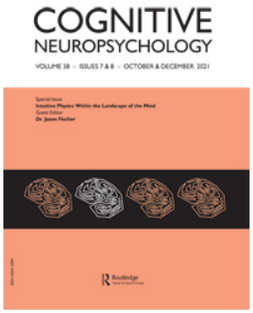

Cognitive Neuropsychology

ISSN: 0264-3294 (Print) 1464-0627 (Online) Journal homepage: www.tandfonline.com/journals/pcgn20

Cochlea to categories: The spatiotemporal dynamics of semantic auditory representations

Matthew X. Lowe, Yalda Mohsenzadeh, Benjamin Lahner, Ian Charest, Aude Oliva & Santani Teng

To cite this article: Matthew X. Lowe, Yalda Mohsenzadeh, Benjamin Lahner, Ian Charest, Aude Oliva & Santani Teng (2021) Cochlea to categories: The spatiotemporal dynamics of semantic auditory representations, Cognitive Neuropsychology, 38:7-8, 468-489, DOI: 10.1080/02643294.2022.2085085

To link to this article: https://doi.org/10.1080/02643294.2022.2085085

COGNITIVE NEUROPSYCHOLOGY
2021, VOL. 38, NOS. 7–8, 468–489
https://doi.org/10.1080/02643294.2022.2085085

Routledge
Taylor & Francis Group

Cochlea to categories: The spatiotemporal dynamics of semantic auditory representations

Matthew X. Lowe[a,b*], Yalda Mohsenzadeh[a,c,d,e*], Benjamin Lahner[a], Ian Charest[f,g], Aude Oliva[a†] and Santani Teng[a,h†]

[a]Computer Science and Artificial Intelligence Lab (CSAIL), MIT, Cambridge, MA, USA; [b]Unlimited Sciences, Colorado Springs, CO, USA; [c]The Brain and Mind Institute, The University of Western Ontario, London, Canada; [d]Department of Computer Science, The University of Western Ontario, London, Canada; [e]Vector Institute for ArtificialIntelligence, Toronto, Ontario,Canada; [f]Département de Psychologie, Université de Montréal, Montréal, Canada; [g]Center for Human Brain Health, University of Birmingham, Birmingham, UK; [h]Smith-Kettlewell Eye Research Institute (SKERI), San Francisco, CA, USA

ABSTRACT

How does the auditory system categorize natural sounds? Here we apply multimodal neuroimaging to illustrate the progression from acoustic to semantically dominated representations. Combining magnetoencephalographic (MEG) and functional magnetic resonance imaging (fMRI) scans of observers listening to naturalistic sounds, we found superior temporal responses beginning ~55 ms post-stimulus onset, spreading to extratemporal cortices by ~100 ms. Early regions were distinguished less by onset/peak latency than by functional properties and overall temporal response profiles. Early acoustically-dominated representations trended systematically toward category dominance over time (after ~200 ms) and space (beyond primary cortex). Semantic category representation was spatially specific: Vocalizations were preferentially distinguished in frontotemporal voice-selective regions and the fusiform; scenes and objects were distinguished in parahippocampal and medial place areas. Our results are consistent with real-world events coded via an extended auditory processing hierarchy, in which acoustic representations rapidly enter multiple streams specialized by category, including areas typically considered visual cortex.

ARTICLE HISTORY
Received 30 July 2021
Revised 31 March 2022
Accepted 25 May 2022

KEYWORDS
Audition; naturalistic sounds; functional magnetic resonance imaging; magnetoencephalography; representational similarity analysis

Introduction

Hearing feels immediate and automatic. Within a few hundred milliseconds, the human auditory system spectrally decomposes incident acoustic energy, then computationally segregates (Kell & McDermott, 2019; Teng et al., 2017; Traer & McDermott, 2016), localizes (Ahveninen et al., 2006, 2014), and semantically categorizes (Charest et al., 2009; De Lucia et al., 2010; Murray et al., 2006) its sources, providing a rich representation of physical objects and events in our surroundings. Despite considerable progress in identifying the neural underpinnings of these transformations (Bizley & Cohen, 2013; McDermott, 2018; Rauschecker & Scott, 2009; Yi et al., 2019), the structural and functional organization of semantically meaningful auditory representation in human listeners remains an open question.

Because semantic classes of sound events are correlated with their acoustic properties (Theunissen & Elie, 2014), the resulting brain responses are themselves correlated (Norman-Haignere & McDermott, 2018) and thus ambiguously interpretable when not controlled (e.g., Engel et al., 2009; Ogg et al., 2020). Additionally, most studies using individual neuroimaging techniques differentiate responses across anatomical regions (Giordano et al., 2013; Kell et al., 2018; Norman-Haignere & McDermott, 2018) or over time (Brodbeck et al., 2018; Daube et al., 2019; Murray et al., 2006; Ogg et al., 2020) but cannot reveal the full spatiotemporal and representational dynamics needed to characterize a processing

CONTACT Santani Teng santani@mit.edu; santani@ski.org 2318 Fillmore St., San Francisco, CA 94115, USA
*Equal contribution.
†Equal senior contribution.
Supplemental data for this article can be accessed online at https://doi.org/10.1080/02643294.2022.2085085.

stream (Bizley & Cohen, 2013; Rauschecker & Scott, 2009). Consequently, it is often equivocal how physical events map to the brain responses they elicit, and whether those responses better reflect stimulus features or higher-level semantics. Here, we address these challenges via similarity-based fusion of human magnetoencephalography (MEG) and functional magnetic resonance imaging (fMRI) (Cichy et al., 2014; Cichy, Pantazis, et al., 2016; Cichy & Oliva, 2020; Henriksson et al., 2019; Khaligh-Razavi et al., 2018; Mohsenzadeh et al., 2019). We combined the spatiotemporal resolution of responses in both imaging modalities within the representational similarity analysis (RSA) framework (Kriegeskorte et al., 2008; Kriegeskorte & Kievit, 2013) and modelled low- (acoustic) and high- (semantic) level representations within those responses, operationalizing representational level as the relative strength of the semantic category vs. acoustic models. In this way we integrated spatiotemporal measures of neural activity with spatiotemporal changes in the representational content of that activity. We found that semantic dominance of neural representations increased systematically over time; that its spatial progression over regions of interest (ROIs) correlated with those ROIs' peak response latencies; and that distinct high-level frontal, temporal, medial and occipital regions rapidly coded distinct semantic categories, dissociable from overall acoustics. Our results suggest that external objects and events, initially transduced as acoustic representations, are categorized along an extended auditory processing hierarchy comprising category-specific streams within and beyond temporal cortex, including typically visual category-selective areas. Further, our data fusion approach is broadly amenable to uncovering the dynamics of auditory coding under a wide variety of stimuli and listening conditions.

Results

Human participants ($n = 16$) listened to 80 different monaural real-world sounds, presented diotically in random order for 500 ms every three seconds, while MEG and fMRI data were acquired in separate sessions. Stimuli were drawn from a large collection comprising sounds from animate (human and animal vocalizations) and inanimate (objects and scenes) sources (Figure S1A–D). To dissociate acoustic from semantic category attributes, we equated root-mean-square intensities and spectrogram correlation distances across categories (see Methods, Figure S1E–F for details). Prior to the fMRI and MEG neuroimaging sessions, participants listened to each sound accompanied by a written description (see Table S1), so that the interpretation of each individual sound was not ambiguous. No explicit category information was provided. Participants were instructed to focus attentively on each sound for the duration of the trial and responded via button press to oddball sounds (200 ms pure tones presented in pairs, separated by 100 ms silence) presented an average of every ten trials. All participants performed at or near ceiling on this vigilance oddball task (mean hit rate 98.44% ± 1.38%, d' = 5.71 ± .57).

MEG-fMRI fusion reveals a spatiotemporally ordered cascade of neural activation

To generate a spatiotemporally unbiased view of neural responses to sounds, we applied whole-brain searchlight-based fMRI-MEG fusion (Cichy et al., 2014; Cichy, Pantazis, et al., 2016; Cichy & Oliva, 2020; Khaligh-Razavi et al., 2018; Mohsenzadeh et al., 2019; Wardle et al., 2020), relating temporal neurodynamics in MEG with spatial BOLD activity patterns in fMRI. In brief, MEG and fMRI brain responses were decoded separately by pairs of conditions, and the resulting decoding matrices then correlated within the RSA framework (Kriegeskorte et al., 2008). To this end, we first applied multivariate pattern analysis (MVPA) to MEG data in trial epochs spanning −200 to +3000 ms relative to stimulus onset. At each time point in the epoch, stimulus conditions were decoded pairwise, with classification accuracy indexing dissimilarity, to construct representational dissimilarity matrices (RDMs) (Cichy et al., 2014; Cichy, Pantazis, et al., 2016; Kriegeskorte & Kievit, 2013). This yielded 3201 MEG RDMs of size 80×80 (1 per millisecond in the epoch; 1 experimental condition per sound) per subject. For fMRI, we computed RDMs for each voxel (Kriegeskorte et al., 2008) via a whole-brain searchlight approach (Cichy, Pantazis, et al., 2016; Haynes & Rees, 2005; Kriegeskorte et al., 2006), using 1—Pearson correlation as the pairwise dissimilarity metric, yielding a total of 40,122 fMRI RDMs (1 per voxel, also of size 80×80) per subject. Finally, we correlated the group-averaged MEG

RDMs at each time point with subject-specific fMRI RDMs (Spearman's ρ; see Equation 2, Methods), then thresholded and cluster-corrected the resulting correlations (cluster-definition threshold $p < 0.001$, cluster size threshold $p < 0.05$; see Figure 1a and Methods for further details). This yielded a series of whole-brain maps representing MEG-fMRI correspondences across time, each serving as a frame in a "movie" spanning the epoch (see Movie S1, Figure 1).

Significant fusion correlations appear in voxels along Heschl's Gyrus and the superior temporal gyrus (STG) starting 55–60 ms post-stimulus onset, spreading laterally along the superior temporal plane by ~80 ms, anteriorly and posteriorly thereafter

