Behavioral/Cognitive

Modality-Independent Coding of Scene Categories in Prefrontal Cortex

Yaelan Jung,[1] Bart Larsen,[2] and Dirk B. Walther[1]

[1]Department of Psychology, University of Toronto, Toronto, Ontario M5S 3G3, Canada and [2]Department of Psychology, University of Pittsburgh, Pittsburgh, Pennsylvania 15260

Natural environments convey information through multiple sensory modalities, all of which contribute to people's percepts. Although it has been shown that visual or auditory content of scene categories can be decoded from brain activity, it remains unclear how humans represent scene information beyond a specific sensory modality domain. To address this question, we investigated how categories of scene images and sounds are represented in several brain regions. A group of healthy human subjects (both sexes) participated in the present study, where their brain activity was measured with fMRI while viewing images or listening to sounds of different real-world environments. We found that both visual and auditory scene categories can be decoded not only from modality-specific areas, but also from several brain regions in the temporal, parietal, and prefrontal cortex (PFC). Intriguingly, only in the PFC, but not in any other regions, categories of scene images and sounds appear to be represented in similar activation patterns, suggesting that scene representations in PFC are modality-independent. Furthermore, the error patterns of neural decoders indicate that category-specific neural activity patterns in the middle and superior frontal gyri are tightly linked to categorization behavior. Our findings demonstrate that complex scene information is represented at an abstract level in the PFC, regardless of the sensory modality of the stimulus.

Key words: cross-modal; fMRI; modality-independent representation; multivoxel pattern analysis; PFC; scene perception

Significance Statement

Our experience in daily life includes multiple sensory inputs, such as images, sounds, or scents from the surroundings, which all contribute to our understanding of the environment. Here, for the first time, we investigated where and how in the brain information about the natural environment from multiple senses is merged to form modality-independent representations of scene categories. We show direct decoding of scene categories across sensory modalities from patterns of neural activity in the prefrontal cortex (PFC). We also conclusively tie these neural representations to human categorization behavior by comparing patterns of errors between a neural decoder and behavior. Our findings suggest that PFC is a central hub for integrating sensory information and computing modality-independent representations of scene categories.

Introduction

Imagine taking a walk on the beach. Your sensory experience would include the sparkle of the sun's reflection on the water, the sound of the crushing waves, and the smell of ocean air. Even though the brain has clearly delineated processing channels for all of these sensory modalities, we still have the integral concept of the beach, which is not tied to particular sensory modalities. What are the neural systems underlying this convergence, which allows our brain to represent the world beyond sensory modalities? Here we show the neural representations of scene information that generalize across different sensory modalities.

Neural mechanisms underlying the perception of visual scenes have been studied extensively for the last two decades, showing a hierarchical structure from posterior to anterior parts of visual cortex. Starting from low-level features, such as orientation, represented in primary visual cortex, the level of representation becomes more abstract, through intermediate-level features represented in V3 and V4 (Nishimoto et al., 2011), to various aspects of scenes represented in high-level visual areas: local elements of a scene in the occipital place area (OPA) (MacEvoy and Epstein, 2007; Dilks et al., 2013), scene geometry and scene content in the parahippocampal place area (PPA) (Epstein and Kanwisher, 1998; Walther et al., 2009), and the embedding of

Received Jan. 26, 2018; revised May 3, 2018; accepted May 26, 2018.

Author contributions: Y.J. wrote the first draft of the paper; Y.J. and D.B.W. edited the paper; D.B.W. designed research; Y.J., B.L., and D.B.W. performed research; Y.J. and D.B.W. analyzed data; Y.J. and D.B.W. wrote the paper.

This work was supported by Natural Sciences and Engineering Research Council Discovery Grant 498390 and Canadian Foundation for Innovation 32896. We thank Michael Mack and Heeyoung Choo for helpful comments on an earlier version of this manuscript.

The authors declare no competing financial interests.

Correspondence should be addressed to Dr. Yaelan Jung, 100 St. George Street, Toronto, Ontario M5S 3G3, Canada. E-mail: yaelan.jung@mail.utoronto.ca.

DOI:10.1523/JNEUROSCI.0272-18.2018

 0270-6474/18/385969-13$15.00/0

a specific scene into real-world topography in the retrosplenial cortex (RSC) and hippocampus (Morgan et al., 2011). Does this abstraction continue beyond the visual domain? To identify representations of scene content beyond this visual processing hierarchy, we here investigate neural activation patterns elicited by visual and auditory scene stimuli.

Previous work has identified neural representations that are not confined to a particular sense. Several brain areas have been shown to integrate signals from more than one sense (Calvert, 2001; Driver and Noesselt, 2008), such as posterior superior temporal sulcus (STS) (Beauchamp et al., 2004), the posterior parietal cortex (Cohen and Anderson, 2004; Molholm et al., 2006; Sereno and Huang, 2006), and the prefrontal cortex (PFC) (Sugihara et al., 2006; Romanski, 2007). Some of these areas show similar neural activity patterns when the same information is delivered from different senses for various stimuli, such as objects (Man et al., 2012), emotions (Müller et al., 2012), or face/voice identities (Park et al., 2010). Despite these observations, little is known about how scene information is processed beyond the sensory modality domain.

In real-world settings, our perception of scenes typically relies on multiple senses. Therefore, we postulate that there should exist a stage of modality-independent representation of scenes, which generalizes information across different modality channels. We hypothesize that PFC may play a role in representing scene categories beyond the modality domain based on previous research showing that PFC shows categorical representations of visual information (Freedman et al., 2001; Walther et al., 2009).

The present study investigates modality-independent scene representations using multivoxel pattern analysis (MVPA) of fMRI data. Two different types of MVPA were performed to define modality-independent representations in the brain. First, we identified brain areas that process both visual and auditory scene information by decoding neural representations of scene categories elicited by scene images and sounds separately. Second, we tested whether these areas represent visual and auditory information with similar neural codes by performing cross-decoding analysis between the two modalities.

After identifying modality-independent representations of scene categories in the brain, we further explored the characteristics of these representations with two additional types of analysis. We first examined whether the neural representations of scene categories in one modality are degraded by conflicting information from the other modality. Second, we tested to what extent scene category representations are related to human behavior and to the physical structure of the stimuli by comparing error patterns. Among the multisensory brain regions that we investigated, only the regions in PFC contain modality-independent representations of scene categories showing both visual and auditory scene representations in similar neural activity patterns.

Materials and Methods

We posit four idealized models of how visual and auditory information can be processed within a brain region: a purely visual model, a purely auditory model, a multimodal model with separate but intermixed neural populations for processing visual and auditory information, and a cross-modal model with neural populations for representing scene category information regardless of sensory modalities (Fig. 1*C*). Experimental conditions and analysis protocols were designed to discriminate between these models (Fig. 1*A*,*B*).

Figure 1*D* shows predicted results for each of the four models. Specifically, we expect that primarily visual and auditory regions will contain neurons dedicated to processing their respective modality exclusively. In these regions, scene categories should be decodable from the corresponding modality condition only, but not across modalities. In multimodal regions, both visual and auditory information should be processed in anatomically collocated but functionally separate neural populations. Therefore, we expect that both image and sound categories can be decoded, but decoding across modalities should not be possible. In cross-modal regions, both image and sound categories should be decodable, and scene category decoding should generalize across modalities.

Participants

Thirteen subjects (18–25 years old; 6 females, 7 males) participated in the fMRI experiment. All participants were in good health with no past history of psychiatric or neurological disorders and reported having normal hearing and normal or corrected-to-normal vision. They gave written informed consent before the experiment began according to the Institutional Review Board of the Ohio State University.

A separate group of 25 undergraduate students from the University of Toronto (18–21 years old; 16 females, 9 males) participated in the behavioral experiment for course credit. All participants had normal hearing and normal or corrected-to-normal vision and gave written informed consent. The experiment was approved by the Research Ethics Board of the University of Toronto.

Stimuli

In the fMRI experiment, 640 color photographs of four scene categories (beaches, forests, cities, and offices) were used. The images have previously been rated as the best exemplars of their categories from a database of ~4000 images that were downloaded from the internet (Torralbo et al., 2013). Images were presented at a resolution of 800 × 600 pixels using a Christie DS+6K-M projector operating at a refresh rate 60 Hz. Images subtended ~21 × 17 degrees of visual angle.

Sixty-four sound clips representing the same four scene categories (beaches, forests, cities, or offices) were used as auditory stimuli. The sound clips were purchased from various commercial sound libraries and edited to 15 s in length. They include auditory scenes from real-world environments (i.e., sounds of waves, seagulls for a beach scene; the sound of office machines and indistinct murmur from human conversations for an office scene). Because of this relatively longer presentation time for each audio exemplar, fewer exemplars were used compared with those in the image condition. Perceived loudness was equated using Replay Gain as implemented in Audacity software (version 2.1.0). In a pilot experiment, the sound clips were correctly identified and rated as highly typical for their categories by 14 naive subjects. Both visual and auditory stimuli are available at the Open Science Framework repository (OSF): DOI https://doi.org/10.17605/OSF.IO/HWXQV.

The same visual and the auditory stimuli were used in the behavioral experiment. In the visual part of the experiment, 400 images were used for practice blocks (key-category association and staircasing), and the other 240 images were used in the main testing blocks. Images were presented on a CRT screen at a resolution of 800 × 600 pixels and subtended ~29 × 22 degrees of visual angle. The resolution of the monitor was 1024 × 768 with a refresh rate at 150 Hz. Images were followed by a perceptual mask, which was generated by synthesizing a mixture of textures reflecting all four scene categories (Portilla and Simoncelli, 2000).

