frontiers in Neuroscience

ORIGINAL RESEARCH
published: 07 June 2016
doi: 10.3389/fnins.2016.00254

# Attention Modulates the Auditory Cortical Processing of Spatial and Category Cues in Naturalistic Auditory Scenes

*Hanna Renvall*[1, 2, 3†], *Noël Staeren*[1†], *Claudia S. Barz*[1, 4, 5], *Anke Ley*[1] *and Elia Formisano*[1, 6*]

*[1] Department of Cognitive Neuroscience, Faculty of Psychology and Neuroscience, Maastricht University, Maastricht, Netherlands, [2] Department of Neuroscience and Biomedical Engineering, Aalto University School of Science, Espoo, Finland, [3] Aalto Neuroimaging, Magnetoencephalography (MEG) Core, Aalto University, Espoo, Finland, [4] Institute for Neuroscience and Medicine, Research Centre Juelich, Juelich, Germany, [5] Department of Psychiatry, Psychotherapy and Psychosomatics, Medical School, RWTH Aachen University, Aachen, Germany, [6] Maastricht Center for Systems Biology (MaCSBio), Maastricht University, Maastricht, Netherlands*

This combined fMRI and MEG study investigated brain activations during listening and attending to natural auditory scenes. We first recorded, using in-ear microphones, vocal non-speech sounds, and environmental sounds that were mixed to construct auditory scenes containing two concurrent sound streams. During the brain measurements, subjects attended to one of the streams while spatial acoustic information of the scene was either preserved (stereophonic sounds) or removed (monophonic sounds). Compared to monophonic sounds, stereophonic sounds evoked larger blood-oxygenation-level-dependent (BOLD) fMRI responses in the bilateral posterior superior temporal areas, independent of which stimulus attribute the subject was attending to. This finding is consistent with the functional role of these regions in the (automatic) processing of auditory spatial cues. Additionally, significant differences in the cortical activation patterns depending on the target of attention were observed. Bilateral planum temporale and inferior frontal gyrus were preferentially activated when attending to stereophonic environmental sounds, whereas when subjects attended to stereophonic voice sounds, the BOLD responses were larger at the bilateral middle superior temporal gyrus and sulcus, previously reported to show voice sensitivity. In contrast, the time-resolved MEG responses were stronger for mono- than stereophonic sounds in the bilateral auditory cortices at ∼360 ms after the stimulus onset when attending to the voice excerpts within the combined sounds. The observed effects suggest that during the segregation of auditory objects from the auditory background, spatial sound cues together with other relevant temporal and spectral cues are processed in an attention-dependent manner at the cortical locations generally involved in sound recognition. More synchronous neuronal activation during monophonic than stereophonic sound processing, as well as (local) neuronal inhibitory mechanisms in the auditory cortex, may explain the simultaneous increase of BOLD responses and decrease of MEG responses. These findings highlight the complimentary role of electrophysiological and hemodynamic measures in addressing brain processing of complex stimuli.

**Keywords: auditory scene analysis, auditory cortex (AC), fMRI BOLD, magnetoencephalography (MEG), auditory attention**

OPEN ACCESS

**Edited by:**
*Elvira Brattico,*
*Aarhus University, Finland*

**Reviewed by:**
*Kimmo Alho,*
*University of Helsinki, Finland*
*Jaakko Kauramäki,*
*Université de Montréal, Canada*

***Correspondence:**
*Elia Formisano*
*e.formisano@maastrichtuniversity.nl*

*[†] These authors have contributed equally to this work.*

**Specialty section:**
*This article was submitted to Auditory Cognitive Neuroscience, a section of the journal Frontiers in Neuroscience*

**Received:** *29 March 2016*
**Accepted:** *23 May 2016*
**Published:** *07 June 2016*

**Citation:**
*Renvall H, Staeren N, Barz CS, Ley A and Formisano E (2016) Attention Modulates the Auditory Cortical Processing of Spatial and Category Cues in Naturalistic Auditory Scenes. Front. Neurosci. 10:254. doi: 10.3389/fnins.2016.00254*

## INTRODUCTION

Overlapping voices, a phone ringing at the background: The auditory signal at our ears usually comprises sounds from several sources. Segregation of a complex sound mixture is a magnificent example of the automatic computational capabilities of our auditory system, likely determined by the interplay between bottom-up processing of the spectral and temporal acoustic elements of the mixture and attentive selection and enhancement of the relevant sounds (Bregman, 1990).

A relevant part of auditory scene analysis relates to processing of spatial information of the sound sources. Vertical and horizontal localization of sounds relies on the direction-dependent modifications of the spectral sound profile, generated by the outer ear and head, and on the timing and sound intensity differences between the ears, respectively; perception of sound motion depends on the dynamic changes of these cues. On the basis of extensive line of studies in primates (e.g., Romanski et al., 1999; Recanzone et al., 2000; Tian et al., 2001; Lomber and Malhotra, 2008; Miller and Recanzone, 2009), a dorsal auditory stream specialized for processing of spatial information has been suggested. Neuroimaging studies in humans generally support this hypothesis. Results on sound localization (Alain et al., 2001; Warren and Griffiths, 2003; Barrett and Hall, 2006; Altmann et al., 2007) and sound motion (Baumgart et al., 1999; Lewis et al., 2000; Pavani et al., 2002; Warren et al., 2002, 2005; Hart et al., 2004; Krumbholz et al., 2005a,b; Getzmann and Lewald, 2010) suggest the involvement of posterotemporal and temporoparietal regions, especially when subjects are actively engaged in sound localization tasks (Zatorre et al., 2002). Furthermore, studies using magnetoencephalography (MEG) and functional magnetic resonance imaging (fMRI) in humans (Salminen et al., 2009; Derey et al., 2016) are consistent with studies in animals (Stecker et al., 2005; Miller and Recanzone, 2009) suggesting the existence of population rate coding in (posterior) auditory areas involved in spatial processing, with populations of neurons broadly tuned to locations in the left and right auditory spatial hemifields.

Postero-temporal auditory regions, however, are not exclusively involved in the spatial analysis of sounds. For example, activation of the planum temporale (PT) has been suggested to reflect integration of spatial and auditory object information, rather than spatial processing *per se* (Zatorre et al., 2002). In addition, manipulating the number of auditory objects within an auditory scene modifies the activation at PT similarly to spatial manipulations (Smith et al., 2010). Furthermore, task-modulated processing of sound location and identity has been demonstrated in the human non-primary auditory cortex (Ahveninen et al., 2006, 2013), suggestive of fine-grained top-down effects on extracting auditory information in real-life.

Knowledge of the auditory scene analysis in the human cortex has mainly been derived from studies applying stimuli with highly-controlled physical properties, necessary to reveal the different stages of processing. Here we take an approach toward natural auditory processing, by examining cortical processing of realistic auditory "mini-scenes" with interspersed spatial cues and different sound attributes. Using ear-insert microphones, vocal, and environmental sounds were recorded and subsequently digitally superimposed. During MEG and fMRI measurements, subjects listened to binaural mini-scenes that either did or did not preserve the original spatial aspects of the sounds (stereophonic vs. monophonic sounds). We then manipulated the top-down processing of the auditory scenes, by directing the subjects' attention either to the voice or environmental excerpts within the sounds while keeping the stimuli unchanged. This design allowed us to examine the relation between the cortical mechanisms of analyzing spatial cues and selecting sound objects from a real-life like scene.