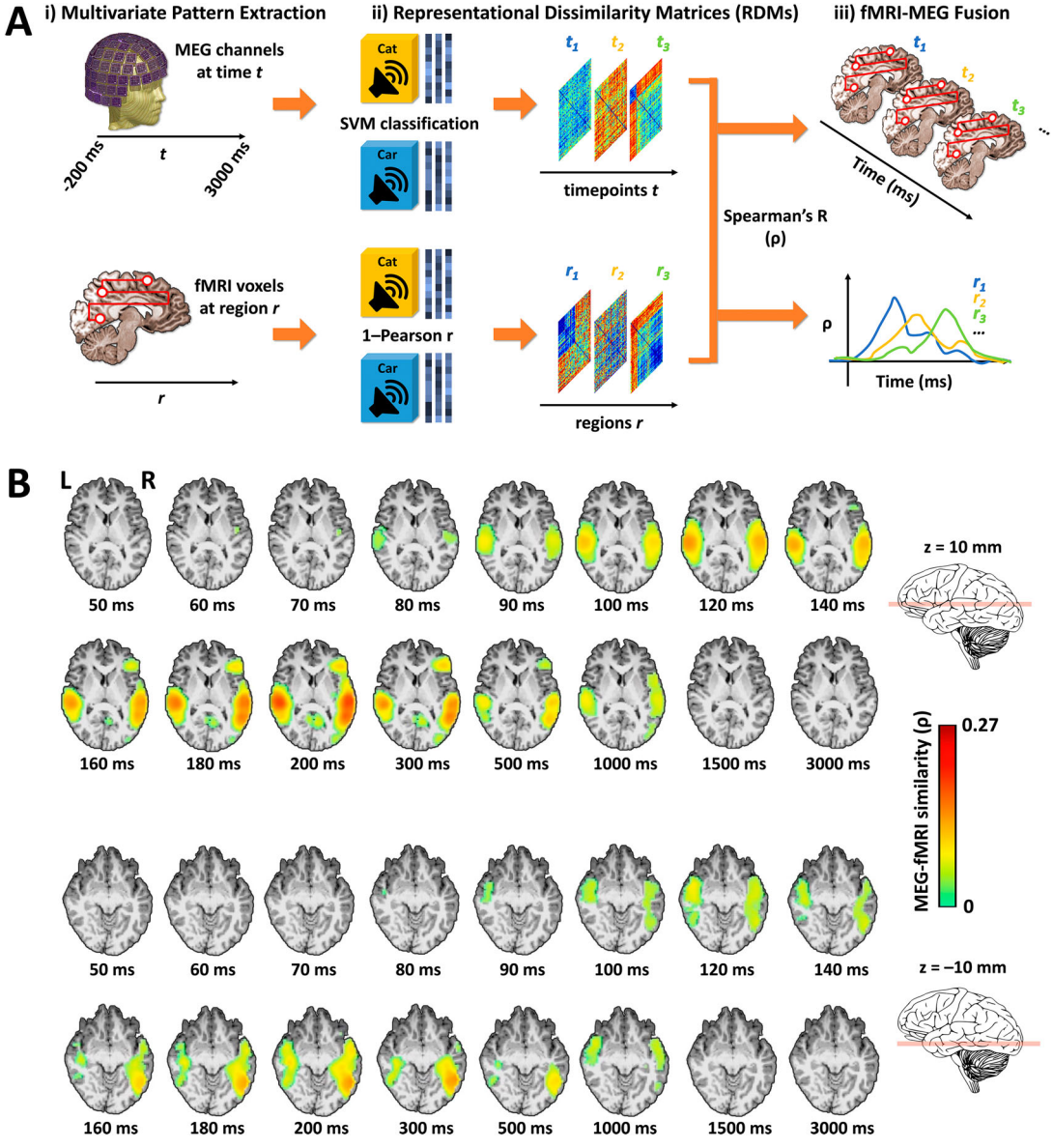

Figure 1. Spatiotemporal propagation of auditory neural responses indexed by MEG-fMRI fusion. (**A**) Overview of MEG-fMRI decoding and RSA-based fusion. (i) Brain responses acquired in separate sessions per subject were extracted for MVPA by time points t (whole-brain MEG sensor values, top) or anatomical regions r (fMRI voxelwise t-values, bottom); *r* may represent successive searchlight volumes, as illustrated here, or anatomical regions of interest, as shown in Figures 2 and 3. (ii) RDMs computed using SVM decoding accuracy (MEG, top) or Pearson correlation distance (fMRI, bottom) to index pairwise dissimilarity between stimulus conditions. (iii) RDMs. (**B**) Main cortical regions involved in the searchlight-based fMRI-MEG fusion. Illustration on two axial slices ($z = 10$ and -10 mm) at select time points. There are significant neural representations beginning over the auditory cortex and spreading towards prefrontal, ventral and medial regions. Colour-coded voxels indicate the strength of MEG-fMRI RDMs correlations (Spearman's ρρ, scaled between 0 and maximal observed value, $n = 16$, cluster definition threshold $p < 0.001$, cluster threshold $p < 0.05$). See the full-brain fusion in Movie S1 [To view this figure in colour, please see the online version of this journal].

to reach pre-frontal, ventral occipitotemporal, and medial regions by ~130–140 ms (see Figure 1b, axial slices at z = 10 and −10 mm) before fading below significance by 1500 ms. Whole-brain fusion thus illustrates an orderly spatiotemporal progression of responses from early sensory cortices to higher-level and extratemporal regions, broadly consistent with processing streams in hierarchical organization (Rauschecker & Scott, 2009).

We further quantified the spatiotemporal distribution of the brain response by repeating the fusion approach in a region-of-interest (ROI) analysis, which yields a single fusion correlation time course for specific cortical regions, rather than each small searchlight volume. While we hypothesized a progression in which MEG-fMRI correspondence appears earliest in core auditory ROIs and travels systematically outward, the nature and organization of this progression remains unclear especially in early regions. We defined primary and nonprimary auditory anatomical ROIs spanning the STG and inferior frontal gyrus (IFG), and functionally defined occipital and temporal ROIs known to be category-selective for faces, objects and scenes in vision as well as audition.

Specifically, the ROIs comprised the primary auditory cortex (PAC, comprising TE1.0 and TE1.1 in posteromedial and middle HG), TE1.2, planum temporale (PT) and planum polare (PP) (Desikan et al., 2006; Morosan et al., 2001; Norman-Haignere et al., 2013). From Pernet et al. (2015), we identified the voice-selective left inferior frontal gyrus (LIFG) region, and a modified temporal voice area (designated TVAx, with voxels overlapping with PT removed). We also selected the Parahippocampal Place Area (PPA, Epstein & Kanwisher, 1998), the Medial Place Area (MPA, Silson et al., 2016), the Fusiform Face Area (Kanwisher et al., 1997; FFA, Grill-Spector et al., 2004), and the Lateral Occipital Complex (Malach et al., 1995; LOC, Grill-Spector et al., 1999). Finally, we defined the Early Visual Cortex (EVC, Lowe et al., 2016; MacEvoy & Epstein, 2011) as a control region (see Methods for details).

Similarly to the whole-brain searchlight analysis, for each participant and ROI, we extracted voxelwise fMRI BOLD responses and computed an RDM from pairwise Pearson correlation distances. We then correlated each group-averaged ROI RDM with subject-specific MEG RDMs over time to generate an fMRI-MEG fusion correlation time course per ROI per subject. As shown in Figure 2 and Table 1, statistically significant clusters (defined using sign-permutation tests, $p < 0.01$ cluster-definition threshold, $p < 0.05$ cluster threshold) were observed in all ROI fusion time courses except EVC. Early primary and nonprimary auditory ROIs (PAC, PT, TE1.2, PP) exhibited similar peak latencies of ~115 ms post-stimulus onset. Peaks in voice-selective regions (TVAx and LIFG) and MPA occurred at ~200 ms, and ~300 ms or later in the other ventral and medial functionally defined regions (FFA, PPA, and LOC). The temporal dynamics of responses in these latter regions were markedly more sustained than in auditory areas, with significant correlations lasting for hundreds of milliseconds beyond the stimulus duration of 500 ms (see Figure 2B inset). We further characterized the timing of the processing cascade by comparing the MEG–fMRI fusion time course first peaks between PAC and the other ROIs (see Table 1, latency difference). Compared with PAC, MEG–fMRI correspondence in representations emerged significantly later (at least 80 ms) in the voice-, face-, object-, and scene-selective regions. These results confirm that sound-evoked neural activity in higher cortical regions occurred later than in the early auditory regions, corroborating and quantifying the forward signal propagation revealed in the whole-brain searchlight analysis. Finally, a brief period of significant fusion correlation reappeared in superior temporal ROIs between ~1050 and ~1300 ms.

Systematic trend from acoustic to semantically dominated, category-specific coding

While the fusion analysis indexes the propagation of spatiotemporal neural response patterns, it is agnostic to their representational content. A hierarchical account would predict that an auditory processing cascade initially encodes acoustic, then progressively higher-level stimulus representations (Bizley & Cohen, 2013; Rauschecker & Scott, 2009).

Figure 2C depicts ROI-wise RDMs and corresponding representational geometry visualized with multidimensional scaling (MDS) plots. How do these representations evolve from lower to higher-level regions, as well as over time? To capture this transition, we operationalized representational levels at two extremes: First, we constructed a *Cochleagram* RDM comprising euclidean distances between

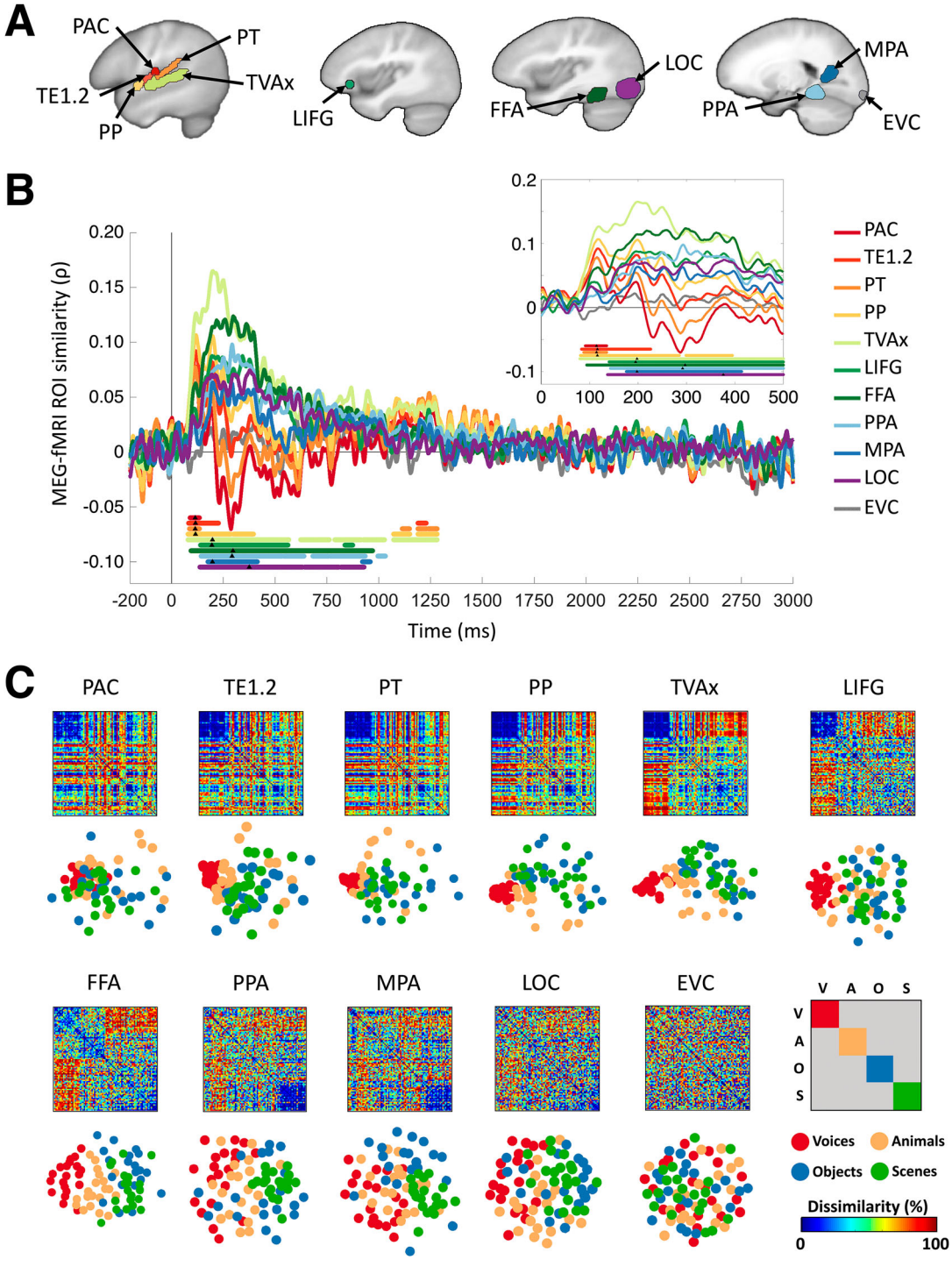

Figure 2. ROI-based MEG-fMRI fusion time courses. (**A**) Eleven regions of interest (ROIs) spanning temporal, frontal, and occipital lobes were selected to examine auditory responses, including the primary auditory cortex (PAC), TE1.2, planum temporale (PT), planum polare (PP), temporal voice area (TVAx), left inferior frontal gyrus (LIFG), fusiform face area (FFA), parahippocampal place area (PPA), medial place area (MPA), lateral occipital area (LOC), and early visual cortex (EVC). (**B**) ROI-specific fusion time courses (colour-coded to match ROI illustrations in **A**) indexing correlation between whole-brain MEG and a single RDM representing each fMRI ROI. Gray vertical line denotes stimulus onset at $t = 0$ ms. Solid colour-coded bars below time courses represent significant time points, observed for all regions except EVC; black triangles indicate respective peak latencies. Inset magnifies 0–500 ms regime (the duration of the stimulus) for clarity. All statistics, $P < 0.01$, $C < 0.05$, 1000 permutations. (**C**) Visualization of category and exemplar selectivity patterns across ROIs. RDMs and their corresponding multidimensional scaling visualization (in the first two dimensions) in the 11 fMRI ROIs. Each of the 80 stimuli is colour-coded by category, as shown in the lower right panel. Dissimilarity scale normalized to maximum within each ROI [To view this figure in colour, please see the online version of this journal].

Table 1. Mean first-peak latency for MEG-fMRI correlation time courses for 10 ROIs and comparison of peak latencies of PAC versus ROIs for interval [−200, 1000] ms.

