Acoustic and higher-level representations of naturalistic auditory scenes in human auditory and frontal cortex

NeuroImage 173 (2018) 472–483

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

Acoustic and higher-level representations of naturalistic auditory scenes in human auditory and frontal cortex

Lars Hausfeld [a,b,*], Lars Riecke [a,b], Elia Formisano [a,b,c]

[a] Department of Cognitive Neuroscience, Faculty of Psychology and Neuroscience, Maastricht University, 6200 MD Maastricht, The Netherlands
[b] Maastricht Brain Imaging Center, 6200 MD Maastricht, The Netherlands
[c] Maastricht Centre for Systems Biology, 6200 MD, Maastricht, The Netherlands

ARTICLE INFO

Keywords:
Audition
Selective attention
Auditory scenes
High-field MRI
Attention networks
MVPA

ABSTRACT

Often, in everyday life, we encounter auditory scenes comprising multiple simultaneous sounds and succeed to selectively attend to only one sound, typically the most relevant for ongoing behavior. Studies using basic sounds and two-talker stimuli have shown that auditory selective attention aids this by enhancing the neural representations of the attended sound in auditory cortex. It remains unknown, however, whether and how this selective attention mechanism operates on representations of auditory scenes containing natural sounds of different categories. In this high-field fMRI study we presented participants with simultaneous voices and musical instruments while manipulating their focus of attention. We found an attentional enhancement of neural sound representations in temporal cortex - as defined by spatial activation patterns - at locations that depended on the attended category (i.e., voices or instruments). In contrast, we found that in frontal cortex the site of enhancement was independent of the attended category and the same regions could flexibly represent any attended sound regardless of its category. These results are relevant to elucidate the interacting mechanisms of bottom-up and top-down processing when listening to real-life scenes comprised of multiple sound categories.

Introduction

The ability to listen selectively in noisy environments is pivotal to our everyday behavior. For example, when sitting in a café, we can listen to the voice of a friend, the tunes of a music ensemble or clanging noises of crockery resulting from simultaneous sound sources. Our brain analyzes these mixtures of sounds - often referred to as auditory scenes - by converting the acoustic signal into source-specific neural representations (auditory-scene-analysis [ASA]; Bregman, 1990). Facing complex scenes, auditory selective attention allows us to follow the sound source we are interested in while disregarding other sources (e.g., attending to our friend's voice while ignoring the musical tunes and clanging noises).

Electrophysiological studies in animals have established that neurons in primary auditory cortex can rapidly adapt their spectro-temporal receptive fields to meet the demands of a current behavioral task (Atiani et al., 2009; Fritz et al., 2003). These top-down modulations are specific to the task-relevant acoustic feature (e.g. the detection of a target tone of a given frequency) and lead to both enhanced processing of that attended feature and suppressed processing of unattended features. In humans, functional MRI (fMRI) studies have shown that attention to spectrally non-overlapping tones within a limited frequency band in a multi-tone scene selectively enhance the tonotopic representation of this target frequency band in primary auditory cortex (AC; Da Costa et al., 2013; Paltoglou et al., 2009; Riecke et al., 2016).

Recently, several studies investigated the cortical responses to mixtures of natural speech and focused on the selective enhancement of attended voices compared to unattended ones. These multi-talker stimuli in combination with electro-corticography (ECoG; Mesgarani and Chang, 2012; Zion Golumbic et al., 2013a, 2013b), magneto-encephalography (MEG; Ding and Simon, 2012a, 2012b; Kerlin et al., 2010) and electro-encephalography (EEG; Horton et al., 2013; O'Sullivan et al., 2015) in humans revealed effects of top-down talker-selective attention on neural responses to the speech streams. In particular, spectrograms (Mesgarani and Chang, 2012) or speech envelopes (Ding and Simon, 2012a; O'Sullivan et al., 2015; Zion Golumbic et al., 2013a, 2013b) reconstructed from recordings were more similar to spectrograms or envelopes of the speech signal from the attended talker compared to the unattended one. These results suggest that the neural representations of

* Corresponding author. Department of Cognitive Neuroscience, Faculty of Psychology and Neuroscience, Maastricht University, P.O. Box 616, 6200 MD Maastricht, The Netherlands.
E-mail address: lars.hausfeld@maastrichtuniversity.nl (L. Hausfeld).

https://doi.org/10.1016/j.neuroimage.2018.02.065
Received 19 December 2017; Received in revised form 12 February 2018; Accepted 28 February 2018
Available online 6 March 2018

List of abbreviations

CII	category information index
AC	auditory cortex
SFG	superior frontal gyrus
MFG	middle frontal gyrus
IFG	inferior frontal gyrus
IFJ	inferior frontal junction
CS	central sulcus
FEF	frontal eye field
AIC	anterior insular cortex
PP	planum polare
HG	Heschl's gyrus
PT	planum temporale
mSTG/S	middle superior temporal gyrus/sulcus
pSTG/S	posterior superior temporal gyrus/sulcus
TPJ	temporo-parietal junction
aIPS	anterior inferior parietal sulcus
pIPS	posterior inferior parietal sulcus
FTS	first transverse gyrus
CSI	circular sulcus of the Insula

attended streams are selectively enhanced also in auditory scenes that contain speech sounds with overlapping acoustic features.

Beyond auditory cortical regions, models of attention and top-down control have emphasized the role of extended fronto-parietal networks (Corbetta et al., 2008; Corbetta and Shulman, 2002), with presumably supramodal nodes in a ventral and dorsal network. In particular, the activity of neurons in ventral prefrontal cortex has been shown to reflect the neural representations of (multiple) visual categories (e.g., Cromer et al., 2010; Freedman et al., 2001) and the task-related processing of complex auditory objects (Cohen et al., 2009; Hill and Miller, 2010). Thus, ventral prefrontal areas may be the source of the observed task-related receptive field changes in auditory cortex (Ahissar et al., 2009; Atiani et al., 2014). Areas in the dorsal network have been linked to auditory stream and figure-ground segregation (Cusack, 2005; Teki et al., 2011; Arnott et al., 2004; Zatorre et al., 2002b). These studies suggest that areas beyond auditory cortex play important roles for auditory selective attention and, more specifically, the enhancement of representations of attended sound categories and objects (Bizley and Cohen, 2013; Griffiths and Warren, 2004; Shinn-Cunningham, 2008).

The aforementioned evidence suggests that when listening to mixtures of simple sounds or speech, neural activity reflects more the activity evoked by the attended stream in isolation than the activity evoked by the ignored stream. However, due to the limited cortical coverage of ECoG recordings and the limited spatial resolution of MEG and EEG, the previous studies leave unclear whether this selective enhancement is confined to specific regions within auditory cortex or extends to non-auditory cortex. Another unresolved question is whether selective attentional enhancement applies to more natural situations in which speech is accompanied by real-life sounds of other categories such as music.

In the present study, we examined whether the selective attention mechanism also operates on representations of different natural sound categories by combining whole-cortex high-field (7 T) fMRI and complex auditory scenes containing natural speech and musical instruments. To enforce that participants paid attention to a given target sound category, they were engaged in within-target category identification tasks. We first determined the network of cortical areas that responds to sounds and subsequently defined regions-of-interest (ROIs), within and beyond temporal cortex. To measure cortical representations of sound categories under different task demands we adopted a measure (here called *category information index* or *CII*; Fig. 1) that has been applied in visual fMRI studies to show that representations of visual categories in object-selective visual cortex are modulated by attention (Peelen and Kastner,

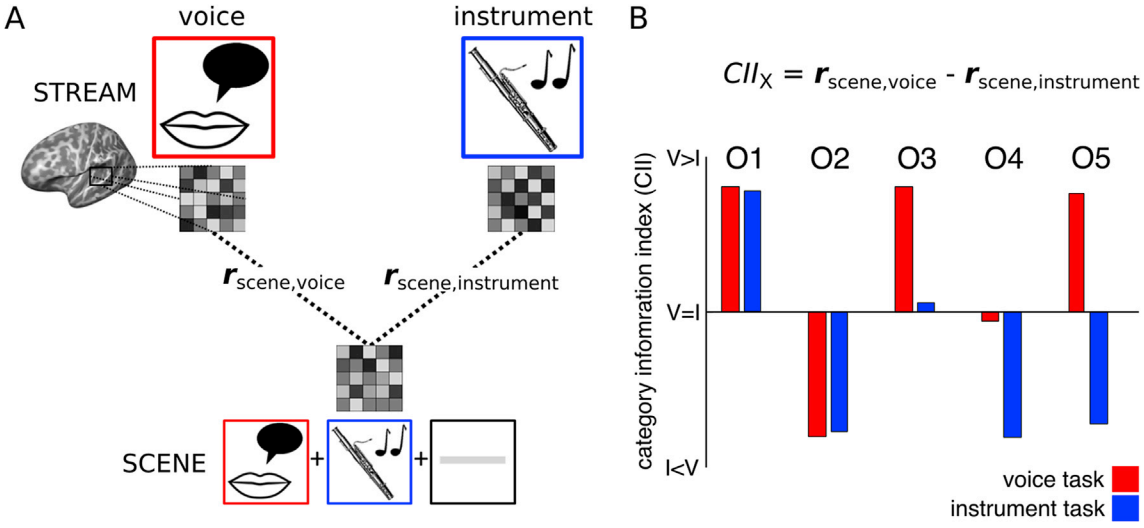

Fig. 1. Overview of Pattern Analysis Approach and Schematic of Possible Outcomes. Panel A visualizes the applied data analysis approach. For a given task (voice, instrument or tone task), the category index (CII) is defined by calculating the similarity between the cortical activation pattern evoked by an auditory scene and the activation patterns evoked by each single stream separately. The CII is computed by subtracting the similarity measure obtained with the instrument-stream pattern from the similarity measure obtained with the voice-stream pattern. Panel B shows possible outcomes for CII. O1 shows a case in which during both tasks the similarity between the activation pattern during the scene presentation and the voice stream is larger than the similarity to the instrument stream. This corresponds to a region that represents voice sounds independent of attention. Outcome O2 is similar to outcome O1 but indicates a region representing instrument sounds. Outcome O3 shows a strong similarity between the voice scene and the voice stream pattern compared to the one observed with the instrument stream (same as O1). In contrast to O1, the observed pattern during the instrument scene is equally similar to both the voice and instrument scene condition. This CII-pattern indicates a region that shows high attentional modulation (i.e., difference between voice and instrument CII) and enhanced representation of the voice sound within the scene but only when attended. O4 is similar to O3 but suggests a region with an enhanced representation for instrument but not voice sounds during scenes and the respective tasks. The results in O5 indicate a region that shows both enhancement to voice sounds during the voice task and enhancement to instrument sounds during the instrument task. This suggests enhanced representations of the attended sound independent of the task and, thus, flexible encoding of the task-relevant sound.

