Received: 24 September 2020 | Revised: 20 November 2020 | Accepted: 25 November 2020

DOI: 10.1002/hbm.25309

RESEARCH ARTICLE

WILEY

Categorizing human vocal signals depends on an integrated auditory-frontal cortical network

Claudia Roswandowitz[1,2] | Huw Swanborough[1,2] | Sascha Frühholz[1,2,3]

[1]Department of Psychology, University of Zurich, Zurich, Switzerland

[2]Neuroscience Center Zurich, University of Zurich and ETH Zurich, Zurich, Switzerland

[3]Center for Integrative Human Physiology (ZIHP), University of Zurich, Zurich, Switzerland

Correspondence
Claudia Roswandowitz, Department of Psychology, University of Zurich, Zurich 8050, Switzerland.
Email: claudia.roswandowitz@uzh.ch

Funding information
Swiss National Science Foundation, Grant/Award Numbers: PP00P1_157409/1, PP00P1_183711/1

Abstract

Voice signals are relevant for auditory communication and suggested to be processed in dedicated auditory cortex (AC) regions. While recent reports highlighted an additional role of the inferior frontal cortex (IFC), a detailed description of the integrated functioning of the AC–IFC network and its task relevance for voice processing is missing. Using neuroimaging, we tested sound categorization while human participants either focused on the higher-order vocal-sound dimension (voice task) or feature-based intensity dimension (loudness task) while listening to the same sound material. We found differential involvements of the AC and IFC depending on the task performed and whether the voice dimension was of task relevance or not. First, when comparing neural vocal-sound processing of our task-based with previously reported passive listening designs we observed highly similar cortical activations in the AC and IFC. Second, during task-based vocal-sound processing we observed voice-sensitive responses in the AC and IFC whereas intensity processing was restricted to distinct AC regions. Third, the IFC flexibly adapted to the vocal-sounds' task relevance, being only active when the voice dimension was task relevant. Forth and finally, connectivity modeling revealed that vocal signals independent of their task relevance provided significant input to bilateral AC. However, only when attention was on the voice dimension, we found significant modulations of auditory-frontal connections. Our findings suggest an integrated auditory-frontal network to be essential for behaviorally relevant vocal-sounds processing. The IFC seems to be an important hub of the extended voice network when representing higher-order vocal objects and guiding goal-directed behavior.

KEYWORDS
auditory-frontal network, DCM, decision-making, fMRI, voice

1 | INTRODUCTION

The human voice conveys various information including linguistic and paralinguistic signals that are of fundamental behavioral relevance for social communication. These vocal sounds compared to other types of sounds are processed in the "temporal voice area" (TVA), a dedicated voice-sensitive brain region along the bilateral auditory cortex (AC) with focus on the superior temporal cortex (STC) (Belin, Zatorre, Lafaille, Ahad, & Pike, 2000; Pernet et al., 2015). To our knowledge, the functional description of the TVA has been almost exclusively

conducted via passive listening tasks to a wide range of vocal-sound objects, including speech and nonspeech compared to a set of nonvocal sound objects (e.g., Belin et al., 2000; Pernet et al., 2015). This common passive task assumes that cortical voice-processing effects are mainly driven by sensory stimulation as bottom-up process. This assumption contrasts a large number of human and animal studies that have shown task-related modulation of sound processing in the AC likely reflecting enhanced representation of task-relevant sound features (for reviews, see Banno et al., 2020; Leonard & Chang, 2014; Sutter & Shamma, 2011). For human vocal sounds, neural task effects in the AC have been described for different aspects of vocal-sound processing including meaningful speech (Bonte, Hausfeld, Scharke, Valente, & Formisano, 2014; Ding & Simon, 2012; Mesgarani & Chang, 2012; O'Sullivan et al., 2019; Rutten, Santoro, Hervais-Adelman, Formisano, & Golestani, 2019) and paralinguistic information such as speaker identity (Schall, Kiebel, Maess, & von Kriegstein, 2015; von Kriegstein, Eger, Kleinschmidt, & Giraud, 2003) and emotional prosody (Frühholz, Ceravolo, & Grandjean, 2012). The question of how cortical processing of the superordinate class of general human vocal sounds including speech and nonspeech signals adapts to varying task requirements is largely unresolved.

Testing the effect of different attentional demands on neural vocal-sound processing, Capilla et al. recorded MEG signals over the superior temporal gyrus (STG) while participants (n = 10) actively categorized vocal from nonvocal sounds and passively listened to the same set of sounds (Capilla, Belin, & Gross, 2013). Although the authors found voice-sensitive responses across different task demands, MEG signals to vocal versus nonvocal sounds (200–250 ms) seemed to differ dependent on the attentional demand with lower relative change in the MEG signal during categorization compared to passive listening (Figure 5, Capilla et al., 2013). Another study using transcranial magnetic stimulation kept the task constant but varied task-relevant sound information (Bestelmeyer, Belin, & Grosbras, 2011). Inhibition of the right mid STG in nine participants revealed decreased categorization performance in a vocal task, but performance was unaffected in a loudness task while listening to the same vocal and nonvocal sounds. Given these limited previous findings, task-specific effects during general vocal-sound processing in the AC remain unclear.

Beside these effects found in the AC, auditory object processing, such as object categorization, is usually accomplished by the interplay of different brain regions including the AC. The ventral auditory processing pathway, connecting the STC with inferior frontal cortex (IFC), has a critical role in auditory object categorization (Bizley & Cohen, 2013; Rauschecker & Tian, 2000; Romanski et al., 1999). Classical neuroanatomical voice models, however, focus on AC regions (Belin, Fecteau, & Bédard, 2004; Maguinness, Roswandowitz, & von Kriegstein, 2018). During voice processing, AC regions are implicated in vocal-feature extraction and more complex aspects of voice processing, such as vocal/nonvocal category or speaker-identity processing. More recent theoretical models incorporated an extended brain system that shares connections with the AC, likely supporting more complex aspects of voice processing (Blank, Wieland, & von Kriegstein, 2014; Roswandowitz, Kappes, et al., 2018; Roswandowitz, Maguinness, & von Kriegstein, 2018). The functional contribution of the extended system has been just recently addressed and one candidate region is the IFC (Blank et al., 2014; Frühholz & Grandjean, 2013), which is functionally interconnected with the AC (Aglieri, Chaminade, Takerkart, & Belin, 2018). In humans, the representation of human vocal speaker identities (Andics, McQueen, & Petersson, 2013; Latinus, Crabbe, & Belin, 2011) and monkey call types (Jiang, Chevillet, Rauschecker, & Riesenhuber, 2018) have been ascribed to the IFC. Even though, sensitivity in the IFC toward the complex category of vocal sounds has been observed (Fecteau et al., 2005, Aglieri et al., 2018; Pernet et al., 2015), the functional contribution of the IFC for behavioral relevant vocal-signal processing remains unclear as previous studies used passive designs and addressed task-based processing only in a post hoc manner.

In the present study, we therefore aimed to systematically investigate if and how varying task requirements during vocal-sound categorizations affect the functional activation patterns within the AC, but also in functional collaboration with the IFC. We performed a task-based event-related fMRI experiment presenting a common set of vocal and nonvocal sounds of two intensity levels (Belin et al., 2000; Capilla et al., 2013). Based on these sounds, participants performed either a voice categorization (i.e., attentional focus on the task-relevant voice dimension) or a loudness categorization task (i.e., attentional focus on the task-relevant loudness dimension) while listening to the same auditory material. We hypothesized that task-relevant categorization of complex vocal sounds during the voice task depends on an integrated auditory-frontal network, while implicit and task-irrelevant vocal-sound processing during the loudness task might reveal neural effects restricted to AC regions. Furthermore, we expected some attention modulation effects in the AC during task-relevant in contrast to task-irrelevant vocal-sound processing.

2 | MATERIALS AND METHODS

2.1 | Participants

We invited 29 healthy volunteers to take part in this experiment (14 females; mean age = 26.10 years, SD = 4.95). The participants reported normal hearing abilities and normal or corrected-to-normal vision. No participant presented a neurological or psychiatric history. All participants gave informed and written consent for their participation in accordance with the ethical and data security guidelines of the University of Zurich. The experiments were approved by the cantonal ethics committee of the Cantone Zurich.