Region of Interest	Peak latency (ms)	Latency difference (ms)	Significance (*P* value)
PAC (TE1.0 & TE1.1)	115 (100, 191)	–	–
TE1.2	117 (111, 207)	2	n.s
Planum Temporale (PT)	115 (103, 197)	0	n.s
Planum Polare (PP)	117 (110, 207)	2	n.s
Temporal Voice Area (TVAx)	199 (188, 251)	84	0.006
Left Inferior Frontal Gyrus (LIFG)	196 (188, 380)	81	0.003
Fusiform Face Area (FFA)	298 (200, 368)	183	0.0006
Parahippocampal Place Area (PPA)	293 (219, 380)	178	0.004
Medial Place Area (MPA)	199 (196, 400)	84	0.001
Lateral Occipital Complex (LOC)	377 (169, 410)	262	0.008

All values are group MEG correlated with and averaged across fMRI subjects ($n = 16$) with 95% confidence intervals in brackets. Latency differences from PAC and significance were determined by bootstrapping (5000 times) the sample of participants (1000 samples). Cluster size threshold $p < 0.05$, significance threshold $p < 0.01$.

stimulus log-frequency cochleagrams (see Equation 1, Methods, and Figure 3 & S1) to model the hypothesized low-level similarity structure of auditory afferents arriving from the cochlea. Second, we devised a *Category* RDM hypothesizing a generalized high-level semantic category selectivity across all four categories: All within-category pairwise distances were set to 0, and all between-category distances to 1. The strength of these representational levels was indexed by computing Spearman correlations between the models and brain data.

While our stimulus set was designed to minimize confounds between these models (see Methods), we would not necessarily expect a given set of fMRI voxels or MEG sensor patterns to correlate only with one and not the other, both due to the intrinsic correlation between sound acoustics and categories (Theunissen & Elie, 2014) and the fact that neurons throughout primary and nonprimary auditory cortex are responsive to low- and higher-level sound properties (King & Nelken, 2009; Staeren et al., 2009). Additionally, our binary approach elides other potential stimulus feature models that may characterize intermediate processing (Giordano et al., 2013; Ogg et al., 2020). However, in a hierarchical processing stream, we would expect their *respective* contributions to vary systematically between regions over time, and our aim was to define endpoints of a continuum of category coding. Thus, we computed the difference between Category and Cochleagram model correlations (Fisher-z normalized to enable direct comparison) with MEG and fMRI patterns. This measure, which we term *Semantic Dominance* (SD), quantifies the relative weight of neural representations: A negative value indicates predominance of acoustic properties, and a positive value indicates predominance of semantic properties. A similar analysis has previously been applied to progressively percept-driven spatial coding along the visual hierarchy (Fischer et al., 2011). We first evaluated the Cochleagram and Category models (and SD) separately on temporal and spatial data, then in a combined fashion guided by the fusion analysis. The model correlation time courses were assessed for statistical significance using permutation-cluster analysis, as with the MEG-fMRI ROI fusion time courses shown in Figure 2.

Cochleagram and Category models both correlated significantly with the whole-brain MEG signal (Figure 3A). The correlation time courses reached significance at similar onset times of ∼80 ms; the period of significant Cochleagram model correlation was shorter and peaked earlier (127 ms; 95% confidence interval 119–178 ms) compared to the Category model, which peaked at 218 (209–259) ms and remained significant through ∼1300 ms. Semantic Dominance was significantly negative from ∼90–150 ms and significantly positive from ∼200–350 ms. These results indicate a rapid onset of frequency-sensitive coding, a broad temporal regime of general semantic category-sensitive coding, and a clear temporal progression in which neural response patterns are more differentiable by semantic category than by their spectra after about 200 ms.

The Cochleagram model correlated significantly with all superior temporal fMRI ROIs (PAC, TE1.2, PT, PP, TVAx), with the PAC correlation significantly higher than the nonprimary ROIs (all $p < 0.05$ Bonferroni-corrected across STG; see Figure 3B). Frontal (LIFG) and visually defined (FFA, PPA, MPA, LOC, EVC) ROIs showed no significant Cochleagram correlation. By contrast, the Category RDM correlated significantly with all ROIs except EVC (Figure 3B).

To quantify the relationship between these results, we next tested for a systematic trend in representational level by computing a Spearman rank correlation between ROIs and their Semantic Dominance score. Our reasoning, adapted from a previous study

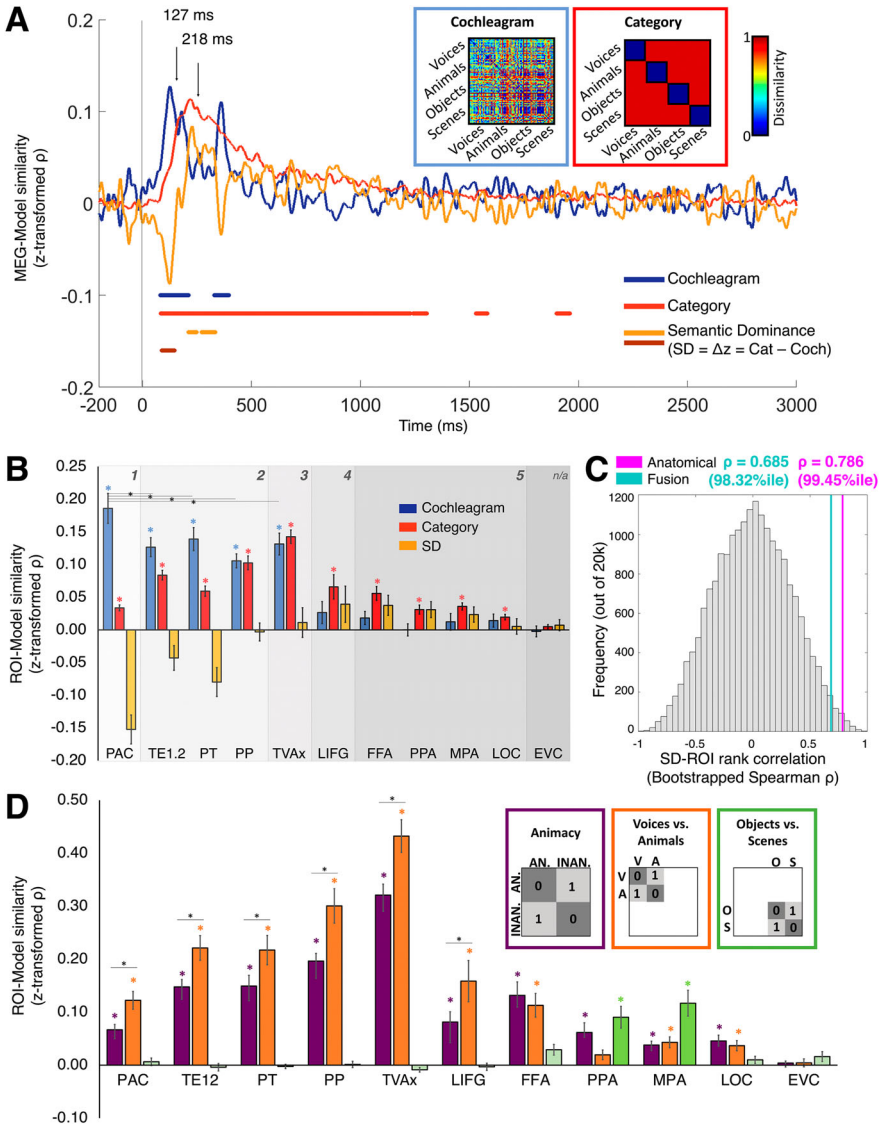

Figure 3. Emerging category vs. acoustic selectivity across time and ROIs. (**A**) Temporal dynamics in MEG responses: Cochleagram (blue) and Category (red) model RDMs shown in insets encode hypothesized similarity structures of acoustic and semantic neural stimulus representations. Pairwise Cochleagram model dissimilarities are indexed by the euclidean distance between stimulus log-frequency spectra; for the Category model, all between-category pairs were assigned a value of 1 and within-category pairs a value of 0. Plot shows Fisher z-normalized correlation time courses between models and whole-brain MEG data. Cochleagram and Category peaks at 127 (CI 119–178) ms and 217 (CI 209–259) ms, respectively. Yellow trace is Semantic Dominance (SD) difference score, $\Delta z = z_{\text{Category}} - z_{\text{Cochleagram}}$. Negative SD = acoustically dominated responses; positive SD = semantic category-dominated responses. Colour-coded bars indicate significant temporal clusters for each trace (dark bar is significantly negative SD, i.e., significantly stronger Cochleagram vs. Category model correlation). (**B**) Model correlations with RDMs computed from fMRI ROI data. Colour code same as in A. Bar plots depict the subject-averaged measure in each ROI; error bars = standard error of the mean (SEM). Numbered gray-shaded partitions indicate ranks assigned to ROIs according to anatomical position. Colour-matched asterisks indicate significantly positive correlation; black horizontal bars and bisecting asterisks indicate significant correlation differences between pairs of STG ROIs. (**C**) Histogram of 20,000 bootstrapped ROI-SD trend rank correlations compared with empirical correlations from our anatomical (magenta; 99.5%ile) and fusion-derived (cyan; 98.3%ile) ROI rank assignments. (**D**) Semantic subcategory distinctions in fMRI ROIs. Subject-averaged bar plots indicate distinction between category animacy (purple), human vs. animal vocalizations (orange), and objects vs. scenes (green). Plotting conventions and statistics same as in B. Black horizontal bars and bisecting asterisks indicate significant correlation differences between models within STG/IFG ROIs. Bonferroni corrected one-tailed t-tests, asterisks indicate significance at $p < 0.05$ [To view this figure in colour, please see the online version of this journal].

of perceptual visuospatial coding (Fischer et al., 2011), was that the ROI ranks reflect our hypothesized ordering schemes, and that a significant positive correlation would indicate a systematic trend along the ranking dimension. We therefore assigned ranks based on independent spatial (functional anatomy) and

temporal (fusion-derived peak latency) criteria. For the spatial analysis, we first differentiated primary auditory cortex based on convergent anatomical, histological and functional criteria (Morosan et al., 2001; Sweet et al., 2005), then ranked non-primary areas progressing along and beyond the supratemporal plane. We assigned PAC a rank of 1; TE1.2, PP, and PT (all adjacent to PAC) a rank of 2; TVAx (farther along the posterolateral STG) a rank of 3; LIFG a rank of 4; and FFA, PPA, MPA and LOC a rank of 5. (EVC was excluded, as no significant model correlations were found in that ROI.) See Figure 3B The trend was significantly positive ($\rho = 0.786$, $p = 0.0071$), indicating that an independent anatomically based ranking closely tracks increasingly category-selective neural representations.

To test our temporally based ordering scheme, we repeated this analysis, this time ordering ROIs guided by their empirically determined fusion peak latencies (as shown in Table 1/Figure 2B). Previous work (Cichy, Pantazis, et al., 2016) suggests that correlation peak, rather than onset, latencies more closely track electrophysiologically measured neuronal dynamics. As before, the correlation ($\rho = 0.685$, $p = 0.0288$) indicates a significant positive trend in Semantic Dominance across ROIs ranked by an independent empirical metric of their fusion correlation peaks; i.e., neural responses become dominated by category-selective patterns over time.

While our rankings were guided by these criteria, they may reflect hidden factors or correlate spuriously with the brain data. Thus, to test our *a priori* anatomical and empirical ROI-fusion-derived ranking schemes against all possible ROI rankings, we compared our ROI-rank-vs.-Semantic-Dominance correlations from those schemes against a 20,000-sample bootstrapped distribution of rank correlations, each computed by assigning each ROI a randomly drawn rank between 1 and 10 (with replacement). The anatomical and fusion-timing-based correlations were higher than 99.5% and 98.3% of all bootstrapped values, respectively (Figure 3C), indicating that our independently computed Semantic Dominance measure closely matches activation patterns in ROIs ranked by anatomical outward progression from PAC, as well as by the temporal sequence of peak fMRI-MEG correspondences in those ROIs.

The SD trend shows that pooled across multiple categories, semantic relative to acoustic selectivity increases over time and space. While a modular organization would predict SD to track category-selective maps and streams (Rauschecker & Scott, 2009), auditory cortex could conceivably be responding to other high-level, non-semantically specific features as early acoustic selectivity wanes. If this were the case, we would expect ROIs to show similar or weaker correlations to subcategorical structures, and for responses to be nonsystematically distributed across ROIs. To disambiguate between these possibilities, we examined the specificity of semantic coding by partitioning the fMRI RDMs into pairwise between- and within-category portions. Category selectivity was indexed by between- minus within-category representational distances, averaged within categories. In this way we parcellated the generic Category model to investigate broad animacy selectivity (Voices & Animals vs. Objects & Scenes) as well as more granular sensitivity to differences between animate (Voices vs. Animals) and inanimate (Objects vs. Scenes) categories. As shown in Figure 3D, Human and animal vocalizations were best differentiated (more than general animacy) in superior temporal and voice-selective ROIs ([Voices vs. Animals] minus Animacy z-transformed correlation differences, one-sample t-test against zero, Bonferroni corrected for 6 superior temporal and frontal ROIs, $p < 0.05$ for all), while inanimate environmental sounds were chiefly differentiated in scene-selective PPA and MPA, indicating functional specialization of auditory responses in these areas.