2011; Peelen et al., 2009). We measured spatial activation patterns within ROIs during presentations of auditory scenes while listeners focused attention on a specific sound category. We also measured patterns during presentations of each individual category alone (i.e., in the absence of the other category). Using the CII we then computed the similarity between the measured patterns. In this way, we determined the task-dependent enhancement of spatially distributed category representations in various human cortical regions. Based on the previous findings, we hypothesized that selective attention to natural sound categories acts by strengthening cortical representations of the attended category (and its defining acoustic features) in comparison to unattended categories (*category-selective enhancement hypothesis*). We predicted that in auditory cortex (AC) the enhancement of sound category representations occurs in distinct regions that have been reported to respond selectively to these different categories (e.g., Alho et al., 2014; Belin et al., 2000; Leaver and Rauschecker, 2010; Norman-Haignere et al., 2015). A different representation of categories has been found in prefrontal cortex of primates in which activity of the same neuronal populations represents various independent visual categories (Cromer et al., 2010; Meyers et al., 2008). Thus, frontal and parietal regions might show either region-specific enhancement effects similar to AC or enhancement of representations of the attended category in the *same* region similar to primate prefrontal cortex. This second alternative would suggest that these cortical regions can encode flexibly various sound categories. We find that independent of listeners' attentional focus, similar networks spanning temporal, frontal and parietal cortex are active during auditory identification tasks. Importantly, our results show that selective attention to a specific sound category in a scene leads to selective enhancement of representations of the attended sound category in temporal and frontal regions. In conformance with our predictions, the location of task-dependent enhancement in auditory cortex is *category-dependent:* for attention to speech, this enhancement is located in middle and posterior superior temporal gyrus and sulcus, whereas for attention to musical instruments, it is confined to right planum polare. In contrast, the location of task-dependent enhancement in frontal cortex is *category-independent*: representations of different sound categories are enhanced by category-specific attention in common frontal regions. Our findings indicate that in AC the category of the attended sound (and its category-defining acoustical features) determines the locus of enhancement. In frontal cortex, fixed regions represent attended sounds regardless of their category. In sum, these results support the category-selective attentional enhancement hypothesis in temporal and frontal cortices. In temporal cortex, category-specific attentional modulations are more spatially separated whereas in frontal cortex the same regions can flexibly encode the sound category as required by ongoing task demands.

Materials and methods

Participants

10 participants (age range: 20–30 years, mean age±s.d.: 23.5 ± 3.4 years; 7 female) took part in this study and had normal or corrected-to-normal vision and reported normal hearing and no language or speech difficulties. None of the participants received training in music theory or for playing a musical instrument for longer than one year, reported to have perfect pitch, was raised bilingually or could speak a tonal language. The study was approved by the local ethics committee at the Faculty of Psychology and Neuroscience (Maastricht University; # ECP_10_05_2012).

Sound stimuli

The stimuli presented in this study (see Hausfeld et al. (2017) for data deposition including (f)MRI data, protocol and stimuli; supplementary audio files 1-6) were either single sounds of a voice, instrument or a pure tone (*stream*) or a mixture of the three sounds (*scene*). Voice stimuli were of same gender (female) and instruments belonged to the same orchestral family (wind instruments).

Supplementary audio related to this article can be found at https://doi.org/10.1016/j.neuroimage.2018.02.065.

Voice stimuli were short utterances (between 800 and 1000 ms) of one word. To reduce the amount of variability between voice stimuli, words were chosen according to semantic and acoustic criteria via the celex database (celex.mpi.nl; Max-Planck Institute for Psycholinguistics, Nijmegen, The Netherlands): (1) word were abstract nouns and consisted of two syllables, (2) syllables were separated by stop consonant/d/,/k/,/t/, or/b/, (3) the second syllable did not include a schwa (ə; e.g., *er, *en, *e), and (4) the word frequency was between 20 and 400 words per million within the INL corpus (Institute of Dutch Lexicology). Words were recorded from two female native Dutch talkers with fundamental frequencies of 230 and 216 Hz (*V1* and *V2*, respectively; analysis done with Praat [Boersma and Weenink, 2015]). In total, 36 words were presented such that two non-overlapping sets of 18 words were presented during fMRI and behavioral testing.

Sounds from two wind instruments (flute and bassoon; *I1* and *I2*, respectively) were created by synthesizing MIDI files with Logic Pro (version 9.1, Apple Inc., Cupertino, CA). Midi files specified sequences of two series of four notes. We synthesized 18 sequences of four notes from two woodwind instruments (flute and bassoon; *I1* and *I2*, respectively) which were semitone transpositions of each other and overlapped in frequency range (flute: G4 [392 Hz], G-sharp4 [415 Hz], A-sharp4 [466 Hz] and B4 [494 Hz]; bassoon: G-sharp4 [415 Hz], A4 [440 Hz], B4 [494 Hz] and C5 [523 Hz]). Fundamental frequencies of tones and voices showed a similar difference (logarithmic scale) between the two sources ($F0_{V1}$-$F0_{V2}$: .063logHz; $F0_{I1}$-$F0_{I2}$: .058logHz). Eighth notes were played at a tempo of 135bpm and led to sequences lasting 1s including attack and decay. Different sequences were created by changing the order of notes.

Tone sounds were 310 Hz pure tones (approximate geometric mean of fundamental frequencies of voice and instrument stimuli) and lasted 1.25s. Tones could either be continuous (*T1*) or were interrupted by gaps (*T2*) which lasted 240 ms in total. The gap sequence appeared at random onsets between 400 and 900 ms after tone onset and consisted of three repetitions of 16 ms up- and down-ramps with a gap of 48 ms in between. To avoid clicks, tones were ramped with 20 ms linear on- and offsets.

Sounds were matched for root-mean-square (RMS); the intensity of tones (based on continuous tone T1) was decreased by $18dB_{RMS}$ to create a soft but clearly audible stimulus. Auditory scenes were created by adding voice, instrument and tone stimuli (Fig. 1) and lasted 1.25s. The tone sound was present during the whole interval; voices and instruments started at 150 ms after sound onset lasting until 100 ms before offset.

Tasks

Participants were asked to perform three tasks. In the voice task participants listened to a mixture of sounds comprising a voice, an instrument and a pure tone (see section 2.2) and were reported whether voice V1 or V2 was presented. In the instrument task, participants listened to the same sound mixtures but now were asked they heard instrument I1 or I2. These within-category tasks served to draw attention to the voice or instrument category. During the tone task, participants were asked to identify whether the presented tone was continuous or interrupted (i.e., gap detection). Participants first underwent behavioral sessions in which they were familiarized with the stimuli and practiced the tasks to achieve an accuracy of 75% or higher for each task. For the main analysis, only data from the voice and instrument task were used. Data during the tone task was acquired only for comparison purposes.

Experimental procedure

Each session consisted of six functional runs. In functional runs, 24 auditory scenes and 6 streams (12 presentations of each voice,

instrument and tone type during scenes and 3 presentations of each task-relevant type during streams) were presented with an ITI of 15s. Thus, 180 trials were presented in total for each task (144 scenes and 36 streams). Sounds were played binaurally in silent gaps via MR-compatible ear-buds (Sensimetrics S14, Sensimetrics Corporation, Malden, MA) and started 125 ms after image acquisition. Participants responded by button presses in silent scanning-free periods to ensure a clear perception of sounds. Auditory scene stimuli presented in the different task conditions were physically the same. To avoid the build-up of order expectations, the probability of the task-relevant type was equalized over two preceding sounds (e.g., during the voice task V1 and V2 appeared with equal probability after previous V1-V1, V1-V2, V2-V1 and V2-V2 presentations).

MRI data acquisition

Brain imaging was performed with a 7 T Siemens scanner (head coil) at the Maastricht Brain Imaging Center (Maastricht, The Netherlands). Nine of ten participants were scanned on three occasions with the first and third session being less than 10 days apart (S3 was scanned on two days; runs were divided into virtual sessions matching those of the other participants).

Anatomical scans were acquired during each session with a T1-weighted MPRAGE sequence for three participants (voxel size: 0.6 mm isotropic; 256 slices; field-of-view [FoV]: 230 mm × 230 mm; repetition time [TR] = 3100 ms; echo time [TE] = 2.52 ms; GRAPPA 3) and corrected for intensity inhomogeneity using a proton-density weighted MPRAGE. For the remaining seven participants, anatomical images were acquired with an MP2RAGE sequence (Marques et al., 2010; voxel size: 0.65 mm isotropic; 240 slices; FoV: 208 mm; TR: 5000 ms; TE: 2.51 ms; GRAPPA 2) and masked with the second inversion contrast.

The 18 functional runs were acquired in three sessions consisting of 6 functional runs and one participant (S3) had two sessions of 10 and 8 runs. In each functional run we collected 154 vol using an echo-planar-imaging (EPI) sequence with multiband 2 acceleration and silent periods of 1400 ms between volume acquisitions (64 slices [S3 with 62 slices]; voxel size: 1.5 mm isotropic; FoV = 204 × 204 mm; TR = 3000 ms; acquisition time = 1600 ms; TE = 19 ms; GRAPPA 3). For correcting EPI distortions two sets of five images were acquired in opposite phase encoding directions during each scanning session.

Data preprocessing

Preprocessing of both functional and anatomical data was performed with BrainVoyager QX (v2.8.4, Brain Innovation, Maastricht, The Netherlands) if not stated otherwise. FMRI data preprocessing consisted of slice-scan-time correction, motion correction, EPI distortion correction, temporal high-pass filtering (7 cycles per run ≈ 0.015 Hz) and spatial smoothing (1.5 mm FWHM). EPI distortions were corrected with the topup algorithm (Andersson et al., 2003; as implemented in FSL [v5.0.6; Smith et al., 2004]). Functional runs were individually aligned to the session offering the best quality of anatomical scans and transformed to Talairach space (Talairach and Tournoux, 1988).