As MEG and fMRI differ in their sensitivity to the underlying neuronal activity in complex auditory (Renvall et al., 2012a) and cognitive tasks (Vartiainen et al., 2011), both imaging modalities were applied in the present study for optimal coverage and for studying the possible discrepancies between the electrophysiological and hemodynamic measures. For example, an earlier MEG study suggested monophonic sounds to elicit stronger auditory cortical activation than pseudostereophonic sounds (Ross et al., 2004), possibly related to more synchronous neuronal activation during monophonic than pseudo-stereophonic sound presentation tracked with MEG. Furthermore, (local) neuronal inhibition in the auditory cortex has been suggested to affect coding of spatial auditory cues (e.g., Fitzpatrick et al., 1997). Thus an increase of local inhibition could lead to simultaneous increase of blood-oxygenation-level-dependent (BOLD) fMRI responses and decrease of MEG responses.

## MATERIALS AND METHODS

### Subjects

We studied, with written informed consent, 10 subjects with MEG (mean ± SEM age 30 ± 1 yrs; four females; nine right-handed and one ambidextrous) and 10 subjects with fMRI (28 ± 4 yrs; four females; nine right-handed and one ambidextrous). Six of the subjects participated in both MEG and fMRI studies. None of the subjects had a history of hearing or neurological impairments, and they were all naïve to the experimental setup. The study received a prior approval by the Ethical Committee of the Faculty of Psychology, University of Maastricht, The Netherlands.

### Experimental Stimuli

High-quality sounds were recorded binaurally using two in-ear microphones (FG-23652-P16, Knowles Electronics, Itasca, Illinois, U.S.A.) and a portable digital recorder (96 kHz, 24-bit, M-Audio MicroTrack 24/96 Pocket Digital Recorder). After recording, sounds were down-sampled to 44.1 kHz/16-bit and low-pass filtered at 10 kHz using Adobe Audition (Adobe Systems, Inc., CA, USA). The environmental sounds comprised, e.g., sounds of tool use, office sounds, and household sounds. The vocalizations were non-speech sounds produced by 12 individuals, and they comprised, e.g., laughing, coughing, sighing, and baby crying. All sounds were fairly stationary, whereas the original vocal sounds had less variability in their degree of channel separation than the environmental sounds (channel intensity difference divided by the summed intensity across channels 0.20 ± 0.02 for voices, 0.38 ± 0.02 for environmental sounds). All original sounds were reported by the participants to be very natural-like and have a clear stereophonic effect. Fifty-nine different stereophonic combinations of environmental and vocal sounds were created by mixing—keeping the two channels separate—recording excerpts that included sounds from 59 different environments and 59 vocal sounds. Before mixing, the waveforms of all original stimuli were first carefully visually inspected and edited/cut in order to have a clear signal onset. Then, after superimposing the vocal and environmental sounds, 25-ms rise times were imposed to the combined sounds to further equalize the onsets amplitude-wise. Finally the average root-mean-square levels of the sounds were matched using MATLAB 7.0.1 (The MathWorks, Inc., Natick, MA, USA). The duration of the sounds varied between 450 and 2635 ms (mean length ± SD 1306 ± 565 ms) including 25-ms rise and fall times. Examples of the original stimuli and their combinations used in the experiment (monophonic and stereophonic) are presented in **Figure 1** (see also Supplementary Files 1–6).

All stimuli in the study were recorded by inserting the microphones to the ear canal of three listeners that did not take part in the fMRI nor MEG measurements. It is known that—because of inter-individual differences in head and external ear shape—non-individualized recordings as used in this study do not produce as good perceptual quality as individualized recordings or use of individualized head-related transfer functions (HRTF). However, well-localized perception has been shown to be relatively independent of the details of HRTFs (Kulkarni and Colburn, 1998), and thus we chose not to record the stimuli individually because of the difficulty of recreating similar, natural complex scenes for each subject. Indeed, each subject reported a clear and natural spatial perception of the stimuli. Given the nature of our recordings, the stimuli contained—in the *Stereophonic* condition—many spatial cues including interaural time and level differences, as well as spectral cues. While this prevents us from making any specific claim on the processing of specific cues, it does reflect auditory scene analysis in naturalistic contexts, where all cues are in fact combined.

The created stereophonic sounds were utilized in the "*Stereophonic*" experimental condition; monophonic version of the same scenes ("*Monophonic*" condition) was created by merging the two audio channels. As demonstrated in **Figure 1**, the overall physical characteristics remained very similar after the transformation. During the experiments, the sounds were delivered to the subjects binaurally at a comfortable listening level through plastic tubes and ear pieces (MEG: ADU audio

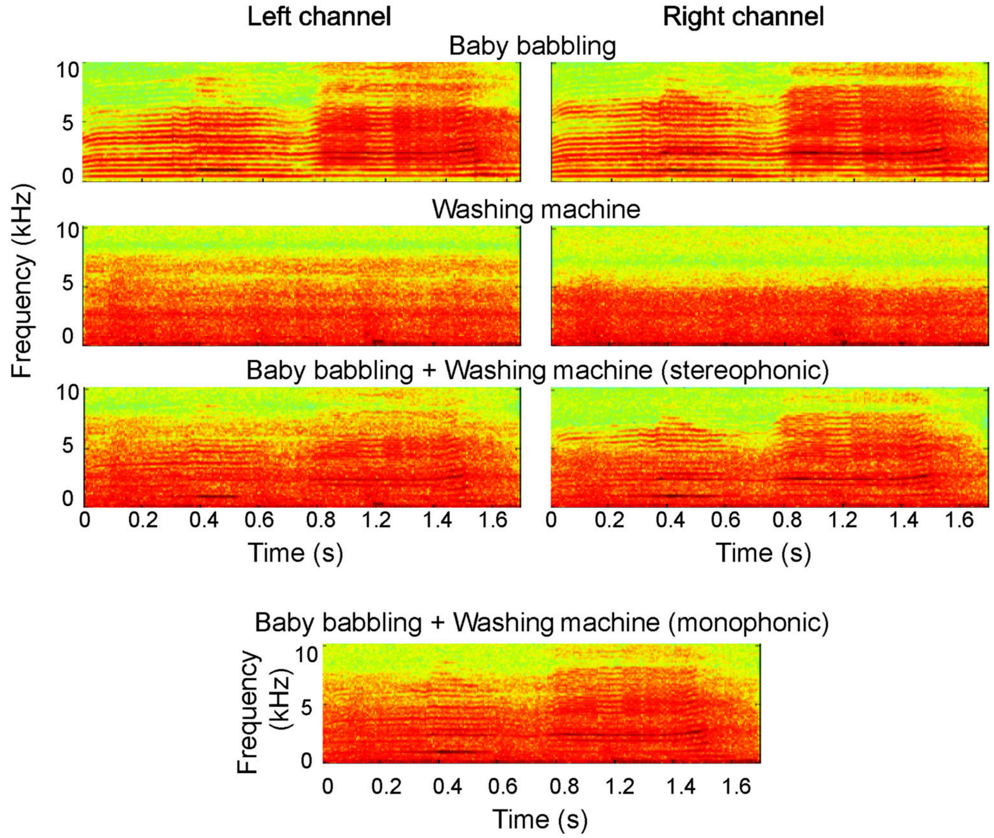

**FIGURE 1 | Examples of the experimental stimuli.** The original sounds ("baby babbling" and "Washing machine") are depicted for both left and right auditory channel (top); the combined sounds used in the experiment, in both stereophonic and monophonic forms, are shown below.

stimulator, KAR Audio, Helsinki, Finland) or via headphones (fMRI; Commander XG, Resonance Technology, Northridge, CA).