In sum, our results show that MEG responses began to correlate significantly with both Cochleagram and Category models at ~80 ms, that the Cochleagram model was dominant until ~200 ms, and that brain responses thereafter were dominated by the Category model. In fMRI responses, Cochlea-dominated coding in PAC, TE1.2, and PT, transitioned after PP and PT to Category-dominated coding, quantified by a significant positive trend in Semantic Dominance over an anatomical or fusion-peak-based ranking of ROIs, and characterized by distinct semantic subcategories coded in different regions.

Discussion

Here we analyzed the spatial, temporal, and representational structure of cortical responses to naturalistic sounds in human listeners—how the auditory system categorizes the acoustic events of the physical world.

We leveraged the temporal resolution of MEG and the spatial resolution of fMRI, fusing them within the RSA framework to track the processing cascade from early auditory cortex to frontal, ventral, and medial occipital regions. The whole-brain searchlight analysis revealed an orderly progression of neural activation originating in bilateral posteromedial temporal lobes and spreading anteriorly and posteriorly to extratemporal regions throughout much of the cortical volume. The timing of MEG-fMRI correspondence in specific regions of interest bears this out concretely, with the earliest peak latencies in superior temporal ROIs (PAC, TE1.2, PT, PP), followed by voice- (LIFG, TVAx) and scene-selective (MPA) regions and reaching ventro-temporal and occipital category-selective regions (FFA, PPA, LOC), but not early visual cortex. The representational content of the response patterns was initially strongly biased toward low-level acoustic features, as indicated by strong correlations with the Cochleagram model of auditory nerve afferents, but increasingly dominated by high-level semantic categories, as indicated by correlations with the Category model RDM. The Semantic Dominance difference score indexed the systematic trend between these two representational extremes over time (significant sign reversal of SD over the MEG epoch) and space (significant positive trend of SD across fMRI ROIs). The ROI rankings modelling the SD trend were independently estimated from two metrics: *a priori* functional-anatomical layout and empirically determined MEG-fMRI fusion time courses. These rankings produced higher trend correlations than almost all possible rankings, supporting the robustness of our convergent temporal and spatial bases for quantifying the evolving nature of the processing stream. Finally, more granular interrogation of category selectivity revealed its spatial specificity, arguing against a generalized, distributed model of categorical sound representation. Taken together, our findings provide converging evidence for large-scale hierarchical organization of the human auditory stream through distinct levels of processing dissociable by space, time, and content.

An integrative approach to identifying attributes of hierarchical coding

A hierarchical processing model posits pathways that, in parallel, transform "low-level" sensory input serially into increasingly complex, abstract "high-level" representations. Consequently, it is reasonable to expect that the physical substrate of these pathways would have sequential properties as well: As the temporal dynamics of representations follow a low-to-high-level sequence, increasing distance from a low-level origin should track with higher-level representations, which are routed to distinct loci. These principles, operationalized for the auditory system (Bizley & Cohen, 2013; Kell et al., 2018; Rauschecker & Scott, 2009), guided our interpretations of the data.

Our integrative approach addresses several methodological and conceptual barriers to resolving conflicting interpretations arising from the body of earlier work. First, the temporal specificity that MEG-fMRI fusion adds to spatially mapped brain responses allows us to directly test spatial hypotheses to which time is intrinsic, such as sequential processing streams carrying evolving levels of representation. In this way, the fusion analysis provides neural data at a level typically unavailable in humans. Even the human intracranial literature, while largely bearing out our results (Nourski, 2017), investigates higher-level processing narrowly with speech (Sahin et al., 2009; Steinschneider et al., 2014; Yi et al., 2019), in patients undergoing surgery for epilepsy, and does not provide the full-brain coverage of fMRI/MEG. Second, beyond enabling the fusion analysis, the RSA framework allows us to operationalize high- and low-level representations of a large naturalistic stimulus set spanning a range of commonly encountered sound classes. Third, the Semantic Dominance difference score analysis allows us to manage the confounds between acoustic and semantic properties of naturalistic sounds, and to test explicitly for spatial and/or temporal trends in their relative contributions to neural response patterns (Fischer et al., 2011). Finally, the richness of a multimodal data set allows us to contextualize our findings against a wider variety of previous work spanning different methods.

Separating levels of representation

How should we interpret the transition from Cochleagram to Category model dominance? Our modelling analyses capture two endpoints of a neural processing cascade transforming a waveform to a semantically organized representation. Other models representing hypothesized processing stages could be operationalized, e.g., as neuronal responses to

different acoustic properties (Giordano et al., 2013; Ogg et al., 2020; Rauschecker & Scott, 2009) components of complex sounds (Teng et al., 2017), anatomical location (Norman-Haignere & McDermott, 2018), or layers of task-optimized neural networks (Kell et al., 2018). The RSA framework and our analysis are adaptable to any of these approaches. Indeed, prior work has mapped multivariate MEG or fMRI responses (separately) to various spectrotemporal properties of natural sounds, but could not specify the extent to which those decoding results show the temporal emergence of object-level semantic labels vs. the evolving grouping of acoustic response properties, as the stimulus set's semantic and acoustic properties were not independent (Ogg et al., 2020) or the data were not temporally resolved (Giordano et al., 2013). As with visual images (Oliva & Torralba, 2001), sounds in the natural world contain some categorical structure in their low-level features (Charest et al., 2009), which complicates the separation of acoustic- vs. category-based neural decoding. We guarded against this confound by preparing the stimulus set to contain minimal category structure so the Category RDM was less likely to spuriously capture acoustic differences (See Methods and Figure S1). It was then additionally normalized relative to the Cochleagram RDM via the Semantic Dominance difference score. This approach allowed us to examine the same MEG or fMRI data for low- and high-level selectivity, and to quantify which selectivity dominates coding in a given area (and how that dominance changes). For example, PT and LIFG were each significantly sensitive to category and acoustic structure (Figure 3B), but to opposite relative extents. PP and TVAx both correlated about equally well with Cochleagram and Category models, suggesting a kind of inflection point: While Cochleagram representations do persist beyond PAC, their strength decreases both (1) across regions relative to PAC, and (2) within regions relative to categorical coding.

Distributed and hierarchical coding within and beyond superior temporal regions

In nonhuman primates, ascending auditory neurons respond to increasingly complex sound properties in distinct regions (Bizley & Cohen, 2013; Kaas & Hackett, 2000; Kaas et al., 1999; Rauschecker & Scott, 2009; Rauschecker & Tian, 2000; Rauschecker et al., 1995), which has been interpreted as an analog to the multistage hierarchy (Felleman & Van Essen, 1991; Cichy, Khosla, et al., 2016) of the visual system. Such a hierarchy in humans has principally been probed in the context of responses to acoustic- to semantic-level speech features (Brodbeck et al., 2018; de Heer et al., 2017; Evans & Davis, 2015; Okada et al., 2010; Overath et al., 2015; Rauschecker & Scott, 2009). However, compared to V1, primary auditory cortex (PAC) responses have undergone more processing (i.e., traversed more preceding synapses) (King & Nelken, 2009) and, in human listeners, are better modelled by intermediate rather than early layers of a task-optimized neural network (Kell et al., 2018). The underlying anatomy itself is more parallelized, with direct medial geniculate projections to PAC as well as non-primary areas (Hackett, 2011; Kaas & Hackett, 2000), and widespread overlapping selectivity for complex sounds along the superior temporal plane (Bizley & Cohen, 2013; King & Nelken, 2009; Staeren et al., 2009). Thus, the strong hierarchical modular view is challenged by distributed accounts of high-level audition (Formisano et al., 2008; Staeren et al., 2009).

Thus, a putative core-belt-parabelt tripartite hierarchy in early human audition remains the subject of ongoing debate. Our fusion time course results did not reveal significant peak latency differences between primary and early nonprimary ROIs (Figure 2, Table 1), but spectral sensitivity was significantly reduced and category sensitivity increased outside PAC (Figure 3B,D). Further, earlier offsets in PAC and PT and extended temporal profiles for TE1.2 and PP (see Figure 2 inset) reflect distinct dynamics for posterior vs. anterior ROIs. Because the fusion analysis combines similarity structures rather than raw signal responses (which also combines noise from both imaging modalities), it may not capture some responses with very low SNR, or which are undetectable to one modality (Cichy & Oliva, 2020). Additionally, our stimuli were controlled for duration, mean intensity, and overall spectral structure, but their complexity and variety could contribute to some temporal smear in response onsets in both neurophysiological and M/EEG studies (Murray et al., 2006; Steinschneider et al., 2014). These factors may also account for the absence of very fast onsets in the fusion signals, e.g., 10–20 ms in response to click trains in medial HG (MEG: Inui et al., 2006; Intracranial: Nourski et al.,

2014). However, our earliest fusion latencies in posteromedial temporal voxels match intracranial neuronal onsets to speech syllables in middle HG (Nourski et al., 2014; Steinschneider et al., 2014), and fusion time courses in STG peak at similar latencies to analogous intracranial responses measured as high-gamma power (Nourski et al., 2014) or modelled as spatiotemporal receptive fields (STRFs) (Hamilton et al., 2018). In the latter work, neurons sensitive to sound onsets (shorter latency, earlier offsets) clustered posteriorly along the STG and neurons with sustained response profiles (slightly longer latency, substantially later offsets) clustered more anteriorly (Hamilton et al., 2018; Nourski et al., 2014). While our analysis may be unable to distinguish onset and peak differences, the overall temporal dynamics are remarkably consistent with those we found for PAC/PT and TE1.2/PP, and suggestive of a functional organization robust to different stimuli, recording methods, and analysis approaches. The overlapping onset and peak timing between PAC, TE1.2, PT, and PP may therefore reflect parallel processing, mediated by direct medial geniculate projections not only to core, but also belt and parabelt areas (Hackett, 2011; Kaas & Hackett, 2000), or multiply branching outputs from HG to other superior temporal areas (Nourski et al., 2014). Together, the evidence suggests a functional hierarchy in superior temporal ROIs whose temporal progression is not monotonically serial, but which has the highest spectral sensitivity (and lowest category sensitivity) in PAC and an anterior-posterior distinction in temporal response profiles as well as semantic dominance (Figure 3). The negative Semantic Dominance score in the nonprimary regions (while still higher than in PAC) further suggests that the response profiles reflect stimulus acoustics more than their category, and thus represent an intermediate stage at which acoustics are being grouped, but semantic categories not yet assigned.

Beyond superior temporal regions, fusion and model analyses show responses further separated by time, space, as well as content. The spatiotemporal progression of fusion responses tended toward progressively higher onset/peak latency and duration farther from PAC (Figures 1 and 2; Table 1), in accordance with the "multiple maps and streams" feature characteristic of a distributed hierarchy (Rauschecker & Scott, 2009). Superior temporal and extratemporal ROIs (LIFG, FFA) showed a clear sensitivity to stimulus animacy and, specifically, human vs. animal vocalizations (Figures 2 and 3D), in contrast to EEG work disputing the specialized processing of human vocalizations (De Lucia et al., 2010) Fusion correlations for LIFG and TVAx both peaked at ∼200 ms, a nearly exact match to frontotemporal voice-selective ERP peaks (Charest et al., 2009) and consistent with the notion of a functionally specialized rapid voice- and speech-processing network (Belin et al., 2000; de Heer et al., 2017; Norman-Haignere et al., 2019; Pernet et al., 2015). Notably, response patterns in FFA, typically considered a visual face-selective region, distinguished animate vs. inanimate and human vs. animal sounds comparably to the classical voice-processing network, with a later fusion peak but onset comparable to the superior temporal ROIs (Figure 2B). Voices unconnected to specific identities typically elicit no or weak FFA activity at best (e.g., de Heer et al., 2017, in which FFA voxels responded to semantic but not spectral or articulatory aspects of speech); however, face and voice regions share functional and structural connectivity (von Kriegstein et al., 2005; Blank et al., 2011), and nonvisual face-related stimuli elicit FFA-colocalized responses in congenitally blind persons (Ratan Murty et al., 2020). Thus, the FFA selectivity (and rapid response compared to its anatomical neighbours) in our results may reflect its role in a person/identity-recognition mechanism with unified coding principles (Yovel & Belin, 2013) rather than voice processing *per se*.