Data analysis

fMRI univariate analysis

To estimate voxel-wise responses to stimuli, a fixed-effects general linear model (GLM; Friston et al., 1994) was computed for each participant by fitting the blood oxygen level-dependent (BOLD) response with predictors coding for the task conditions (voice, instrument and tone), sound types (scene and stream) and confound predictors coding for motion (3 translations and 3 rotation predictors). Group-level statistics for overall activation and activation differences (Fig. 2, see Supplementary Fig. 1) was performed with individual beta estimates and voxel-wise paired *t*-tests after projecting activation maps from the individual onto the group aligned surface via cortex-based alignment (Goebel et al., 2006). Activation maps and contrasts were corrected for multiple comparisons with false-discovery rate (FDR; Benjamini and Hochberg, 1995; Genovese et al., 2002). For higher sensitivity, univariate contrasts between tasks were corrected with surface-based cluster-size thresholding (Forman et al., 1995) using a liberal initial threshold of $p < .01$.

Definition of region-of-interest

ROIs on the superior temporal plane were individually defined according to anatomical criteria from Kim et al. (2000); (similar to Moerel et al., 2013). We divided the superior temporal plane into three areas: Heschl's gyrus (HG), planum polare (PP) and planum temporale (PT). Other ROIs were defined in frontal cortex (superior frontal gyrus, SFG; middle frontal gyrus, MFG; inferior frontal gyrus, IFG; inferior frontal junction, IFJ; frontal eye field, FEF; anterior insular cortex, AIC; and left central sulcus, CS, covering primary motor and somatosensory regions M1 and S1), temporal cortex (middle portion of superior temporal gyrus/sulcus, mSTG/S; posterior portion of superior temporal gyrus/sulcus, pSTG/S) and parietal cortex (anterior inferior parietal sulcus, aIPS; posterior inferior parietal sulcus, pIPS; temporo-parietal junction, TPJ) based on previous studies and anatomical landmarks.

HG was defined as the first transverse gyrus (FTS) on the temporal plane. Medially, the HG border was defined by the circular sulcus of the Insula (CSI). Its antero-medial border was defined by the FTS and the

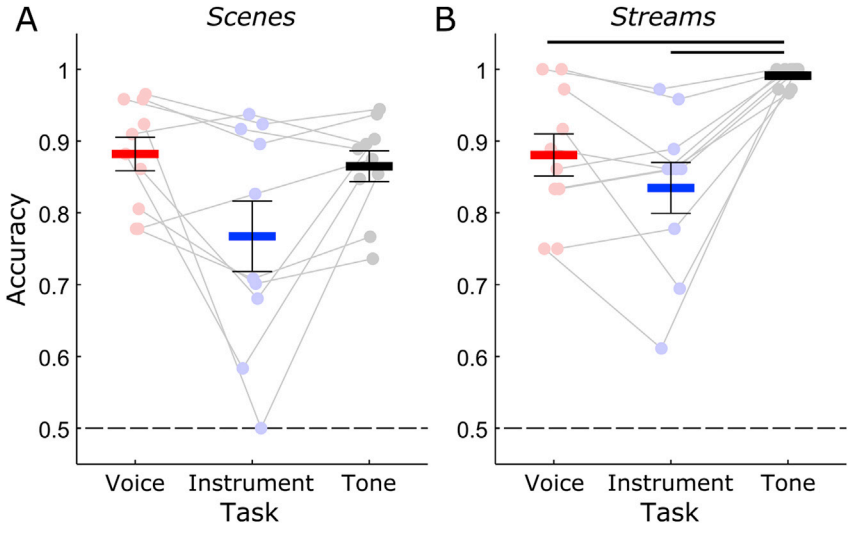

Fig. 2. Behavioral Performance during FMRI. Connected data points show individual participants' behavioral accuracy during the three tasks, for the scene condition (panel A) and the single-stream condition (panel B). Thick red, blue and black lines show the average accuracy during the voice, instrument and tone task, respectively (error bars indicate ±s.e.m.). Upper lines denote significant differences between conditions (corrected for multiple comparisons; $p_{FWE} < .05$). Dashed lines represent chance performance.

postero-lateral border by the Heschl's sulcus or – if present – by the intermediate sulcus. Second, PP was confined by the CSI (medial), the HG ROI (posterior) and the rim of the STG (lateral). The PT region was defined by the HG ROI (anterior) the deepest point of the Sylvian fissure (medial), the rim of the STG (lateral) and the most posterior point of the temporal plane.

Based on anatomical landmarks and previous studies we defined regions in frontal, temporal and parietal cortex. We narrowed the extent of the these initial region definitions using functional activation maps (FDR-corrected, $q < .05$) obtained during the tone task (scenes) to create the individual ROIs (Fig. 5) in frontal (SFG, MFG, IFG, IFJ, FEF, AIC, and left MC), parietal (aIPS, pIPS and TPJ), temporal cortex (mSTG/S and pSTG/S). SFG and TPJ were identified as deactivated areas whereas the other regions included vertices showing activation.

Multivariate (category information) analysis

We tested for information content of patterns by adapting a *Category Information* measure which is defined by correlations of spatial activation patterns (Peelen and Kastner, 2011; Peelen et al., 2009). We defined the *category information index* (CII) as the difference between similarities of voxel activation patterns (Fig. 1A). To avoid an influence of session-specific effects on this measure, CII was computed by using two independent sets of data from different fMRI sessions. This led to three different divisions of two (set 1; data from 1 session) and four runs (set 2; data from 2 sessions). We computed r_{XY} which denoted the Pearson correlation between activation patterns of parameter estimates for *scenes* of condition *X* from set 1 with patterns of parameter estimates of stream conditions *Y* from set 2. Responses to scene and stream conditions were estimated from independent GLMs over 2 and 4 runs, respectively. Before further analysis these correlation coefficients were Fisher transformed ($z = 0.5\ln[(1 + r)/(1\text{-}r)]$) and averaged across the three session combinations.

The CII was defined with respect to two templates: activation patterns of voice stream and instrument stream conditions. It was computed by subtracting the instrument-category from voice-category pattern similarity: CIIX = zXsce-Vstr - zXsce-Istr. Importantly, CIIvoice, CIIinstrument and CIItone were based on pattern comparisons with the same physical stimuli but different task demands. Determining the CII of the tone scene condition revealed response biases towards one of the two categories when attention was directed to a simple tone that did not overlap with fundamental frequencies or higher harmonics of the voice and instrument stimuli. The CII for two activation patterns was computed from independent datasets and GLMs and, hence, generalized across measurements and sound items. Possible outcomes and interpretations of the CII are described in Fig. 1B. The statistical tests of CII for each task were without directional hypothesis (i.e., two-sided), whereas for task modulation we tested whether the CII during the voice task was larger than the CII during the instrument task (i.e., one-sided; see outcomes O3, O4, O5 in Fig. 1).

To ensure that the CII did not depend on our chosen similarity measure, we computed additionally the CII_{euc} based on Euclidean distance: $CII_{eucX} = 1 - d\,(Xsce_n, Vstr_n) - (1 - d\,(Xsce_n, Vstr_n)) = d\,(Xsce_n, Istr_n) - d\,(Xsce_n, Vstr_n)$ where $d\,(g_n, h_n)$ denotes the Euclidean distance of vectors g and h each normalized to unit length. This led to similar outcomes (see supplementary Figs. 3, 4, and 6) indicating that these results did not depend on the choice for a particular similarity measure.

Results

Behavioral results and cortical network activity during task performance

The behavioral performance (accuracy) during MRI acquisition (Fig. 2) showed that participants performed well above chance level, suggesting that they paid attention to the task-relevant sound (voice task: $.88 \pm .02$ [$.88 \pm .03$]; instrument task: $.77 \pm .05$ [$.83 \pm .04$]; tone task: $.87 \pm .02$ [$.99 \pm .01$]; for scene [stream] sounds; mean ± s.e.m.). Performance in the scene conditions was lower compared to stream conditions ($t(9) = 4.28$, $p < .01$, paired *t*-test; pooled across tasks), reflecting increased processing effort during auditory scenes (note that this difference was mostly driven by the tone task). The performance for all three tasks in the scene conditions was approximately equal (voice vs. instrument: $t(9) = 2.46$, $p_{FWE} = .11$,; voice vs. tone: $t(9) = 1.02$, $p_{FWE} > .5$; instrument vs. tone: $t(9) = -2.31$, $p_{FWE} = .14$) but with a tendency for higher performance during the voice vs. instrument task, and a larger variability for the instrument task. Similarly, our results suggested that performance was similar during the voice and instrument task in stream conditions ($t(9) = 1.69$, $p_{FWE} = .37$). This indicates that task difficulty was similar for the experimental conditions used to compute the CII.

Fig. 3 shows the cortical activation in response to sound stimuli during scene and stream presentations for each experimental task as determined by a general-linear model (GLM). The network for auditory and task-related processing was characterized by activation on the temporal plane (incl. PT, HG and PP), superior temporal gyrus and sulcus (incl. mSTG/S and pSTG/S), inferior parietal sulcus (aIPS and pIPS) and frontal cortex (IFG, IFJ, MFG, FEF and AIC). Sites of deactivation were found in SFG, TPJ and precuneus. In addition, areas presumably related to behavioral responses (i.e. button presses) were found in right M1 and S1 and bilateral premotor areas. These networks were observed during both scene and stream presentations and were similar between tasks: contrasting activation levels of tasks did not reveal differences for auditory scenes (voxel-wise paired *t*-tests, $q > .10$). Thus, this voxel-by-voxel analysis of activation levels did not provide evidence for the category-selective enhancement hypothesis. However, we found trends for higher activation for the instrument task compared to voice and tone task in left IFG, higher activation during the voice compared to the instrument task in left pSTS, and higher activation in right TPJ for the tone task compared to voice and instrument task (Fig. 4).

In sum, the behavioral results and activation maps indicate that sensory and cognitive processing demands were similar across tasks and suggest that further results were not strongly influenced by differences in task difficulty or cognitive strategies. As earlier studies have shown that the representation of sound categories may be reflected in spatial activation patterns rather than single-voxel activation levels (e.g., Staeren et al., 2009), we next examined category-selective enhancement of spatial activation patterns representing specific categories.

Fig. 5 shows the overlap of individual ROIs (15 in left hemisphere and 14 in right hemisphere) and their extent which were defined based on anatomical landmarks and functional activation measured during the tone task (average activation of ROIs for scene and stream sounds are presented in Supplementary Figs. 1 and 2, respectively [Insert Supplementary Figs. 1 and 2 about here]).