Procedure and experimental design

fMRI experiment. The fMRI experiment consisted of three conditions: the image condition, the sound condition, and the mixed condition, in which both images and sounds were presented concurrently. Participants' brains were scanned during 12 experimental runs, four runs for each condition. Each run started with the instruction asking participants to attend, for the duration of the run, to either images (image runs and half of the mixed runs) or sounds (sound runs and the other half of the mixed runs). In the analysis, we combined the data across the two attention manipulation conditions because this attention manipulation was not the main purpose of the present study, and it did not influence the decoding accuracy in most of the ROIs (except lateral occipital complex [LOC], middle temporal gyrus [MTG], and superior parietal gyrus [SPG] for visual categories).

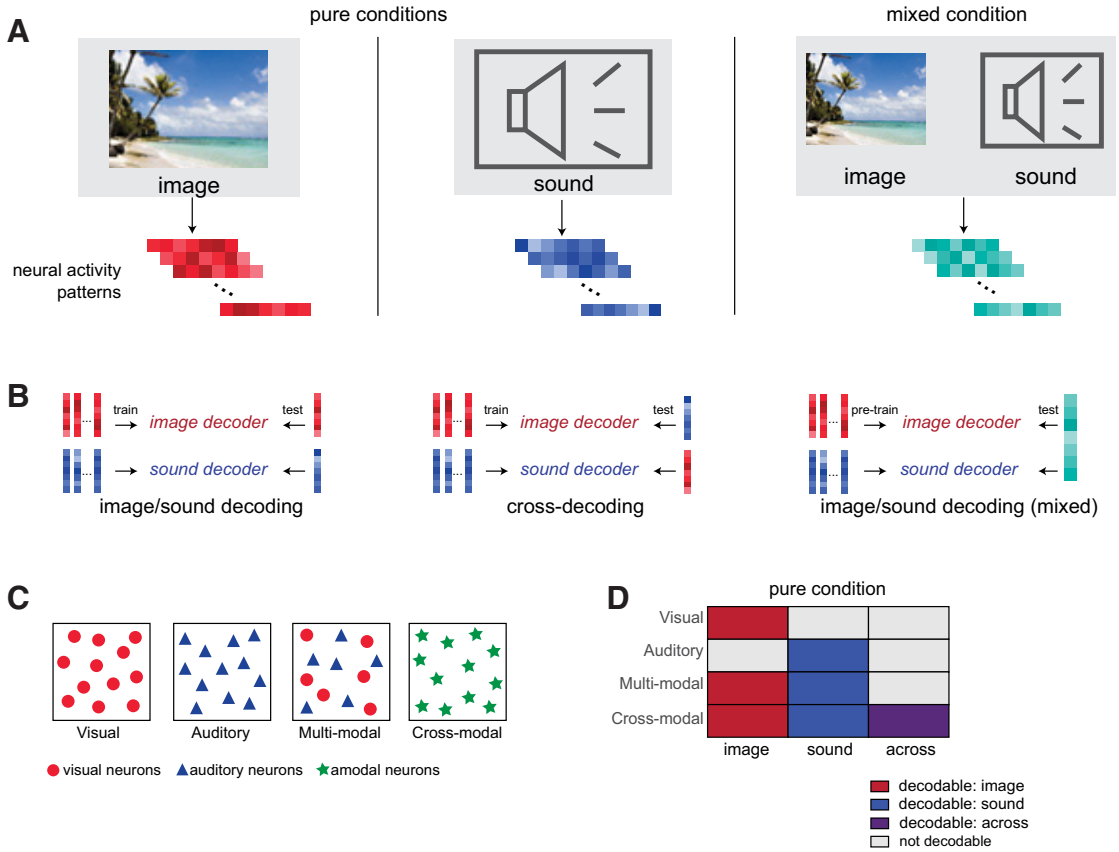

Figure 1. ***A***, Illustration of the image, sound, and mixed conditions. In the pure image and sound conditions, either images or sounds from four different scene categories were presented while neural activity patterns of participants were recorded. In the mixed condition, both images and sounds were presented simultaneously, but they were always from different categories (i.e., a beach images with city sounds). ***B***, Multivariate analysis of fMRI data for decoding image and sound categories, cross-decoding between images and sounds, and decoding images and sounds from the mixed condition. ***C***, Models for separate brain areas dedicated to visual, auditory, multimodal, and cross-modal processing. ***D***, Predictions for decoding performance of each model in different conditions: full color represents decodable; gray represents not decodable.

Runs contained eight blocks, two for each scene category, interleaved with 12.5 s fixation periods to allow for the hemodynamic response to return to baseline levels. The beginning and the end of a run also included a fixation period of 12.5 s. The order of blocks within runs and the order of runs were counterbalanced across participants. Mixed runs were only presented after at least two pure image and sound runs. Stimuli were arranged into eight blocks of 15 s duration. During image blocks, participants were shown 10 color photographs of the same scene category for 1.5 s each. During sound blocks, they were shown a blank screen with a fixation cross and a 15 s sound clip was played using Sensimetrics S14 MR-compatible in-ear noise-canceling headphones at ~70 dB. During mixed blocks, participants were shown images and played a sound clip of a different scene category at the same time. A fixation cross was presented throughout each block, and subjects were instructed to maintain fixation. Each run lasted 3 min 52.5 s.

fMRI data acquisition and preprocessing. Imaging data were recorded on a 3 Tesla Siemens MAGNETOM Trio MRI scanner with a 12-channel head coil at the Center for Cognitive and Behavioral Brain Imaging at Ohio State University. High-resolution anatomical images were acquired with a 3D-MPRAGE sequence with sagittal slices covering the whole brain; inversion time = 930 ms, TR = 1900 ms, TE = 4.68 ms, flip angle = 9°, voxel size = 1 × 1 × 1 mm, matrix size = 224 × 256 × 160 mm. Functional images for the main experiment were recorded with a gradient echo, EPI sequence with a volume TR of 2.5 s, a TE of 28 ms, and a flip angle of 78 degrees; 48 axial slices with 3 mm thickness were recorded without gap, resulting in an isotropic voxel size of 3 × 3 × 3 mm.

fMRI data were motion corrected to one EPI image (the 72nd volume of the 10th run), followed by spatial smoothing with a Gaussian kernel with 2 mm FWHM and temporal filtering with a high-pass filter at 1/400 Hz. Data were normalized to percentage signal change by subtracting the mean of the first fixation period in each run and dividing by the mean across all runs. The effects of head motion (6 motion parameters) and scanner drift (second degree polynomial) were regressed out using a GLM. The residuals of this GLM analysis were averaged over the duration of individual blocks, resulting in 96 brain volumes that were used as input for MVPA. Preprocessing was performed using AFNI (Cox, 1996).

Behavioral experiment. The behavioral experiment consisted of two parts: a visual and an auditory part. The order of the two parts was randomized for each participant.

The visual part consisted of two phases: practice and testing. Participants performed two types of practice: one for the key-category association and the other for the fast image presentation. During the first block of practice, photographs of natural scenes were presented for 200 ms (stimulus onset asynchrony [SOA]), immediately followed by a perceptual mask for 500 ms. Participants were asked to press one of four keys ('a,' 's,' 'k,' 'l') within 2 s of stimulus onset. They received acoustic feedback (a short beep) when they made an error. Assignment of the keys to the four scene categories (beaches, forests, cities, offices) was randomized for each participant. After participants achieved 90% accuracy in this key practice phase, they completed four additional practice blocks (40 trials each), during which the SOA was linearly decreased to 26.7 ms. The main testing phase consisted of six blocks (40 trials each) of four alternative-forced-choice (4AFC) scene categorization with a fixed SOA of 26.7 ms and without feedback.

In the auditory part, participants listened to scene soundscapes of 15 s in length. They indicated their categorization decision by pressing one of four keys (same key assignment as in the visual part). To make the task difficulty comparable with the image categorization task, we overlaid pure-tone noise onto the original sound clips. Noise consisted of 30 ms snippets of pure tones, whose frequency was randomly chosen between

50 and 2000 Hz with 3 ms of fade-in and fade-out (linear ramp). Based on a pilot experiment, we set the relative volume of the noise stimulus to 9 times the volume of the scene sounds. To familiarize participants with the task, they first performed 4 practice trials. Participants were asked to respond with the key corresponding to the correct category, starting from 8 s after the onset of the sound and without an upper time limit. Participants were encouraged not to deliberate on the answer but to respond as quickly and as accurately as possible.

Data analysis and statistical analysis

Defining ROIs. ROIs in visual cortex were defined using functional localizer scans, which were performed at the end of the same session as the main experiment. Retinotopic areas in early visual cortex were identified using the meridian stimulation method (Kastner et al., 1998). The vertical and horizontal meridians of the visual field were stimulated alternatingly with wedges with flickering checkerboard pattern. Boundaries between visual areas were outlined on the computationally flattened cortical surface. The boundary between V1 and V2 was identified as the first vertical meridian activity, the boundary between V2 and V3 as the first horizontal meridian, and the boundary between V3 and V4 (lower bank of the calcarine fissure only) as the second vertical meridian. To establish the anterior boundary of V4, we stimulated the upper and lower visual field in alternation with flickering checkerboard patterns. The anterior boundary of V4 was found by ensuring that both upper and lower visual field are represented in V4. Participants maintained central fixation during the localizer scan.

To define high-level visual areas, we presented participants with blocks of images of faces, scenes, objects, and scrambled images of objects. fMRI data from this localizer were preprocessed the same way as the main experiment data, but spatially smoothed with a 4 mm FWHM Gaussian filter. Data were further processed using a GLM (3dDeconvolve in Afni) with regressors for all four image types. ROIs were defined as contiguous clusters of voxels with significant contrast ($q < 0.05$, corrected using false discovery rate [FDR]) of the following: scenes > (faces and objects) for PPA, RSC (Epstein and Kanwisher, 1998), and OPA (Dilks et al., 2013); and objects > (scrambled objects) for LOC (Malach et al., 1995). Voxels that could not be uniquely assigned to one of the functional ROIs were excluded from the analysis.