2.2 | Experimental design

The stimulus material consisted of 500 ms sound clips including 70 vocal human and 70 nonvocal nonhuman sounds (Capilla et al., 2013). Vocal sounds included 27 speech-like (e.g., syllables,

vowel, pseudo-words, words) and 43 nonspeech vocalizations (e.g., yawn, laughter, giggle, sighs) of different emotional prosody. Vocal and nonvocal sounds had similar mean f0 (vocal: 270 ± 92 Hz; mean ± *SD*), nonvocal: 281 ± 163 Hz) and a comparable variation in f0 (vocal: 39.9 ± 37.9 Hz, nonvocal: 42. 1 ± 52.5 Hz). The harmonic-to-noise-ration was higher for vocal (13.5 ± 6.8 dB) than for nonvocal (6.3 ± 8.8 dB) sounds (for statistical comparisons, see Capilla et al., 2013). Human voices included different female and male speakers of different age and were all unfamiliar to participants. Nonvocal sounds included 18 animal vocalizations (e.g., dog bark, neigh), 28 artificial sounds (e.g., bell, telephones, music instruments), and natural sounds (e.g., waterfall, wind, door closing). Each sound presentation was preceded by a 500 ms fixation cross and followed by a jittered blank 3,550–5,000 ms gap before the onset of the next stimulus. During the fMRI experiment, sounds were presented using OptoActive II active noise canceling headphones actively reducing MR EPI gradient noise (Optoacoustics Ltd., Mazor, Israel). Visual task instructions were presented to participants via a back-projection screen mounted on top of the head-coil. The experiment consisted of two separate runs, each including one of two tasks (see below) as similarly used in a previous study (Capilla et al., 2013). Each sound clip was presented twice during each run, one time with a lower intensity (60 dB SPL) and one time with a higher intensity (75 dB SPL). Each run thus consisted of 280 trials. In one run, participants were asked to perform a two-choice task (right hand: index and middle finger; button assignment counterbalanced across participants) on the stimuli to decide whether they heard a vocal or a nonvocal sound ("voice task" or "VOICE") irrespective of the sound intensity. In the other run, participants were asked to perform a two-choice task to decide whether they heard a high or low intensity sound ("loudness task" or "LOUD") irrespective of the vocal/nonvocal category. Responses were given on a button box. Sequence of runs was counterbalanced across participants and each run had a duration of 11.30 min. Each run was preceded by either 10 training trials of the voice task or the loudness task to familiarize the participants with the respective tasks. The training trials were discarded from further analyses. During the voice task, participants' attention was on the vocal-sound dimension and we tested task-relevant vocal categorization that requires the abstraction of physical sound features. Therefore, we refer to this task as the higher-order and more complex task. During the loudness task, decisions were based on the physical sound features without abstraction. The loudness task allowed us to contrast vocal versus nonvocal sound trials while participants' attention was on the intensity dimension and thus to test for task-irrelevant vocal-sound processing. We used Cogent 2000 implemented in MATLAB (version 9.5, The MathWorks, Inc.) to present stimuli and to record responses.

2.3 | Image acquisition

Functional and structural brain data were recorded on a 3 T-Philips Ingenia by using a standard 32-channel head coil. High-resolution structural MRI was acquired by using T1-weighted scans (TR 7.91 s, TE 3.71 ms, voxel size $0.57mm^3$; in-plane resolution 256×251). Functional whole-brain images were recorded continuously with a T2*-weighted echo-planar pulse sequence (TR 1.65 s, TE 30 ms, FA 88°; in-plane resolution 128×128 voxels, voxel size $1.71 \times 1.71 \times 3.5$ mm; gap 0.4 mm; 19 slices, sequential ordering and ascending acquisition). For T1 stabilization, five initial "dummy" scans were acquired and then discarded by the scanner. We used partial volume acquisition of slices rotated ~30° nose-up to the anterior commissure (AC)–posterior commissure (PC) plane. The partial volume covered the STC (including superior temporal gyrus/sulcus) and IFC (including inferior frontal gyrus/sulcus). For each run, we acquired 470 volumes, resulting in a total of 940 volumes per participant. Scanning time per run was 775.5 s.

2.4 | Data preprocessing

Preprocessing and statistical analyses of functional images were performed with the Statistical Parametric Mapping software (SPM12, Welcome Department of Cognitive Neurology, London; fil.ion.ucl.ac.uk/spm) implemented in MATLAB (version 9.5, The MathWorks, Inc., MA). Functional data were first manually realigned to the AC–PC axis, followed by motion correction using rigid-body transformation with the first scan as the reference scan of the functional images. Each participant's structural image was coregistered to the mean functional image and then segmented using the standard parameters of the CAT12 toolbox implemented in SPM12 to allow estimation of normalization parameters. Using the resulting parameters, we spatially normalized the anatomical and functional images to the Montreal Neurological Institute (MNI) stereotactic space. The functional images for the main experiment were resampled into $1.7mm^3$ voxels. All functional images were spatially smoothed with a 6 mm full-width half-maximum isotropic Gaussian kernel.

2.5 | Single-subject and group analysis

For the first-level analysis, we used a general linear model (GLM), and all trials were modeled with a stick function aligned to the onset of each stimulus, which was then convolved with a standard hemodynamic response function. For each task block, we modeled trials with correct responses for vocal and nonvocal sounds separately for low and high-intensity presentations resulting in four regressors, one regressor for incorrect responses for vocal and one for incorrect nonvocal sounds and one regressor for the training trials. Thus, in total, the GLM included 14 regressors, 7 for each task run. We also included six motion correction parameters as regressors of no interest to account for signal changes not related to the conditions of interest.

Contrast images of the eight main event types (four for each task: correct trials vocal sounds of high intensity, correct trials vocal sounds of low intensity, correct trials nonvocal sounds of high intensity, correct trials nonvocal sounds of low intensity) were then taken to a random-effects, group-level analysis to investigate the neural

processing for vocal and nonvocal sounds during both tasks, and for high- and low-intensity sounds during the loudness task. All group results were thresholded at a combined voxel threshold of $p < .005$ corrected for multiple comparisons at a cluster level of $k = 65$. This combined voxel and cluster threshold corresponds to $p = .05$ corrected at the cluster level and was determined by the 3DClustSim algorithm implemented in the AFNI software (afni.nimh.nih.gov/afni; version AFNI_18.3.01) including the recent extension to estimate the (spatial) autocorrelation function according to the median estimated smoothness of the residual images. The cluster extent threshold of $k = 65$ was the maximum value for the minimum cluster size across contrasts of the voice and loudness experiments. Functional activations were anatomically labeled according to the probabilistic cytoarchitectonic maps implemented in the SPM Anatomy Toolbox (version 2.2c, Eickhoff et al., 2005, 2007) and the Harvard-Oxford Cortical Structural Atlas (Desikan et al., 2006) implemented in FSLeyes (https://zenodo.org/record/3937147#.X7ZPUy2ZOuU).

2.6 | Dynamic causal modeling of effective brain connectivity

We used the DCM12 toolbox implemented in SPM12 to model the information flow in the bilateral auditory-frontal network during vocal-sound processing under different attention and task demands. We defined individual regions of interest (ROIs) of each seed region within bilateral AC and IFC. For that, we identified individual peaks within a sphere of 6.8 mm radius to the group peak activations found for the contrast vocal against nonvocal sounds for the voice task, which revealed group peak activations in the left mid STC (mSTC) (MNI xyz [−62 −10 −2]), right posterior STC (pSTC) [66 −16 2], left IFG, pars orbitalis (IFGorb) [−38 33 −2], and right IFG, pars triangularis (IFGtri) [49 17 25]. Individual time-series were extracted from all voxels within a sphere of 3.4 mm centered on the individual maximum and exceeding the threshold of $p_{uncorrected} < .05$ (see Table S1 for individual peak coordinates and number of voxels extracted for each ROI). The signal-time course was quantified as the first eigenvariate representing the most typical time course across voxels included in each ROI. Using a dynamic causal modeling (DCM) specific slice time correction procedure, acquisition delay for each ROI was accounted for by estimating the time of signal acquisition in each ROI across the time interval of the TR.

We used a two-step procedure to define the most likely model explaining the empirical data. First, we defined the most plausible driving inputs (C matrix) to the four-node network. We therefore permuted through any configuration ($n = 16$) of possible inputs. While vocal-sound trials across both tasks were potential driving inputs to the STC regions, only vocal-sound trials during the voice tasks could drive activity in the frontal ROIs as this region was only responsive during task-relevant vocal-sound processing. For the winning model, we determined the significance ($p < .05$) of the driving inputs using posterior probability of the driving effects. Only the driving inputs to the STC region of all vocal-sound trials were significant.

In a second step, we then took the significant driving input to the STC as a constraint to permute across the possible model space of modulations of connections by the experimental conditions (B matrix). Concerning the B matrix, we defined connectivity modulations from the predominant response of the seed and target region. For example, the connectivity from STC regions could be modulated by all voice trials, while connectivity from the frontal regions could be only modulated by vocal-sound trials during the voice task. Based on these restrictions, we created three different families of models based on the intrinsic connection between regions (A matrix). For the A matrix, we allowed any possible connections between regions, but without bilateral connections of the frontal regions given missing strong evidence for direct frontal interactions during auditory-object processing and discrimination. The first model family is referred to as "forward models" ($n = 64$) with potential modulations of bilateral STC and modulation of STC-to-IFC connections. The family of "backward models" ($n = 64$) was identical to the forward models, but including modulation of the IFC-to-STC connections. The "bidirect models" ($n = 896$) allowed any directional modulation of connections.

For each model family, we estimated DCM for any possible combination of parameters in the B matrix, ranging from no modulation of connectivity to full modulation of any connection as defined in the A matrix. To define the winning model, we first compared the evidence for the model families and defined the winning family based on Bayesian model selection by using a fixed-effect approach. This approach assumes the same model architecture across participants (Stephan, Penny, Daunizeau, Moran, & Friston, 2009) and has been used in a previous study on auditory processing (Chennu et al., 2016). Second, within the winning family, we used Bayesian model averaging (BMA) to create a weighted average of all families within the winning family by taking the model evidence of each model into account. The resulting posterior estimates for each parameter in the A, B, and C matrix and the weighted average model were tested for significance by using a t test against "0" with resulting p-values adjusted for multiple comparisons using FDR-correction.

Next to the definition of generic DCM models for general and task-relevant vocal-sound processing, we also defined a set of DCM control models for task-irrelevant vocal-sound processing. These control models were identical to the generic DCM models, with the exception that we replaced the condition of vocal- and nonvocal sound trials during the voice task with vocal and nonvocal-sound trials during the loudness task. Using this adaption, we permuted through the same space of the three model families for the modulation of connections (B matrix). This analysis with the DCM control models was done to determine the specificity of the generic DCM models for task-relevant vocal-sound processing.