As the behavioral responses collected during the brain measurements (see below) were too few for making statistical inferences between conditions, the stimuli were behaviorally tested with two tasks in 16 subjects outside the fMRI/MEG experiment. In the first task, the sounds were presented alone in a random order. The subjects were asked to respond with a button press, as fast and accurately as possible, whether the presented sound was a voice or environmental sound. In the second task, the stimuli were the superimposed sounds used in the actual neuroimaging experiments; the stereophonic and monophonic sounds were used in different test runs. The subjects were instructed to attend to either the voice or environmental sound excerpt of the combined sound and respond with a button press in case the attended sound part was repeated. The order of runs (mono- or stereophonic, attention to voice or environmental sound) was counterbalanced across subjects.

## fMRI Experiment and Signal Analysis

A 2 × 2 block design with auditory space ("*Stereophonic*" vs. "*Monophonic*") and task (attention to the voice excerpt in the combined sound, "*Voice*," vs. attention to the environmental sound excerpt, "*Environment*") as factors was used. The experiment consisted of two functional runs during which auditory scenes from the four different conditions were presented in a block design. Each of the runs (22 min each) included nine blocks per condition and four target blocks (see below); the sequence of conditions was randomized. Each block consisted of four TRs (TR = 4640 ms, total block duration 18.5 s), and one auditory mini-scene was presented for each TR. The blocks were separated by a fixation period of three TRs. Every block was preceded by a cue presented at the fixation point, indicating the attention condition ("*Voice*" or "*Environment*"). Subjects were instructed to respond with a button press if the attended sound part was the same in two consecutive auditory scenes. This occurred in 10% of the cases ("target blocks"; altogether two target blocks per condition and four target blocks per run). The response hand was alternated across subjects. Imaging was performed with a three Tesla Siemens Allegra (head setup) at the Maastricht Brain Imaging Center. In each subject, two runs of 282 volumes were acquired with a T2*-weighted gradient-echo planar imaging (EPI) sequence (TR = 4640 ms, voxel size = 2.5 × 2.5 × 2.5 mm$^3$, TE = 30 ms, FOV 256 × 256; matrix size 96 × 96, 32 slices covering the cortex). Anatomical images (1 × 1 × 1 mm$^3$) were collected between the functional runs using a 3D-MPRAGE T1-weighted sequence. To reduce the effect of scanner noise, the sounds were presented during silent periods using a clustered volume EPI technique that allowed for presentation of auditory stimuli in silence between subsequent volume acquisitions (van Atteveldt et al., 2004; Riecke et al., 2007; Staeren et al., 2009).

Functional and anatomical images were analyzed with BrainVoyager QX (Brain Innovation, Maastricht, The Netherlands). Pre-processing consisted of slice scan-time correction (using sinc interpolation), linear trend removal, temporal high-pass filtering to remove nonlinear drifts of seven or less cycles per time course, and three-dimensional motion correction. Temporal low-pass filtering was performed using a Gaussian kernel with FWHM of two data points. Functional slices were co-registered to the anatomical data, and both data were normalized to Talairach space (Talairach and Tournoux, 1988).

Statistical analysis of the fMRI data was based on the general linear modeling (GLM) of the time series. For each subject, a design matrix was formed using a predictor for each experimental condition ("*Stereophonic-Voice*," "*Stereophonic-Environment*," "*Monophonic-Voice*," "*Monophonic-Environment*") and for the target blocks. The predicted time courses were adjusted for the hemodynamic response delay by convolution with a canonical hemodynamic response function (sum of two gamma functions).

Cortex-based realignment was performed for aligning the functional time series of individual subjects and to perform random effect group-based statistics (Goebel et al., 2006). Statistical maps were thresholded and corrected for multiple comparisons (alpha = 0.05) on the basis of cluster-level statistical threshold estimation performed on the cortical surface data (Forman et al., 1995; Goebel et al., 2006).

## MEG Experiment and Signal Analysis

The sounds were presented with an interstimulus interval (from offset to onset) of 1500 ms in four separate ∼6-min runs ("*Stereophonic-Voice*," "*Stereophonic-Environment*," "*Monophonic-Voice*," "*Monophonic-Environment*"). Before each run, the subject was indicated verbally which stimulus attribute to attend to. The order of runs was counterbalanced across subjects.

Subjects were instructed to respond with a button press in case the attended sound attribute was the same in two consecutive sounds (10% of the cases). The target sounds were excluded from the analysis. The response hand was alternated across subjects.

The auditory evoked fields were recorded in a magnetically shielded room using a whole-head MEG systems with 275 axial gradiometers (VSM/CTF Systems Inc., Port Coquitlam, Canada; six subjects), and a 306-channel Vectorview$^{\text{TM}}$ device (Elekta Neuromag, Helsinki, Finland; four subjects). Head-position-indicator coils were attached to the scalp, and their positions were measured with a three-dimensional digitizer; the head coordinate frame was anchored to the two ear canals/periauricular points and the nasion. The head position was determined by feeding current to the marker coils and measuring their positions with respect to the sensory array.

The MEG signals were low-pass filtered at 300 Hz and digitized at 1200 Hz with the VSM/CTF Systems device, and band-pass filtered at 0.03–200 Hz and digitized at 600 Hz with the Vectorview system. The signals were averaged from 200 ms before the stimulus onset to 1000 ms after it, setting as baseline the 200-ms interval immediately preceding the stimulus onset. The averaged signals were digitally low-pass filtered at 40 Hz. The horizontal and vertical electro-oculograms were recorded to discard data contaminated by eye blinks and movements.

For source analysis, the head was modeled as a homogeneous spherical volume conductor. The model parameters were optimized for the intracranial space obtained from MR images that were available for all subjects. The neurophysiological responses were analyzed by first segregating the recorded sensor-level signals into spatiotemporal components, by means of manually-guided multi-dipole current modeling (equivalent current dipole, ECD; Hämäläinen et al., 1993). The analysis was conducted separately for each subject using Elekta Neuromag (Elekta Oy) software package, following standard procedures (Salmelin et al., 1994; Hansen et al., 2010). The parameters of an ECD represent the location, orientation, and strength of the current in the activated brain area. The ECDs were identified by searching for systematic local changes that persist for tens of milliseconds in the measured magnetic field pattern. ECD model parameters were then determined at those time points at which the magnetic field pattern was clearly dipolar. Only ECDs explaining more than 85% of the local field variance during each dipolar response peak were accepted in the multidipole model. Based on this criterion, 2–4 spatiotemporal components were selected into the individual subjects' models. The analysis was then extended to the entire time period, and all MEG channels were taken into account: The previously found ECDs were kept fixed in orientation and location while their strengths were allowed to change.