Anatomically defined ROIs were extracted using a probability atlas in AFNI (DD Desai MPM) (Destrieux et al., 2010): MTG, superior temporal gyrus (STG), STS, angular gyrus (AG), SPG, intraparietal sulcus (IPS), medial frontal gyrus (MFG), superior frontal gyrus (SFG), and inferior frontal gyrus (IFG) with pars opercularis, pars orbitalis, and pars triangularis. Anatomical masks for auditory cortex (ACX) and its subdivisions (planum temporale [PT], posteromedial Heschl's gyrus, middle Heschl's gyrus, anterolateral Heschl's gyrus, and planum polare) were made available by Sam Norman-Haignere (Norman-Haignere et al., 2013). After nonlinear alignment of each participants' anatomical image to MNI space using AFNI's 3dQwarp function, the inverse of the alignment was used to project anatomical ROI masks back into original subject space using 3dNwarpApply. All decoding analyses, including for the anatomically defined ROIs, were performed in original subject space.

MVPA. For each participant, we trained a linear support vector machine (SVM; using LIBSVM) (Chang and Lin, 2011) to assign the correct scene category labels to the voxel activations inside an ROI based on the fMRI data from all runs except one. The SVM decoder then produced predictions for the labels of the data in the left-out run. This leave-one-run-out (LORO) cross-validation procedure was repeated with each run being left out in turn, thus producing predicted scene category labels for all runs. Decoding accuracy was assessed as the fraction of blocks with correct category labels. Group-level statistics was computed over all 13 participants using one-tailed *t* tests, determining whether decoding accuracy was significantly above chance level (0.25). Significance of the *t* test was adjusted for multiple comparisons using FDR (Westfall and Young, 1993).

To curb overfitting of the classifier to the training data, we reduced the dimensionality of the neural data by selecting a subset of voxels in each ROI. Voxel selection was performed by ranking voxels in the training data according to the *F* statistics of a one-way ANOVA of each voxel's activity with scene category as the main factor (Pereira et al., 2009). We determined the optimal number of voxels by performing a separate LORO cross-validation within the training data. For pure image and sound conditions, the training data of each cross-validation fold were used (nested cross-validation). In the cross-decoding and the mixed condition, the entire training data were used for voxel selection because training and test data were completely separate in these conditions. Using the training data, we performed LORO cross-validation analyses with the number of selected voxels varying from 100 to 1000 (step size of 100). We included voxels according to decreasing rank order of their *F* statistics. We compared the decoding performance across different voxel numbers and determined the optimal number of voxels, which showed the best decoding performance within the training data. Once we decided the optimal number of voxels, we trained the classifier using the entire training set, selecting the optimal number of voxels. Optimal voxel numbers varied by ROI and participant, showing an overall average of 107.125 across all ROIs and participants.

Error correlations. Category label predictions of the decoder were recorded in a confusion matrix, whose rows indicate the category of the stimulus, and whose columns represent the category predictions by the decoder. Diagonal elements indicate correct predictions, and off-diagonal elements represent decoding errors. Neural representations of scene categories were compared with human behavior by correlating the error patterns (the off-diagonal elements of the confusion matrices) between neural decoding and behavioral responses (Walther et al., 2012). Statistical significance of the correlations was established nonparametrically against the null distribution of all error correlations that were obtained by jointly permuting the rows and columns of one of the confusion matrices in question (24 possible permutations of four labels). Error correlations were considered as significant when none of the correlations in the null set exceeded the correlation for the correct ordering of category labels ($p < 0.0417$).

To assess the similarity between neural representations and the physical characteristics of the stimuli, we constructed simple computational models of scene categorization based on low-level stimulus features. Scene images were filtered with a bank of Gabor filters with four different orientations at four scales, averaged in a 3×3 grid. Images were categorized based on the resulting 144-dimensional feature vector in a 16-fold cross-validation, using a linear SVM, resulting in a classification accuracy of 85.8% (chance: 25%).

Physical properties of the sounds were assessed using a cochleagram, which mimics the biomechanics of the human ear (Meddis et al., 1990; Wang and Brown, 2006). The cochleagrams of individual sound clips were integrated over their duration and subsampled to 128 frequency bands, resulting in a biomechanically realistic frequency spectrum. The activation of the frequency bands was used as input to a linear SVM, which predicted scene categories of sounds in a 16-fold cross-validation. The classifier accurately categorized 57.8% of the scene sounds (chance: 25%). Error patterns from the computational analyses of images and sounds were correlated with those obtained from the neural decoder.

Error patterns of human observers were obtained from the behavioral experiment. Mean accuracy was 76.6% for the visual task (SD 12.35%, mean reaction time = 885.6 ms) and 78.1% for the auditory task (SD 6.86%, mean reaction time = 7.84 s). Behavioral errors were recorded in confusion matrices, separately for images and sounds. Rows of the confusion matrix indicate the true category of the stimulus, and columns indicate participants' responses. Individual cells contain the relative frequency of the responses indicated by the columns to stimuli indicated by the rows. Group-mean confusion matrices were compared with confusion matrices derived from neural decoding.

Searchlight analysis. To explore representations of scene categories outside of predefined ROIs, we performed a searchlight analysis. We defined a cubic "searchlight" of $7 \times 7 \times 7$ voxels ($21 \times 21 \times 21$ mm). The searchlight was centered on each voxel in turn (Kriegeskorte et al., 2006), and LORO cross-validation analysis was performed within each searchlight location using a linear SVM classifier (CosmoMVPA Toolbox) (Oosterhof et al., 2016). Decoding accuracy as well as the full confusion matrix at a given searchlight location were assigned to the central voxel.

For the group analysis, we first coregistered each participant's anatomical brain to the MNI 152 template using a diffeomorphic transformation as calculated by AFNI's 3dQWarp. We then used the same transformation parameters to register individual decoding accuracy maps to MNI space using 3dNWarpApply, followed by spatial smoothing with a 4 mm FWHM Gaussian filter. To identify voxels with decodable categorical information at the group level, we performed one-tailed t tests to test whether decoding accuracy at each searchlight location was above chance (0.25). After thresholding at $p < 0.01$ (one-tailed) from the t test, we conducted a cluster-level correction for multiple comparisons. We used AFNI's 3dClustSim to conduct α probability simulations for each participant. The estimated smoothness parameters computed by 3dFWHMx were used to conduct the cluster simulation. In the simulations, a p value of 0.01 was used to threshold the simulated data before clustering and a corrected α of 0.001 was used to determine the minimum cluster size. The average of the minimum cluster sizes across all the participants was 35 voxels.

Results

Decoding scene categories of images and sounds

To assess neural representations of scene categories from images and sounds, we performed MVPA for each ROI. A linear SVM using LIBSVM (Chang and Lin, 2011) was trained to associate neural activity patterns with category labels and then tested to determine whether a trained classifier can predict the scene category in a LORO cross-validation.

Figure 2 illustrates decoding accuracy in each condition for various brain areas (for the results of statistical tests, see Table 1). As shown in previous studies (Walther et al., 2009, 2011; Kravitz et al., 2011; Park et al., 2011; Choo and Walther, 2016), both early visual areas V1-V4 and high-level visual areas, including the PPA, RSC, OPA, and LOC, show category-specific scene representations. We were also able to decode the scene categories from activity elicited by sounds of the respective natural environments in auditory cortex (ACX) as well as its anatomical subdivisions (Figs. 2, 3).

Unlike previous reports showing that auditory content can be decoded from early visual cortex (Vetter et al., 2014; Paton et al., 2016), we did not find representations of auditory scene information in V1-V4. However, we were able to decode auditory scene categories in higher visual areas: the OPA (30.5%, $t_{(12)} = 1.966$, $q = 0.036$, $d = 1.36$) and the RSC (31.3%, $t_{(12)} = 1.803$, $q = 0.048$, $d = 0.5$). Intriguingly, we could also decode scene categories from images in ACX with a decoding accuracy of 29.8% ($t_{(12)} = 1.910$, $q = 1.91$, $d = 0.53$).

Having found modality-specific representations of scene categories in visual and auditory cortices, we aimed to identify scene representations in areas that are not limited to a specific sensory modality. We could decode categories of both visual and auditory scenes in several temporal and parietal regions (Fig. 2): the MTG, the STG, the SPG, and the AG. In the STS, we could decode scene categories only from images, not from sounds. Although previous studies have suggested that the IPS is involved in audiovisual processing (Calvert et al., 2001), we could not decode visual or auditory scene categories in the IPS.