Task-dependent enhancement of sound category representations in temporal cortex

We used the CII to assess cortical representations of auditory categories and their modulation by category-specific attention within each of the individual ROIs. The CII was based on activation patterns evoked by the same auditory scenes but varying task conditions (i.e., attention to voice, instrument) and template patterns evoked by isolated sounds of a voice or an instrument (i.e., without distractor) (see section 2.7.3). To test whether the CII was affected by ROI size, we performed the same analysis with 500 random subsamples of 100, 200 and 400 vertices within each ROI and found that the CII with the original number of vertices was within the inner quartile range of the CII obtained with random subsamples (left hemisphere: 38.6–54.2 [lowest percentile – highest percentile], right hemisphere: 36.4–60.6; lowest and highest percentile denote the minimum and maximum percentile of the original results across the CII types [i.e., voice task, instrument task, voice-instrument task]). This suggests that the estimated CII did not depend on the size of the ROIs.

In line with our hypothesis for temporal cortex, we found a positive

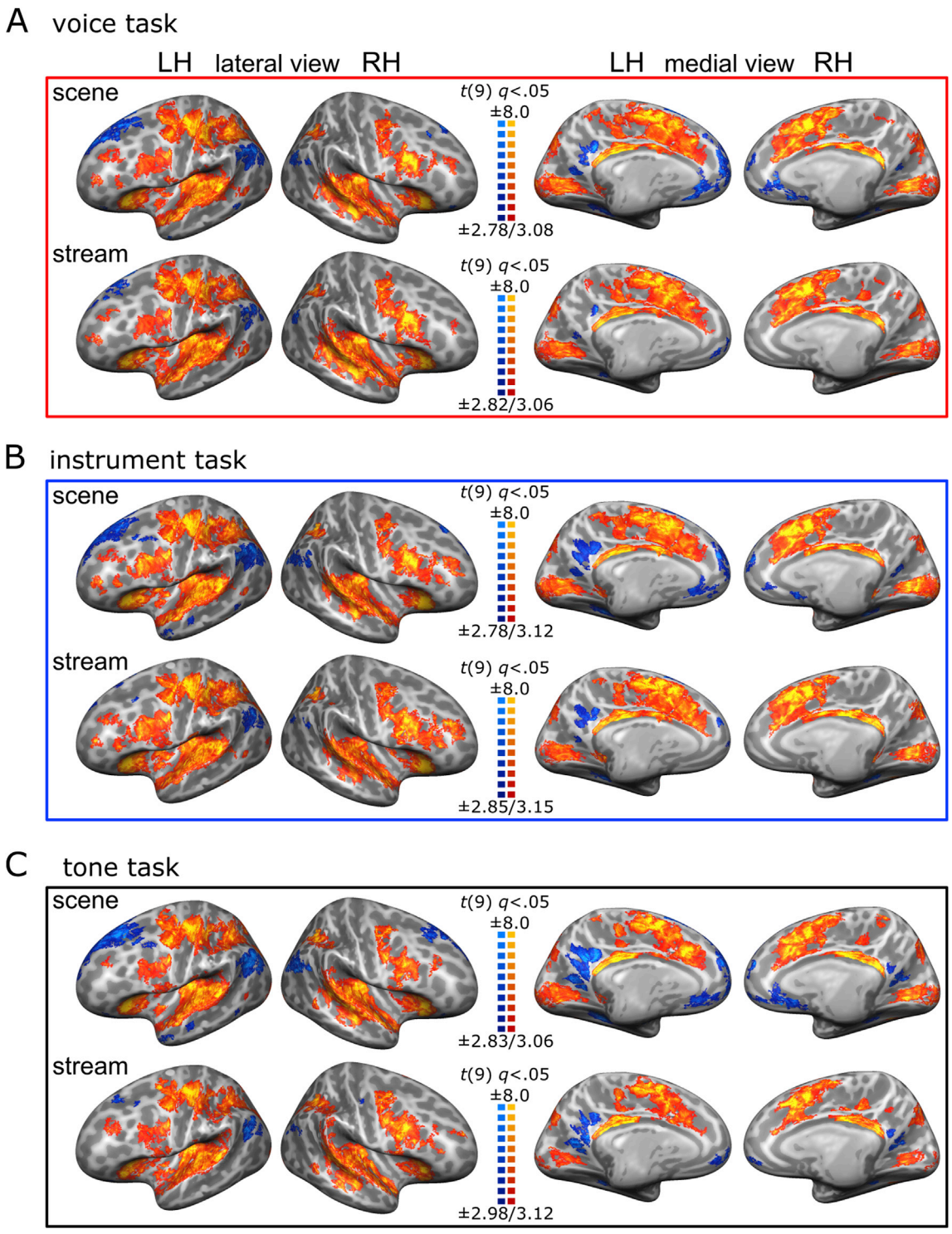

Fig. 3. Cortical Activation during Task Execution. Group maps of cortical activation and deactivation (compared to baseline) are shown on average cortical surfaces separately for each of the three tasks (panels A–C: voice, instrument and tone task, respectively) for each stimulus condition (top: scene and bottom: streams). Maps are corrected for multiple comparisons (FDR; $q < .05$) within each hemisphere (left threshold for LH, right for RH).

CII for the voice task (Fig. 6; see Supplementary Fig. 3 for CII_{euc}) in left and right mSTG/S and pSTG/S. This shows that - when attending to the voice category-the similarity between the scene activation pattern and the voice template pattern is higher than the similarity between the scene activation pattern and the instrument template pattern (paired *t*-test, two-tailed, multiple comparison corrected by false-discovery rate [FDR; Benjamini and Hochberg, 1995], $q < .05$). In contrast to the voice task, the CII during the instrument task (and tone task) was not different from zero in these regions. Moreover, CII was higher during the voice task compared to the instrument task (paired *t*-test, one-tailed, $q < .05$) indicating that category representations were modulated by task. These results provide strong evidence for enhancement of voice category representations in left and right STG/STS under voice-selective attention in auditory scenes. Interestingly, we found that the enhancement in right and left STG/S was category-specific, i.e. it was only observed when participants attended to the voice but not to the instrument ($|t(9)| \leq 2.39$,

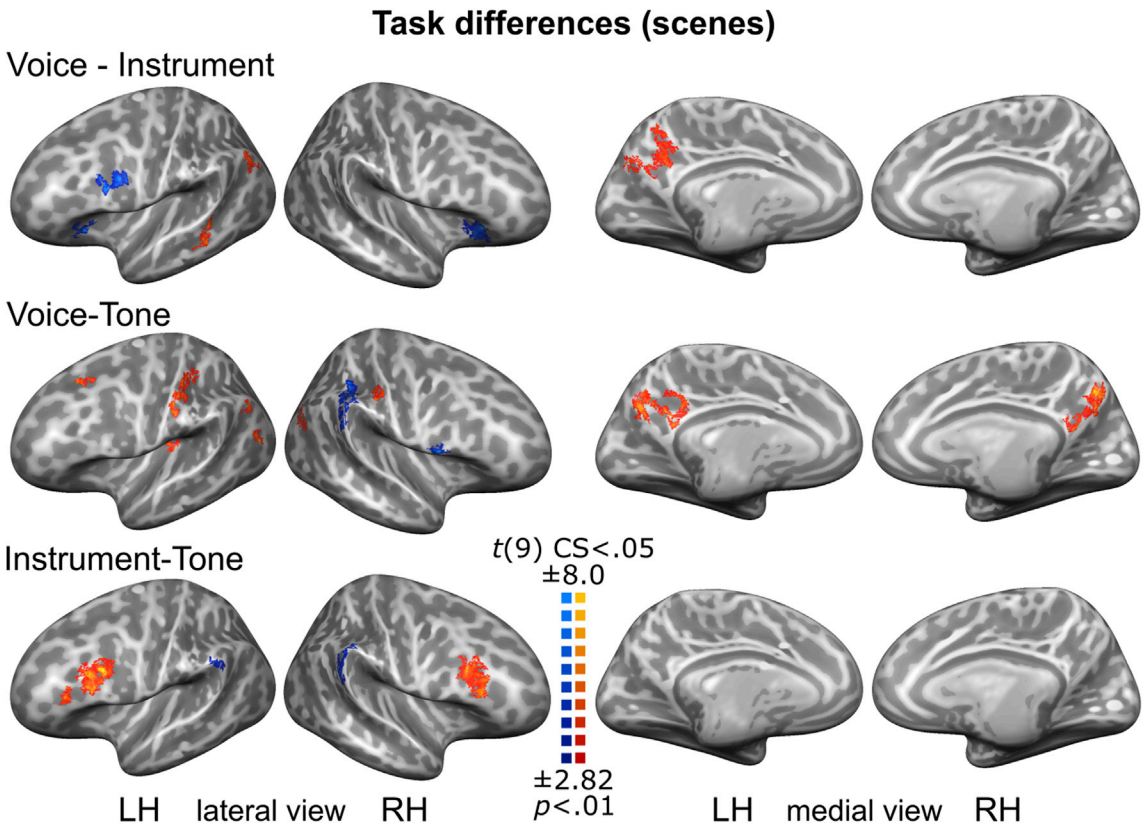

Fig. 4. Cortical Activation Differences between Tasks for Auditory Scenes. Group maps of activation differences between the three tasks (voice and instrument: upper row; voice and tone: middle row; instrument and tone: lower row) in the scene condition (i.e., given fixed acoustic input). Difference maps are corrected for multiple comparisons using cluster-size thresholding (initial threshold $p < .01$ and cluster-size probability CS < .05) for each hemisphere.

$p_{FDR} > .420$) This indicates that natural sound mixtures containing voice-specific acoustics were actively processed in middle and posterior STG/S according to behavioral demands.

The opposite pattern was observed for the instrument task in PP and PT: these regions showed a negative CII when participants attended to instrument sounds and this effect was right-lateralized (Fig. 6; right vs. left hemisphere: PP: $t(9) = 2.56$, $p = .031$, uncorrected, PT: $t(9) = 5.38$, $p_{FDR} = .006$; paired t-test, two-tailed). Moreover, these regions showed no significant CII when listeners focused on the voice category ($|t(9)| \leq 2.40$, $p_{FDR} > 0.169$). In right PP, the enhanced representation was specific to the instrument category; thus when listeners performed the instrument task, representations of the auditory scene were rendered similar to the representation evoked by the isolated instrument sound. This indicates that sound mixtures containing instrument-specific acoustics were actively processed in right PP, analogous to voices in STG/S. In both right PP and PT, the CII was lower during the instrument compared to the voice task showing that activation patterns in these regions are modulated by task demands (paired t-test, one-tailed, $q < .05$). The pattern of results in right PT was similar to that in right PP but, in addition, showed a trend towards positive CII during the voice task ($t(9) = 2.40$, $p > 0.040$, uncorrected) which indicated flexible task-related sound encoding that evokes representations more similar to the voice template pattern when performing the voice task and more similar to the instrument template pattern when performing the instrument task. These differential effects of task-dependent category-selective processing in right PP for instruments and right pSTG for voices are reflected in an interaction of region of interest and task (ROI × task interaction: $F(1,9) = 19.93$, $p = .002$; repeated-measures ANOVA).