For optimizing the accuracy of the spatial fits, the orientation and location of the ECDs were estimated in each individual in the condition with the strongest signals in the time windows of the main experimental effects suggested by the sensor level data. To avoid spurious interactions between close-by sources of the 100-ms and sustained responses with similar current orientations, the dipoles modeled during the sustained responses were used to explain also the 100-ms responses. The variability in the signal-to-noise ratios between conditions was very small: Visual inspection and the calculated goodness-of-fit values obtained by comparing the original data and the data predicted by the fitted sources showed that the same sources explained well the responses in the other conditions as well.

The ECD source waveforms were analyzed for the late sustained (>300 ms) responses that were hypothesized to show the main experimental effects. Two measures on the strength of the response were obtained in each stimulus condition: (i) average over a 50-ms time window during the rising slope of the sustained response (360–410 ms; later in the text referred to as time window A), and (ii) average over a 100-ms time window centered at each individual's peak response, determined in the condition showing overall strongest signal (i.e., monophonic voice) and then applied in all experimental conditions (time window B). The response strengths were statistically tested using paired *t*-tests (two-sided, Bonferroni-corrected for multiple comparisons).

# RESULTS

## Behavioral Experiment

When presented separately, both environmental sounds and voices were recognized with high accuracy (99 ± 1%), whereas the reaction times were significantly longer for the environmental sounds than voices (931 ± 260 vs. 819 ± 250 ms, $P < 0.01$). When the sounds were superimposed, the sounds continued to be well recognizable, although subjects made more errors for the stereophonic than monophonic sounds (1.9 vs. 0.6%, $P < 0.01$) irrespective of whether attention was focused on the voice or environmental sound part. Reaction times did not significantly differ between the attended monophonic sounds (attention to environmental sounds 896 ± 260 ms vs. attention to voices 846 ± 290 ms, $P = 0.06$). In the stereophonic condition, reaction times were prolonged for environmental sounds but not for voices compared with the monophonic sounds (attention to environmental sounds 961 ± 270 ms, $P < 0.01$; attention to voices 840 ± 270 ms, $P = 0.8$; Attentional × Spatial condition interaction $P = 0.05$).

## fMRI Results

Listening to the auditory mini-scenes induced extensive activations at the superior temporal cortex bilaterally, including the Heschl's gyrus and surrounding regions at the superior temporal gyrus and sulcus (see **Figure 2A**). Additional activation was found in the left middle temporal gyrus (MTG), left inferior frontal gyrus (IFG), and bilateral inferior parietal lobule (IPL). The overall activation pattern was largely common to all experimental conditions.

### "Stereophonic" vs. "Monophonic" Scenes

To examine the brain regions especially involved in the processing of spatial cues, we first compared the activation to *"Stereophonic"* vs. *"Monophonic"* scenes grouped across the two attention conditions. We observed statistically significantly higher BOLD responses in the *"Stereophonic"* than *"Monophonic"* condition (see **Figure 2B**) bilaterally in the posterior-lateral STG regions. In the left hemisphere (LH), this region was located at the adjacency of the temporal-parietal border; in the right hemisphere (RH), an additional cluster was detected along the STS.

We further dissected the "*Stereophonic*" vs. *"Monophonic"* contrast by analyzing the two attention conditions separately, i.e., the two orthogonal contrasts' "*Stereophonic-Environment*" vs. *"Monophonic-Environment"* and "*Stereophonic-Voice*" vs. "*Monophonic-Voice*." In the right posterior STG, these two contrasts were independently significant (marked with yellow in **Figure 2C**). Similarly, in the LH, clusters with significantly different responses in one of the two contrasts were interspersed, but in none of these locations both contrasts were independently significant. Besides these common clusters, activations specific to the different sound attributes the subjects were attending to were detected. When listeners attended to the environmental excerpts in the sounds, significant activation differences between "*Stereophonic*" and *"Monophonic"* conditions were found in the left PT and in the left IFG. Conversely, when listeners attended to the vocal excerpts in the sounds, significant activation differences between conditions were found in the left middle STG and in the right posterior and middle STS. These latter clusters resemble regions reported to be selectively activated for voices in previous studies (e.g., Belin et al., 2000; Bonte et al., 2014).

To test these observations statistically, interaction maps were calculated (**Figure 2D**). Of the regions for which any individual contrast was significant (blue or red regions in **Figure 2C**), only

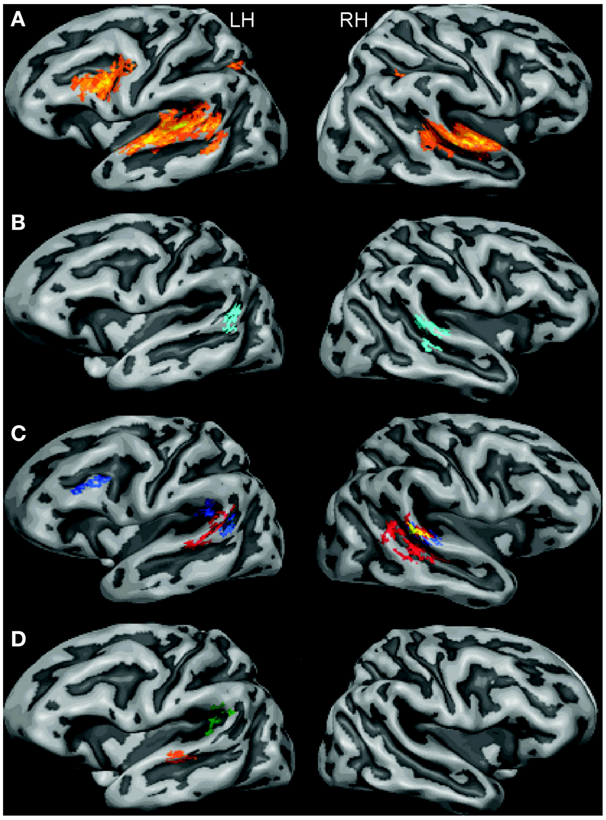

**FIGURE 2 | Results from cortex-based aligned random effect analysis in stereophonic and monophonic experimental conditions (LH, left hemisphere; RH, right hemisphere). (A)** All conditions vs. baseline (F-map; the color ranges from yellow to reddish with increasing *f*-value). **(B)** "*Stereophonic*" vs. "*Monophonic*" stimuli (merged attention conditions; light blue: *Stereophonic* > *Monophonic*). **(C)** "*Stereophonic*" vs. "*Monophonic*" stimuli in the *"Voice"* condition (red: *Stereophonic-Voice* > *Monophonic-Voice*) and in the *"Environment"* condition (blue: *Stereophonic-Environment* > *Monophonic-Environment*); the overlapping area common for both these contrasts is highlighted in yellow. **(D)** The statistically significant interaction maps of the activations depicted in **(C)**. Green: "*Stereophonic-Environment*"—"*Monophonic-Environment*" > "*Stereophonic-Voice*"—"*Monophonic-Voice*"; orange: "*Stereophonic-Voice*"—*Monophonic-Voice*" > "*Stereophonic-Environment*"—"*Monophonic-Environment*."

the regions in the left PT and in the left middle STG survived a rigorous statistical threshold ($p < 0.05$, corrected), whereas homologous activity in the right STS did not.