2.7 | Behavioral analysis

We performed repeated-measures analysis of variances (ANOVAs) including subject as a random effect to compare behavioral differences between the voice and loudness task. We also compared vocal

and nonvocal decisions of the voice task and between low and high-intensity decisions during the loudness task. As behavioral measures, we used accuracy rates and reaction times (RTs). RT data were based only on correct trials, and accuracy data were based on the ratio of correct trials compared to all trials. Next to the frequentist analyses, for each ANOVA we computed Bayes Factors (BF) to estimate the likelihood of rejecting the alternative hypothesis (H_1) compared to the null hypothesis (H_0) using the function *anovaBF* implemented in R with subject as random factor. BF of 1–3.2 indicate anecdotal, 3.2–10 substantial, 10–100 strong, and BF > 100 decisive evidence against H_0 (Kass & Raftery, 1995). Behavioral analyses were carried out in R (Team RDC, 2019) using RStudio (version 1.1.456).

3 | RESULTS

3.1 | Comparable behavioral performance for the voice and loudness task

We first compared the overall performance across both tasks by comparing the mean RTs and accuracy levels for recognizing vocal and nonvocal sound trials in the voice task, with the mean RTs and accuracy level for high- and low-intensity sound trials during the loudness task (Figure 1). Participants performed the voice and loudness tasks with high accuracy (ACC_{VOICE} = 92.91 ± 7.43%, ACC_{LOUD} = 90.59 ± 6.23%, mean ± *SD*). The accuracy rates were comparable between both tasks ($F_{(1,28)}$ = 3.644, p = .066, n = 29, BF = 1.09) but RTs were significantly faster ($F_{(1,28)}$ = 10.57, p = .003, n = 29, BF = 11.39) for the loudness task (RT_{LOUD} = 901 ± 223 ms) in comparison to the voice task (RT_{VOICE} = 983 ± 199 ms).

We next compared the performance between conditions within each task. Accuracy rates ($F_{(1,28)}$ = 3.97, p = .056, n = 29, BF = 1.71) and RTs ($F_{(1,28)}$ = 2.65, p = .115, n = 29, BF = 0.79) did not differ between vocal (ACC_{vc} = 90.44 ± 13.72%; RT_{vc} = 1,000 ± 222 ms) and nonvocal sound (ACC_{nvc} = 95.37 ± 3.28%; RT_{nvc} = 966 ± 190 ms) trials in the voice task. For the loudness task, high-intensity sound trials (ACC_{hi} = 88.92 ± 9.19%; RT_{hi} = 866.62 ± 201 ms) were responded to faster than low-intensity sound trials (ACC_{lo} = 92.27 ± 7.06%; RT_{lo} = 934.67 ± 252 ms) ($F_{(1,28)}$ = 13.58, p = .001, n = 29, BF = 28.32), but there was no difference in accuracy rates ($F_{(1,28)}$ = 2.87, p = .102, n = 29, BF = 1.01).

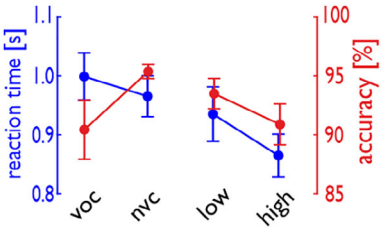

FIGURE 1 Behavioral performance during the voice and loudness task. Reaction times (blue) and accuracy rate (red) for the voice task (voc = vocal, nvc = nonvocal) and the loudness task (low = low intensity, high = high intensity)

3.2 | General task-based voice processing in auditory and frontal cortex

When contrasting vocal against nonvocal trials during the voice and loudness task, we found extensive activation in bilateral AC covering subregions of primary, secondary, and higher-level AC (Figure 2a, Table 1a). We refer to these activations as general task-based TVAs. The peaks of these activations were located in bilateral Te3 at the level of the left mSTC and right pSTC. We additionally found increased activity in the IFC, with peaks located in the IFGorb and IFGtri. We also performed a control analysis by contrasting vocal against nonvocal trials during high (mid panel; left IFGorb/IFGtri, left mSTC/pSTC, right IFGtri, right mSTC/pSTC; Table 1b) and low intensity trials (left IFGorb/IFGtri, left mSTC/pSTC, right IFGtri, right mSTC/pSTC; Table 1c). This revealed almost identical neural activity as found for the general vocal against nonvocal trials contrast, with the exception of no activity found in the right IFGorb.

Next, we checked whether these voice-sensitive BOLD responses obtained with our active task design overlap with the ones obtained with the original passive-listening design from previous studies (Figure 2b). To do so, we visually compared activation maps of our voice task with the t-map created by Pernet et al. based on 218 participants (2015). The spatial extent of our task-based AC patches overlap with the classical anterior, mid, and posterior TVA and the task-based IFC patches with the anterior (IFG, pars orbitalis) and mid frontal voice area (FVA) (IFG, pars triangularis) found during passive vocal-sound listening (Aglieri et al., 2018; Pernet et al., 2015). The posterior FVA (precentral gyrus as part of the motor cortex) that was responsive during passive listening was not active in our task-based categorization design. The precentral gyrus is part of the dorsal ("Where") stream we speculate that the dorsal IFC (precentral gyrus) is suppressed during task-engaged vocal-sound categorization, which is likely supported by the ventral ("What") stream including the IFG, pars orbitalis, and triangularis.

3.3 | Task relevance of the voice dimension modulates frontal but not temporal voice activity

We next tested if the task-relevance of vocal sounds modulates the auditory-frontal network. Contrasting vocal against nonvocal trials only during the voice task (i.e., participants' attention was on the voice dimension and vocal-sounds were task-relevant), we found similar activity in bilateral mSTC and pSTC as well as in bilateral IFGorb and IFGtri (Figure 3a, upper panel; Table 1d) as for the general vocal versus nonvocal contrast performed before. The same contrast of vocal against nonvocal sounds during the loudness task (i.e., participants' attention was on the intensity dimension and vocal-sounds were task irrelevant), revealed a similarly extended neural activity in bilateral

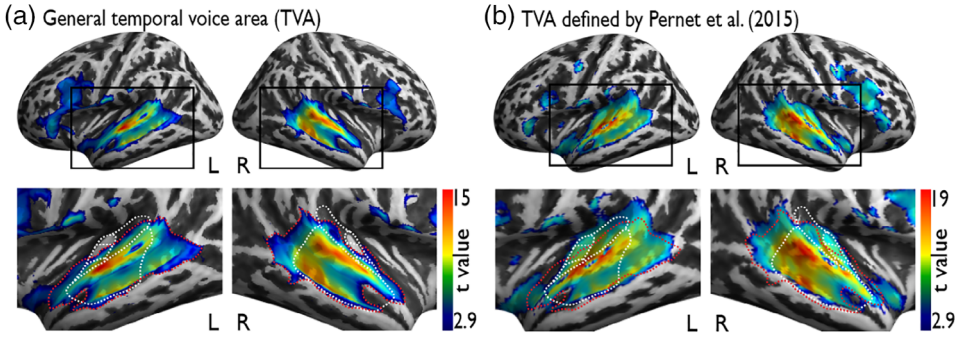

FIGURE 2 Neural activation for task-based and passive vocal-sound processing. (a) The general task-based voice-sensitive cortex (temporal voice area [tTVA]) revealed by the contrast vocal versus nonvocal sound trials during the voice and loudness task (red dashed line indicates the TVA defined by our task-based design) in the temporal lobe covers large parts of primary (Te1.0–1.2), secondary, and higher-level auditory cortex (AC) (Te3) as indicated by the white dashed line. (b) T-map of the TVA as reported by Pernet et al. (2015) resulting from passive listening to vocal and nonvocal sounds. Red dashed line indicates the tTVA

mSTC and pSTC, but not in the IFC (Figure 3a, middle panel; Table 1e). Furthermore, an interaction contrasts for vocal against nonvocal sound trials during the voice as compared to the loudness task revealed a single peak of activity in left IFGorb (Figure 3a, lower panel; Table 1f).

3.4 | Processing intensity information elicits AC activity outside voice-sensitive regions

To test if neural intensity processing overlaps with voice-sensitive AC activity, we contrasted high against low-intensity trials during the loudness task, and found bilateral activity in the AC that was centered in the right planum temporale (PTe) and left Heschl's Gyrus (Figure 3b; Table 2a). Because the extended cluster found for this contrast overlapped with the general definition of the task-based voice-sensitive cortex according to the vocal versus nonvocal trials contrast during the voice task, we further conducted interaction analysis. Intensity specific activation was located in bilateral PTe (interaction contrast: loudness task [high > low] > voice task [vocal>nonvocal]) and the voice task specific activation peaked in higher-level AC in the STC (interaction contrast: voice task [vocal > nonvocal] > Loudness task [high > low]) (Figure 3b; Table 2b,c).

3.5 | Voice-sensitive regions and task difficulty

For each participant, we performed first-level analyses with RT for each trial as covariate of no interest (Figure 4; Table 3). This analysis was done to assess if the different RTs in the voice and loudness task were associated with certain neural activity during vocal-sound processing. Based on the first-level contrasts accounting for RT performance, we then performed second-level analyses for our contrasts of main interest as described above. We found almost identical activation patterns in the AC and IFC when generally contrasting vocal against nonvocal trials (Table 3a). This neural pattern was found for the voice task (left IFGorb, left mSTC, left pSTC, right IFGtri, right mSTC, right pSTC, Table 3b) and for the loudness task (bilateral mSTC, pSTC, Table 3c). The only exception was that we did not find left IFGorb activity based on the interaction contrast of vocal against nonvocal trials specifically for the voice task (Table 3d).