For control, we also calculated the CII for the tone task. In temporal cortex, it was not different from zero ($|t(9)| \leq 1.59$, $p_{FDR} > .596$), which shows that activation patterns evoked by auditory scenes during the tone task were neither more similar to the voice-template pattern nor more similar to the instrument-template pattern (see Supplementary Figs. 5 and 6).

Task modulation of representations in frontal cortex

Fig. 7 depicts the results for ROIs in prefrontal cortex (see Supplementary Fig. 4 for CII_{euc}). Overall, the CII values were on average smaller than in temporal cortex, which may be related to higher response variability. Analyzing for category information, we find that left IFG and right SFG showed positive CII for activation patterns evoked by auditory scenes during the voice task (paired t-test, two-tailed, $q < .05$). For the instrument task (and tone task), the CII was not different from zero (Fig. 5). These results show task-dependent activation patterns in frontal regions that are more similar to voice-template patterns compared to instrument-template patterns when listeners perform a voice task (but not instrument task or tone task). We did not find a significant negative CII during any task in frontal ROIs.

However, in prefrontal regions (i.e. SFG, MFG, IFG, IFJ, AIC), we consistently found a positive CII for activation patterns evoked by scenes for the voice task and a negative CII for patterns for the instrument task. Furthermore, to test the consistency of CII in prefrontal regions we pool the statistical outcomes of these ROIs via partial conjunction of hypotheses (Benjamini and Heller, 2008), which provide probabilities for observing at least x out of n significant tests. The consistency of CII results is reflected in the left hemisphere (voice: $z_3 = 1.81$, $p_3 = .035$, tests that at least 3 ROIs are positive [statistical values for at least 1 and 2 ROIs are $z_1 = 3.28$ and $z_2 = 2.44$]; instrument: $z_1 = 1.87$, $p_1 = .031$; Stouffer's method) and right hemisphere (voice: $z_3 = 2.10$, $p_3 = .018$ [statistical values for at least 1 and 2 ROIs are $z_1 = 3.62$ and $z_2 = 2.57$]; instrument: $z_1 = 2.13$, $p_1 = .017$). This tendency of a positive CII during the voice task and negative CII during the instrument task is reflected in differences of CII between the voice and instrument task. Specifically, we found a

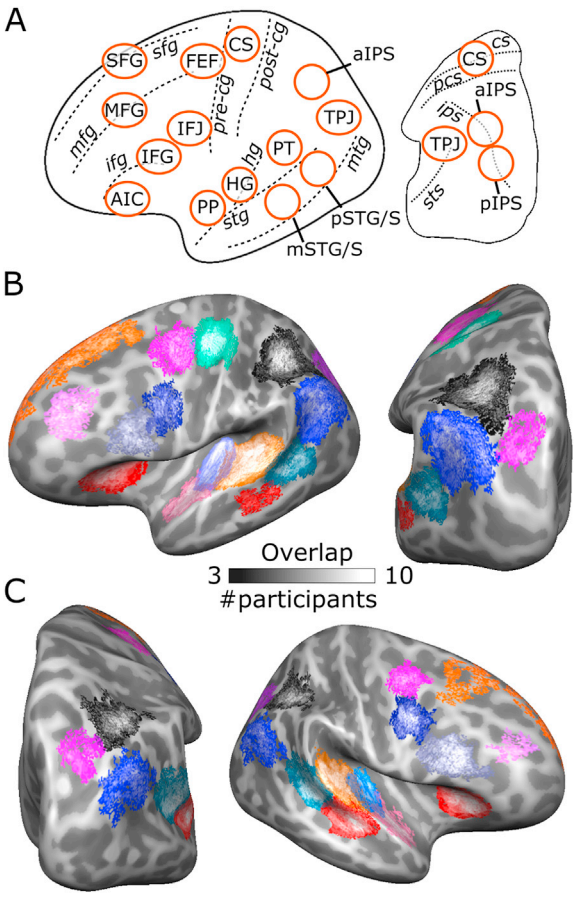

Fig. 5. Overview of Regions-of-Interest. Panel A shows a schematic overview of the analyzed ROIs, which were defined based on anatomical landmarks and functional responses to the tone task. ROIs were defined individually. Their overlap across participants is visualized in panels B and C on participants' average cortical surface obtained from cortex-based alignment for the left and right hemisphere, respectively. See section 2.7.2 for abbreviations.

higher CII for the voice task compared to the instrument task in left and right IFG and MFG as well as right SFG and MFG (paired *t*-test, one-tailed, $q < .05$). The remaining prefrontal regions showed a similar tendency (partial conjunction of hypothesis in left hemisphere: $z_4 = 2.66$, $p_4 = .004$; right hemisphere: $z_4 = 2.34$, $p_4 = .010$, Stouffer's method). For the tone task, the CII in frontal cortex was not different from zero ($|t(9)| = 2.28$, $p_{FDR} > .596$). This suggests that activation patterns evoked by auditory scenes during the gap detection task were as similar to the voice template pattern as to the instrument template pattern.

Altogether, these results show that the similarity of activation patterns with regard to the voice and instrument template patterns in frontal cortex were modulated by the listener's task.

Category information in parietal cortex and motor-control regions

Applying the category information analysis to parietal and motor-control regions (CS; including right M1 and S1) neither revealed a significant CII ($|t(9)| \leq 2.15$, $p_{FDR} > .17$) nor differences of CII during the voice or instrument task ($|t(9)| \leq 1.78$, $p_{FDR} > .18$) (Supplementary Fig. 5 [Insert Supplementary Fig. 3 about here]; see Supplementary Fig. 6 for CII_{euc}). Nevertheless, the parietal and motor-control regions were active during the tasks (Fig. 3, Supplementary Figs. 1 and 2). These observations indicate that these regions contributed to the processing of auditory scenes and task execution, but that the processing in these regions was independent of the current task demand (note that button presses indicated sound identity but not category in our study).

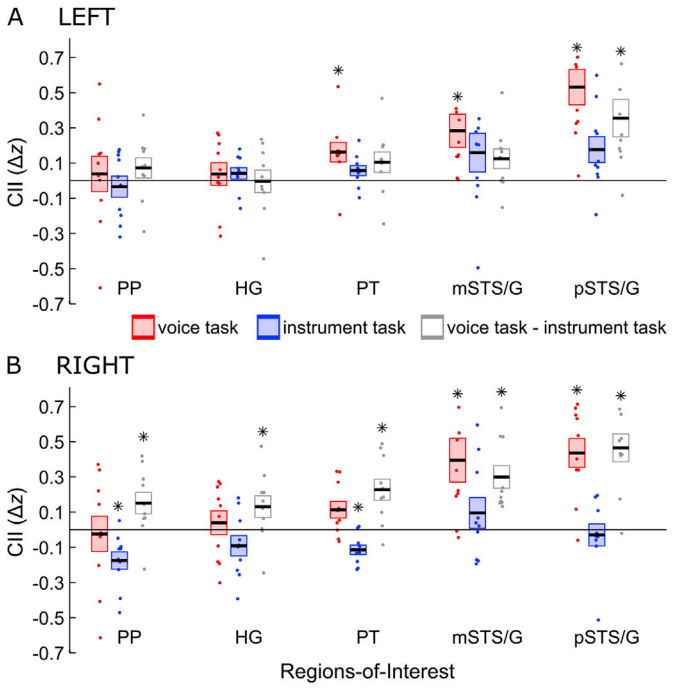

Fig. 6. Category Information Index for Regions in Temporal Cortex. Bar graphs in panel A show the CII for the voice and instrument task, separately for five ROIs in the left temporal cortex. Panels B depicts data obtained from the same ROIs in the right cerebral hemisphere. Boxes and thick black lines denote ±s.e.m. and mean across participants, respectively. Single points show individual CII for each of the two tasks and the difference of the voice task and the instrument task. Asterisks denote significant CII or a significant CII difference between the voice task and the instrument task (*$q < .05$, FDR-corrected across 15 [14] ROIs in LH [RH]).

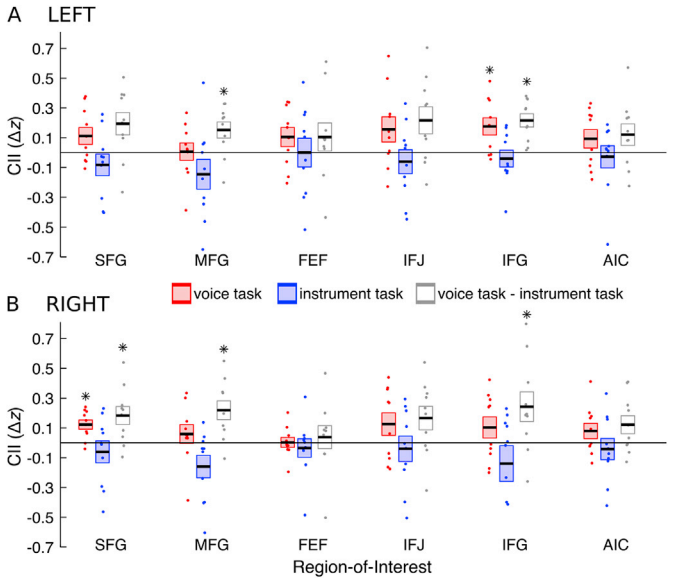

Fig. 7. Category Information Index for Regions in Frontal Cortex. Same as Fig. 4, but for six ROIs in frontal cortex.

Discussion

Previous evidence has shown that cortical representations of attended sounds are enhanced compared to unattended sounds in auditory scenes consisting of basic sounds (Da Costa et al., 2013; Riecke et al., 2016) or speech (Ding and Simon, 2012a; Mesgarani and Chang, 2012). The

underlying processes are still not fully understood but selective spectro-temporal filtering induced by task-related plasticity of neuronal receptive fields seems to play a crucial role (Atiani et al., 2009; Fritz et al., 2007; Lakatos et al., 2013). Here, we extend earlier studies and examined whether attentional selective cortical enhancement also occurs in naturalistic scenes with real-life sounds of different categories. In addition, by exploiting the high localization capability of high-field MRI in humans, we could examine attentional modulation simultaneously in multiple cortical regions.