### "Voice" vs. "Environment" Scenes

To examine the brain regions specifically affected by the applied attentional manipulation, we compared the activations to the scenes grouped across the stereophonic and monophonic conditions. We observed significantly higher BOLD responses for the "*Environment*" condition in a largely left-lateralized network including posterior STG, posterior STS/MTG, and the dorsolateral prefrontal cortex (DLPFC). Bilateral activation of the posterior parietal cortex (PPC) and the left precentral gyrus (PrG) were also observed. No region showed increased activation for "*Voice*" condition compared with the "*Environment*" condition (See Supplementary Results). When analyzing the stereo- and monophonic conditions separately, i.e., *"Stereophonic-Environment"* vs. *"Stereophonic-Voice"* and *"Monophonic-Environment"* vs. *"Monophonic-Voice*," a generally similar pattern of overall activation differences was observed.

## MEG Results

The initial sensor-level analysis revealed that all stimuli evoked strong responses bilaterally over the temporal areas, peaking at ~50, ~100, and at ~250–700 ms after the sound onset. In agreement with previous studies (for a review, see Hari, 1990), the prominent 100-ms responses were explained by two ECDs, one in the left (8 out of 10 subjects) and one in the right (10/10 subjects) supratemporal auditory cortex (individual source locations indicated by white dipoles in **Figure 3**). The same sources explained adequately also the 50-ms responses and the sustained activity peaking >300 ms.

In both hemispheres, another source with more variable location and direction of current flow over subjects was needed to explain the responses at ~250 ms (locations indicated by black dipoles in **Figure 3**; 8/10 and 10/10 subjects in the LH and RH, respectively), in line with earlier auditory MEG studies applying naturalistic sounds (Renvall et al., 2012a,b).

The activation strengths and latencies were fairly similar in all stimulus conditions (see **Figure 3**). We first compared the activation to "*Stereophonic*" vs. "*Monophonic*" scenes grouped across the two attention conditions. In the RH, sustained responses peaking at ~465 ± 20 ms (time window B, mean ± SEM over subjects) were stronger in the *"Monophonic"* than "*Stereophonic*" condition ($P < 0.04$). The responses did not statistically significantly differ between conditions during the response rise time 360–410 ms (time window A) in the RH ($P = 0.09$), nor in either of the tested analysis time windows in the LH (time window A, $P = 0.07$; time window B, LH peak ~535 ± 30 ms, $P = 0.18$).

When the two attention conditions were analyzed separately (**Figure 4**), the "*Monophonic*" scenes produced stronger activation than *"Stereophonic"* scenes only when the subjects were attending to voices ("*Stereophonic-Voice*" vs. "*Monophonic-Voice*": RH time window B, $P < 0.03$; LH time window A, $P < 0.02$). The 250-ms responses did not differ between the stimulus conditions.

Finally, we examined the effects of the applied attentional manipulation by comparing the activations to scenes grouped across the stereophonic and monophonic conditions ("*Voice*" vs. "*Environment*" scenes). For this comparison, no significant effect of condition was detected in either hemisphere.

## DISCUSSION

The present study investigated cortical processing of complex, naturalistic auditory mini-scenes with (stereophonic), or without (monophonic) spatial acoustic information. In particular, modulating the focus of subjects' attention between superimposed human voices and other natural

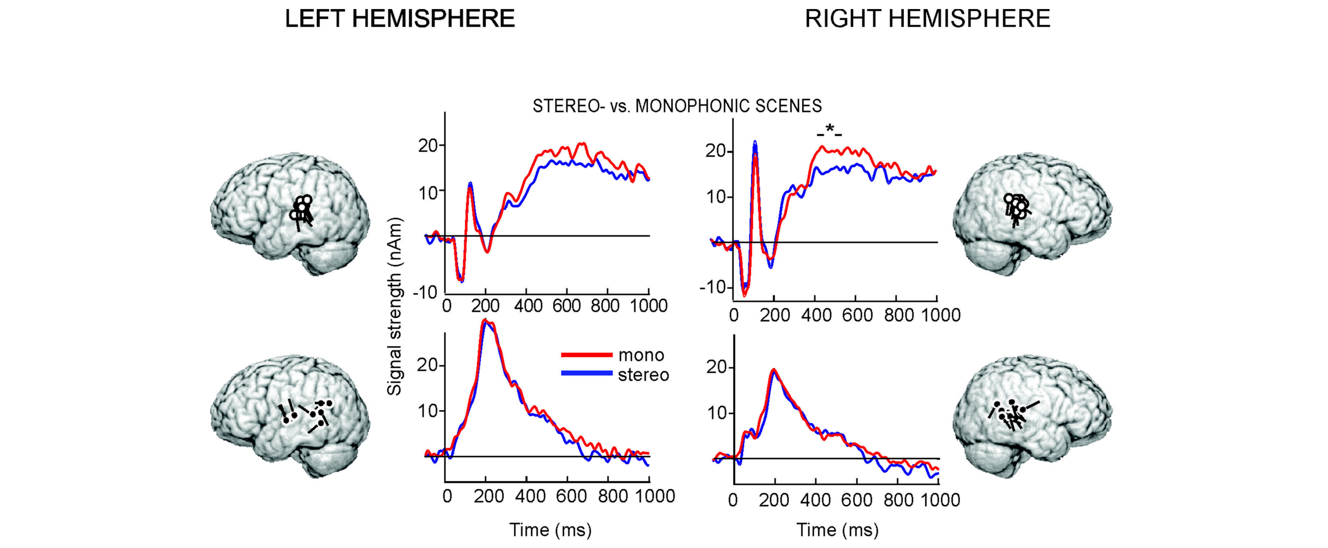

**FIGURE 3 | MEG group-level data.** The locations and orientations of individual ECDs used to model the 100-ms, sustained (white dots), and 250-ms (black dots) responses superimposed on an average brain, and the corresponding averaged ECD time courses from −200 to 1000 ms with respect to the stimulus onset, grouped across the two attention conditions. The asterisk indicates the 100-ms time window with statistically significant difference ($P < 0.04$) between the conditions.