3.6 | Dynamic causal modeling reveals integrated auditory-frontal network for voice discrimination

Having established that vocal-sound processing elicits some common neural activity in bilateral AC irrespective of the sound's task relevance, but critically differ in terms of frontal activity, we then estimated the neural information flow in this auditory-frontal network using DCM.

We first determined the input model that was significantly driving the neural network activity (Figure 5a, upper panel; Table 4). The winning input model (posterior probability [postP] = 100%) consisted of input of all vocal-sound trials for both task-relevant (voice task) and task-irrelevant (loudness task) sound processing to the bilateral AC regions, while the vocal-sound trials drove input to the bilateral frontal regions only during task-relevant processing (voice task). Within this winning model, we tested for the significance of the input and we only found significant effects of the all voice condition to the AC (left: posterior estimate [postE] = 0.851, $p < .001$; right: postE = 0.330, $p < .001$; all p-values FDR corrected), but not of the vocal-sound trials during the voice task to the frontal regions. This winning input model with significant input modulations (i.e., driving input of all vocal-sound trials to bilateral AC) was then taken to the next step of estimating the modulation of connections by the task relevance of vocal sounds (Figure 5a, lower panel).

By permuting across the entire model space including three major modal families, we found that the winning model family was the one including bidirectional connections (Figure 5b, upper panel) with a postP = 100%. We then performed BMA in the winning family to determine significant modulation of connections (Figure 5c, upper

TABLE 1 Peak coordinates of functional contrasts for the main analysis. The table includes peak activations for contrasting vocal with nonvocal sound trials (a) across both tasks, (b) for high intensity sounds, (c) low-intensity sound, and (d,e) for each task separately; (f) reports peak activations for the interaction analysis. Functional activations are thresholded at voxel level $p = .005$ and cluster level $k = 65$, resulting in a combined $p = .05$ corrected at the cluster level

			MNI		
Region	**Cluster size**	**Z value**	**x**	**y**	**z**
(a) Vocal > nonvocal					
L mSTC	5,012	Inf	−60	−10	−2
L pSTC		Inf	−63	−22	−2
L IFGorb		5.61	−39	31	−4
L IFGtri		4.65	−50	17	25
R pSTC	3,472	Inf	61	−25	1
R mSTC		Inf	61	0	−4
R IFGtri	686	4.35	49	19	25
R IFGorb		3.22	44	31	−4
R cerebellum	117	3.52	10	−81	−36
(b) High intensity: vocal > nonvocal					
L mSTC	2,870	Inf	−60	−10	−2
L pSTC		6.84	−61	−24	0
L IFGorb		4.98	−38	33	−2
R pSTC	2,629	7.67	61	−25	1
R mSTC		7.40	61	0	−4
L IFGtri	416	3.68	−50	17	25
R IFGtri	235	3.62	51	21	25
(c) Low intensity: vocal > nonvocal					
L mSTC	2,530	Inf	−60	−10	−2
L pSTC		6.38	−63	−22	−2
L IFGorb		3.21	−41	29	−5
R pSTC	2,620	Inf	61	−25	1
R mSTC		7.24	61	0	−4
L IFGtri	137	3.00	−48	17	25
R IFGtri	71	2.94	46	19	24
(d) VOICE task: vocal > nonvocal					
L mSTC	4,738	Inf	−60	−10	−2
L pSTC		6.65	−63	−22	−2
L IFGorb		5.85	−38	33	−2
L IFGtri		4.38	−48	19	24
R pSTC	2,919	Inf	59	−30	1
R mSTC		7.75	61	0	−4
R IFGtri	922	4.89	49	17	25
R IFGorb		3.62	44	31	−4
R cerebellum	104	3.32	10	−81	−34
(e) LOUD task: vocal > nonvocal					
L mSTC	2,054	Inf	−60	−10	−2
L pSTC		6.52	−63	−22	−2
R pSTC	2,317	7.37	61	−25	1
R mSTC		6.88	61	0	−4
(f) Interaction: VOICE task (vocal > nonvocal) > LOUD task (vocal > nonvocal)					
L IFGorb	121	3.20	−43	41	−5
		3.15	−50	38	1
		2.74	−50	33	−7

Abbreviations: DCM, dynamic causal modeling; IFG, inferior frontal gyrus; MNI, Montreal Neurological Institute; mSTC; mid STC; pSTC, posterior STC; STC, superior temporal cortex.

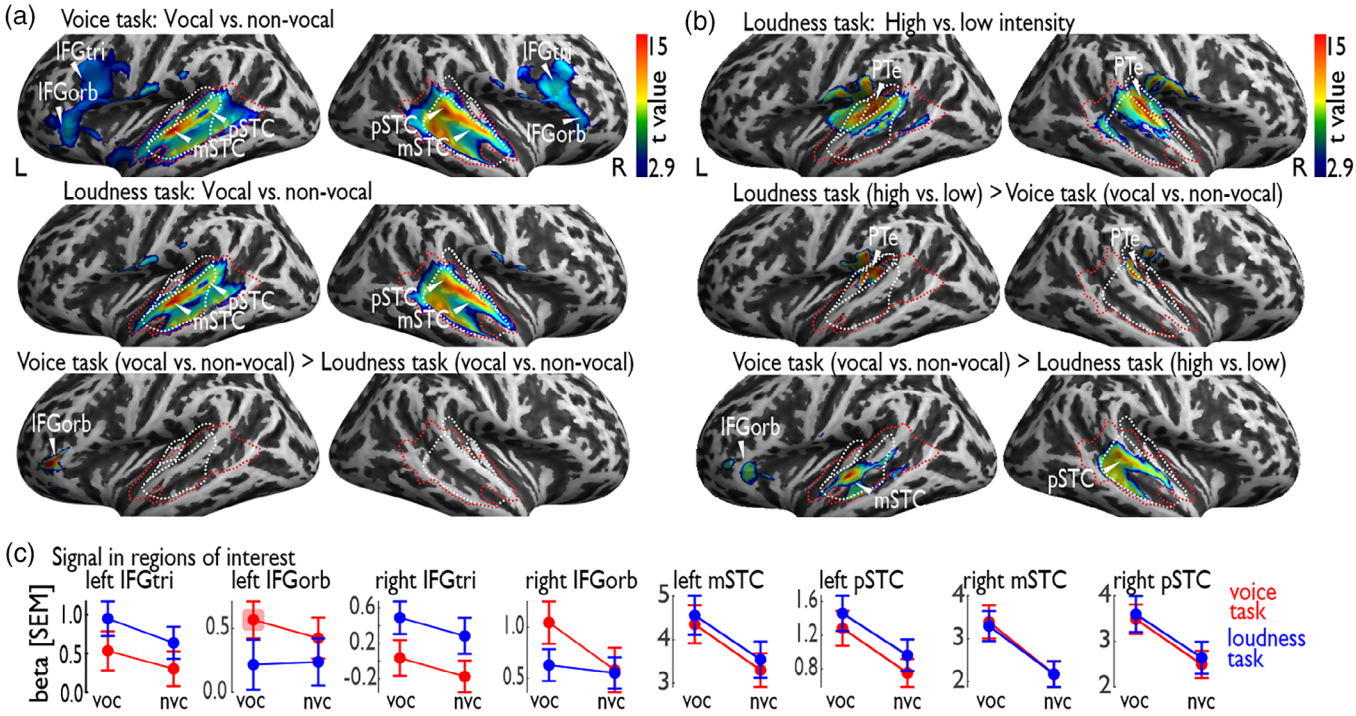

FIGURE 3 Neural activation for vocal and nonvocal sound trials during the two tasks. (a) Contrasting vocal against nonvocal sound trials separately for task-relevant (voice task) and task-irrelevant (loudness task) voice processing revealed extended bilateral activity in the auditory cortex (AC) with peaks in superior temporal cortex (mSTC) and posterior STC (pSTC). Additional activity was found in the inferior frontal cortex (IFC) for vocal trials in the voice task (upper panel) located in bilateral IFGorb and IFGtri. An interaction contrast revealed specific activity in left IFGorb for vocal versus nonvocal sound trials in the voice compared to the loudness task. (b) Functional activity for high versus low-intensity trials during loudness task (upper panel), as well as the interaction contrasts defining specific activity to high against low-intensity trials during the loudness task (mid panel) and the vocal against nonvocal trials during the voice task (lower panel). (c) Signal estimates in regions of interest in 3 mm sphere around peak coordinates. The left IFGorb showed an interaction effect for vocal-sound trials during the voice task, while a similar pattern of activity was found in the right IFGorb, this activation did not survive the cluster threshold. All activations thresholded at voxel $p = .005$ and a cluster extent of $k = 65$ (corresponds to $p = .05$ correct at the cluster level)

panel; Table 5). We found significant bilateral positive modulations of connections between ipsilateral AC and IFC for task-relevant processing (voice task) as well as significant positive modulation of the contralateral forward connection from the AC to the contralateral IFC. Furthermore, we found a negative modulation of the connection from left IFGorb to right pSTC. A general effect of connectivity modulation was found for the connection of left-to-right AC, which was modulated during both tasks.

We performed the same DCM analysis based on the winning input model, but including modulation of the connections only during task-irrelevant vocal-sound processing (loudness task) (Figure 5b, lower panel; Table 6). We again found the bidirectional model family as the winning family, but with postP = 98.34%. BMA furthermore did not lead to any significant modulation of connections between regions (Figure 5c, lower panel).