We hypothesized that selectively attending to a sound category in a naturalistic auditory scene enhances the representation of the attended category and its defining acoustic features. To assess sound representations and test for category-selective enhancement, we used the CII measure (Peelen and Kastner, 2011; Peelen et al., 2009). This showed that, for a subset of ROIs in temporal and frontal cortex, activation patterns measured during selective listening to voices (instruments) in a natural auditory scene were more similar to the patterns evoked by isolated voices (instruments) than those evoked by isolated instruments (voices), respectively. These results thus provide evidence for the hypothesized category-selective enhancement in temporal and frontal cortex. Our results further suggest that specific regions in auditory cortex enhance representations of a specific sound category (i.e., category-dependent enhancement) whereas common regions in frontal cortex enhance the representation of any attended category (i.e., category-independent enhancement). We suggest that this enhancement arises from the neural mechanisms described above, which remains to be studied with invasive methods.

Task-dependent enhancement of natural sound categories in auditory cortex

In temporal cortex, we found that, when attending to a voice in natural scenes, the enhancement of its representations was most prominent in middle and posterior STG/S in both hemispheres. When attending to the instrument the representation of the instrument category was enhanced in right PP. In contrast, representations in STG/S showed only little or no enhancement for the instrument category when attending to the instrument. Likewise, representations in right PP did not show enhancement of the voice category when voices were attended. These findings indicate that when selectively listening to auditory scenes containing natural sound categories, the representation of the attended category in temporal cortex is enhanced at sites that depend on the current task.

The regions in auditory cortex in which we observed task-related enhancement have been shown to respond particularly to either voice/speech sounds or fine spectral variations inherent to sounds of musical instruments. Sites important for voice and speech sounds processing have repeatedly been allocated to STG and STS (Belin et al., 2000; Overath et al., 2015). These sites were established by contrasting activations evoked by speech sounds and various spectral and temporal manipulations thereof or artificial stimuli matched with respect to acoustic properties of speech sounds. Similarly, previous studies suggested a critical role of PP for processing music sounds (Angulo-Perkins et al., 2014; Leaver and Rauschecker, 2010; Norman-Haignere et al., 2015). These studies found that anterior STG/PP was more activated for music sounds compared to human speech, vocalizations and animal sounds with evidence for right-hemispheric lateralization (Angulo-Perkins et al., 2014; Leaver and Rauschecker, 2010) similar to studies proposing right-lateralized processing of sounds with high spectral resolution (Schönwiesner and Zatorre, 2009; Zatorre et al., 2002a). In another study, Norman-Haignere et al. (2015) defined underlying components of MRI activation and found that a specific component that was related to instrumental and vocal music was located in PP (another component was related to voice and speech sounds and located in STG). Taken together, these results suggest a preference for voice and music sounds (including the category's acoustic properties) in STG/S and PP, respectively. Our results are in line with these previous results and critically extend them by showing a similar regional preference when selectively listening to mixtures of natural sounds (see also Bishop and Miller, 2009; Kerlin et al., 2010). In addition, our results suggest a right hemispheric bias in PP and PT when attending to instruments embedded in noisy situations. We speculate that the site of enhancement during selective attention in auditory cortex is determined by the acoustic properties that define the category.

In right PT, we found that activation patterns were more similar to the voice template during the voice task and more similar to the instrument template during the instrument task, which is opposed to left PT, for which we did not observe a modulation of activation patterns by task. This finding is partly in line with the proposal that PT does not only process spatial information but that it rather integrates spatial and spectro-temporal information to segregate auditory streams when multiple sounds are presented simultaneously (Griffiths and Warren, 2002; Smith et al., 2010; Zatorre et al., 2002b). Furthermore, our results strongly suggest that activation patterns in (right) PT are modulated by task. We speculate that this is achieved by upstream areas (e.g., IFG or IPS) signaling the current behavioral demand which in turn might modulate different subparts of PT (Galaburda and Sanides, 1980; Hickok and Saberi, 2012).

Modulation of representation in frontal cortex by task

In frontal cortex, especially in SFG, MFG and IFG we found that category enhancement was modulated by the attended category. In contrast to findings in temporal cortex, these regions jointly showed a positive category index when participants attended to voices and a negative index when they attended to instruments.

We expected category enhancement in IFG due to its anatomical connections with auditory belt and parabelt regions (Hackett, 2011; Kaas and Hackett, 2000; Rauschecker and Tian, 2000; Romanski and Averbeck, 2009) and its roles in the representation of (visual) categories (e.g., Cromer et al., 2010; Freedman et al., 2001) and task-related auditory processing (Atiani et al., 2014; Cohen et al., 2009; Hill and Miller, 2010). Indeed, we found that frontal activation patterns are modulated by sound category-specific attention: activation patterns evoked by scenes while attending to voices were found to be more similar to patterns evoked by isolated voice sounds and, vice-versa, activation patterns while attending to instruments were more similar to patterns evoked by isolated instrument sounds (note that the CII difference between tasks was significant, but not the CII of single tasks). This suggests that IFG encodes sound categories flexibly according to the current task demands, and is in line with electrophysiological studies in monkeys which showed that neuronal populations in ventral frontal cortex represent multiple visual categories (Cromer et al., 2010; Freedman et al., 2001). However, compared to these studies our design contained a *within-category* task on two simultaneously presented objects, whereas the aforementioned studies focused on category differences in a *between-category* task on separately presented objects (Cohen et al., 2009; Cromer et al., 2010; Freedman et al., 2001). Thus, the latter tasks required the monkeys to optimize neural processing related to the detection of category differences, which might have facilitated the observed category effects. In contrast, our task required human participants to extract the category of interest from the scene and then perform within-category identification. Interestingly, Tsunada and Cohen (2014) found that category selectivity of speech sounds is more pronounced in lateral belt compared to ventral frontal cortex and suggested that ventral frontal cortex might represent auditory categories to a lesser degree than areas in auditory cortex. This matches our observation that frontal regions showed less category enhancement compared to temporal regions. The enhancement was not restricted to IFG but included SFG and MFG as well. The observed task modulation within fixed regions indicates that task-related cognitive processes interact with abstract sound representations in frontal cortex. In addition, our results of flexible and task-dependent representation of sounds in prefrontal cortex within complex, noisy scenes support

previous evidence showing noise-robust categorical representations of syllables (Du et al., 2014) as well as identity processing of speakers (Latinus et al., 2011) and sound sources belonging to different categories (Giordano et al., 2014).

Overall, these findings suggest that in frontal cortex the same regions flexibly represent task-relevant sound categories and potential task-specific cognitive processing. In other words, while listening selectively to auditory scenes, frontal regions represent the task-relevant sound and its processing. We speculate that this may occur because of attentional filtering in auditory cortex, which enhances task-relevant and suppresses task-irrelevant sounds. A non-mutually exclusive interpretation would be that prefrontal regions send category- and task-specific signals to auditory areas to enhance and suppress the relevant and irrelevant sounds, respectively. Please note that the suggested flexible representation of task-relevant sounds refers to the activation of the pre-defined ROIs. It is possible that within these ROIs, separate interspersed voxels (and neuronal populations) encode voices and instruments and underlie our pattern-analysis results.

Part of the findings in frontal cortex might be explained by the experimental procedure, which contained similar tasks in the scene and stream conditions. However, we did not find differences in difficulty between voice and instrument tasks (but note the observed tendency for higher performance during the voice vs. instrument task). In addition, the similarity of activation maps and similar activation levels in ROIs between voice and instrument tasks indicates similar cognitive and sensory processes involved in the two tasks. Further studies are required to estimate to what extent acoustic features, task or intermediate representations are reflected in activation patterns in frontal cortex.

Mechanisms of auditory selective attention

Our results suggest that selective attention strengthens representations of attended sound sources in temporal cortex at sites that are specific to the current sound content *and* behavioral demands. In the visual domain, attention effects were first observed in higher level sensory regions that are thought to bias the processing of downstream areas to promote features of the attended objects (Buffalo et al., 2010; see also Ahissar et al., 2009). Following this idea, attentional enhancement for natural sounds with high spectral complexity and temporal dynamics should incorporate regions reflecting complex features represented higher in the auditory processing hierarchy. Additional enhancement should be observed in primary areas when sound categories also differ among them in terms of low-level acoustic features. In the present study, we found that selective listening affected processing in higher auditory areas like pSTG/S and PT. However, we did not observe attention-specific effects in primary auditory areas as found in animal studies (Atiani et al., 2014, 2009; David et al., 2009; Fritz et al., 2003; Niwa et al., 2012). We explain this null finding with the highly dynamic acoustic properties of the natural sounds which make it difficult to differentiate the sounds by single low-level acoustic features. Which acoustic features our participants exploited to perform the sound identification task remains unclear. As the pitch differed both across and within categories (see Supplementary Fig. 7A), it is possible that listener selectively attended to pitch to segregate and identify the sources. However, sounds also differed in other acoustic dimensions (e.g. spectro-temporal modulations; Supplementary Fig. 7B), which likely provided additional cues for performing the task.

Similar to earlier studies on auditory streaming (Cusack, 2005; Hill et al., 2011), we found high activation in anterior and posterior IPS, which together with FEF, is closely linked to the dorsal fronto-parietal attention network, but no evidence for differences in activation levels or category-specific activation pattern between the voice and instrument task. These results suggest a task-general role of the IPS in coordinating attention and/or structuring perceptual organization of complex sensory input (such as mixtures of natural sounds) rather than representing acoustic input or attended sounds (Cusack, 2005; Hill et al., 2011; Shomstein and Yantis, 2006).

Our finding that attended sound sources are strengthened could hint towards an enhancement of the representations of the attended sound source, a suppression of the non-attended sound sources or the two processes happening in parallel. Recent studies using high temporal precision data found that reconstructed sounds resembled more the attended sound source compared to the distractor (Ding and Simon, 2012a; Horton et al., 2013; Mesgarani and Chang, 2012; Zion Golumbic et al., 2013a, 2013b). However, the observed enhancement could be explained as well by enhancement of the attended source, suppression of the unattended source or both. In primary auditory cortex, the facilitation of neural responses to task-relevant frequencies and inhibition of neurons tuned away from this frequency (Atiani et al., 2009; David et al., 2009; Fritz et al., 2003) suggests that the processes take place in parallel for low-level acoustic features (Lakatos et al., 2013). To what extent excitatory vs. inhibitory processes contribute when attending to complex natural sound objects remains to be investigated (Bizley and Cohen, 2013).

To better estimate the magnitude of the enhancement we computed the CII on isolated sounds to estimate the intra-session reproducibility of activation patterns (reasoning that the enhancement should not exceed the pattern similarity of same conditions). We found that the enhancement and the CII on isolated sounds were of similar magnitude compared to the CII on auditory scenes for many regions in higher auditory and frontal cortex (see Supplementary Fig. 5).