Interestingly, only extratemporal regions, typically considered part of the visual ventral stream, distinguished between inanimate (scene and object) categories of sounds, with representations consistently correlating with semantic properties but only weakly or not at all to acoustic properties (Figure 3B,D). Traditionally sensory-specific areas have long been implicated in cross-sensory processing (Jung et al., 2018; Kim & Zatorre, 2011; von Kriegstein et al., 2005; Smith & Goodale, 2015; Vetter et al., 2014); however, in this case, the regions were activated in the absence of any visual stimulation or association with a visual cue, ruling out a multisensory integrative or paired associative process. Auditory objects and scenes have resisted easy analogical transfer from their visual counterparts, both as fundamental concepts (Griffiths & Warren, 2004) and as experimental constructs, due to differences in the spatiotemporal structures of images vs. sounds (Teng et al., 2017).

Consequently, could these category-specific representations in visually selective regions simply result from visual mental imagery? Substantial work documents the involvement of early visual cortex in mental imagery (Kosslyn et al., 1995, 2003); participants instructed to imagine sound scene categories without visual stimulation generate category-specific EVC representations (Vetter et al., 2014). Yet we found no evidence for category selectivity specifically, or MEG-fMRI correspondence generally, at any time points in EVC (Figures 2 and 3, Movie S1). Still, imagery may occur without EVC activation, generating higher-order representations (Reddy et al., 2010). These representations emerge much later compared to feedforward signals. Cued visual imagery of faces and houses becomes decodable in brain dynamics after ~400 ms, peaking at ~1000 ms (Dijkstra et al., 2018); even brief, highly overlearned stimuli like auditory letter phonemes elicit visual imagery responses peaking at ~400 ms or later (Raij, 1999; Raij et al., 2000). By contrast, MEG-fMRI fusion latencies in most of the ROIs in our study peaked before 300 ms, with LOC peak latency at 377 ms; the outermost bounds of the 95% confidence intervals were 410 ms for LOC and 400 ms for MPA (Figure 2B, Table 1). In fact, MPA peaked relatively early, at ~200 ms, a time scale for rapid auditory scene-object distinction on par with that of voice processing. The stimuli themselves were 500 ms in duration; a semantic understanding and resultant unprompted crossmodal imagery would be unlikely to arise and peak while the sound was still being presented to the listener. The fusion analysis did show a brief re-emergence of activity in nonprimary STG regions from ~1100–1300 ms (Figure 2), but only the Category model correlated significantly with whole-brain MEG at that point. Given the regions and temporal regime (over 500 ms after offset of a 500 ms stimulus), this may reflect imagery related to the task instructions to attend carefully to each stimulus (Dijkstra et al., 2018).

Our overall results are therefore consistent with the accumulating evidence for a functional primary/nonprimary distinction in humans, inconsistent with an imagery account of higher-level representations in traditionally visual areas, and extend previous work by tracing out a distributed acoustic-to-semantic processing hierarchy from primary auditory, nonprimary auditory, and multiple high-level extratemporal cortical regions.

Limitations and future directions

RSA-based M/EEG-fMRI fusion is constrained by some fundamental and practical limitations (Cichy & Oliva, 2020). For example, as described above, its central assumption of isomorphism between neuromagnetic and BOLD responses is most likely to reveal aspects of neural signals accessible to both imaging modalities, and to raise the combined signal-to-noise threshold relative to each modality alone. In our study, we mitigated these limitations by testing our hypotheses via a combination of unimodal and multimodal analyses (Figures 2 and 3). Additionally, the correlation of whole-brain MEG data with a single static map of fMRI data means that, theoretically, the dynamics of similar representations in two different regions cannot be distinguished. However, this issue can be overcome with carefully designed experimental paradigms and stimulus sets: Our present study specifically maximized differences in representations of the stimulus set, both between low- and high-level patterns as well as different category-specific patterns. Finally, care must be taken when formulating and interpreting model RDMs, which may index features other than those intended, as with other modelling approaches (e.g., Daube et al., 2019; Norman-Haignere & McDermott, 2018).

These constraints notwithstanding, the fusion approach, primarily applied to visual systems to date, has proven a powerful way to elucidate spatio-temporal features of visual processing such as V1-Inferior-Temporal (IT) timing and evolution (Cichy et al., 2014; Cichy, Pantazis, et al., 2016), scene layout encoding timing in occipital place area (OPA) (Henriksson et al., 2019), ventral-dorsal dynamics (Cichy, Pantazis, et al., 2016), task and attentional modulations (Hebart et al., 2018; Salmela et al., 2018), dissociable object size and animacy selectivity (Khaligh-Razavi et al., 2018), model- or behaviourally based similarity (Cichy, Khosla, et al., 2016, 2019), and feedforward-feedback interplay (Mohsenzadeh et al., 2018) when viewing visual objects. The versatility of its framework is readily generalizable to outstanding questions in auditory neuroscience (Cichy & Teng, 2017). For example, in the present study, we interrogated the generalized structure governing the neural representation of auditory category information; future variations on that paradigm could examine, e.g., task-contingent modulations of timing

and stimulus representation in responses, representations drawn from behavioural tasks, or complex manipulations within a single domain such as speech (e.g., O'Sullivan et al., 2019), in which spatio-temporally resolved data has previously been available only via rare, invasive clinical procedures (Nourski, 2017; Sahin et al., 2009; Yi et al., 2019). Similarly, in probing the proposed dual "what" and "where" auditory pathways (Bizley & Cohen, 2013; Rauschecker & Tian, 2000), multilevel response patterns to a large stimulus set that includes spatial as well as semantic manipulations could identify distinct and overlapping components of the processing streams from complex real-world stimuli, including dynamics of crosstalk between streams. Further, the attentional, grouping, and segregation processes mediating auditory scene analysis (Shamma et al., 2011; Teng et al., 2017) could be parcellated with fine-grained spatial and temporal resolution. In all these cases, RSA-based M/EEG-fMRI fusion provides a powerful, conceptually straightforward integrative framework especially well suited for rich, naturalistic stimulus sets that probe multiple dimensions of the physical world.

Methods

Participants

Sixteen right-handed, healthy volunteers with normal or corrected-to-normal vision and no hearing impairments (8 male, age: Mean ± s.d. = 28.25 ± 5.95 years) participated in the experiment. The study was conducted in accordance with the Declaration of Helsinki and approved by the MIT Committee on the Use of Human Experimental Subjects. All participants completed one MRI and one MEG session. All participants provided written consent for each of the sessions.

Stimuli

An initial set of 200 naturalistic monaural sounds was resampled to 44.1 kHz, root-mean-square normalized, and trimmed to 500 ms duration, including 10 ms linear rise and fall times. We computed cochleagrams for each sound using Matlab code (Brown & Cooke, 1994; Ma, 2008) that emulates the filtering properties of the human cochlea by transforming time-domain signals into a time–frequency representation of the average firing rate of auditory nerve fibers. Specifically, each waveform was divided into 1-ms bins and passed through a gammatone filterbank (64 sub-bands, centre frequencies 0–20,000 Hz). We then selected 80 sounds, 20 in each of four semantic categories, minimizing within-category repetition (i.e., of multiple sounds from the same objects or animals). To test for categorical structure in low-level features, we first created a matrix comprising all pairwise euclidean distances between stimulus cochleagrams. Each distance d between stimuli p and q was computed as the square root of absolute squared differences, summed across frequency and time bins m and n:

$$d(p, q) = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} |q_{ij} - p_{ij}|^2} \quad (1)$$

Categorical membership distances were then computed by contrasting the average between- and within-category matrix subsections. These distances were compared to a null distribution of category membership distances (corrected significance threshold $p < 0.05$), generated by randomly shuffling rows and columns of the matrix 10,000 times, each time computing the pseudo-category membership distances under the null hypothesis that the category labels are interchangeable. Using this procedure, the final stimulus set was adjusted such that no significant categorical differences were observed in the pattern of pairwise distances. A similar test on a spectrogram-based matrix revealed no significant categorical structure. A 2-d MDS-based visualization of the sounds' representational geometry is shown in Supplemental Figure S1. The full 80×80 pairwise euclidean distance matrix, normalized to the range of distances, was used as the Cochleagram model for model-brain correlation analyses seen in Figure 3.

Experimental design and procedure

Participants were familiarized with the sounds prior to the neuroimaging sessions. Each sound was accompanied simultaneously with a written description, such as "a horse neighing", "a trumpet playing", "a male shouting angrily", and "howling wind through a city" (see Table S1). Participants

could repeat playing a sound until they were familiar with it. This procedure was completed twice during a monitored setting, and once prior to each experimental session. During the neuroimaging experiment, lighting was dimmed, and participants were instructed to keep their eyes closed at all times. Each sound was presented diotically through earphones (i.e., the same waveform in both channels). In detail, for each MEG and fMRI session, participants completed 16 (MEG) or 12–14 (fMRI) runs, each lasting 330 s. Each sound was presented once in each run in random order, and sounds were randomly interleaved with twenty null (no sound) trials and ten oddball (target detection; monotone double beep) trials. Each sound trial (including oddball trials) consisted of a 500 ms sound presentation followed by 2500 ms of silence preceding the next trial. Optseq2 (https://surfer.nmr.mgh.harvard.edu/optseq/; Greve, 2002) was used to generate, optimize, and jitter the presentation of all trials, including null and oddball-detection trials, and therefore some trials contained extended periods of silence preceding the next sound. During the neuroimaging experiment, participants were instructed to press a button in response to the target (oddball) so that their focus was maintained during the entire sound trial. Null and oddball trials were excluded from the main analysis.

Meg acquisition

We acquired continuous MEG signals from 306 channels (204 planar gradiometers, 102 magnetometers, Elekta Neuromag TRIUX, Elekta, Stockholm) at a sampling rate of 1 kHz, filtered between 0.03 and 330 Hz. Raw data were preprocessed using spatiotemporal filters (Maxfilter software, Elekta, Stockholm) and then analyzed using Brainstorm (Tadel et al., 2011). MEG trials were extracted with 200 ms baseline and 3 s post-stimulus (i.e., 3,201 ms length), the baseline mean of each channel was removed, and data were temporally smoothed with a low-pass filter of 30 Hz. A total of 16 trials per condition was obtained for each session and participant.

Multivariate analysis of MEG data

To determine the amount of sound information contained in MEG signals, we employed multivariate analysis using linear support vector machine classifiers (SVM; http://www.csie.ntu.edutw/~cjlin/libsvm/; Chang & Lin, 2011). The decoding analysis was conducted independently for each participant and session. For each time point in the trial, MEG data were arranged in the form of 306-dimensional measurement vectors, yielding N pattern vectors per time point and stimulus condition (sound). We used supervised learning, with a leave-one-out cross-validation approach, to train the SVM classifier to pairwise decode any two conditions. For this, we first randomly assigned the trials to $N = 4$ trial subgroups and sub-averaged the trials within each subgroup. For each time point and pair of conditions, N − 1 pattern vectors comprised the training set and the remaining Nth pattern vectors the testing set, and the performance of the classifier to separate the two conditions was evaluated. The process was repeated 100 times with random reassignment of the data to the subgroups; the overall decoding accuracy of the classifier (chance level 50%) was the mean of those 100 iterations. The decoding accuracy was then assigned to a matrix with rows and columns indexing the conditions classified. The matrix is symmetric across the diagonal, with the diagonal undefined. This procedure yielded one 80×80 representational dissimilarity matrix (RDM) of decoding accuracies constituting a summary of representational dissimilarities for each time point. Iterating across all time points in the epoch yielded a total of 3,201 MEG RDMs.

fMRI acquisition

Magnetic resonance imaging (MRI) was conducted at the Athinoula A. Martinos Imaging Center at the MIT McGovern Institute, using a 3 T Siemens Trio scanner (Siemens, Erlangen, Germany) with a 32-channel phased-array head coil. We acquired structural images using a standard T_1-weighted sequence (176 sagittal slices, FOV = 256 mm^2, TR = 2530 ms, TE = 2.34 ms, flip angle = 9°). For the experimental task, we conducted 12–14 runs per participant in which 456 volumes were acquired for each run (gradient-echo EPI sequence: TR = 750 ms, TE = 30 ms, flip angle = 54°, FOV read = 210 mm, FOV phase = 100%, ascending acquisition, ga$p$ = 10%, resolution = 3 mm isotropic, slices = 44). For the localizer task, two runs were acquired for each participant with

440 volumes per run (gradient-echo EPI sequence: TR = 1000 ms, TE = 30 ms, flip angle = 54°, FOV read = 210 mm, FOV phase = 100%, ascending acquisition, gap = 10%, resolution = 3 mm isotropic, slices = 44).