4 | DISCUSSION

Using an event-related fMRI experiment, we tested voice-signal processing under different task requirements. Our study revealed several key findings. First, task-based vocal-sound processing revealed similar cortical effects in the mid/posterior AC and ventral IFC as found during passive listening to vocal compared to nonvocal auditory objects (Pernet et al., 2015). However, previous studies determined only scarcely and in a post hoc manner which activations in the neural auditory-frontal network were directly linked to context-dependent vocal-sound processing (Aglieri et al., 2018; Pernet et al., 2015). Therefore, second, by (a) comparing vocal-sound and intensity categorization and (b) modulating the task relevance of the vocal-sound dimension, we found differential involvement of bilateral frontal regions. IFC regions were only active when vocal sounds were relevant to the task. In contrast, the IFC was not responsive when vocal-sounds were task-irrelevant and when the task was to categorize feature-based sound information, that is, intensity. Third, using directional connectivity modeling we showed that only task-relevant but not task-irrelevant voice-signal processing involved dynamic bidirectional modulations between auditory and inferior frontal brain areas.

Our findings emphasize an integrated network architecture for voice-signal processing, especially in situations when vocal sounds are relevant for the ongoing behavioral goal. We identified central involvements of the AC and the IFC and suggest distinct functions of

TABLE 2 Peak coordinates of functional activations for the high- and low-intensity trials. The table includes peak activations for contrasting (a) high- and low-intensity trials during the loudness task, and (b,c) for contrasting intensity against vocal-sound processing and vice versa. Functional activations are thresholded at voxel level $p = .005$ and cluster level $k = 65$, resulting in a combined $p = .05$ corrected at the cluster level

			MNI		
Region	**Cluster size**	**Z value**	**x**	**y**	**z**
(a) LOUD task: high > low					
R planum temporale	2,244	5.94	49	−25	10
L Heschl's gyrus/planum temporale	2,143	5.58	−46	−25	8
(b) Interaction: VOICE task (vocal > nonvocal) > LOUD task (high > low)					
L mSTC	573	5.64	−60	−10	−2
R pSTC	861	4.55	56	−25	0
L IFGorb	137	3.57	−36	31	−2
(c) Interaction: LOUD task (high > low) > VOICE task (vocal > nonvocal)					
L planum temporale	359	4.34	−50	−30	15
R planum temporale	287	4.01	51	−27	12

Abbreviations: DCM, dynamic causal modeling; IFG, inferior frontal gyrus; MNI, Montreal Neurological Institute; mSTC; mid STC; pSTC, posterior STC; STC, superior temporal cortex.

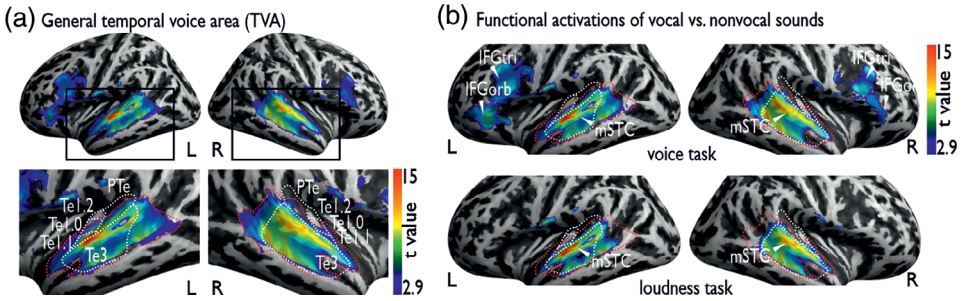

FIGURE 4 Neural activation including reactions time as covariate. (a) Task-based voice-sensitive auditory cortical regions by contrasting vocal against nonvocal sound trials across both tasks. (b) Functional activations for comparing vocal against nonvocal sound trials within the voice task (upper panel) and the loudness task (lower panel). All activations thresholded at voxel $p = .005$ and a cluster extent of $k = 65$ (corresponds to $p = .05$ correct at the cluster level)

both. Within the network, the AC seems to be obligatory and relatively independent of task requirements whereas the IFC becomes significant only when vocal objects are task relevant by modulating goal-directed processing. Our findings suggest the involvement of an extended voice-processing network comprising the IFC to accomplish attentional and goal-directed processing of vocal sounds. Thus, voice processing seems to be based on specific AC and IFC nodes, presumable located bilaterally in the middle/posterior STC and IFC, pars orbitalis (BA 47) and IFC, pars triangularis (BA 45) and might be considered as anatomical structures of the ventral auditory pathway (Friederici, 2015; Frühholz & Grandjean, 2013). That our categorization task, similar to data from passive listening experiments (Pernet et al., 2015), recruited a bilateral rather than an unilateral network may be explained by the nature of our diverse vocal stimulus set including sounds that are characterized by fast-changing temporal features (e.g., vowels) with a left hemispheric and slow-changing spatial features (e.g., syllables, prosodic vocal nonspeech sounds) with a right hemispheric processing preference (Poeppel, 2003; Zatorre, Belin, & Penhune, 2002).

Our findings reveal vocal-sound sensitive responses in secondary AC and IFC overlapping with the classical TVA and FVA found for passive voice listening (Belin et al., 2000; Pernet et al., 2015). With our study design, modulating the relevance of vocal sounds for the ongoing task, we went one step further and showed voice-sensitive AC responses to vocal sounds being task relevant but also when vocal sounds were the to-be-ignored dimension. Our finding parallels studies using neuroimaging methods with high temporal resolution suggesting robust representation of task-relevant and task-irrelevant, unattended speech signals in the AC, although with lower precision of sounds outside the focus of attention (Capilla et al., 2013; Ding & Simon, 2012; Hausfeld, Riecke, Valente, & Formisano, 2018; Mesgarani & Chang, 2012). Our results on largely task-independent processing in the AC support the notion that the AC is fundamental for vocal-sound processing and thus likely constitutes the key region of the core-voice system (Maguinness et al., 2018; Roswandowitz, Kappes, et al., 2018; Roswandowitz, Maguinness, & von Kriegstein, 2018).

Sound processing in the AC seems to be particularly sensitive to the vocal dimension and ignores acoustic variations (Agus, Paquette, Suied, Pressnitzer, & Belin, 2017; Bestelmeyer et al., 2011). Also in our study, feature-based sound intensity elicited activity outside the vocal-sound sensitive AC in bilateral PTe, which seems concerned

TABLE 3 Peak coordinates of functional activations for the analysis with reaction time as covariate. The table includes peak activations for contrasting vocal with nonvocal sound trials (a) across both tasks or (b,c) for each task separately; (d) for the interaction analysis. Functional activations are thresholded at voxel level $p = .005$ and cluster level $k = 65$, resulting in a combined $p = .05$ corrected at the cluster level

			MNI		
Region	**Cluster size**	**Z value**	**x**	**y**	**z**
(a) Vocal > nonvocal					
L mSTC	4,596	Inf	−60	−8	0
L pSTC		Inf	−63	−18	5
L IFGorb		5.12	−38	33	−2
L IFGtri		5.00	−50	17	24
R pSTC	3,361	Inf	61	−25	1
R mSTC		Inf	61	0	−4
R IFGtri	477	4.11	49	19	25
R IFGorb		3.32	52	26	7
(b) VOICE task: vocal > nonvocal					
L mSTC	4,251	Inf	−60	−8	0
L pSTC		7.04	−61	−22	0
L IFGorb		5.58	−36	33	−2
L IFGtri		4.73	−50	17	24
R pSTC	2,806	7.83	58	−32	3
R mSTC		7.33	61	0	−4
R IFGtri	648	4.66	49	17	25
R IFGorb		3.23	47	31	−2
(c) LOUD task: vocal > nonvocal					
L mSTC	1,951	Inf	−60	−8	0
L pSTC		7.14	−61	−22	1
R pSTC	2,242	7.00	59	−24	1
R mSTC		6.66	61	0	−5
(d) Interaction: VOICE task (vocal > nonvocal) > LOUD task (vocal > nonvocal)					
	—	—	—	—	—

Abbreviations: DCM, dynamic causal modeling; IFG, inferior frontal gyrus; MNI, Montreal Neurological Institute; mSTC; mid STC; pSTC, posterior STC; STC, superior temporal cortex.

with feature-based acoustic analyses and the formation of spectro-temporal sound patterns before object identification (Griffiths & Warren, 2002). Furthermore, during voice processing the spatial extend of AC response was comparable to high- and low-intensity sounds pointing to acoustic invariant vocal-sound processing and thus, in turn, suggests object-like processing at the level of the AC (Ding & Simon, 2012; Rutten et al., 2019).

The specific modulation of the IFC by changing task demands is in line with findings on gradually increasing sensitivity toward attention from primary, secondary, up to frontal regions (Atiani et al., 2014; Nourski, 2017; Rauschecker & Tian, 2000). The IFC responded selectively to the more abstract vocal-sound dimension in contrast to feature-based sound processing. In marked contrast to the AC, the IFC flexibly adapted to the sound's behavioral relevance. The IFC was responsive when processing was overt and task relevant and not during task-irrelevant processing and this response was insensitive to differences in sound intensity. This finding may explain the heterogeneous IFC finding when passively listening to vocal sounds (Aglieri et al., 2018; Pernet et al., 2015). Pernet et al. (2015) noted that only 15–20% of their 218 participants showed voice-sensitive IFC responses. Because task engagement is not experimentally controlled during passive listening and inter-individual differences, some participants might attend and overtly process highly salient vocal sounds leading to IFC response and others not.