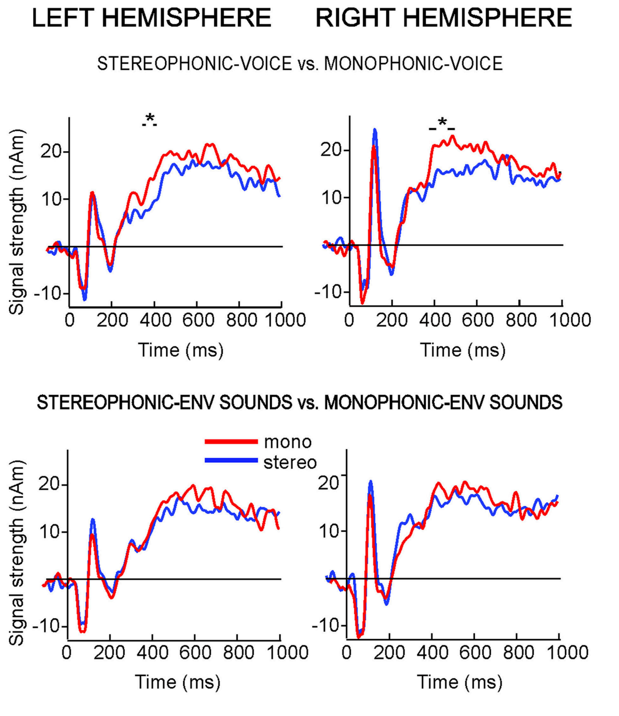

**FIGURE 4 | Time courses of the ECDs used to model the 100-ms and sustained responses to stereo- vs. monophonic sounds in both attentional conditions.** (**Top**: attention to voice; **bottom**: attention to environmental sounds). The asterisks indicate the 50-ms (LH) and 100-ms time windows (RH) with statistically significant differences between the conditions.

sounds enabled us to study auditory cortical mechanisms related to selecting sound objects from real-life-like auditory scenes.

Listening to stereophonic scenes resulted in a robust increase of BOLD response in the right and the left posterior auditory cortex. This increased activation was present irrespective of whether the subjects were attending to voices or environmental sounds. The location of task-independent activation corresponded to the posterior portion of planum temporale, a site compatible with cytoarchitectonic area Tpt (Galaburda and Sanides, 1980; Sweet et al., 2005), area STA (Rivier and Clarke, 1997; Wallace et al., 2002), or area Te3 (Morosan et al., 2005). These locations are also in agreement with previous functional neuroimaging studies that have investigated sound localization and motion using sounds presented in isolation (Hart et al., 2004; Krumbholz et al., 2005a; Callan et al., 2013; Derey et al., 2016). Our results with real-life-like auditory stimuli support the role of these areas in processing auditory spatial information, and, in line with a "where" auditory processing stream, point to their role in the analysis of spatial cues also within complex auditory scenes. Processing of spatial acoustic cues in these areas seems to be rather automatic and only marginally influenced by subjects' focus of attention, which may be particularly relevant for efficient localization of relevant sounds. It is worth noting that our experimental task did not explicitly require listeners to localize the sounds, which further highlights the obligatory nature of the observed effects.

Besides sound localization, spatial acoustic cues within complex auditory scenes contribute to sound stream segregation and formation (Bregman, 1990). Thus, any effects related to the focus of auditory attention during the listening task may reflect cortical processing mechanisms devoted to integration of spatial cues with other spectral and temporal cues, with the ultimate goal of segregating and grouping relevant sound objects in a scene. The present results are consistent with this view. The fMRI BOLD responses in the left STG and the right STS, as well as the late sustained MEG responses from ∼360 ms onwards bilaterally, were affected by the spatial sound manipulation when voice attributes within the stimuli were attended to. The observed effect is unlikely to be related to greater demands in attending to stereophonic vs. monophonic voice sounds since the reaction times to stereophonic and monophonic sounds were similar for voices but differed for environmental sounds. Moreover, the anatomical locations of the BOLD responses resemble the so-called "voice-sensitive" regions reported in previous studies (Belin et al., 2000).

Interestingly, the BOLD responses at these areas were larger for the stereophonic sounds, whereas the MEG responses showed the opposite effect with stronger responses to monophonic sounds. The MEG results are in concordance with an earlier observation of decreased auditory steady-state responses for pseudo-stereophonic vs. monophonic sounds (Ross et al., 2004). MEG evoked responses are highly sensitive to the synchronicity of neuronal activation, and the current results are likely to reflect a reduced convergence of temporally coincident neural firing during the stereophonic stimulation with different acoustic inputs to the two ears compared with the monophonic stimulus presentation. In addition, active neuronal inhibition in the auditory cortex (Fitzpatrick et al., 1997) could result in a simultaneous increment of BOLD responses and decrement of MEG responses.

Furthermore, we observed an effect of spatial acoustic cues on the BOLD response at the left IFG and PT when attention was directed to environmental sounds. PT has previously been associated with processing of tool sounds (Lewis et al., 2005), which in fact constitute a large subset of our experimental sounds. In contrast, our MEG experiment did not reveal a time window nor source areas with similar effect. This discrepancy between fMRI and MEG responses can be at least partly related to MEG's suboptimal sensitivity to spatially extended frontal activations (Tarkiainen et al., 2003).

Although consistent with previous studies, our interpretations above are not univocal. In our scenes, voices were located somewhat more centrally for mimicking typical communicational situations, while the environmental sounds were more variable in their original left-right channel difference. However, as the MEG responses differed between mono- and stereophonic conditions only when attending to voices, the effect is unlikely to be related to purely acoustical differences between the attended sounds, as such differences between the conditions were actually slightly larger for the environmental than voice sounds. Further studies could verify whether the experimental effects that were dependent on the focus of attention, indeed, reflect mainly the different sensitivity of the outlined regions to the distinct sound categories, or relate to the variation of different acoustic cues within the auditory scene as well.

When the subjects focused their attention on the "Environment" excerpts of the sounds irrespective of the spatial aspects of the sound, a robust increase of fMRI BOLD in the left temporal, left frontal, and bilateral parietal areas was detected compared with attending to "Voices," whereas neither fMRI nor MEG responses highlighted areas or time windows with greater activity for "Voices." This result can be interpreted in the light of ecological validity and/or acoustic properties of the corresponding stimuli. Attending to the environmental sounds within our scenes may have required additional top-down signaling from frontal and parietal areas for overriding or counteracting the automatic allocation of attention to vocalized sounds. This interpretation is supported by the longer reaction times to environmental sounds in the behavioral experiment.

Here we used measures of neural activation that are most frequently used in neurophysiological and hemodynamic non-invasive brain mapping, namely, MEG evoked responses and fMRI BOLD signals. Other measures could be more sensitive to such fine-grained changes in brain activations as examined in the present study. Indeed oscillatory activity, especially in the high-gamma range (>75 Hz) measured intracortically from the auditory cortices (Mesgarani and Chang, 2012; Zion Golumbic et al., 2013) has been demonstrated to track the envelope of attended speech within multi-talker settings.

Several modeling studies have recently successfully mapped different tempo-spectral characteristics of acoustic stimuli to their corresponding neural representations (Pasley et al., 2012; Mesgarani et al., 2014; Santoro et al., 2014), suggesting strong correspondences between the auditory input and cortical reactivity. However, in real life the input to the ears is often a complicated mixture from several sound sources, thus requiring auditory cortical areas to use more sophisticated computational scene-analysis strategies for sound localization than the pure physical cue extraction (Młynarski and Jost, 2014). Accordingly, the current results show that spatial auditory cues—already within rather simple, but naturalistic auditory stimuli—are processed together with other relevant temporal and spectral cues, and that the related cortical processing is attention- and stimulus-dependent.