Next, we examined whether PFC showed category-specific representations for both visual and auditory scene information. Previous studies have found strong hemispheric specialization in PFC (Gaffan and Harrison, 1991; Slotnick and Moo, 2006; Goel et al., 2007). We therefore analyzed functional activity in PFC separately by hemisphere. We were able to decode visual scene categories significantly above chance from the left IFG, pars opercularis, the right IFG, pars triangularis, and in both hemispheres from the MFG and the SFG (Figs. 2, 3; Table 1). The categories of

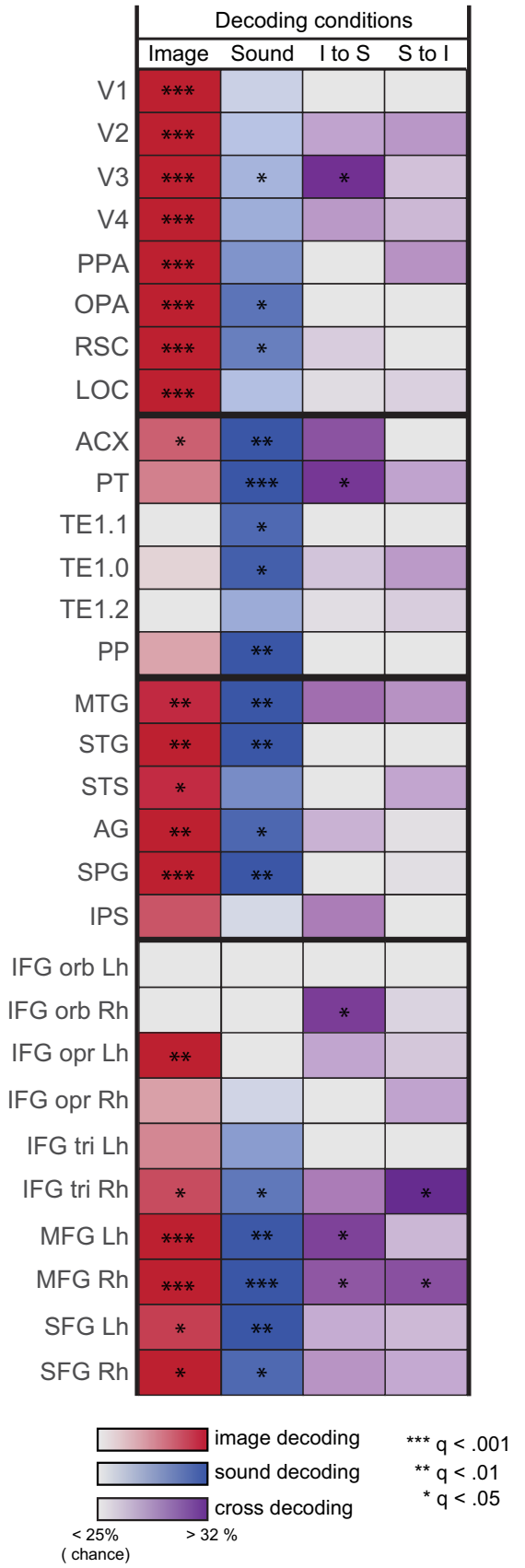

Figure 2. Decoding accuracy in each ROI (rows) for each condition (columns) are illustrated with the degree of color saturation. Red represents accuracy for decoding scene categories from images. Blue represents accuracy for decoding scene categories from sounds. Purple represents accuracy for cross-decoding. Significance of the one-sample t tests (one-tailed) was adjusted for multiple comparisons using FDR. *$q < 0.05$, **$q < 0.01$, ***$q < 0.001$.

Table 1. Statistical results of decoding performance in each ROI for each type of decoding analysis[a]

	Pure condition				Mixed condition	
ROI	Image decoding	Sound decoding	Image to sound	Sound to image	Image decoding	Sound decoding
V1	4.90, 0.001, 1.36	0.41, 0.345, 0.11	−0.48, 0.679, 0.13	−0.12, 0.679, 0.03	9.28, <0.001, 2.57	−0.13, 0.551, 0.04
V2	8.34, <0.001, 2.31	0.62, 0.272, 0.17	0.87, 0.200, 0.24	1.01, 0.200, 0.28	6.65, <0.001, 1.85	−0.64, 0.733, 0.18
V3	3.43, 0.005, 0.95	0.90, 0.193, 0.25	2.67, 0.020, 0.74	0.46, 0.326, 0.13	6.34, <0.001, 1.76	−1.43, 0.910, 0.40
V4	4.22, 0.001, 1.17	1.01, 0.166, 0.28	1.00, 0.284, 0.28	0.59, 0.284, 0.16	9.15, <0.001, 2.54	2.25, 0.022, 0.62
PPA	7.26, <0.001, 2.01	1.46, 0.085, 0.40	−0.33, 0.627, 0.09	1.10, 0.294, 0.30	9.51, <0.001, 2.64	0.87, 0.200, 0.24
OPA	8.10, <0.001, 2.25	1.97, 0.036, 0.55	−0.30, 0.617, 0.08	−0.13, 0.617, 0.04	9.29, <0.001, 2.58	0.00, 0.500, 0.00
RSC	5.77, <0.001, 1.60	1.80, 0.048, 0.50	0.34, 0.704, 0.10	−0.55, 0.704, 0.15	6.35, <0.001, 1.76	1.80, 0.048, 0.50
LOC	8.58, <0.001, 2.38	0.70, 0.250, 0.19	0.12, 0.454, 0.03	0.28, 0.454, 0.08	9.98, <0.001, 2.77	−0.12, 0.545, 0.03
ACX	1.91, 0.040, 0.53	4.17, 0.001, 1.16	1.97, 0.072, 0.55	−0.49, 0.682, 0.13	−3.00, 0.994, 0.83	4.80, <0.001, 1.33
ACX PT	1.45, 0.087, 0.40	4.74, 0.000, 1.31	2.53, 0.026, 0.70	0.88, 0.199, 0.24	−0.88, 0.801, 0.24	5.30, <0.001, 1.47
ACX TE1.1	0.00, 0.500, 0.00	2.23, 0.045, 0.62	−1.41, 0.908, 0.39	−0.38, 0.908, 0.11	−0.45, 0.670, 0.12	3.30, 0.006, 0.92
ACX TE1.0	0.26, 0.399, 0.07	2.57, 0.024, 0.71	0.42, 0.340, 0.12	0.98, 0.340, 0.27	0.55, 0.296, 0.15	7.94, <0.001, 2.20
ACX TE1.2	−2.86, 0.993, 0.79	1.06, 0.310, 0.29	0.12, 0.455, 0.03	0.33, 0.455, 0.09	−0.48, 0.679, 0.13	3.26, 0.007, 0.91
ACX PP	0.95, 0.181, 0.26	3.08, 0.010, 0.85	−0.15, 0.614, 0.04	−0.30, 0.614, 0.08	−0.61, 0.724, 0.17	4.16, 0.001, 1.15
MTG	2.69, 0.010, 0.75	3.28, 0.007, 0.91	1.56, 0.146, 0.43	1.09, 0.148, 0.30	0.96, 0.357, 0.27	0.37, 0.358, 0.10
STG	3.35, 0.003, 0.93	4.26, 0.001, 1.18	−0.72, 0.756, 0.20	−0.43, 0.756, 0.12	−0.52, 0.693, 0.14	3.36, 0.006, 0.93
STS	2.67, 0.020, 0.74	1.59, 0.069, 0.44	0.00, 0.500, 0.00	0.84, 0.418, 0.23	2.62, 0.022, 0.73	1.47, 0.084, 0.41
AG	4.18, 0.001, 1.16	2.28, 0.021, 0.63	0.67, 0.464, 0.19	0.09, 0.464, 0.03	2.38, 0.035, 0.66	1.58, 0.070, 0.44
SPG	4.81, <0.001, 1.33	2.96, 0.006, 0.82	−0.21, 0.581, 0.06	0.10, 0.581, 0.03	8.72, 0.000, 2.42	−1.26, 0.885, 0.35
IPS	2.07, 0.060, 0.58	0.25, 0.405, 0.07	1.37, 0.196, 0.38	−0.72, 0.756, 0.20	1.17, 0.264, 0.32	0.53, 0.303, 0.15
IFG pars orbitalis LH	−1.15, 0.864, 0.32	−0.15, 0.864, 0.04	−0.11, 0.935, 0.03	−1.63, 0.935, 0.45	−0.24, 0.595, 0.07	1.11, 0.287, 0.31
IFG pars orbitalis RH	−2.85, 0.993, 0.79	−0.10, 0.993, 0.03	2.41, 0.033, 0.67	0.27, 0.397, 0.07	1.94, 0.054, 0.54	1.74, 0.054, 0.48
IFG pars opercularis LH	3.64, 0.003, 1.01	−0.06, 0.525, 0.02	0.85, 0.348, 0.24	0.40, 0.348, 0.11	1.50, 0.086, 0.41	1.45, 0.086, 0.40
IFG pars opercularis RH	1.00, 0.337, 0/28	0.33, 0.375, 0.09	0.00, 0.500, 0.00	0.88, 0.398, 0.24	0.37, 0.721, 0.10	−0.89, 0.804, 0.25
IFG pars triangularis LH	1.35, 0.108, 0.37	1.30, 0.108, 0.36	−0.69, 0.749, 0.19	−0.63, 0.749, 0.17	0.33, 0.375, 0.09	1.28, 0.223, 0.36
IFG pars triangularis RH	2.19, 0.040, 0.61	1.91, 0.040, 0.53	1.38, 0.096, 0.38	2.96, 0.012, 0.82	0.58, 0.285, 0.16	1.72, 0.112, 0.48
MFG LH	5.01, <0.001, 1.39	2.88, 0.007, 0.80	2.29, 0.041, 0.63	0.61, 0.275, 0.17	0.98, 0.349, 0.27	−0.09, 0.537, 0.03
MFG RH	5.87, <0.001, 1.69	4.30, 0.001, 1.19	1.89, 0.042, 0.52	2.02, 0.042, 0.56	3.08, 0.010, 0.85	1.33, 0.104, 0.37
SFG LH	2.39, 0.017, 0.66	3.49, 0.004, 0.97	0.77, 0.288, 0.21	0.57, 0.288, 0.16	1.30, 0.219, 0.36	0.69, 0.252, 0.19
SFG RH	2.98, 0.011, 0.83	2.22, 0.023, 0.62	1.06, 0.223, 0.29	0.79, 0.223, 0.22	1.82, 0.094, 0.50	−2.74, 0.991, 0.76

[a] *t* value (12 degrees of freedom), *q* value (*p* corrected for multiple comparison), Significant result ($q < 0.05$), Cohen's *d*. One-sample *t* test (one-tailed) was performed to test whether decoding accuracy was above chance (25%). Significance of the tests was adjusted using FDR.

scene sounds were decodable in the right IFG, pars triangularis, as well as the MFG and SFG in both hemispheres (Figs. 2, 3; Table 1).