In addition, we found that attention to a pure tone did not show category enhancement for voice or instrument sounds, especially in temporal cortex (see Supplementary Fig. 5). These two findings exemplify the strong influence of behavioral demand on cortical representations and suggest that paying attention to one sound is as if almost only the single-stream sound was presented. We speculate that this was due to unresolved stream formation for the voice or instrument sound: Paying attention to one of the natural categories might have been necessary to form the stream which then led to the observed category enhancement of the respective sound (Cusack et al., 2004; Shamma et al., 2011). However, another option is that a stream selection process enhanced the neural representation of the (existing) attended stream (Desimone and Duncan, 1995; Kastner and Ungerleider, 2000; Shinn-Cunningham, 2008).

To conclude, this study investigated the effects of auditory selective attention in human listeners presented with naturalistic auditory scenes. Varying the focus of attention to different sound categories in the same auditory scenes led to selective enhancement of the cortical representation encoding the attended category. In auditory cortex, the regions that contained enhanced representations depended on the attended category whereas in frontal cortex the same regions showed enhanced representations independent of the attended category. These findings shed new light on the extent and magnitude of task-related top-down modulations in auditory cortex and suggest crucial roles of frontal areas for auditory selective attention.

Data deposition

The fMRI data, experimental protocol and stimuli are available at Zenodo under Creative Commons Attribution 4.0 License (Hausfeld et al., 2017; https://doi.org/10.5281/zenodo.832994).

Funding

This work was supported by Maastricht University and the Netherlands Organisation for Scientific Research (NWO; VENI grant 451-17-033 to L.H., VICI grant 435-12-002 to E.F.).

Acknowledgements

We thank Peter de Weerd, Niels Disbergen and Martha Shiell for

discussions on the experimental setup, Federico De Martino for support with data acquisition, Milene Bonte, Inge Timmers and Tahnée Engelen for sound creation, Giancarlo Valente and João Correia for support on statistical testing.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.neuroimage.2018.02.065.

References

Ahissar, M., Nahum, M., Nelken, I., Hochstein, S., 2009. Reverse hierarchies and sensory learning. Philos. Trans. R. Soc. B Biol. Sci. 364, 285–299. https://doi.org/10.1098/rstb.2008.0253.

Alho, K., Rinne, T., Herron, T.J., Woods, D.L., 2014. Stimulus-dependent activations and attention-related modulations in the auditory cortex: a meta-analysis of fMRI studies. Hear. Res. 307, 29–41. https://doi.org/10.1016/J.HEARES.2013.08.001.

Andersson, J.L.R., Skare, S., Ashburner, J., 2003. How to correct susceptibility distortions in spin-echo echo-planar images: application to diffusion tensor imaging. Neuroimage 20, 870–888. https://doi.org/10.1016/S1053-8119(03)00336-7.

Angulo-Perkins, A., Aubé, W., Peretz, I., Barrios, F.A., Armony, J.L., Concha, L., 2014. Music listening engages specific cortical regions within the temporal lobes: differences between musicians and non-musicians. Cortex 59, 126–137. https://doi.org/10.1016/j.cortex.2014.07.013.

Arnott, S.R., Binns, M.A., Grady, C.L., Alain, C., 2004. Assessing the auditory dual-pathway model in humans. Neuroimage 22, 401–408. https://doi.org/10.1016/j.neuroimage.2004.01.014.

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. https://doi.org/10.1016/j.neuron.2014.02.029.

Atiani, S., Elhilali, M., David, S.V., Fritz, J.B., Shamma, S.A., 2009. Task difficulty and performance induce diverse adaptive patterns in gain and shape of primary auditory cortical receptive fields. Neuron 61, 467–480. https://doi.org/10.1016/j.neuron.2008.12.027.

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

Benjamini, Y., Heller, R., 2008. Screening for partial conjunction hypotheses. Biometrics 64, 1215–1222. https://doi.org/10.1111/j.1541-0420.2007.00984.x.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B 57, 289–300. https://doi.org/10.2307/2346101.

Bishop, C.W., Miller, L.M., 2009. A multisensory cortical network for understanding speech in noise. J. Cogn. Neurosci. 21, 1790–1804. https://doi.org/10.1162/jocn.2009.21118.

Bizley, J.K., Cohen, Y.E., 2013. The what, where and how of auditory-object perception. Nat. Rev. Neurosci. 14, 693–707. https://doi.org/10.1038/nrn3565.

Boersma, P., Weenink, D., 2015. Praat: Doing Phonetics by Computer.

Bregman, A.S., 1990. Auditory Scene Analysis: the Perceptual Organization of Sound. MIT Press.

Buffalo, E.A., Fries, P., Landman, R., Liang, H., Desimone, R., 2010. A backward progression of attentional effects in the ventral stream. Proc. Natl. Acad. Sci. 107, 361–365. https://doi.org/10.1073/pnas.0907658106.

Cohen, Y.E., Russ, B.E., Davis, S.J., Baker, A.E., Ackelson, A.L., Nitecki, R., 2009. A functional role for the ventrolateral prefrontal cortex in non-spatial auditory cognition. Proc. Natl. Acad. Sci. 106, 20045–20050. https://doi.org/10.1073/pnas.0907248106.

Corbetta, M., Patel, G., Shulman, G.L., 2008. The reorienting system of the human brain: from environment to theory of mind. Neuron 58, 306–324. https://doi.org/10.1016/j.neuron.2008.04.017.

Corbetta, M., Shulman, G.L., 2002. Control of goal-directed and stimulus-driven attention in the brain. Nat. Rev. Neurosci. 3, 215–229. https://doi.org/10.1038/nrn755.

Cromer, J.A., Roy, J.E., Miller, E.K., 2010. Representation of multiple, independent categories in the primate prefrontal cortex. Neuron 66, 796–807. https://doi.org/10.1016/j.neuron.2010.05.005.

Cusack, R., 2005. The intraparietal sulcus and perceptual organization. J. Cogn. Neurosci. 17, 641–651. https://doi.org/10.1162/0898929053467541.

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. J. Exp. Psychol. Hum. Percept. Perform. 30, 643–656. https://doi.org/10.1037/0096-1523.30.4.643.

Da Costa, S., van der Zwaag, W., Miller, L.M., Clarke, S., Saenz, M., 2013. Tuning in to sound: frequency-selective attentional filter in human primary auditory cortex. J. Neurosci. 33, 1858–1863. https://doi.org/10.1523/JNEUROSCI.4405-12.2013.

David, S.V., Mesgarani, N., Fritz, J.B., Shamma, S.A., 2009. Rapid synaptic depression explains nonlinear modulation of spectro-temporal tuning in primary auditory cortex by natural stimuli. J. Neurosci. 29, 3374–3386. https://doi.org/10.1523/JNEUROSCI.5249-08.2009.

Desimone, R., Duncan, J., 1995. Neural mechanisms of selective visual attention. Annu. Rev. Neurosci. 18, 193–222. https://doi.org/10.1146/annurev.ne.18.030195.001205.

Ding, N., Simon, J.Z., 2012a. Emergence of neural encoding of auditory objects while listening to competing speakers. Proc. Natl. Acad. Sci. 109, 11854–11859. https://doi.org/10.1073/pnas.1205381109.

Ding, N., Simon, J.Z., 2012b. Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. J. Neurophysiol. 107, 78–89. https://doi.org/10.1152/jn.00297.2011.

Du, Y., Buchsbaum, B.R., Grady, C.L., Alain, C., 2014. Noise differentially impacts phoneme representations in the auditory and speech motor systems. Proc. Natl. Acad. Sci. U. S. A. 111, 7126–7131. https://doi.org/10.1073/pnas.1318738111.

Forman, S.D., Cohen, J.D., Fitzgerald, M., Eddy, W.F., Mintun, M.A., Noll, D.C., 1995. Improved assessment of significant activation in functional magnetic resonance imaging (fMRI): use of a cluster-size threshold. Magn. Reson. Med. 33, 636–647. https://doi.org/10.1002/mrm.1910330508.

Freedman, D.J., Riesenhuber, M., Poggio, T., Miller, E.K., 2001. Categorical representation of visual stimuli in the primate prefrontal cortex. Science 291 (80), 312–316. https://doi.org/10.1126/science.291.5502.312.

Friston, K.J., Holmes, A.P., Worsley, K.J., Poline, J.-P., Frith, C.D., Frackowiak, R.S.J., 1994. Statistical parametric maps in functional imaging: a general linear approach. Hum. Brain Mapp. 2, 189–210. https://doi.org/10.1002/hbm.460020402.

Fritz, J., Shamma, S., Elhilali, M., Klein, D., 2003. Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex. Nat. Neurosci. 6, 1216–1223. https://doi.org/10.1038/nn1141.

Fritz, J.B., Elhilali, M., David, S.V., Shamma, S.A., 2007. Auditory attention—focusing the searchlight on sound. Curr. Opin. Neurobiol. 17, 437–455. https://doi.org/10.1016/j.conb.2007.07.011.

Galaburda, a, Sanides, F., 1980. Cytoarchitectonic organization of the human auditory cortex. J. Comp. Neurol. 190, 597–610. https://doi.org/10.1002/cne.901900312.

Genovese, C.R., Lazar, N.A., Nichols, T., 2002. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. Neuroimage 15, 870–878. https://doi.org/10.1006/nimg.2001.1037.

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. https://doi.org/10.1016/j.cortex.2014.06.005.

Goebel, R., Esposito, F., Formisano, E., 2006. Analysis of functional image analysis contest (FIAC) data with brainvoyager QX: from single-subject to cortically aligned group general linear model analysis and self-organizing group independent component analysis. Hum. Brain Mapp. 27, 392–401. https://doi.org/10.1002/hbm.20249.

Griffiths, T.D., Warren, J.D., 2004. What is an auditory object? Nat. Rev. Neurosci. 5, 887–892. https://doi.org/10.1038/nrn1538.

Griffiths, T.D., Warren, J.D., 2002. The planum temporale as a computational hub. Trends Neurosci. 25, 348–353. https://doi.org/10.1016/S0166-2236(02)02191-4.

Hackett, T.A., 2011. Information flow in the auditory cortical network. Hear. Res. 271, 133–146. https://doi.org/10.1016/j.heares.2010.01.011.

Hausfeld, L., Riecke, L., Formisano, E., 2017. Data from: Acoustic and Higher-level Representations of Naturalistic Auditory Scenes in Human Auditory and Frontal Cortex. https://doi.org/10.5281/zenodo.832994.