## AUTHOR CONTRIBUTIONS

HR and NS contributed equally to this work. HR, NS, and EF designed the experiment. HR and NS recorded the stimuli and acquired the fMRI and MEG data. CB and AL ran and analyzed the behavioral experiment. HR, NS, and EF analyzed the imaging data. HR, EF, and NS wrote the paper.

## FUNDING

The authors declare no conflicting interests. Funding for the present research was contributed to EF from the Netherlands Organization for Scientific Research (NWO, Vernieuwingsimpuls VIDI 452-04-330, VICI 453-12-002) and from the Province of Limburg and to HR from the Academy of Finland (grants #127401 and #277655).

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fnins.2016.00254

# REFERENCES

Ahveninen, J., Huang, S., Nummenmaa, A., Belliveau, J. W., Hung, A. Y., Jääskeläinen, I. P., et al. (2013). Evidence for distinct human auditory cortex regions for sound location versus identity processing. *Nat. Commun.* 4, 2585. doi: 10.1038/ncomms3585

Ahveninen, J., Jääskeläinen, I. P., Raij, T., Bonmassar, G., Devore, S., Hämäläinen, M., et al. (2006). Task-modulated "what" and "where" pathways in human auditory cortex. *Proc. Natl. Acad. Sci. U.S.A.* 103, 14608–14613. doi: 10.1073/pnas.0510480103

Alain, C., Arnott, S. R., Hevenor, S., Graham, S., and Grady, C. L. (2001). "What" and "where" in the human auditory system. *Proc. Natl. Acad. Sci. U.S.A.* 98, 12301–12306. doi: 10.1073/pnas.211209098

Altmann, C. F., Bledowski, C., Wibral, M., and Kaiser, J. (2007). Processing of location and pattern changes of natural sounds in the human auditory cortex. *Neuroimage* 35, 1192–1200. doi: 10.1016/j.neuroimage.2007.01.007

Barrett, D. J., and Hall, D. A. (2006). Response preferences for "what" and "where" in human non-primary auditory cortex. *Neuroimage* 32, 968–977. doi: 10.1016/j.neuroimage.2006.03.050

Baumgart, F., Gaschler-Markefski, B., Woldorff, M. G., Heinze, H. J., and Scheich, H. (1999). A movement-sensitive area in auditory cortex. *Nature* 400, 724–726. doi: 10.1038/23390

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

Bonte, M., Hausfeld, L., Scharke, W., Valente, G., and Formisano, E. (2014). Task-dependent decoding of speaker and vowel identity from auditory cortical response patterns. *J. Neurosci.* 34, 4548–4557. doi: 10.1523/JNEUROSCI.4339-13.2014

Bregman, A. S. (1990). *Auditory Scene Analysis.* Cambridge, MA: MIT Press.

Callan, A., Callan, D. E., and Ando, H. (2013). Neural correlates of sound externalization. *Neuroimage* 66, 22–27. doi: 10.1016/j.neuroimage.2012.10.057

Derey, K., Valente, G., de Gelder, B., and Formisano, E. (2016). Opponent coding of sound location (azimuth) in planum temporale is robust to sound level variations. *Cereb. Cortex.* 26, 450–464. doi: 10.1093/cercor/bhv269

Fitzpatrick, D. C., Batra, R., Stanford, T. R., and Kuwada, S. (1997). A neuronal population code for sound localization. *Nature* 388, 871–874.

Forman, S. D., Cohen, J. D., Fitzgerald, M., Eddy, W. F., Mintun, M. A., and 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. doi: 10.1002/mrm.1910330508

Galaburda, A., and Sanides, F. (1980). Cytoarchitectonic organization of the human auditory cortex. *J. Comp. Neurol.* 190, 597–610. doi: 10.1002/cne.901900312

Getzmann, S., and Lewald, J. (2010). Shared cortical systems for processing of horizontal and vertical sound motion. *J. Neurophysiol.* 103, 1896–1904. doi: 10.1152/jn.00333.2009

Goebel, R., Esposito, F., and 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. doi: 10.1002/hbm.20249

Hämäläinen, M., Hari, R., Ilmoniemi, R. J., Knuutila, J., and Lounasmaa, O. V. (1993). Magnetoencephalography – theory, instrumentation, and applications to noninvasive studies of the working human brain. *Rev. Mod. Phys.* 65, 413–497. doi: 10.1103/RevModPhys.65.413

Hansen, P. C., Kringelbach, M. L., and Salmelin, R. (2010). *MEG - An Introduction to Methods.* New York, NY: Oxford University Press.

Hari, R. (1990). "The neuromagnetic method in the study of the human auditory cortex," in *Auditory Evoked Magnetic Fields and Electric Potentials,* eds F. Grandori, M. Hoke, and G. L. Romani (Basel: Karger), 222–282.

Hart, H. C., Palmer, A. R., and Hall, D. A. (2004). Different areas of human non-primary auditory cortex are activated by sounds with spatial and nonspatial properties. *Hum. Brain Mapp.* 21, 178–190. doi: 10.1002/hbm.10156

Krumbholz, K., Schonwiesner, M., Rubsamen, R., Zilles, K., Fink, G. R., and von Cramon, D. Y. (2005a). Hierarchical processing of sound location and motion in the human brainstem and planum temporale. *Eur. J. Neurosci.* 21, 230–238. doi: 10.1111/j.1460-9568.2004.03836.x

Krumbholz, K., Schonwiesner, M., von Cramon, D. Y., Rubsamen, R., Shah, N. J., Zilles, K., et al. (2005b). Representation of interaural temporal information from left and right auditory space in the human planum temporale and inferior parietal lobe. *Cereb. Cortex* 15, 317–324. doi: 10.1093/cercor/bhh133

Kulkarni, A., and Colburn, H. S. (1998). Role of spectral detail in sound-source localization. *Nature* 396, 747–749. doi: 10.1038/25526

Lewis, J. W., Beauchamp, M. S., and DeYoe, E. A. (2000). A comparison of visual and auditory motion processing in human cerebral cortex. *Cereb. Cortex* 10, 873–888. doi: 10.1093/cercor/10.9.873

Lewis, J. W., Brefczynski, J. A., Phinney, R. E., Janik, J. J., and DeYoe, E. A. (2005). Distinct cortical pathways for processing tool versus animal sounds. *J. Neurosci.* 25, 5148–5158. doi: 10.1523/JNEUROSCI.0419-05.2005

Lomber, S. G., and Malhotra, S. (2008). Double dissociation of 'what' and 'where' processing in auditory cortex. *Nat. Neurosci.* 11, 609–616. doi: 10.1038/nn.2108

Mesgarani, N., and Chang, E. F. (2012). Selective cortical representation of attended speakers in multi-talker speech perception. *Nature* 485, 233–236. doi: 10.1038/nature11020

Mesgarani, N., Cheung, C., Johnson, K., and Chang, E. F. (2014). Phonetic feature encoding in human superior temporal gyrus. *Science* 343, 1006–1010. doi: 10.1126/science.1245994