Although the temporal, parietal, and PFC all showed both visual and auditory scene representations, this does not necessarily imply that these areas process scene information beyond the sensory modality domain. Neural representations of scene categories in cross-modal regions should not merely consist of coexisting populations of neurons with visually and auditorily triggered activation patterns (Fig. 1*C*, the multimodal model); the voxels in these ROIs should be activated equally by visual and auditory inputs if they represent the same category. In other words, if the neural activity pattern elicited by watching a picture of a forest reflects the scene category of forest, then this neural representation should be similar to that elicited by listening to forest sounds (Fig. 1*C*, the cross-modal model). We aimed to explicitly examine whether scene category information in the prefrontal areas transcends sensory modalities using cross-decoding analysis between the image and sound conditions.

Cross-modal decoding

For the cross-decoding analysis, we trained the decoder using the image labels from the image runs and then tested whether it could correctly predict the categories of scenes presented as sounds in the sound runs. We also performed the reverse analysis, training the decoder on the sound runs and testing it on the image runs.

Cross-decoding from images to sounds succeeded in the MFG in both hemispheres and in the right IFG, pars orbitalis. The right MFG and the right IFG, pars triangularis, showed significant decoding accuracy for cross-decoding from sounds to images (Figs. 2, 3). However, cross-decoding was not possible in either direction anywhere in sensory cortices or temporal and parietal cortices, which have significant decoding of both image and sound categories. Although V3 and the PT showed significant decoding in the cross-decoding analysis of images to sounds (Figs. 2, 3; Table 1), it is hard to interpret these findings as equivalent to those in prefrontal regions because these ROIs only show significant decoding of either image (V3) or sound categories (PT) in the straight decoding analysis. These results therefore suggest that only prefrontal areas contain modality-independent representations of scene categories with similar neural activity patterns from visual and auditory scene information.

Presenting images and sounds concurrently

We explored the characteristics of cross-modal regions using an interference condition, in which images and sounds from incongruous categories were simultaneously presented. If a population of neurons encodes a scene category independently of sensory modality, then we should see a degradation of the category representation in the presence of a competing signal from the other modality. If, on the other hand, two separate but intermixed populations of neurons encode the visual and auditory categories, respectively, then we should be able to still decode the category from at least one of the two senses.

To decode scene categories from this mixed condition, we created an image and a sound decoder by training separate classifiers with data from the image-only and the sound-only conditions. We then tested these decoders with neural activity patterns from the mixed condition, using either image or sound labels as ground truth. As the training and the test data are from separate sets of runs, cross-validation was not needed for this analysis.

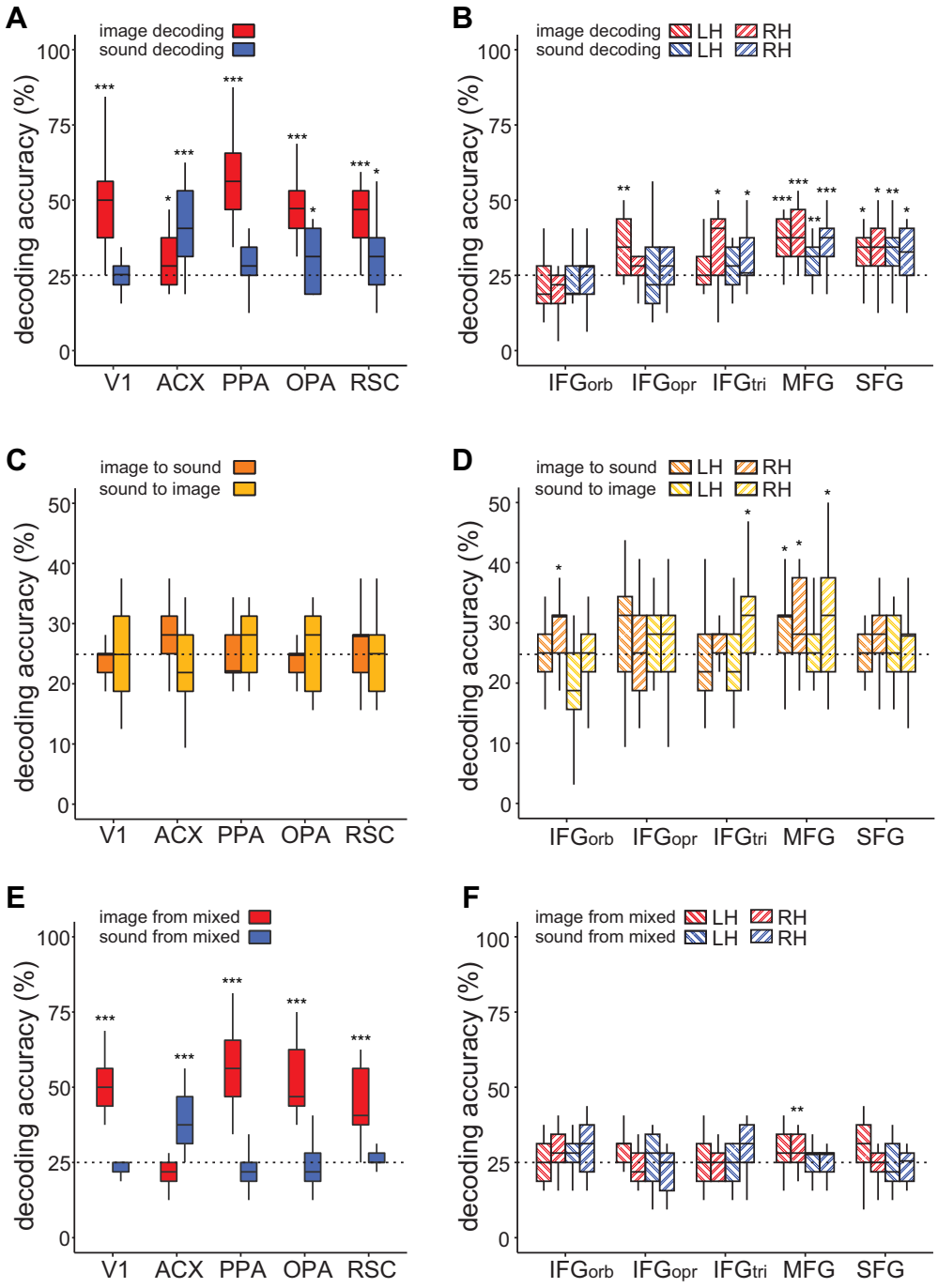

Figure 3. Decoding accuracy in representative ROIs (***A***, ***C***, ***E***: modality-specific areas; ***B***, ***D***, ***F***: prefrontal areas) for each condition (***A***, ***B***: image/sound condition; ***C***, ***D***: cross-modal decoding; ***E***, ***F***: decoding in the mixed condition). Each box spans the first quartile to the third quartile. Middle line indicates the median point. Whiskers above and below the box represent the location of the minimum and maximum. Significance of the one-sample *t* tests (one-tailed) versus chance (25%) was adjusted for multiple comparisons using FDR. $*q < 0.05$, $**q < 0.01$, $***q < 0.001$.

We were able to decode visual and auditory scene categories from the respective sensory brain areas, even in the presence of conflicting information from the other modality. In temporal and parietal ROIs, we could decode scene categories, but those ROIs were no longer multimodal; they only represented scene categories in either the visual or auditory domain but no longer both (Fig. 3*E*,*F*; Table 1). These findings suggest that these ROIs contain separate but intermixed neural populations for visual and auditory information. For ROIs in PFC, we found that conflicting audiovisual stimuli interfered heavily with representations of scene categories (Fig. 3*E*,*F*). Scene categories could no longer be decoded in PFC from either modality, except for visual scenes in the right MFG. Presumably, this breakdown of the decoding of scene categories is due to the conflicting information from the two sensory modalities arriving at the same cross-modal populations of neurons.

Error correlations

To further explore the characteristics of the neural representations of scene categories, we compared the patterns of errors from the neural decoders with those from human behavior as well as with the physical attributes of the stimuli. If representations of

scenes in a certain brain region are used directly for categorical decisions, then error patterns from this ROI should be similar to errors made in behavioral categorization (Walther et al., 2009). However, in early stages of neural processing, scene representations might reflect the physical properties of the scene images or sounds. In this case, the error patterns of the decoders should resemble the errors that a classifier would make solely based on low-level physical properties.

To assess similarity of representations, we correlated the patterns of errors (off-diagonal elements of the confusion matrices; see Materials and Methods) between the neural decoders, physical structure of the stimuli, and human behavior. Statistical significance of the correlations was established with nonparametric permutation tests. Here we considered error correlations to be significant when none of the correlations in the null set exceeded the correlation of the correct ordering of the categories ($p < 0.0417$).

Behavioral errors from image categorization were not correlated with the errors derived from image properties ($r = -0.458$, $p = 0.083$), suggesting that behavioral judgment of scene categories was not directly driven by low-level physical differences between the images. There was, however, a positive error correlation between the auditory task and physical properties of sounds ($r = 0.407$, $p < 0.0417$).