Hickok, G., Saberi, K., 2012. Redefining the Functional Organization of the Planum Temporale Region: Space, Objects, and Sensory–motor Integration. Springer, New York, NY, pp. 333–350. https://doi.org/10.1007/978-1-4614-2314-0_12.

Hill, K.T., Bishop, C.W., Yadav, D., Miller, L.M., 2011. Pattern of BOLD signal in auditory cortex relates acoustic response to perceptual streaming. BMC Neurosci. 12, 85. https://doi.org/10.1186/1471-2202-12-85.

Hill, K.T., Miller, L.M., 2010. Auditory attentional control and selection during cocktail party listening. Cereb. Cortex 20, 583–590. https://doi.org/10.1093/cercor/bhp124.

Horton, C., D'Zmura, M., Srinivasan, R., 2013. Suppression of competing speech through entrainment of cortical oscillations. J. Neurophysiol. 109, 3082–3093. https://doi.org/10.1152/jn.01026.2012.

Kaas, J.H., Hackett, T.A., 2000. Subdivisions of auditory cortex and processing streams in primates. Proc. Natl. Acad. Sci. U. S. A. 97, 11793–11799. https://doi.org/10.1073/pnas.97.22.11793.

Kastner, S., Ungerleider, L.G., 2000. Mechanisms of visual attention in the human cortex. Annu. Rev. Neurosci. 23, 315–341. https://doi.org/10.1146/annurev.neuro.23.1.315.

Kerlin, J.R., Shahin, A.J., Miller, L.M., 2010. Attentional gain control of ongoing cortical speech representations in a “cocktail party”. J. Neurosci. 30, 620–628. https://doi.org/10.1523/JNEUROSCI.3631-09.2010.

Kim, J.-J., Crespo-Facorro, B., Andreasen, N.C., O'Leary, D.S., Zhang, B., Harris, G., Magnotta, V.A., 2000. An MRI-based parcellation method for the temporal lobe. Neuroimage 11, 271–288. https://doi.org/10.1006/nimg.2000.0543.

Lakatos, P., Musacchia, G., O'Connel, M.N., Falchier, A.Y., Javitt, D.C., Schroeder, C.E., 2013. The spectrotemporal filter mechanism of auditory selective attention. Neuron 77, 750–761. https://doi.org/10.1016/j.neuron.2012.11.034.

Latinus, M., Crabbe, F., Belin, P., 2011. Learning-induced changes in the cerebral processing of voice identity. Cereb. Cortex 2, 1–9. https://doi.org/10.1093/cercor/bhr077.

Leaver, A.M., Rauschecker, J.P., 2010. Cortical representation of natural complex sounds: effects of acoustic features and auditory object category. J. Neurosci. 30, 7604–7612. https://doi.org/10.1523/JNEUROSCI.0296-10.2010.

Marques, J.P., Kober, T., Krueger, G., van der Zwaag, W., Van de Moortele, P.-F., Gruetter, R., 2010. MP2RAGE, a self bias-field corrected sequence for improved segmentation and T1-mapping at high field. Neuroimage 49, 1271–1281. https://doi.org/10.1016/j.neuroimage.2009.10.002.

Mesgarani, N., Chang, E.F., 2012. Selective cortical representation of attended speaker in multi-talker speech perception. Nature 485, 233–236. https://doi.org/10.1038/nature11020.

Meyers, E.M., Freedman, D.J., Kreiman, G., Miller, E.K., Poggio, T., 2008. Dynamic population coding of category information in inferior temporal and prefrontal cortex. J. Neurophysiol. 100, 1407–1419. https://doi.org/10.1152/jn.90248.2008.

Moerel, M., De Martino, F., Santoro, R., Ugurbil, K., Goebel, R., Yacoub, E., Formisano, E., 2013. Processing of natural sounds: characterization of multipeak spectral tuning in human auditory cortex. J. Neurosci. 33, 11888–11898. https://doi.org/10.1523/JNEUROSCI.5306-12.2013.

Niwa, M., Johnson, J.S., O'Connor, K.N., Sutter, M.L., 2012. Active engagement improves primary auditory cortical neurons' ability to discriminate temporal modulation. J. Neurosci. 32, 9323–9334. https://doi.org/10.1523/JNEUROSCI.5832-11.2012.

Norman-Haignere, S., Kanwisher, N.G., McDermott, J.H., 2015. Distinct cortical pathways for music and speech revealed by hypothesis-free voxel decomposition. Neuron 88, 1281–1296. https://doi.org/10.1016/j.neuron.2015.11.035.

O'Sullivan, J.A., Power, A.J., Mesgarani, N., Rajaram, S., Foxe, J.J., Shinn-Cunningham, B.G., Slaney, M., Shamma, S.A., Lalor, E.C., 2015. Attentional selection in a cocktail party environment can Be decoded from single-trial EEG. Cereb. Cortex 25, 1697–1706. https://doi.org/10.1093/cercor/bht355.

Overath, T., McDermott, J.H., Zarate, J.M., Poeppel, D., 2015. The cortical analysis of speech-specific temporal structure revealed by responses to sound quilts. Nat. Neurosci. 18, 903–911. https://doi.org/10.1038/nn.4021.

Paltoglou, A.E., Sumner, C.J., Hall, D.A., 2009. Examining the role of frequency specificity in the enhancement and suppression of human cortical activity by auditory selective attention. Hear. Res. 257, 106–118. https://doi.org/10.1016/j.heares.2009.08.007.

Peelen, M.V., Kastner, S., 2011. A neural basis for real-world visual search in human occipitotemporal cortex. Proc. Natl. Acad. Sci. https://doi.org/10.1073/pnas.1101042108.

Peelen, M.V., Fei-Fei, L., Kastner, S., 2009. Neural mechanisms of rapid natural scene categorization in human visual cortex. Nature 460, 94–97. https://doi.org/10.1038/nature08103.

Rauschecker, J.P., Tian, B., 2000. Mechanisms and streams for processing of “what” and “where” in auditory cortex. Proc. Natl. Acad. Sci. 97, 11800–11806. https://doi.org/10.1073/pnas.97.22.11800.

Riecke, L., Peters, J.C., Valente, G., Kemper, V.G., Formisano, E., Sorger, B., 2016. Frequency-selective attention in auditory scenes recruits frequency representations throughout human superior temporal cortex. Cereb. Cortex. https://doi.org/10.1093/cercor/bhw160 bhw160.

Romanski, L.M., Averbeck, B.B., 2009. The primate cortical auditory system and neural representation of conspecific vocalizations. Annu. Rev. Neurosci. 32, 315–346. https://doi.org/10.1146/annurev.neuro.051508.135431.

Schönwiesner, M., Zatorre, R.J., 2009. Spectro-temporal modulation transfer function of single voxels in the human auditory cortex measured with high-resolution fMRI. Proc. Natl. Acad. Sci. 106, 14611–14616. https://doi.org/10.1073/pnas.0907682106.

Shamma, S.A., Elhilali, M., Micheyl, C., 2011. Temporal coherence and attention in auditory scene analysis. Trends Neurosci. 34, 114–123. https://doi.org/10.1016/j.tins.2010.11.002.

Shinn-Cunningham, B.G., 2008. Object-based auditory and visual attention. Trends Cogn. Sci. 12, 182–186. https://doi.org/10.1016/j.tics.2008.02.003.

Shomstein, S., Yantis, S., 2006. Parietal cortex mediates voluntary control of spatial and nonspatial auditory attention. J. Neurosci. 26, 435–439. https://doi.org/10.1523/JNEUROSCI.4408-05.2006.

Smith, K.R., Hsieh, I.-H., Saberi, K., Hickok, G., 2010. Auditory spatial and object processing in the human planum temporale: No evidence for selectivity. J. Cogn. Neurosci. 22, 632–639. https://doi.org/10.1162/jocn.2009.21196.

Smith, S.M., Jenkinson, M., Woolrich, M.W., Beckmann, C.F., Behrens, T.E.J., Johansen-Berg, H., Bannister, P.R., De Luca, M., Drobnjak, I., Flitney, D.E., Niazy, R.K., Saunders, J., Vickers, J., Zhang, Y., De Stefano, N., Brady, J.M., Matthews, P.M., 2004. Advances in functional and structural MR image analysis and implementation as FSL. Neuroimage 23, S208–S219. https://doi.org/10.1016/j.neuroimage.2004.07.051.

Staeren, N., Renvall, H., De Martino, F., Goebel, R., Formisano, E., 2009. Sound categories are represented as distributed patterns in the human auditory cortex. Curr. Biol. 19, 498–502. https://doi.org/10.1016/j.cub.2009.01.066.

Talairach, J., Tournoux, P., 1988. Co-planar Stereotaxic Atlas of the Human Brain: 3-Dimensional Proportional System - an Approach to Cerebral Imaging. Thieme Medical Publishers, New York, NY. https://doi.org/10.1604/9783137117018.

Teki, S., Chait, M., Kumar, S., von Kriegstein, K., Griffiths, T.D., 2011. Brain bases for auditory stimulus-driven figure-ground segregation. J. Neurosci. 31, 164–171. https://doi.org/10.1523/JNEUROSCI.3788-10.2011.

Tsunada, J., Cohen, Y.E., 2014. Neural mechanisms of auditory categorization: from across brain areas to within local microcircuits. Front. Neurosci. 8, 161. https://doi.org/10.3389/fnins.2014.00161.

Zatorre, R.J., Belin, P., Penhune, V.B., 2002a. Structure and function of auditory cortex: music and speech. Trends Cogn. Sci. 6, 37–46. https://doi.org/10.1016/S1364-6613(00)01816-7.

Zatorre, R.J., Bouffard, M., Ahad, P., Belin, P., 2002b. Where is “where” in the human auditory cortex? Nat. Neurosci. 5, 905–909. https://doi.org/10.1038/nn904.

Zion Golumbic, E., Cogan, G.B., Schroeder, C.E., Poeppel, D., 2013a. Visual input enhances selective speech envelope tracking in auditory cortex at a “cocktail party”. J. Neurosci. 33, 1417–1426. https://doi.org/10.1523/JNEUROSCI.3675-12.2013.

Zion Golumbic, E.M., Ding, N., Bickel, S., Lakatos, P., Schevon, C.A., McKhann, G.M., Goodman, R.R., Emerson, R., Mehta, A.D., Simon, J.Z., Poeppel, D., Schroeder, C.E., 2013b. Mechanisms underlying selective neuronal tracking of attended speech at a “cocktail party”. Neuron 77, 980–991. https://doi.org/10.1016/j.neuron.2012.12.037.