The IFC was proposed as a candidate region of the extended voice network, but its function for the behaviorally relevant class of vocal objects was unclear so far. Previous investigations in nonhuman primates suggested that the IFC houses acoustically invariant representation of behaviorally relevant higher-order auditory information including conspecific vocalization (Gifford, MacLean, Hauser, & Cohen, 2005) and human speech sounds (Cohen, Theunissen, Russ, & Gill, 2007; Lee, Russ, Orr, & Cohen, 2009). Similarly, in humans when categorizing speaker identities (Andics et al., 2013; Latinus et al., 2011; Zäske, Awwad Shiekh Hasan, & Belin, 2017) and monkey vocalization (Jiang et al., 2018), the IFC has been associated with category-selective responses that was independent of acoustical sound changes. We here extend previous findings on the IFC to the class of vocal-sound objects and systematically assessed the

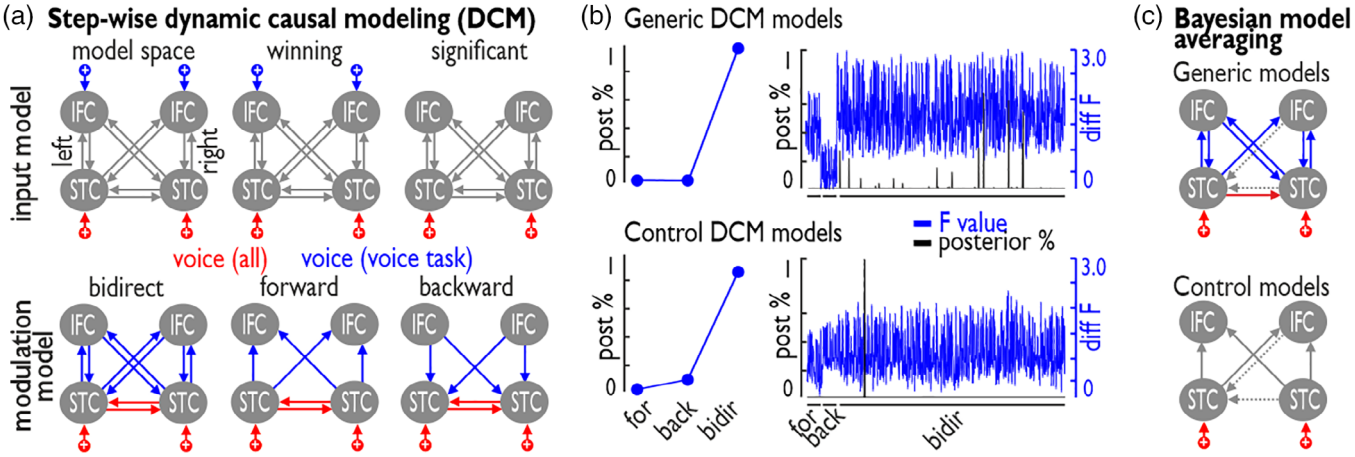

FIGURE 5 Dynamic causal modeling (DCM) of brain activity during task-based vocal-sound processing. (a) Two-step procedure for the DCM modeling, including a first estimation of the most likely input model (upper panel) based on permuting through the model space (left panel), determining the wining model (mid panel), and determining significant driving inputs (right panel). In the second step, using the significant driving inputs as fixed parameters, we permuted through all possible modulations of connections including three major families (bidirect, forward, and backward). The generic model takes voice trials during the voice and loudness task (all voice) and the task-relevant voice trials during the voice task (voice task). (b) The bidirectional family achieved the highest posterior probability for the generic DCM models (upper panel) and the control DCM models (lower panel). The control models take voice trials during voice and loudness task (all voice) and the voice trials during task-irrelevant sound processing during the loudness task. The right panel shows the posterior probability (black) and the relative log-evidence (blue) of all models. (c) Bayesian model averaging revealed significant modulation of connections by all voice trails during the voice and loudness task (red) and by the voice trials during the voice task (blue) between bilateral auditory cortex (AC) and inferior frontal cortex (IFC) for the generic DCM models (upper panel); no such significant modulation was found for the control DCM models (lower panel). Light gray lines indicate significant intrinsic connections with positive connections as full and negative connections as dashed lines

TABLE 4 Estimated parameters for the input model estimation. The input model estimation for the generic DCM models revealed the model with significant input only to the AC regions. Significant input marked with bold numbers; the first number indicates the estimated posterior value of the parameter; number in brackets indicates the FDR-corrected significance level derived from the posterior probability values. The condition "Voice (all)" represents vocal-sound trials across both tasks; the condition "Voice (VOICE tasks)" represents vocal-sound trials only during the voice task

	Condition	
Region	Voice (all)	Voice (VOICE task)
L IFGorb	0	0
R IFGtri	0	0
L mSTC	**0.95** (1.000)	0
R pSTC	**0.33** (0.994)	0

Abbreviations: AC, auditory cortex; DCM, dynamic causal modeling; IFG, inferior frontal gyrus; mSTC; mid STC; pSTC, posterior STC; STC, superior temporal cortex.

functionality of the IFC dependent on the sound's task relevance. In line with previous studies, the IFC response was insensitive to low-level acoustic sound manipulation. Rather the IFC was responsive when the task was to process more complex and abstract vocal information. Although our study design allows no interpretation about how selective our IFC regions encoded vocal objects, we suppose that similar regions in the IFC represent also other complex auditory objects, dependent on the task-relevant category (Giordano et al., 2014; Hausfeld et al., 2018).

We found IFC activations in both hemispheres, with the most specific response in the left IFG pars orbitalis. A common function ascribed to the left IFG pars orbitalis (BA 47), an anterior extension of the IFG pars triangularis (BA 45), is related to semantic language processes (Friederici, 2015; Goucha & Friederici, 2015). Even though our experimental stimulus set comprised speech sounds, of the 27 speech sounds only 2 were meaningful words besides 18 syllables and 7 vowels. Thus, we consider the present speech sounds as rather speech-like without semantic meaning and suggest that the present IFC response encodes human vocalization that is relatively independent of language-based semantic processes. Another note on IFC activity for task-based processing concerns its potential relation to behavioral data and task difficulty. RTs during the loudness task were faster compared to the voice task raising the question whether IFC effects were biased by different task difficulties. We think this is unlikely given the comparable accuracy levels in the voice and loudness task; with slightly better performance in the voice task. Furthermore, the additional analysis accounting for RT differences between tasks revealed similar patterns of voice processing in the AC and IFC. The only activation that seemed sensitive to RTs was the left IFG pars orbitalis resulting from the interaction contrast. The left IFG pars orbitalis, however, had a similar level of peak activity compared to the analysis not accounting for RT, but the cluster size (k = 41) did not survive the general cluster-level threshold (k = 65). Another possible explanation for different RTs during the loudness and voice task might

TABLE 5 Estimated parameters for the generic DCM models. The table includes estimated parameters values and FDR-corrected significance level (in brackets) for the (a) C matrix entries after Bayesian model averaging, (b) intrinsic connectivity A matrix, (c) modulation of connections by the "Voice (all)" condition, and (d) modulation of connections by the "Voice (VOICE task)" condition. Columns represent the seed region and rows represent the target region. Significant parameters marked as bold numbers

(a) C matrix				
	Voice (all)	**VOICE (VOICE task)**		
L IFGorb	—	—		
R IFGtri	—	—		
L mSTC	**0.575** (>.001)	—		
R pSTC	**0.302** (>.001)	—		
(b) A matrix				
	Seed			
Target	L IFGorb	R IFGtri	L mSTC	R pSTC
L IFGorb	—	—	**0.056** (.015)	−0.017 (.419)
R IFGtri	—	—	**0.076** (.012)	**0.047** (.047)
L mSTC	−0.026 (.231)	**−0.073** (.012)	—	**−0.211** (0.001)
R pSTC	0.019 (.149)	−0.027 (.213)	0.026 (.457)	—
(c) B matrix: voice (all)				
	Seed			
Target	L IFGorb	R IFGtri	L mSTC	R pSTC
L IFGorb	—	—	—	—
R IFGtri	—	—	—	—
L mSTC	—	—	—	0.135 (.476)
R pSTC	—	—	**0.402** (.042)	—
(d) B matrix: voice (VOICE task)				
	Seed			
Target	L IFGorb	R IFGtri	L mSTC	R pSTC
L IFGorb	—	—	**0.346** (.001)	**0.189** (.002)
R IFGtri	—	—	**0.009** (.002)	**0.446** (.001)
L mSTC	**0.329** (.022)	0.080 (.145)	—	—
R pSTC	**0.239** (.019)	**0.235** (.042)	—	—

Abbreviations: AC, auditory cortex; DCM, dynamic causal modeling; IFG, inferior frontal gyrus; mSTC; mid STC; pSTC, posterior STC; STC, superior temporal cortex.

arise from different processing time windows for low-level intensity (∼100 ms) (Näätänen & Picton, 1987) and higher-level vocal/nonvocal decisions (∼150–200 ms) (Capilla et al., 2013; Charest et al., 2009).

Together with findings in human (Andics et al., 2013; Latinus et al., 2011) and nonhuman primates (Cohen et al., 2007; Gifford et al., 2005; Lee et al., 2009; Russ, Lee, & Cohen, 2007), we here propose that the more ventral part of the IFC becomes an important node of the extended voice network when vocal information represents a higher-order and more abstract category and when attention is on the auditory object for task-relevant processing. Given the high accuracy rates of the voice task, our study design is not sensitive to differentiate between attentional processes or a direct contribution of the IFC to accurate object categorizations. Future studies modulating the task difficulty by, for example, presenting sound continuums between object categories may clarify whether the IFC also modulates the accuracy of categorical decisions.