Multivariate analysis of fMRI data

fMRI data were processed and analyzed using BrainVoyager QX 2.8 (Brain Innovation, Maastricht, the Netherlands). Data preprocessing included slice acquisition time correction, 3D motion correction, temporal filtering (linear trend removal and high-pass filtering set at 3 cycles/run), and Talairach space transformation (Talairach & Tournoux, 1988). For each participant and session separately, data were realigned and slice-time corrected, and then coregistered to the T_1 structural scan acquired in the first MRI session. fMRI data was not spatially smoothed. We then modelled the fMRI responses to the 80 sounds with a general linear model (GLM). The onsets and durations of each stimulus presentation (excluding null and oddball trials, which were omitted from further analysis after preprocessing) were entered into the GLM as regressors and convolved with a hemodynamic response function. Movement parameters entered the GLM as nuisance regressors. For each of the 80 conditions, we converted GLM parameter estimates into t-values (z-scored) by contrasting each condition estimate against the implicitly modelled baseline.

Next, we conducted a searchlight analysis to reveal similarity structures in locally constrained fMRI activity patterns. For each voxel v in the brain, we extracted fMRI patterns in its local vicinity (a 4-voxel radius) and calculated condition-wise dissimilarity (1 − Pearson's R), resulting in an 80×80 fMRI representational dissimilarity matrix (fMRI RDM). Repeating this analysis voxelwise across the whole brain yielded 40,122 RDMs.

Representational similarity analysis and MEG–fMRI searchlight/ROI fusion

To relate neuronal temporal dynamics with their spatial loci, we used representational similarity analysis (Cichy et al., 2014; Cichy, Pantazis, et al., 2016; Kriegeskorte et al., 2008; Kriegeskorte & Kievit, 2013) to link MEG and fMRI signal patterns. The basic idea is that if two sounds are similarly represented in MEG patterns, they should also be similarly represented in fMRI patterns. Our overall approach was to index pairwise (dis)similarity between sounds via SVM decoding accuracy for MEG (Guggenmos et al., 2018); Pearson correlation distance for fMRI (Cichy & Oliva, 2020; Kriegeskorte & Kievit, 2013); euclidean distance for acoustic models (Ogg et al., 2020; Sainburg et al., 2020); assigned binary membership for categorical selectivity; and between-minus-within item averages for subcategories. Correspondence between fMRI, MEG, or model representational dissimilarity matrices (RDMs) is then assessed by Spearman rank-order correlation:

$$\rho_{Spear} = r_{R(X),R(Y)} \tag{2}$$

where ρ is the Spearman correlation, r is the Pearson correlation coefficient, and $R(X)$ and $R(Y)$ are the respective rank-transformed values in the RDMs being compared. In representational (dis)similarity space, MEG and fMRI patterns thus become directly comparable; correlating them yields a spatiotemporally resolved map of neural representations, depictable as a series of spatial maps over time or, equivalently, a series of time courses over anatomical regions (see Figure 1A for illustration).

To compute whole-brain MEG-fMRI fusion maps, for each time-specific group-averaged MEG RDM, we calculated the similarity (Spearman's ρ) to each subject-specific fMRI searchlight's fMRI RDM, yielding a 3-D map of MEG–fMRI fusion correlations. Repeating this procedure across the whole brain for each millisecond yielded a MEG–fMRI correspondence "movie" indexing spatiotemporal neural dynamics of cortical auditory responses (see Movie S1).

To compute MEG-fMRI fusion time courses for ROIs, we extracted voxel responses within each ROI and computed the condition-specific pairwise Pearson distances to create a single fMRI ROI RDM per subject. Next, we averaged ROI RDMs over subjects and then computed Spearman's ρ correlations between each group-averaged ROI RDM and subject-specific whole-brain MEG RDMs over time to yield fMRI-MEG fusion time courses for each ROI per subject. Final time courses were computed by averaging across subjects.

Note that the whole-brain analysis averaged subject-specific fMRI data with the group average

MEG RDMs, while the ROI fusion analysis averaged subject-specific MEG data with the group average fMRI RDMs. Given that the same set of subjects are available from MEG and fMRI data, both these approaches are valid. In each case the correlations are bounded by the noise ceiling in the subject-specific data. Thus, the best practice is to group-average the data in the noisier modality (Cichy et al., 2014; Cichy, Pantazis, et al., 2016; Mohsenzadeh et al., 2019).

Regions of interest

Middle Heschyl's gyrus (HG; TE1.0), posteromedial HG (TE1.1), anterolateral HG (TE1.2), the planum polare (PP), and the planum temporale (PT) were first identified from anatomical volume masks derived from probabilistic volume maps (Desikan et al., 2006; Morosan et al., 2001; Norman-Haignere et al., 2013). The primary auditory cortex (PAC) was defined as a subdivision of HG including the middle and posteromedial branches (TE1.0 and TE1.1). The temporal voice area (TVA) was identified from probabilistic volume maps of voice-sensitive areas (vocal > non-vocal contrast) along the human superior temporal gyrus (Pernet et al., 2015) and was further restricted to exclude voxels overlapping with PAC, PP, PT, and TE1.2; to reflect this exclusion it is referred to as TVAx here. The left inferior frontal gyrus was identified from spatial coordinates provided by Pernet et al. (2015) and converted to Talairach space using the Yale BioImage Suite Package (Lacadie et al., 2008). For details on the functional localization of these ROIs, see Supplemental Figure S2.

For our visually defined ROIs, data from an independent functional visual localizer was analyzed using a general linear model (GLM), accounting for hemodynamic response lag (Friston et al., 1994). Each participant took part in two runs of the visual localizer task. A 7.1-min functional localizer consisting of photographs of various scenes, faces, common objects, and tile-scrambled images was used to localize the parahippocampal place area (PPA), medial place area (MPA), lateral occipital complex (LOC), fusiform face area (FFA) and an area of early visual cortex (EVC). PPA was defined as a region in the collateral sulcus and parahippocampal gyrus (Epstein & Kanwisher, 1998) whose activation was higher for scenes than for faces and objects (false discovery rate, $q < 0.05$; this threshold applies to all functional regions localized in individual observers; identified in fifteen participants). MPA (Medial Place Area; see Silson et al., 2016) was a functionally defined region overlapping with retrosplenial cortex– posterior cingulate–medial parietal cortex whose activations were higher for scenes than for faces and objects (identified in 15 participants). In accordance with Grill-Spector et al. (2000), LO, a subdivision of the lateral occipital complex (LOC), was defined as a region in the lateral occipital cortex near the posterior inferotemporal sulcus, with activation higher for objects than scrambled objects (identified in 15 participants). The fusiform face area (FFA) was identified as a region in the extrastriate cortex whose activations were higher for faces than scenes or objects (Kanwisher et al., 1997) (identified in 14 participants). Finally, a control region, early visual cortex (EVC), was identified as an area of peak activity over the posterior occipital cortex (contrast: Scrambled images > objects; identified in 15 participants) (Lowe et al., 2016; MacEvoy & Epstein, 2011). Following the identification of these functionally defined regions within participants, probabilistic maps were created using the BrainVoyager QX VOI analysis tool to evaluate the spatial consistency of each region across participants, and converted to volume masks over an averaged brain in Talairach space.

Visualization of category structure using multidimensional scaling

To visualize underlying patterns contained within the complex high-dimensional structure of the 80×80 MEG decoding matrix, we used multidimensional scaling (MDS) (Kruskal & Wish, 1978) to plot the data into a two-dimensional space of the first two dimensions of the solution, such that similar conditions were grouped together and dissimilar conditions far apart. MDS is an unsupervised method to visualize the level of similarity contained in a distance matrix (correlation distance in Figure 2C; euclidean distance between cochleagrams in Figure S1), whereby conditions are automatically assigned coordinates in space so that distances between conditions are preserved.

Quantification of relative model fits and trends

To quantify the relative strengths of the Cochleagram and Category model fits, and operationalize the trends of those fits over space and time, we first

linearized the RDM model fit estimates by applying a Fisher-z transformation to the correlation coefficients. We could thus directly compare the Δ*z* difference scores (*Semantic Dominance;* SD) between Category and Cochleagram fits within each ROI in the fMRI data, as well as the whole-brain MEG data (Fischer et al., 2011; Rosenthal & Rosnow, 1991). Performing this analysis within data sets (ROIs or whole-brain MEG) held various factors constant, such as number of voxels, local SNR, etc., that would otherwise confound direct comparisons. The resulting relative difference score was then tested for trends across time (MEG) and space (ROIs). For whole-brain MEG, we compared the SD time course against zero, testing for significance using the methods described above.

For ROI-based analyses, we assigned each ROI a rank reflecting its position in a hypothesized sequence, then computed a Spearman rank correlation ρ between ROIs and their SD Δz-score to assess the direction and significance of the trend toward neural coding dominated by categorical vs. acoustic (cochleagram) information ($\alpha = 0.05$). The SD trend analysis was conducted with two ROI rank assignments based on hypothesized spatial and temporal criteria. First, to assign ranks based on functional anatomy, we differentiated primary auditory cortex based on convergent anatomical, histological and functional criteria (Morosan et al., 2001; Sweet et al., 2005), then ranked non-primary areas progressing from PAC along and beyond the supratemporal plane. We assigned PAC a rank of 1; TE1.2, PP, and PT (all adjacent to PAC) a rank of 2; TVAx (farther along the posterolateral STG) a rank of 3; LIFG a rank of 4; and FFA, PPA, MPA and LOC a rank of 5. EVC was excluded, as no significant MEG-fMRI fusion time points or RDM model correlations were found in that ROI. To test our temporally based hypothesis, we repeated the analysis, this time assigning ROI ranks guided by MEG-fMRI fusion peak latencies. Finally, to assess how well our anatomical and fusion-guided rank correlations tracked the evolution of Semantic Dominance compared to all possible such correlations, we generated a 20,000-sample bootstrapped distribution of ROI-Semantic Dominance rank correlations, assigning ROIs a randomly drawn rank from 1 to 10 (with replacement) for each sample and comparing the empirical correlations against that distribution.

Statistical testing

Nonparametric statistical tests were used to assess significance. To obtain a permutation distribution of maximal cluster size, we randomly shuffled the sign of participant-specific data points 1000 times, averaged the permuted data across participants each time, and determined 4-D clusters by spatial and temporal contiguity at the cluster-definition threshold (cluster definition threshold $p < 0.001$; cluster size threshold at $p < 0.05$). Storing the maximal cluster statistic (size of cluster with each voxel equally weighted) for each permutation sample yielded a distribution of the maximal cluster size under the null hypothesis. We report clusters as significant if they were greater than the 95% threshold constructed from the maximal cluster size distribution (i.e., cluster size threshold at $P = 0.05$). A similar statistical test method was used for time-series data (fMRI-MEG ROI fusion), but clusters were determined in 1D at a cluster-definition threshold of $p < 0.01$ and cluster-size threshold of $p < 0.05$.

For statistical assessments of time-series peak and onset latencies, we performed bootstrapping tests. To estimate an empirical distribution over peak and onset latencies of time courses, the subject-specific ROI fusion time series were bootstrapped (5000 times) and the 95% confidence interval was defined on the empirical distribution. For peak-to-peak latency comparisons, we obtained the 1000 bootstrapped samples of two peaks and rejected the null hypothesis if the 95% confidence interval of the peak latency differences did not include zero.

ROI-based analysis in functionally defined auditory regions

Independent functional data was used to localize three auditory regions of interest. This data was taken from the final run of each fMRI session for each participant, and these runs were not included in the experimental analysis. Primary auditory cortex (PAC) was identified bilaterally in fourteen participants as a region of peak activity (contrast: Oddball > null) located over Heschl's gyrus. Following Pernet and colleagues (2015), the posterior cluster of the temporal voice area (TVA) was identified bilaterally in twelve participants as a region of peak activity

(contrast: Vocal > non-vocal) located over the middle/posterior superior temporal sulcus. Similarly, a voice area within the extended voice processing network was identified in nine participants as a region of peak activity (contrast: Vocal > non-vocal) over the left inferior frontal gyrus (LIFG). Probabilistic mapping was then used to create volume masks over an averaged Talairach brain using BrainVoyager QX.

Acknowledgment

The study was conducted at the Athinoula A. Martinos Imaging Center, MIBR, MIT. We thank Dimitrios Pantazis for helpful discussion, and Michele Winter for help with stimulus selection and processing.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was funded by an Office of Naval Research Vannevar Bush Faculty Fellowship to A.O. [grant number N00014-16-1-3116]. S.T. was supported by an NIH training grant to Smith-Kettlewell Institute [grant number T32EY025201] and by the National Institute on Disability, Independent Living, and Rehabilitation Research [grant number 90RE5024-01-00].

