Cortex*. 2020:128: 1–21.

Zhang D, Chen Y, Hou X, Wu YJ. Near-infrared spectroscopy reveals neural perception of vocal emotions in human neonates. *Hum Brain Mapp*. 2019:40:2434–2448.