Beside the distributed activity pattern in bilateral AC and IFC, we finally determined the functional connectivity between these regions of the voice-processing network. The AC and IFC are anatomically and functionally connected during auditory processing as shown in nonhuman primates (Medalla & Barbas, 2014; Plakke & Romanski, 2014), and a similar network is suggested for humans (Rocchi et al., 2020). We here, for the first time, characterize the direction of functional modulation of auditory-frontal connections during task-based vocal-sound processing and systematically modulated the relevance of vocal sounds for the ongoing task. In line with our local response findings, a functional connection between the AC and IFC was only apparent when higher-order vocal objects but not feature-based intensity information were processed. During vocal-object processing, irrespective of the sounds' task relevance, the secondary AC received the sensory input, likely for an initial perceptual analysis (von Kriegstein & Giraud, 2004; Warren, Scott, Price, & Griffiths, 2006; Zäske et al., 2017). Within the network, the AC interacted with higher-level IFC regions by significantly modulated ipsilateral and contralateral connections, but only when attention is drawn toward the higher-order vocal-sound category. The auditory-frontal connection might support the transition of stimulus-specific spectro-temporal sound information to frontal regions for the proper mapping

TABLE 6 Estimated parameters for the control DCM models. The table shows estimated parameters values and FDR-corrected significance level (in brackets) for the (a) C matrix entries after Bayesian model averaging, (b) intrinsic connectivity A matrix, (c) modulation of connections by the "Voice (all)" condition, and (d) modulation of connections by the "Voice (LOUD task)" condition. Columns represent the seed region and rows represent the target region. Significant parameters marked as bold numbers

(a) C matrix				
	Voice (all)	**Voice (LOUD task)**		
L IFGorb	—	—		
R IFGtri	—	—		
L mSTC	**0.810** (>.001)	—		
R pSTC	**0.296** (.014)	—		
(b) A matrix				
	Seed			
Target	L IFGorb	R IFGtri	L mSTC	R pSTC
L IFGorb	—	—	**0.088** (.008)	**0.061** (.008)
R IFGtri	—	—	**0.114** (.001)	**0.073** (.025)
L mSTC	−0.056 (.130)	**−0.066** (.025)	—	**−0.157** (.008)
R pSTC	−0.035 (.156)	−0.027 (.277)	0.029 (.496)	—
(c) B matrix: voice (all)				
	Seed			
Target	L IFGorb	R IFGtri	L mSTC	R pSTC
L IFGorb	—	—	—	—
R IFGtri	—	—	—	—
L mSTC	—	—	—	0.377 (.226)
R pSTC	—	—	0.633 (.065)	—
(d) B matrix: voice (LOUD task)				
	Seed			
Target	L IFGorb	R IFGtri	L mSTC	R pSTC
L IFGorb	—	—	—	0.464 (.101)
R IFGtri	—	—	0.056 (.786)	0.068 (.786)
L mSTC	0.225 (.361)	0.030 (.928)	—	—
R pSTC	0.236 (.226)	0.019 (.978)	—	—

Abbreviations: DCM, dynamic causal modeling; IFG, inferior frontal gyrus; mSTC; mid STC; pSTC, posterior STC; STC, superior temporal cortex.

between attended sensory input, internal task-relevant object representation, and goal-directed response (Miller & Cohen, 2001). The DCM analysis also revealed significant reversed positive modulations of the ipsilateral and partly contralateral frontal-auditory connections. We speculate that positive top-down modulations of the AC improve sound quality for the ongoing task by more fine-tuned spectro-temporal analysis (Mesgarani & Chang, 2012; Rutten et al., 2019). Besides the task-specific modulation of AC–IFC connection for voice processing when attention was directed toward the vocal-sound dimension, the connection from left to right AC was modulated for voice processing independent of the task. General voice processing thus might involve the interaction of bilateral AC, such that fast changing auditory information analyzed by the left AC is shared with the right AC for the integration with slow changing voice features (Zatorre & Belin, 2001).

Given the top-down modulation of the AC during task-relevant but not during task-irrelevant vocal-sound processing, it seems surprising that we see no differential BOLD activation in the AC dependent on the sound's task relevance. As mentioned earlier, previous studies using MEG/EEG and invasive recording reported neural representation to be enhanced for the attended stream (∼150 ms), but also unattended sounds could be robustly decoded from the brain signal, especially in an early time window (70-110 ms) (Hausfeld et al., 2018). The hierarchical decomposition model of auditory scene analysis proposes an initial general perceptual grouping of attended and unattended sounds with subsequent selective attention-driven processes of the task-relevant sound stream (Cusack, Decks, Aikman, & Carlyon, 2004). Given the lower temporal resolution inherent in functional-imaging studies, compared to electrophysiological and single cell recordings, the differentiation of an early perceptual and later attentional processing phase within the AC might be hardly detectable with conventional fMRI experiments (Lee, Grady, Habak, Wilson, & Moscovitch, 2011). Future studies using methods with higher temporal resolution or even invasive recordings can further teas apart time-sensitive attention-modulated mechanisms in the AC and on the network level that is critical for a comprehensive functional understanding of goal-directed vocal-sound processing.

5 | CONCLUSION

Taken together, our findings suggest that the AC alone is not sufficient to master behaviorally relevant processing of human vocal sounds. With our findings, we shed light on the functionality of the IFC in corporation with the AC and suggest the IFC as a relevant node of the extended voice system. Within the auditory-frontal network, the AC seems largely invulnerable to varying levels of task engagement whereas the role of the IFC is modulatory in situations requiring goal-directed judgments of abstract auditory objects such as vocal sounds. Thus, goal-directed voice-signal processing is not restricted to a single area, but to a collection of defined brain areas that are dynamically integrated into a functional neural network.

ACKNOWLEDGMENTS

This study was supported by the Swiss National Science Foundation (SNSF PP00P1_157409/1 and PP00P1_183711/1 to S. F.).

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Claudia Roswandowitz and **Sascha Frühholz**: Contributed to designing the experiments, data acquisition, data analysis, and writing the manuscript. **Huw Swanborough**: Contributed to the analysis of the data and writing the manuscript.

DATA AVAILABILITY STATEMENT

All data are available from the corresponding authors upon reasonable request.

ETHICS STATEMENT

All experimental procedures were approved by the cantonal ethics committee of the Cantone Zurich.

PATIENT CONSENT STATEMENT

All participants gave written informed consent.

ORCID

Claudia Roswandowitz https://orcid.org/0000-0003-1688-6494

REFERENCES

Aglieri, V., Chaminade, T., Takerkart, S., & Belin, P. (2018). Functional connectivity within the voice perception network and its behavioural relevance. *NeuroImage*, *183*, 356–365.

Agus, T. R., Paquette, S., Suied, C., Pressnitzer, D., & Belin, P. (2017). Voice selectivity in the temporal voice area despite matched low-level acoustic cues. *Scientific Reports*, *7*, 11526.

Andics, A., McQueen, J. M., & Petersson, K. M. (2013). Mean-based neural coding of voices. *NeuroImage*, *79*, 351–360.

Atiani, S., David, S. V., Elgueda, D., Locastro, M., Radtke-Schuller, S., Shamma, S. A., & Fritz, J. B. (2014). Emergent selectivity for task-relevant stimuli in higher-order auditory cortex. *Neuron*, *82*, 486–499.

Belin, P., Fecteau, S., & Bédard, C. (2004). Thinking the voice: Neural correlates of voice perception. *Trends in Cognitive Sciences*, *8*, 129–135.

Belin, P., Zatorre, R. J., Lafaille, P., Ahad, P., & Pike, B. (2000). Voice-selective areas in human auditory cortex. *Nature*, *403*, 309–312.

Bestelmeyer, P. E., Belin, P., & Grosbras, M. H. (2011). Right temporal TMS impairs voice detection. *Current Biology*, *21*, R838–R839.

Bizley, J. K., & Cohen, Y. E. (2013). The what, where and how of auditory-object perception. *Nature Reviews. Neuroscience*, *14*, 693–707.

Blank, H., Wieland, N., & von Kriegstein, K. (2014). Person recognition and the brain: Merging evidence from patients and healthy individuals. *Neuroscience and Biobehavioral Reviews*, *47*, 717–734.

Bonte, M., Hausfeld, L., Scharke, W., Valente, G., & Formisano, E. (2014). Task-dependent decoding of speaker and vowel identity from auditory cortical response patterns. *The Journal of Neuroscience*, *34*, 4548–4557.

Capilla, A., Belin, P., & Gross, J. (2013). The early spatio-temporal correlates and task independence of cerebral voice processing studied with MEG. *Cerebral Cortex*, *23*, 1388–1395.

Charest, I., Pernet, C. R., Rousselet, G. A., Quiñones, I., Latinus, M., Fillion-Bilodeau, S., … Belin, P. (2009). Electrophysiological evidence for an early processing of human voices. *BMC Neuroscience*, *10*, 127.

Chennu, S., Noreika, V., Gueorguiev, D., Shtyrov, Y., Bekinschtein, T. A., & Henson, R. (2016). Silent expectations: Dynamic causal modeling of cortical prediction and attention to sounds that weren't. *Journal of Neuroscience*, *36*, 8305–8316.