References

Ahveninen, J., Jaaskelainen, I. P., Raij, T., Bonmassar, G., Devore, S., Hamalainen, M., Levanen, S., Lin, F.-H., Sams, M., Shinn-Cunningham, B. G., Witzel, T., & Belliveau, J. W. (2006). Task-modulated "what" and "where" pathways in human auditory cortex. *Proceedings of the National Academy of Sciences*, *103*(39), 14608–14613. https://doi.org/10.1073/pnas.0510480103

Ahveninen, J., Kopčo, N., & Jääskeläinen, I. P. (2014). Psychophysics and neuronal bases of sound localization in humans. *Hearing Research*, *307*, 86–97. https://doi.org/10.1016/j.heares.2013.07.008

Belin, P., Zatorre, R. J., Lafaille, P., Ahad, P., & Pike, B. (2000). Voice-selective areas in human auditory cortex. *Nature*, *403*(6767), 309–312. https://doi.org/10.1038/35002078

Bizley, J. K., & Cohen, Y. E. (2013). The what, where and how of auditory-object perception. *Nature Reviews Neuroscience*, *14*(10), 693–707. https://doi.org/10.1038/nrn3565

Blank, H., Anwander, A., & von Kriegstein, K. (2011). Direct structural connections between voice- and face-recognition areas. *Journal of Neuroscience*, *31*(36), 12906–12915. https://doi.org/10.1523/JNEUROSCI.2091-11.2011

Brodbeck, C., Hong, L. E., & Simon, J. Z. (2018). Rapid transformation from auditory to linguistic representations of continuous speech. *Current Biology*, *28*(24), 3976–3983.e5. https://doi.org/10.1016/j.cub.2018.10.042

Brown, G. J., & Cooke, M. (1994). Computational auditory scene analysis. *Computer Speech & Language*, *8*(4), 297–336. https://doi.org/10.1006/csla.1994.1016

Chang, C.-C., & Lin, C.-J. (2011). Libsvm. *ACM Transactions on Intelligent Systems and Technology*, *2*(3), 1–27. https://doi.org/10.1145/1961189.1961199

Charest, I., Pernet, C. R., Rousselet, G. A., Quiñones, I., Latinus, M., Fillion-Bilodeau, S., Chartrand, J.-P., & Belin, P. (2009). Electrophysiological evidence for an early processing of human voices. *BMC Neuroscience*, *10*(1), 1–11. https://doi.org/10.1186/1471-2202-10-127.

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, *6*(1), 27755. https://doi.org/10.1038/srep27755

Cichy, R. M., Kriegeskorte, N., Jozwik, K. M., van den Bosch, J. J. F., & Charest, I. (2019). The spatiotemporal neural dynamics underlying perceived similarity for real-world objects. *NeuroImage*, *194*, 12–24. https://doi.org/10.1016/j.neuroimage.2019.03.031

Cichy, R. M., & Oliva, A. (2020). A M/EEG-fMRI fusion primer: Resolving human brain responses in space and time. *Neuron*, *107*(5), 772–781. https://doi.org/10.1016/j.neuron.2020.07.001

Cichy, R. M., Pantazis, D., & Oliva, A. (2014). Resolving human object recognition in space and time. *Nature Neuroscience*, *17*(3), 455–462. https://doi.org/10.1038/nn.3635

Cichy, R. M., Pantazis, D., & Oliva, A. (2016). Similarity-based fusion of MEG and fMRI reveals spatio-temporal dynamics in human cortex during visual object recognition. *Cerebral Cortex*, *26*(8), 3563–3579. https://doi.org/10.1093/cercor/bhw135

Cichy, R. M., & Teng, S. (2017). Resolving the neural dynamics of visual and auditory scene processing in the human brain: A methodological approach. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1714)), 20160108. https://doi.org/10.1098/rstb.2016.0108

Daube, C., Ince, R. A. A., & Gross, J. (2019). Simple acoustic features Can explain phoneme-based predictions of cortical responses to speech. *Current Biology*, *29*(12), 1924–1937.e9. https://doi.org/10.1016/j.cub.2019.04.067

de Heer, W. A., Huth, A. G., Griffiths, T. L., Gallant, J. L., & Theunissen, F. E. (2017). The hierarchical cortical organization of human speech processing. *The Journal of Neuroscience*, *37*(27), 6539–6557. https://doi.org/10.1523/JNEUROSCI.3267-16.2017

De Lucia, M., Clarke, S., & Murray, M. M. (2010). A temporal hierarchy for conspecific vocalization discrimination in humans. *Journal of Neuroscience*, *30*(33), 11210–11221. https://doi.org/10.1523/JNEUROSCI.2239-10.2010

Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., Buckner, R. L., Dale, A. M., Paul Maguire, R.,

Hyman, B. T., Albert, M. S., & Killiany, R. J. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into Gyral based regions of interest. *NeuroImage*, *31*(3), 968–980. https://doi.org/10.1016/j.neuroimage.2006.01.021

Dijkstra, N., Mostert, P., de Lange, F. P., Bosch, S., & van Gerven, M. A. J. (2018). Differential temporal dynamics during visual imagery and perception. *eLife*, *7*, e33904. https://doi.org/10.7554/eLife.33904

Engel, L. R., Frum, C., Puce, A., Walker, N. A., & Lewis, J. W. (2009). Different categories of living and non-living sound-sources activate distinct cortical networks. *NeuroImage*, *47*(4), 1778–1791. https://doi.org/10.1016/j.neuroimage.2009.05.041

Epstein, R., & Kanwisher, N. (1998). A cortical representation of the local visual environment. *Nature*, *392*(6676), 598–601. https://doi.org/10.1038/33402

Evans, S., & Davis, M. H. (2015). Hierarchical organization of auditory and motor representations in Speech Perception: Evidence from searchlight similarity analysis. *Cerebral Cortex*, *25*(12), 4772–4788. https://doi.org/10.1093/cercor/bhv136

Felleman, D. J., & Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, *1*(1), 1–47.

Fischer, J., Spotswood, N., & Whitney, D. (2011). The emergence of perceived position in the visual system. *Journal of Cognitive Neuroscience*, *23*(1), 119–136. https://doi.org/10.1162/jocn.2010.21417

Formisano, E., De Martino, F., Bonte, M., & Goebel, R. (2008). "Who" is saying "what"? Brain-based decoding of human voice and speech. *Science*, *322*(5903), 970–973.

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. *Human Brain Mapping*, *2*(4), 189–210. https://doi.org/10.1002/hbm.460020402

Giordano, B. L., McAdams, S., Zatorre, R. J., Kriegeskorte, N., & Belin, P. (2013). Abstract encoding of auditory objects in cortical activity patterns. *Cerebral Cortex*, *23*(9), 2025–2037. https://doi.org/10.1093/cercor/bhs162

Greve, D. (2002). *Optseq Home Page*. http://surfer.nmr.mgh.harvard.edu/optseq.

Griffiths, T. D., & Warren, J. D. (2004). What is an auditory object? *Nature Reviews Neuroscience*, *5*(11), 887–892. https://doi.org/10.1038/nrn1538

Grill-Spector, K., Knouf, N., & Kanwisher, N. (2004). The fusiform face area subserves face perception, not generic within-category identification. *Nature Neuroscience*, *7*(5), 555–562. https://doi.org/10.1038/nn1224

Grill-Spector, K., Kushnir, T., Edelman, S., Avidan, G., Itzchak, Y., & Malach, R. (1999). Differential processing of objects under various viewing conditions in the human lateral occipital complex. *Neuron*, *24*(1), 187–203. https://doi.org/10.1016/S0896-6273(00)80832-6

Grill-Spector, K., Kushnir, T., Hendler, T., & Malach, R. (2000). The dynamics of object-selective activation correlate with recognition performance in humans. *Nature Neuroscience*, *3*(8), 837–843. https://doi.org/10.1038/77754

Guggenmos, M., Sterzer, P., & Cichy, R. M. (2018). Multivariate pattern analysis for MEG: A comparison of dissimilarity measures. *NeuroImage*, *173*, 434–447.

Hackett, T. A. (2011). Information flow in the auditory cortical network. *Hearing Research*, *271*(1-2), 133–146. https://doi.org/10.1016/j.heares.2010.01.011

Hamilton, L. S., Edwards, E., & Chang, E. F. (2018). A spatial map of onset and sustained responses to speech in the human superior temporal gyrus. *Current Biology*, *28*(12), 1860–1871.e4. https://doi.org/10.1016/j.cub.2018.04.033

Haynes, J.-D., & Rees, G. (2005). Predicting the orientation of invisible stimuli from activity in human primary visual cortex. *Nature Neuroscience*, *8*(5), 686–691. https://doi.org/10.1038/nn1445

Hebart, M. N., Bankson, B. B., Harel, A., Baker, C. I., & Cichy, R. M. (2018). The representational dynamics of task and object processing in humans. *Elife*, *7*, e32816. https://doi.org/10.7554/eLife.32816

Henriksson, L., Mur, M., & Kriegeskorte, N. (2019). Rapid invariant encoding of scene layout in human OPA. *Neuron*, *103*(1), 161–171.e3. https://doi.org/10.1016/j.neuron.2019.04.014

Inui, K., Okamoto, H., Miki, K., Gunji, A., & Kakigi, R. (2006). Serial and parallel processing in the human auditory cortex: A magnetoencephalographic study. *Cerebral Cortex*, *16*(1), 18–30. https://doi.org/10.1093/cercor/bhi080

Jung, Y., Larsen, B., & Walther, D. B. (2018). Modality-Independent coding of scene categories in prefrontal cortex. *The Journal of Neuroscience*, *38*(26), 5969–5981. https://doi.org/10.1523/JNEUROSCI.0272-18.2018

Kaas, J. H., & Hackett, T. A. (2000). Subdivisions of auditory cortex and processing streams in primates. *Proceedings of the National Academy of Sciences*, *97*(22), 11793–11799. https://doi.org/10.1073/pnas.97.22.11793

Kaas, J. H., Hackett, T. A., & Tramo, M. J. (1999). Auditory processing in primate cerebral cortex. *Current Opinion in Neurobiology*, *9*(2), 164–170.

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for face perception. *The Journal of Neuroscience*, *17*(11), 4302–4311. https://doi.org/10.1523/JNEUROSCI.17-11-04302.1997

Kell, A. J. E., & McDermott, J. H. (2019). Invariance to background noise as a signature of non-primary auditory cortex. *Nature Communications*, *10*(1), 3958. https://doi.org/10.1038/s41467-019-11710-y

Kell, A. J. E., Yamins, D. L. K., Shook, E. N., Norman-Haignere, S. V., & McDermott, J. H. (2018). A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron*, *98*(3), 630–644.e16. https://doi.org/10.1016/j.neuron.2018.03.044

Khaligh-Razavi, S.-M., Cichy, R. M., Pantazis, D., & Oliva, A. (2018). Tracking the spatiotemporal neural dynamics of real-world object size and animacy in the human brain.

Journal of Cognitive Neuroscience, *30*(11), 1559–1576. https://doi.org/10.1162/jocn_a_01290

Kim, J.-K., & Zatorre, R. J. (2011). Tactile–auditory shape learning engages the lateral occipital complex. *Journal of Neuroscience*, *31*(21), 7848–7856. https://doi.org/10.1523/JNEUROSCI.3399-10.2011

King, A. J., & Nelken, I. (2009). Unraveling the principles of auditory cortical processing: Can we learn from the visual system? *Nature Neuroscience*, *12*(6), 698–701. https://doi.org/10.1038/nn.2308

Kosslyn, S. M., Ganis, G., & Thompson, W. L. (2003). Mental imagery: Against the nihilistic hypothesis. *Trends in Cognitive Sciences*, *7*(3), 109–111. https://doi.org/10.1016/S1364-6613(03)00025-1

Kosslyn, S. M., Thompson, W. L., Kim, I. J., & Alpert, N. M. (1995). Topographical representations of mental images in primary visual cortex. *Nature*, *378*(6556), 496–498. https://doi.org/10.1038/378496a0

Kriegeskorte, N., Goebel, R., & Bandettini, P. (2006). Information-based functional brain mapping. *Proceedings of the National Academy of Sciences*, *103*(10), 3863–3868. https://doi.org/10.1073/pnas.0600244103

Kriegeskorte, N., & Kievit, R. A. (2013). Representational geometry: Integrating cognition, computation, and the brain. *Trends in Cognitive Sciences*, *17*(8), 401–412. https://doi.org/10.1016/j.tics.2013.06.007

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Neuroscience*, *2*(1), 4–5. https://doi.org/10.3389/neuro.01.016.2008

Kruskal, J. B., & Wish, M. (1978). *Multidimensional scaling (No. 11)*. Sage Publications.