Miller, L. M., and Recanzone, G. H. (2009). Populations of auditory cortical neurons can accurately encode acoustic space across stimulus intensity. *Proc. Natl. Acad. Sci. U.S.A.* 106, 5931–5935. doi: 10.1073/pnas.0901023106

Młynarski, W., and Jost, J. (2014). Statistics of natural binaural sounds. *PLoS ONE* 9:e108968. doi: 10.1371/journal.pone.0108968

Morosan, P., Schleicher, A., Amunts, K., and Zilles, K. (2005). Multimodal architectonic mapping of human superior temporal gyrus. *Anat. Embryol.* 210, 401–406. doi: 10.1007/s00429-005-0029-1

Pasley, B. N., David, S. V., Mesgarani, N., Flinker, A., Shamma, S. A., Crone, N. E., et al. (2012). Reconstructing speech from human auditory cortex. *PLoS Biol.* 10:e1001251. doi: 10.1371/journal.pbio.1001251

Pavani, F., MacAluso, E., Warren, J. D., Driver, J., and Griffiths, T. D. (2002). A common cortical substrate activated by horizontal and vertical sound movement in the human brain. *Curr. Biol.* 12, 1584–1590. doi: 10.1016/S0960-9822(02)01143-0

Recanzone, G. H., Guard, D. C., Phan, M. L., and Su, T. K. (2000). Correlation between the activity of single auditory cortical neurons and sound-localization behavior in the macaque monkey. *J. Neurophysiol.* 83, 2723–2739.

Renvall, H., Formisano, E., Parviainen, T., Bonte, M., Vihla, M., and Salmelin, R. (2012a). Parametric merging of MEG and fMRI reveals spatiotemporal differences in cortical processing of spoken words and environmental sounds in background noise. *Cereb. Cortex*, 22, 132–143. doi: 10.1093/cercor/bhr095

Renvall, H., Staeren, N., Siep, N., Esposito, F., Jensen, O., and Formisano, E. (2012b). Of cats and women: temporal dynamics in the right temporoparietal cortex reflect auditory categorical processing of vocalizations. *Neuroimage* 62, 1877–1883. doi: 10.1016/j.neuroimage.2012.06.010

Riecke, L., van Opstal, A. J., Goebel, R., and Formisano, E. (2007). Hearing illusory sounds in noise: sensory-perceptual transformations in primary auditory cortex. *J. Neurosci.* 27, 12684–12689. doi: 10.1523/JNEUROSCI.2713-07.2007

Rivier, F., and Clarke, S. (1997). Cytochrome oxidase, acetylcholinesterase, and NADPH-diaphorase staining in human supratemporal and insular cortex: evidence for multiple auditory areas. *Neuroimage* 6, 288–304. doi: 10.1006/nimg.1997.0304

Romanski, L. M., Tian, B., Fritz, J., Mishkin, M., Goldman-Rakic, P. S., and Rauschecker, J. P. (1999). Dual streams of auditory afferents target multiple domains in the primate prefrontal cortex. *Nat. Neurosci.* 2, 1131–1136. doi: 10.1038/16056

Ross, B., Herdman, A. T., Wollbrink, A., and Pantev, C. (2004). Auditory cortex responses to the transition from monophonic to pseudo-stereo sound. *Neurol. Clin. Neurophysiol.* 2004, 18.

Salmelin, R., Hari, R., Lounasmaa, O. V., and Sams, M. (1994). Dynamics of brain activation during picture naming. *Nature* 368, 463—465. doi: 10.1038/368463a0

Salminen, N. H., May, P. J., Alku, P., and Tiitinen, H. (2009). A population rate code of auditory space in the human cortex. *PLoS ONE* 4:e7600. doi: 10.1371/journal.pone.0007600

Santoro, R., Moerel, M., De Martino, F., Goebel, R., Ugurbil, K., Yacoub, E., et al. (2014). Encoding of natural sounds at multiple spectral and temporal resolutions in the human auditory cortex. *PLoS Comput. Biol.* 10:e1003412. doi: 10.1371/journal.pcbi.1003412

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

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

Stecker, G. C., Harrington, I. A., and Middlebrooks, J. C. (2005). Location coding by opponent neural populations in the auditory cortex. *PLOS Biol.* 3:e78. doi: 10.1371/journal.pbio.0030078

Sweet, R. A., Dorph-Petersen, K. A., and Lewis, D. A. (2005). Mapping auditory core, lateral belt, and parabelt cortices in the human superior temporal gyrus. *J. Comp. Neurol.* 491, 270–289. doi: 10.1002/cne.20702

Talairach, J., and Tournoux, P. (1988). *Co-Planar Stereotactic Atlas of the Human Brain.* Stuttgart: Thieme.

Tarkiainen, A., Liljeström, M., Seppä, M., and Salmelin, R. (2003). The 3D topography of MEG source localization accuracy: effects of conductor model and noise. *Neuroimage* 114, 1977–1992. doi: 10.1016/S1388-2457(03)00195-0

Tian, B., Reser, D., Durham, A., Kustov, A., and Rauschecker, J. P. (2001). Functional specialization in rhesus monkey auditory cortex. *Science* 292, 290–293. doi: 10.1126/science.1058911

van Atteveldt, N., Formisano, E., Goebel, R., and Blomert, L. (2004). Integration of letters and speech sounds in the human brain. *Neuron* 43, 271–282. doi: 10.1016/j.neuron.2004.06.025

Vartiainen, J., Liljeström, M., Koskinen, M., Renvall, H., and Salmelin, R. (2011). Functional magnetic resonance imaging blood oxygenation level-dependent signal and magnetoencephalography evoked responses yield different neural functionality in reading. *J. Neurosci.* 31, 1048–1058. doi: 10.1523/JNEUROSCI.3113-10.2011

Wallace, M. N., Johnston, P. W., and Palmer, A. R. (2002). Histochemical identification of cortical areas in the auditory region of the human brain. *Exp. Brain Res.* 143, 499–508. doi: 10.1007/s00221-002-1014-z

Warren, J. D., and Griffiths, T. D. (2003). Distinct mechanisms for processing spatial sequences and pitch sequences in the human auditory brain. *J. Neurosci.* 23, 5799–5804.

Warren, J. D., Zielinski, B. A., Green, G. G., Rauschecker, J. P., and Griffiths, T. D. (2002). Perception of sound-source motion by the human brain. *Neuron* 34, 139–148. doi: 10.1016/S0896-6273(02)00637-2

Warren, J. E., Wise, R. J., and Warren, J. D. (2005). Sounds do-able: auditory-motor transformations and the posterior temporal plane. *Trends Neurosci.* 28, 636–643. doi: 10.1016/j.tins.2005.09.010

Zatorre, R. J., Bouffard, M., Ahad, P., and Belin, P. (2002). Where is 'where' in the human auditory cortex? *Nat. Neurosci.* 5, 905–909. doi: 10.1038/nn904

Zion Golumbic, E. M., Ding, N., Bickel, S., Lakatos, P., Schevon, C. A., McKhann, G. M., et al. (2013). Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party." *Neuron* 77, 980–991. doi: 10.1016/j.neuron.2012.12.037

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.