Cohen, Y. E., Theunissen, F., Russ, B. E., & Gill, P. (2007). Acoustic features of rhesus vocalizations and their representation in the ventrolateral prefrontal cortex. *Journal of Neurophysiology*, *97*, 1470–1484.

Cusack, R., Decks, J., Aikman, G., & Carlyon, R. P. (2004). Effects of location, frequency region, and time course of selective attention on auditory scene analysis. *Journal of Experimental Psychology. Human Perception and Performance*, *30*, 643–656.

Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., … Hyman, B. T. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage*, *31*, 968–980.

Ding, N., & Simon, J. Z. (2012). Emergence of neural encoding of auditory objects while listening to competing speakers. *Proceedings of the National Academy of Sciences of the United States of America*, *109*, 11854–11859.

Eickhoff, S. B., Paus, T., Caspers, S., Grosbras, M.-H., Evans, A. C., Zilles, K., & Amunts, K. (2007). Assignment of functional activations to probabilistic cytoarchitectonic areas revisited. *NeuroImage*, *36*, 511–521.

Eickhoff, S. B., Stephan, K. E., Mohlberg, H., Grefkes, C., Fink, G. R., Amunts, K., & Zilles, K. (2005). A new SPM toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data. *NeuroImage*, *25*, 1325–1335.

Fecteau, S., Armony, J. L., Joanette, Y., & Belin, P. (2005). Sensitivity to voice in human prefrontal cortex. *Journal of Neurophysiology*, *94*, 2251–2254.

Friederici, A. D. (2015). White-matter pathways for speech and language processing. In *Handbook of clinical neurology*, 129, 177–186. Amsterdam, Netherlands: Elsevier.

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

Frühholz, S., & Grandjean, D. (2013). Processing of emotional vocalizations in bilateral inferior frontal cortex. *Neuroscience and Biobehavioral Reviews*, *37*, 2847–2855.

Gifford, G. W., MacLean, K. A., Hauser, M. D., & Cohen, Y. E. (2005). The neurophysiology of functionally meaningful categories: Macaque ventrolateral prefrontal cortex plays a critical role in spontaneous categorization of species-specific vocalizations. *Journal of Cognitive Neuroscience*, *17*, 1471–1482.

Giordano, B. L., Pernet, C., Charest, I., Belizaire, G., Zatorre, R. J., & Belin, P. (2014). Automatic domain-general processing of sound source identity in the left posterior middle frontal gyrus. *Cortex*, *58*, 170–185.

Goucha, T., & Friederici, A. D. (2015). The language skeleton after dissecting meaning: A functional segregation within Broca's area. *NeuroImage*, *114*, 294–302.

Griffiths, T. D., & Warren, J. D. (2002). The planum temporale as a computational hub. *Trends in Neurosciences*, *25*, 348–353.

Hausfeld, L., Riecke, L., Valente, G., & Formisano, E. (2018). Cortical tracking of multiple streams outside the focus of attention in naturalistic auditory scenes. *NeuroImage*, *181*, 617–626.

Jiang, X., Chevillet, M. A., Rauschecker, J. P., & Riesenhuber, M. (2018). Training humans to categorize monkey calls: Auditory feature- and category-selective neural tuning changes. *Neuron*, *98*, 405–416.e4.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795.

Latinus, M., Crabbe, F., & Belin, P. (2011). Learning-induced changes in the cerebral processing of voice identity. *Cerebral Cortex*, *21*, 2820–2828.

Lee, J. H., Russ, B. E., Orr, L. E., & Cohen, Y. (2009). Prefrontal activity predicts monkeys' decisions during an auditory category task. *Frontiers in Integrative Neuroscience*, *3*, 16.

Lee, Y., Grady, C. L., Habak, C., Wilson, H. R., & Moscovitch, M. (2011). Face processing changes in Normal aging revealed by fMRI adaptation. *Journal of Cognitive Neuroscience*, *23*, 3433–3447.

Leonard, M. K., & Chang, E. F. (2014). Dynamic speech representations in the human temporal lobe. *Trends in Cognitive Sciences*, *18*, 472–479.

Maguinness, C., Roswandowitz, C., & von Kriegstein, K. (2018). Understanding the mechanisms of familiar voice-identity recognition in the human brain. *Neuropsychologia*, *116*, 179–193.

Medalla, M., & Barbas, H. (2014). Specialized prefrontal "auditory fields": Organization of primate prefrontal-temporal pathways. *Frontiers in Neuroscience*, *8*, 77.

Mesgarani, N., & Chang, E. F. (2012). Selective cortical representation of attended speaker in multi-talker speech perception. *Nature*, *485*, 233–236.

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, *24*, 167–202.

Näätänen, R., & Picton, T. (1987). The N1 wave of the human electric and magnetic response to sound: A review and an analysis of the component structure. *Psychophysiology*, *24*, 375–425.

Nourski, K. V. (2017). Auditory processing in the human cortex: An intracranial electrophysiology perspective. *Laryngoscope Investigative Otolaryngology*, *2*, 147–156.

O'Sullivan, J., Herrero, J., Smith, E., Schevon, C., McKhann, G. M., Sheth, S. A., … Mesgarani, N. (2019). Hierarchical encoding of attended auditory objects in multi-talker speech perception. *Neuron*, *104*, 1195–1209.e3.

Pernet, C. R., McAleer, P., Latinus, M., Gorgolewski, K. J., Charest, I., Bestelmeyer, P. E. G., … Belin, P. (2015). The human voice areas: Spatial organization and inter-individual variability in temporal and extra-temporal cortices. *NeuroImage*, *119*, 164–174.

Plakke, B., & Romanski, L. M. (2014). Auditory connections and functions of prefrontal cortex. *Frontiers in Neuroscience*, *8*, 199.

Poeppel, D. (2003). The analysis of speech in different temporal integration windows: Cerebral lateralization as 'asymmetric sampling in time'. *Speech Communication*, *41*, 245–255.

Rauschecker, J. P., & Tian, B. (2000). Mechanisms and streams for processing of "what" and "where" in auditory cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *97*, 11800–11806.

Rocchi, F., Oya, H., Balezeau, F., Billig, A. J., Kocsis, Z., Jenison, R., … Petkov, C. I. (2020). Common fronto-temporal effective connectivity in humans and monkeys. *bioRxiv* 2020.04.03.024042.

Romanski, L. M., Tian, B., Fritz, J., Mishkin, M., Goldman-Rakic, P. S., & Rauschecker, J. P. (1999). Dual streams of auditory afferents target multiple domains in the primate prefrontal cortex. *Nature Neuroscience*, *2*, 1131–1136.

Roswandowitz, C., Maguinness, C., & von Kriegstein, K. (2018). Deficits in voice-identity processing: Acquired and developmental phonagnosia. In P. S. Frühholz Belin (Ed.), *The Oxford handbook of voice perception*, Oxford, England: Oxford University Press.

Roswandowitz, C., Kappes, C., Obrig, H., & Von Kriegstein, K. (2018). Obligatory and facultative brain regions for voice-identity recognition. *Brain*, *141*, 234–247.

Russ, B. E., Lee, Y.-S., & Cohen, Y. E. (2007). Neural and behavioral correlates of auditory categorization. *Hearing Research*, *229*, 204–212.

Rutten, S., Santoro, R., Hervais-Adelman, A., Formisano, E., & Golestani, N. (2019). Cortical encoding of speech enhances task-relevant acoustic information. *Nature Human Behaviour*, *3*, 974–987.

Schall, S., Kiebel, S. J., Maess, B., & von Kriegstein, K. (2015). Voice identity recognition: Functional division of the right STS and its behavioral relevance. *Journal of Cognitive Neuroscience*, *27*, 280–291.

Stephan, K. E., Penny, W. D., Daunizeau, J., Moran, R. J., & Friston, K. J. (2009). Bayesian model selection for group studies. *Drug and Alcohol Dependence*, *46*, 1004–1017.

Sutter, M. L., & Shamma, S. A. (2011). The relationship of auditory cortical activity to perception and behavior. In *The auditory cortex* (pp. 617–641). Boston, MA: Springer US.

Team RDC. (2019). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing Retrieved from https://www.R-project.org

von Kriegstein, K., Eger, E., Kleinschmidt, A., & Giraud, A. L. (2003). Modulation of neural responses to speech by directing attention to voices or verbal content. *Cognitive Brain Research*, *17*, 48–55.

von Kriegstein, K., & Giraud, A. L. (2004). Distinct functional substrates along the right superior temporal sulcus for the processing of voices. *NeuroImage*, *22*, 948–955.

Warren, J. D., Scott, S. K., Price, C. J., & Griffiths, T. D. (2006). Human brain mechanisms for the early analysis of voices. *NeuroImage*, *31*, 1389–1397.

Zäske, R., Awwad Shiekh Hasan, B., & Belin, P. (2017). It doesn't matter what you say: fMRI correlates of voice learning and recognition independent of speech content. *Cortex*, *94*, 100–112.

Zatorre, R. J., & Belin, P. (2001). Spectral and temporal processing in human auditory cortex. *Cerebral Cortex*, *11*, 946–953.

Zatorre, R. J., Belin, P., & Penhune, V. B. (2002). Structure and function of auditory cortex: Music and speech. *Trends in Cognitive Sciences*, *6*, 37–46.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Roswandowitz C, Swanborough H, Frühholz S. Categorizing human vocal signals depends on an integrated auditory-frontal cortical network. *Hum Brain Mapp*. 2021;42:1503–1517. https://doi.org/10.1002/hbm.25309