For the image condition, errors from neural decoders were similar to those from image structure in early visual cortex and similar to human behavioral errors in high-level visual areas (Fig. 4*A*); in early visual cortex, decoding errors were positively correlated with image structure (V1: $r = 0.746$, $p < 0.0417$; V2: $r = 0.451$, $p = 0.083$) but not with behavioral errors (V1: $r = -0.272$, $p = 0.929$; V2: $r = -0.132$, $p = 0.333$). Negative error correlations with image behavior in these areas are due to image behavior being negatively correlated with image structure. V3 and V4 showed no significant correlation with stimulus structure (V3: $r = -0.171$, $p = 0.292$; V4: $r = 0.076$, $p = 0.667$) or behavior (V3: $r = 0.356$, $p = 0.125$; V4: $r = 0.250$, $p = 0.125$). In high-level scene-selective areas, decoding errors were positively correlated with image behavior (PPA: $r = 0.570$, $p < 0.0417$; RSC: $r = 0.637$, $p < 0.0417$; not in OPA: $r = 0.183$, $p = 0.292$), but not with the error patterns representing image structure (PPA: $r = 0.0838$, $p = 0.333$; RSC: $r = -0.230$, $p = 0.292$; OPA: $r = 0.099$, $p = 0.458$).

Errors from the neural decoders in SPG were positively correlated with image behavior ($r = 0.404$, $p = < 0.0417$) but not with image structure ($r = 0.220$, $p = 0.167$). The errors from MTG, STS, and AG also showed high correlation with image behavior errors but did not reach significance in the permutation test (MTG: $r = 0.360$, $p = 0.083$; STS: $r = 0.348$, $p = 0.083$; AG: $r = 0.552$, $p = 0.083$; Fig. 4*A*). Finally, in PFC, errors from the image decoders in the right MFG and SFG show positive correlation with image behavior (right MFG: $r = 0.377$, $p < 0.0417$; right SFG, $r = 0.338$, $p < 0.0417$). The left hemisphere of these ROIs also showed positive correlations but not significantly (left MFG: $r = 0.309$, $p = 0.083$; left SFG: $r = 0.212$, $p = 0.208$). On the other hand, the left IFG, pars opercularis, and the right IFG, pars triangularis, showed no error correlation at all with either image behavior or structure (Fig. 4*A*).

In the sound condition, error patterns from sound structure as well as sound behavior were positively correlated with decoding errors from ACX, even though the permutation test did not reach significant level (with sound structure: $r = 0.438$, $p = 0.083$; with sound behavior: $r = 0.46$; $p = 0.125$). Four of the five anatomical subdivisions of ACX showed positive correlation

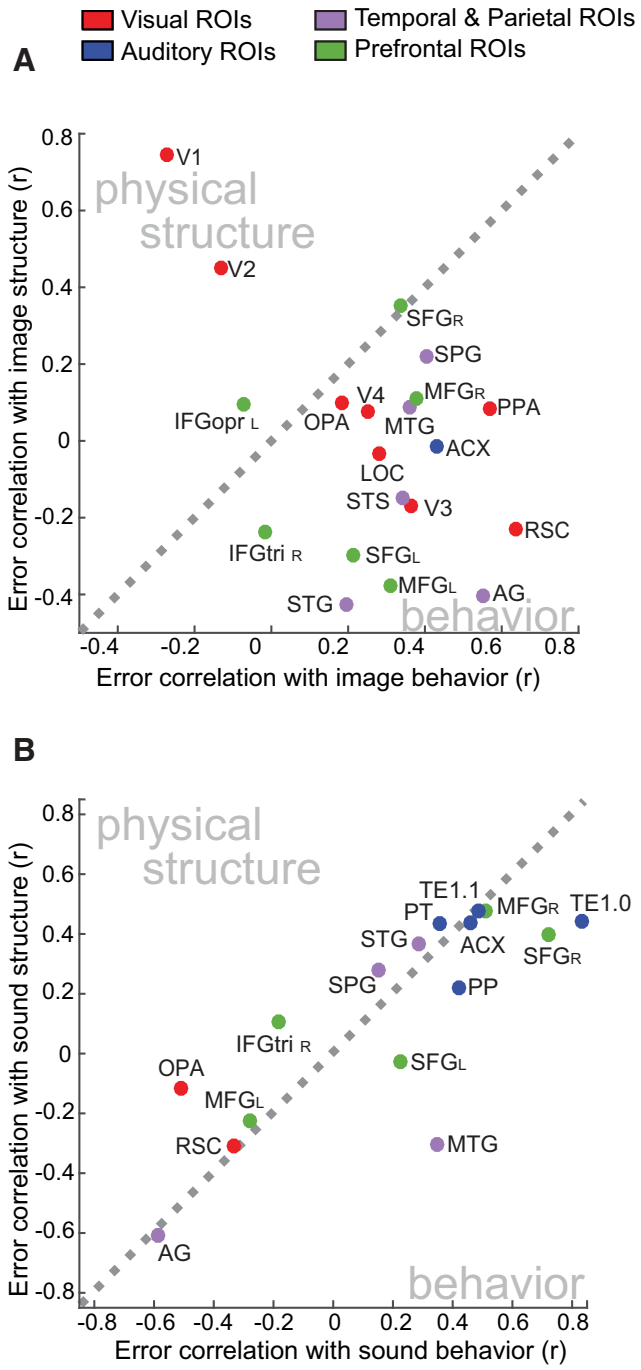

Figure 4. Error correlations of the neural decoder with behavior and stimuli structure in the image (***A***) and sound (***B***) conditions. Diagonal axes indicate the points where the error correlations with behavior and with stimuli structure are equal. In the image condition, we see a clear progression from V1 through V2–4 to higher-level visual areas (red), moving from strong error correlation with image structure to strong error correlation with visual behavior. Error patterns from decoding image categories from PFC (green) are most similar to visual behavior. In the sound condition, all ROIs are close to the diagonal because sound structure and sound behavior error patterns are significantly correlated with each other. We see ACX and its subdivisions (blue) high along the diagonal, indicating strong similarity with both sound structure and sound behavior. OPA and RSC (red) show low error correlation with either structure or behavior. Prefrontal ROIs (green) are distributed along the diagonal, with right MFG and right SFG showing high and left MFG showing low error correlations.

with sound structure (TE1.1: $r = 0.478$, $p < 0.0417$; TE1.0: $r = 0.443$, $p < 0.0417$; PT: $r = 0.419$, $p = 0.167$; PP: $r = 0.219$, $p = 0.0125$). These areas also showed positive error correlations with human behavior for sounds (TE1.1: $r = 0.50$, $p = 0.083$; TE1.0: $r =$

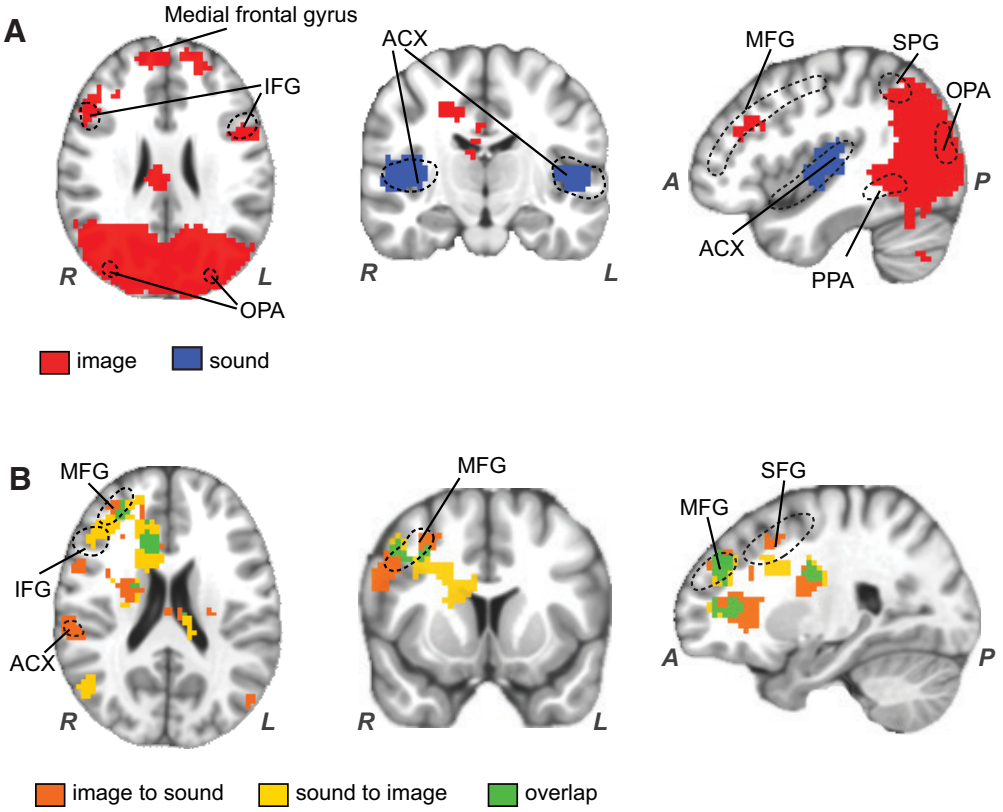

Figure 5. Neural decoding accuracy maps for image and sound categories (***A***) and for cross-decoding analysis (***B***) (thresholded at $p < 0.01$). A cube of 342 voxels ($7 \times 7 \times 7$ voxels) was used in the searchlight analysis. ***A***, Searchlight locations with above-chance decoding of images (red) and sounds (blue) (thresholded at $p < 0.01$ with a cluster-based multiple comparison correction). MNI coordinates of the sections shown: $x = -42, y = 15, z = 26$. ***B***, Searchlight locations with above-chance level cross-decoding accuracy for image to sound (orange), sound to image (yellow), and both (green) (thresholded at $p < 0.01$ with a cluster-based multiple comparison correction). MNI coordinates of the sections shown: $x = -24, y = -11, z = 21$.

0.83, $p < 0.0417$; PT: $r = 0.347$, $p = 0.167$; PP: $r = 0.419$, $p = 0.125$; Fig. 4*B*).