Lacadie, C. M., Fulbright, R. K., Rajeevan, N., Constable, R. T., & Papademetris, X. (2008). More accurate Talairach coordinates for neuroimaging using non-linear registration. *Neuroimage*, *42*(2), 717–725. https://doi.org/10.1016/j.neuroimage.2008.04.240

Lowe, M. X., Gallivan, J. P., Ferber, S., & Cant, J. S. (2016). Feature diagnosticity and task context shape activity in human scene-selective cortex. *Neuroimage*, *125*, 681–692. https://doi.org/10.1016/j.neuroimage.2015.10.089

Ma, N. (2008). *Cochleagram representation of sound*. Resources. http://staffwww.dcs.shef.ac.uk/people/N.Ma/resources/ratemap/.

MacEvoy, S. P., & Epstein, R. A. (2011). Constructing scenes from objects in human occipitotemporal cortex. *Nature Neuroscience*, *14*(10), 1323–1329. https://doi.org/10.1038/nn.2903

Malach, R., Reppas, J. B., Benson, R. R., Kwong, K. K., Jiang, H., Kennedy, W. A., Ledden, P. J., Brady, T. J., Rosen, B. R., & Tootell, R. B. (1995). Object-related activity revealed by functional magnetic resonance imaging in human occipital cortex. *Proceedings of the National Academy of Sciences*, *92*(18), 8135–8139. https://doi.org/10.1073/pnas.92.18.8135

McDermott, J. H. (2018). Audition. In J. T. Wixted (Ed.), *Stevens' handbook of experimental psychology and cognitive neuroscience* (pp. 1–57). John Wiley & Sons, Inc.

Mohsenzadeh, Y., Mullin, C., Lahner, B., Cichy, R. M., & Oliva, A. (2019). Reliability and generalizability of similarity-based fusion of MEG and fMRI data in human ventral and dorsal visual streams. *Vision (Basel)*, *3*(8), 1–18. https://doi.org/10.3390/vision3010008

Mohsenzadeh, Y., Qin, S., Cichy, R. M., & Pantazis, D. (2018). Ultra-Rapid serial visual presentation reveals dynamics of feedforward and feedback processes in the ventral visual pathway. *Elife*, *7*, e36329. https://doi.org/10.7554/eLife.36329

Morosan, P., Rademacher, J., Schleicher, A., Amunts, K., Schormann, T., & Zilles, K. (2001). Human primary auditory cortex: Cytoarchitectonic subdivisions and mapping into a spatial reference system. *Neuroimage*, *13*(4), 684–701. https://doi.org/10.1006/nimg.2000.0715

Murray, M. M., Camen, C., Andino, S. L. G., Bovet, P., & Clarke, S. (2006). Rapid brain discrimination of sounds of objects. *Journal of Neuroscience*, *26*(4), 1293–1302. https://doi.org/10.1523/JNEUROSCI.4511-05.2006

Norman-Haignere, S., Kanwisher, N., & McDermott, J. H. (2013). Cortical pitch regions in humans respond primarily to resolved harmonics and are located in specific tonotopic regions of anterior auditory cortex. *Journal of Neuroscience*, *33*(50), 19451–19469. https://doi.org/10.1523/JNEUROSCI.2880-13.2013

Norman-Haignere, S. V., Kanwisher, N., McDermott, J. H., & Conway, B. R. (2019). Divergence in the functional organization of human and macaque auditory cortex revealed by fMRI responses to harmonic tones. *Nature Neuroscience*, *22*(7), 1057–1060. https://doi.org/10.1038/s41593-019-0410-7

Norman-Haignere, S. V., & McDermott, J. H. (2018). Neural responses to natural and model-matched stimuli reveal distinct computations in primary and nonprimary auditory cortex. *PLOS Biology*, *16*(12), e2005127. https://doi.org/10.1371/journal.pbio.2005127

Nourski, K. V. (2017). Auditory processing in the human cortex: An intracranial electrophysiology perspective. *Laryngoscope Investigative Otolaryngology*, *2*(4), 147–156. https://doi.org/10.1002/lio2.73

Nourski, K. V., Steinschneider, M., McMurray, B., Kovach, C. K., Oya, H., Kawasaki, H., & Howard, M. A. 3rd. (2014). Functional organization of human auditory cortex: Investigation of response latencies through direct recordings. *Neuroimage*, *101*, 598–609. https://doi.org/10.1016/j.neuroimage.2014.07.004.

Ogg, M., Carlson, T. A., & Slevc, L. R. (2020). The rapid emergence of auditory object representations in cortex reflect central acoustic attributes. *Journal of Cognitive Neuroscience*, *32*(1), 111–123. https://doi.org/10.1162/jocn_a_01472

Okada, K., Rong, F., Venezia, J., Matchin, W., Hsieh, I.-H., Saberi, K., Serences, J. T., & Hickok, G. (2010). Hierarchical organization of human auditory cortex: Evidence from acoustic

invariance in the response to intelligible speech. *Cerebral Cortex*, *20*(10), 2486–2495. https://doi.org/10.1093/cercor/bhp318

Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, *42*(3), 145–175. https://doi.org/10.1023/A:1011139631724

O'Sullivan, J., Herrero, J., Smith, E., Schevon, C., McKhann, G. M., Sheth, S. A., Mehta, A. D., & Mesgarani, N. (2019). Hierarchical encoding of attended auditory objects in multi-talker speech perception. *Neuron*, *104*(6), 1195–1209.e3. https://doi.org/10.1016/j.neuron.2019.09.007

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. *Nature Neuroscience*, *18*(6), 903–911. https://doi.org/10.1038/nn.4021

Pernet, C. R., McAleer, P., Latinus, M., Gorgolewski, K. J., Charest, I., Bestelmeyer, P. E. G., Watson, R. H., Fleming, D., Crabbe, F., Valdes-Sosa, M., & Belin, P. (2015). The human voice areas: Spatial organization and inter-individual variability in temporal and extra-temporal cortices. *Neuroimage*, *119*, 164–174. https://doi.org/10.1016/j.neuroimage.2015.06.050

Raij, T. (1999). Patterns of brain activity during visual imagery of letters. *Journal of Cognitive Neuroscience*, *11*(3), 282–299. https://doi.org/10.1162/089892999563391

Raij, T., Uutela, K., & Hari, R. (2000). Audiovisual integration of letters in the human brain. *Neuron*, *28*(2), 617–625. https://doi.org/10.1016/S0896-6273(00)00138-0

Ratan Murty, N. A., Teng, S., Beeler, D., Mynick, A., Oliva, A., & Kanwisher, N. (2020). Visual experience is not necessary for the development of face-selectivity in the lateral fusiform gyrus. *Proceedings of the National Academy of Sciences*, *117*(37), 23011–23020. https://doi.org/10.1073/pnas.2004607117

Rauschecker, J. P., & Scott, S. K. (2009). Maps and streams in the auditory cortex: Nonhuman primates illuminate human speech processing. *Nature Neuroscience*, *12*(6), 718–724. https://doi.org/10.1038/nn.2331

Rauschecker, J. P., & Tian, B. (2000). Mechanisms and streams for processing of "what" and "where" in auditory cortex. *Proceedings of the National Academy of Sciences*, *97*(22), 11800–11806. https://doi.org/10.1073/pnas.97.22.11800

Rauschecker, J. P., Tian, B., & Hauser, M. (1995). Processing of complex sounds in the macaque nonprimary auditory cortex. *Science*, *268*(5207), 111–114.

Reddy, L., Tsuchiya, N., & Serre, T. (2010). Reading the mind's eye: Decoding category information during mental imagery. *Neuroimage*, *50*(2), 818–825. https://doi.org/10.1016/j.neuroimage.2009.11.084

Rosenthal, R., & Rosnow, R. L. (1991). *Essentials of behavioral research: Methods and data analysis*. McGraw-Hill.

Sahin, N. T., Pinker, S., Cash, S. S., Schomer, D., & Halgren, E. (2009). Sequential processing of lexical, grammatical, and phonological information within Broca's area. *Science*, *326*(5951), 445–449. https://doi.org/10.1126/science.1174481

Sainburg, T., Thielk, M., & Gentner, T. Q. (2020). Finding, visualizing, and quantifying latent structure across diverse animal vocal repertoires. *PLOS Computational Biology*, *16*(10), e1008228. https://doi.org/10.1371/journal.pcbi.1008228

Salmela, V., Salo, E., Salmi, J., & Alho, K. (2018). Spatiotemporal dynamics of attention networks revealed by representational similarity analysis of EEG and fMRI. *Cerebral Cortex*, *28*, 549–560. https://doi.org/10.1093/cercor/bhw389

Shamma, S. A., Elhilali, M., & Micheyl, C. (2011). Temporal coherence and attention in auditory scene analysis. *Trends in Neurosciences*, *34*(3), 114–123. https://doi.org/10.1016/j.tins.2010.11.002

Silson, E. H., Steel, A. D., & Baker, C. I. (2016). Scene-Selectivity and retinotopy in medial parietal cortex. *Frontiers in Human Neuroscience*, *10*, 412. https://doi.org/10.3389/fnhum.2016.00412

Smith, F. W., & Goodale, M. A. (2015). Decoding visual object categories in early somatosensory cortex. *Cerebral Cortex*, *25*(4), 1020–1031. https://doi.org/10.1093/cercor/bht292

Staeren, N., Renvall, H., De Martino, F., Goebel, R., & Formisano, E. (2009). Sound categories are represented as distributed patterns in the human auditory cortex. *Current Biology*, *19*(6), 498–502. https://doi.org/10.1016/j.cub.2009.01.066

Steinschneider, M., Nourski, K. V., Rhone, A. E., Kawasaki, H., Oya, H., & Howard, M. A. 3rd. (2014). Differential activation of human core, non-core and auditory-related cortex during speech categorization tasks as revealed by intracranial recordings. *Frontiers in Neuroscience*, *8*, 240. https://doi.org/10.3389/fnins.2014.00240

Sweet, R. A., Dorph-Petersen, K.-A., & Lewis, D. A. (2005). Mapping auditory core, lateral belt, and parabelt cortices in the human superior temporal gyrus. *The Journal of Comparative Neurology*, *491*(3), 270–289. https://doi.org/10.1002/cne.20702

Tadel, F., Baillet, S., Mosher, J. C., Pantazis, D., & Leahy, R. M. (2011). Brainstorm: A user-friendly application for MEG/EEG analysis. *Computational Intelligence and Neuroscience*, *2011*, 1–13. https://doi.org/10.1155/2011/879716

Talairach, J., & Tournoux, P. (1988). *Co-planar stereotaxic atlas of the human brain* (Vol. 270). Theime.

Teng, S., Sommer, V. R., Pantazis, D., & Oliva, A. (2017). Hearing scenes: A neuromagnetic signature of auditory source and reverberant space separation. *eNeuro*, *4*, 1–15. https://doi.org/10.1523/ENEURO.0007-17.2017

Theunissen, F. E., & Elie, J. E. (2014). Neural processing of natural sounds. *Nature Reviews Neuroscience*, *15*(6), 355–366. https://doi.org/10.1038/nrn3731

Traer, J., & McDermott, J. H. (2016). Statistics of natural reverberation enable perceptual separation of sound and space. *Proceedings of the National Academy of Sciences*, *113*(48), E7856–E7865. https://doi.org/10.1073/pnas.1612524113

Vetter, P., Smith, F. W., & Muckli, L. (2014). Decoding sound and imagery content in early visual cortex. *Current Biology*, *24*(11), 1256–1262. https://doi.org/10.1016/j.cub.2014.04.020

von Kriegstein, K., Kleinschmidt, A., Sterzer, P., & Giraud, A.-L. (2005). Interaction of face and voice areas during speaker recognition. *Journal of Cognitive Neuroscience*, *17*(3), 367–376. https://doi.org/10.1162/0898929053279577

Wardle, S. G., Taubert, J., Teichmann, L., & Baker, C. I. (2020). Rapid and dynamic processing of face pareidolia in the human brain. *Nature Communications*, *11*(1), 4518. https://doi.org/10.1038/s41467-020-18325-8

Yi, H. G., Leonard, M. K., & Chang, E. F. (2019). The encoding of speech sounds in the superior temporal gyrus. *Neuron*, *102*(6), 1096–1110. https://doi.org/10.1016/j.neuron.2019.04.023

Yovel, G., & Belin, P. (2013). A unified coding strategy for processing faces and voices. *Trends in Cognitive Sciences*, *17*(6), 263–271. https://doi.org/10.1016/j.tics.2013.04.004