None of the temporal or parietal ROIs showed significant error correlations of decoding sounds with sound structure or behavior (Fig. 4*B*). In PFC, we found that errors from the right SFG positively correlate with both sound structure ($r = 0.397$, $p < 0.0417$) and sound behavior ($r = 0.721$, $p < 0.0417$). The right MFG also showed high correlations with sound behavior but not significantly (right MFG: $r = 0.486$, $p = 0.125$; Fig. 4*B*). In the left hemisphere, neither MFG nor SFG showed significant correlation with sound behavior (left MFG: $r = -0.279$, $p = 0.375$; left SFG: $r = 0.22$, $p = 0.208$).

Whole-brain analysis

To explore representations of scene categories beyond the predefined ROIs, we performed a whole-brain searchlight analysis with a size of $7 \times 7 \times 7$ voxels ($21 \times 21 \times 21$ mm) cubic searchlight. The same LORO cross-validation analysis for image and sound conditions as well as the same two cross-decoding analyses as for the ROI-based analysis were performed at each searchlight location using a linear SVM classifier, followed by a cluster-level correction for multiple comparisons. For each decoding condition, we found several spatial clusters with significant decoding accuracy. Some of these clusters confirmed the predefined ROIs; others revealed scene representations in unexpected regions beyond the ROIs.

For the decoding of image categories, we found a large cluster of 14,599 voxels with significant decoding accuracy for decoding scene images, spanning most of visual cortex. In accordance with the ROI-based analysis, we also found clusters in PFC, overlapping with the MFG in both hemispheres as well as the right SFG and the right IFG. For a complete list of clusters, see Figure 5*A* and Table 2. Decoding of sound categories produced a large cluster in each hemisphere, which overlapped with auditory cortices (Fig. 5*A*; Table 2).

Even though we were able to find ROIs that allowed for decoding of both images and sounds, we could not find any searchlight locations where this was possible. This may be due to spatial smoothing introduced by the spatial extent of the searchlight volume as well as the alignment to the standard brain.

We found several significant clusters in the right PFC that allowed for cross-decoding between images and sounds (Fig. 5*B*). The image-to-sound condition produced clusters with significant decoding accuracy in the right MFG, IFG, and SFG as well as right MFG, and right MTG. The sound-to-image condition resulted in clusters in the right MFG, SFG, and MTG (Fig. 5*B*; Table 3). We found four compact clusters that allowed for cross-decoding in both directions in the right PFC, overlapping with the right MFG and cingulate gyrus (102 voxels), right MFG (95 voxels; 22 voxels), and right SFG (80 voxels).

We compared error patterns from the neural decoders to stimulus properties and human behavior using the same searchlight cube of $7 \times 7 \times 7$ voxels ($21 \times 21 \times 21$ mm). In each searchlight location, the error patterns of the decoder were recorded in a confusion matrix for image and sound categories separately and compared with the error patterns of corresponding stimuli structure and behavior error patterns (for details, see Materials and Methods). In this error pattern analysis, we only

Table 2. Clusters identified in the searchlight analysis for decoding of image and sound scene categories (thresholded at $p < 0.01$)

	Peak (MNI coordinates)					
Decoding condition	*x*	*y*	*z*	Accuracy (%)	Volume (μl)	Description
Image	−10.4	84.9	10	58.36	386,329	Middle occipital gyrus, cuneus, precuneus, superior occipital gyrus, parhippocampal gyrus, MTG, inferior temporal gyrus, superior parietal lobule, right cingulate gyrus
	−13.4	−42.9	40	34.8	10,188	Right SFG, right MFG
	43.1	−4.3	31	36.25	3890	Left IFG, left MFG
	−49.1	−28.1	28	34.74	2355	Right MFG
	−1.5	−45.9	−2	33.49	1985	Right anterior cingulate
Sound	−43.1	10.6	7	37.61	9818	Right STG
	49.1	16.5	4	35.13	9315	Left STG

Table 3. Clusters identified in the searchlight analysis for cross-decoding of scene images and sounds (thresholded at $p < 0.01$)

	Peak (MNI coordinates)					
Decoding condition	*x*	*y*	*z*	Accuracy (%)	Volume (μl)	Description
Image to sound	−19.3	−34	7	35.68	16,010	Right SFG, right MFG, right MFG
	4.5	19.5	16	32.05	3255	Left thalamus, left lateral dorsal nucleus
	−31.2	−42.9	25	33.35	2884	Right MFG, right SFG
	19.3	55.2	−41	31.46	2673	Left cerebellar tonsil
	−25.3	7.6	28	32.25	2593	Right cingulate gyrus
	−52	−10.2	25	31.93	2567	Right IFG
	−49.1	61.1	1	32.23	2038	Right MTG,
	−58	31.4	22	32.48	1879	Right inferior parietal lobule
	49.1	73	19	31.79	1799	Left MTG
	−40.1	4.6	−29	32	979	Right fusiform gyrus, right MTG
	−37.2	61.1	−29	30.80	953	Right declive
Sound to image	−25.3	−42.9	31	34	29,056	Right SFG, right MFG
	10.4	19.5	25	31.66	1985	Left cingulate gyrus
	4.5	90.8	13	32.92	1747	Left middle occipital gyrus, left cuneus
	−40.1	40.3	−14	32.17	1455	Right fusiform gyrus, right parahippocampal gyrus
	−49.1	64.1	22	32.88	1455	Right MTG

included searchlight locations with significant decoding accuracy for the corresponding modality condition: the image condition in a comparison with image structure/behavior and the sound condition with sound structure/behavior (thresholded at $p < 0.01$; Fig. 5). Clusters in visual cortex, overlapping with V1-V4, showed significant error correlations with image properties (Fig. 6*A*; Table 4). Error patterns from searchlight locations in the SPG and the parahippocampal gyrus, overlapping with the PPA, correlated with errors from human behavior for image categorization. In general, we observed a posterior-to-anterior trend, with voxels in the posterior (low-level) visual regions more closely matched to stimulus properties and with voxels more anterior (high-level) visual regions more closely related to behavior.

Clusters in bilateral STG, overlapping with ACX, showed significant error correlations with sound properties and behavioral errors for sound categorization. Within this cluster, we see the same posterior-to-anterior trend, with posterior voxels being more closely related to sound properties and more anterior voxels being more closely related to behavioral categorization of scene sounds (Fig. 6*B*; Table 4).

Discussion

The present study investigated where and how scene information from different sensory domains forms modality-independent representations of scene categories. We have found that both visual and auditory stimuli of the natural environment elicit representations of scene categories in subregions of PFC. These neural representations of scene categories generalize across sensory modalities, suggesting that scene representations in PFC reflect scene categories not constrained to a specific sensory domain. To our knowledge, our study is the first to demonstrate a neural representation of scenes at such an abstract level.

Three distinct characteristics support the idea that neural representations of scene categories in PFC are distinct from those in modality-specific areas, such as the visual or the auditory cortices or other multisensory areas. First, both image and sound categories could be decoded from the same areas in PFC. Thus, it can be inferred that neural representations of scene categories in PFC are not limited to a specific sensory modality channel. Second, the representations in PFC could be cross-decoded from one modality to the other, showing that the category-specific neural activity patterns were similar across the sensory modalities. Third, when subjects were presented with incongruous visual and auditory scene information simultaneously, it was no longer possible to decode scene categories in PFC, whereas modality-specific areas as well as multimodal areas still carried the category-specific neural activity patterns. This result shows that inconsistent information entering through the two sensory channels in the mixed condition interferes, preventing the formation of scene categories in PFC.

Although scene categories could be decoded from both images and sounds in several ROIs in the temporal and parietal lobes, cross-decoding across sensory modalities was not possible there, suggesting that neural representations elicited by visual stimuli were not similar to those elicited by auditory stimuli. Further supporting the idea that visual and auditory representations are separate but intermixed in these regions, decoding of scene categories from the visual or auditory domain was still possible in the presence of a conflicting signal from the other domain. These findings suggest that, even though information from both visual

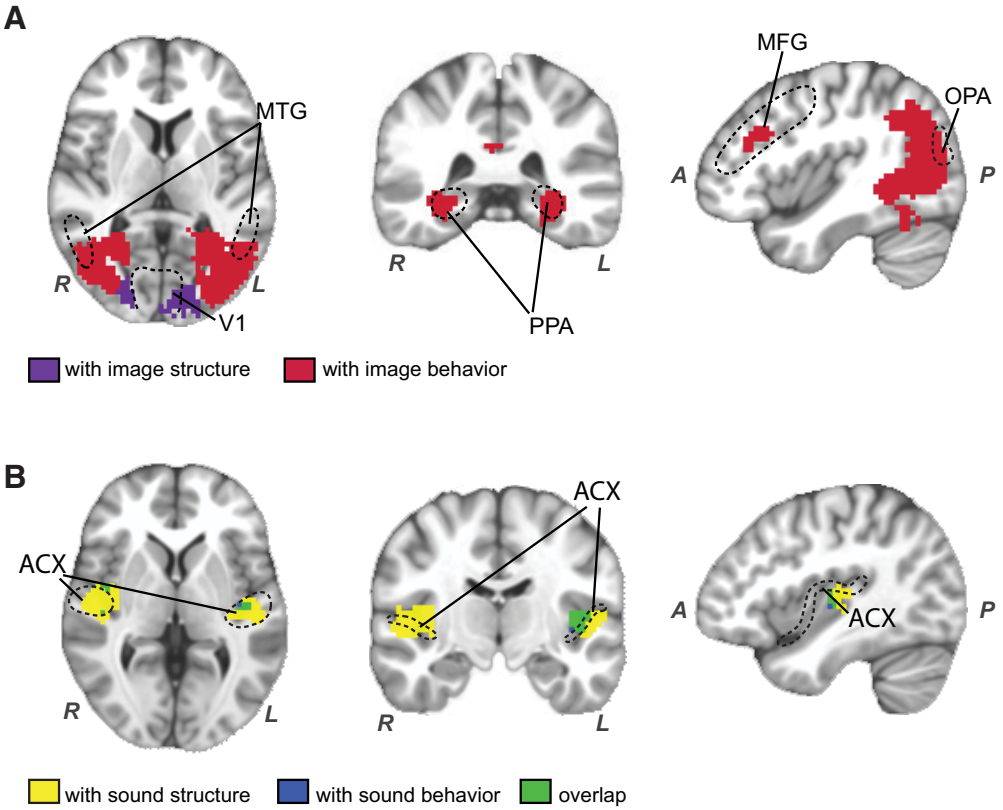

Figure 6. Searchlight maps for error correlations in the image (***A***) and sound (***B***) condition (thresholded at $p < 0.05$). A cube of 342 voxels ($7 \times 7 \times 7$ voxels) was used in the searchlight analysis, and only searchlight locations with significant decoding accuracy were included ($p < 0.01$; Fig. 5*A*). ***A***, Purple represents searchlight locations with significant error correlations with image structure. Magenta represents searchlight locations with significant error correlations with image behaviors. MNI coordinates of the sections shown: $x = -42, y = 37, z = 7$. ***B***, Yellow represents searchlight locations with significant error correlations with sound structure. Blue represents searchlight locations with significant error correlations with sound behavior. Green represents the overlap. MNI coordinates of the sections shown: $x = 46, y = 11, z = 5$.

Table 4. Clusters identified in the searchlight analysis for error correlations in the image and the sound conditions (thresholded at $p < 0.05$)

	Peak (MNI coordinates)						
Decoding condition	*x*	*y*	*z*	Accuracy (%)	Volume (μl)	Description	ROI with overlap
Image condition							
With image structure	19.3	87.9	4	0.706	11,776	Middle occipital gyrus, inferior occipital gyrus, lingual gyrus, cuneus	V1, V2, V3, V4
	−19.3	87.9	10	0.583	2038	Right middle occipital gyrus	
With image behavior	10.4	81.9	28	0.789	168,726	MTG, precuneus, parahippocampal gyrus, inferior temporal gyrus, middle occipital gyrus	PPA, OPA, RSC, LOC
	−49.1	−10.2	28	0.561	2223	Right IFG, right MFG	Right MFG
Sound condition							
With sound structure	−37.2	4.6	−5	0.543	7489	Right STG	Right ACX
	58	16.5	10	0.482	4631	Left STG	Left ACX
With sound behavior	40	13.5	4	0.603	941	Left STG	Left ACX

and auditory stimuli is present in these regions (Calvert et al., 2001; Beauchamp et al., 2004), scene information is computed separately for each sensory modality, unlike in PFC. The discrimination between multimodal and truly cross-modal representations is not possible with the univariate analysis techniques used in those previous studies.

Analysis of decoding errors demonstrated that the category representations in the visual areas have a hierarchical organization. In the early stage of processing, categorical representations are formed based on the physical properties of visual inputs, whereas in the later stage, the errors of neural decoders correlate with human behavior, confirming previous findings, which mainly focused on scene-selective areas (Walther et al., 2009, 2011; Choo and Walther, 2016). Significant error correlation between human behavior and the neural decoders in prefrontal areas confirms that this hierarchical organization is extended to PFC, beyond the modality-specific areas PPA, OPA, and RSC.

Intriguingly, no similar hierarchical structure of category representations was found in the auditory domain. Both types of errors, the errors representing the physical properties and those from human behavior, were correlated to the errors of neural decoder in the ACX. This difference between the visual and the auditory domain might reflect the fact that much auditory processing occurs in subcortical regions, before the information arrives in ACX. Thus, if auditory scene processing is relying on a

hierarchical neural architecture, it might not be easily detectable with fMRI. A recent study by Teng et al. (2017) showed evidence suggesting a potential structure for auditory scene processing, finding that different types of auditory features in a scene, reverberant space, and source identity are processed at different times. Further investigation with time-resolved recording techniques, such as MEG/EEG in combination with fMRI, as well as with computational modeling (Cichy and Teng, 2016) are needed for a better understanding of the neural mechanism of auditory scene processing.

Our findings show a distinction of cross-modal versus multimodal neural representations of real-world scenes in prefrontal areas versus temporal and parietal areas. However, this classification of brain areas might not necessarily hold for all situations and all types of stimuli. Indeed, previous fMRI studies as well as our findings show that brain regions traditionally considered to be sensory-specific process information from other modalities (Vetter et al., 2014; Smith and Goodale, 2015; Paton et al., 2016). For instance, Vetter et al. (2014) showed that auditory content of objects can be decoded from early visual cortex, suggesting cross-modal interactions in modality-specific areas. Using more complex stimuli at the level of scene category, our data show that auditory content can be decoded from high-level scene-selective areas (RSC and OPA). Furthermore, we also found that ACX represents visual information. Especially, one subregion of ACX, the PT, showed significant decoding accuracy in cross-modal decoding analysis as well as relatively high decoding accuracy in the image decoding analysis. These findings lead to a host of further questions for future research, such as how these visual and auditory areas are functionally connected, whether the multisensory areas mediate this interaction between the visual and auditory areas by sending feedback signals, or whether these cross-modal representations can influence or interfere with perceptual sensitivity in each sensory domain.

The whole-brain searchlight analysis confirmed the findings of our ROI-based analysis. In the image and sound decoding analyses, we found clusters with significant decoding accuracy in the visual and auditory areas as well as in the temporal, parietal, and prefrontal regions. Furthermore, the clusters in the prefrontal areas showed significant accuracy in the cross-decoding analysis, whereas the clusters in other modality-specific or multimodal areas did not, supporting the view that only representations in the PFC transcend sensory modalities. In the analysis of decoding errors, we observed that the errors of the image decoders were significantly correlated with human categorization behavior in scene-selective areas PPA and RSC as well as in the SPG, consistent with previous work by our group (Walther et al., 2009, 2011; Choo and Walther, 2016).

Previous studies addressing the integration of audiovisual information to form modality-independent representations have used univariate analysis (Downar et al., 2000; Beauchamp et al., 2004) or correlations of content-specific visual and auditory information in the brain (Hsieh et al., 2012). These methods do not distinguish between coactivation from multiple senses and modality-independent processing. Recent studies using MVPA have shown that visual and auditory information about objects (Man et al., 2012) or emotions (Peelen et al., 2010; Müller et al., 2012) evokes similar neural activity patterns across different senses, suggesting that stimulus content is represented independently of sensory modality at later stages of sensory processing. Unlike the present study, however, these studies report that areas in temporal or parietal cortex are involved in this multimodal integration. One reason for this difference could be that real-world scenes are more variable in their detailed sensory representation, typically including multiple visual and auditory cues. Furthermore, we here consider representations of scene categories as opposed to object identity (Man et al., 2012). Our results indicate that generalization across sensory modalities at the level of scene categories occurs only in PFC. The same brain regions have been found to be involved in purely visual categorization and category learning (Freedman et al., 2001; Miller and Cohen, 2001; Wood and Grafman, 2003; Meyers et al., 2008; Mack et al., 2013). This discrepancy between object identity and scene categorization might also explain different findings of cross-modal representations in visual areas.

In a recent review, Grill-Spector and Weiner (2014) suggested that the ventral temporal cortex contains a hierarchical structure for visual categorization, which has the more exemplar-specific representations in posterior areas, but the more abstract representations in anterior areas of the ventral temporal cortex. Several studies have suggested that such abstraction is tightly related to how we represent concepts in the brain by showing amodal representations across words and images (Devereux et al., 2013; Fairhall and Caramazza, 2013). Adding to this growing body of literature, we found that the posterior-to-anterior hierarchy of levels of abstraction extends to the PFC, which represents scene categories beyond the sensory modality domain. The abstraction and generalization across sensory modalities are likely to contribute to the efficiency of cognition by representing similar concepts in a consistent manner, even when the physical signal might be delivered via different sensory channels (Huth et al., 2016).

References

Beauchamp MS, Lee KE, Argall BD, Martin A (2004) Integration of auditory and visual information about objects in superior temporal sulcus. Neuron 41:809–823. CrossRef Medline

Calvert GA (2001) Crossmodal processing in the human brain: insights from functional neuroimaging studies. Cereb Cortex 11:1110–1123. CrossRef Medline

Calvert GA, Hansen PC, Iversen SD, Brammer MJ (2001) Detection of audio-visual integration sites in humans by application of electrophysiological criteria to the BOLD effect. Neuroimage 14:427–438. CrossRef Medline

Chang CC, Lin CJ (2011) LIBSVM: a library for support vector machines. ACM Trans Intell Syst Technol 2:27. CrossRef

Choo H, Walther DB (2016) Contour junctions underlie neural representations of scene categories in high-level human visual cortex. Neuroimage 135:32–44. CrossRef Medline

Cichy RM, Teng S (2016) Resolving the neural dynamics of visual and auditory scene processing in the human brain: a methodological approach. Philos Trans R Soc Lond B Biol Sci 372:1714. CrossRef Medline

Cohen YE, Andersen RA (2004) Multimodal spatial representations in the primate parietal lobe. In: Crossmodal Space Crossmodal Attention (Spence C, Driver J, eds), pp 99–121. New York, NY: Oxford University Press.

Cox RW (1996) AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. Comput Biomed Res 29:162–173. CrossRef Medline

Destrieux C, Fischl B, Dale A, Halgren E (2010) Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. Neuroimage 53:1–15. CrossRef Medline

Devereux BJ, Clarke A, Marouchos A, Tyler LK (2013) Representational similarity analysis reveals commonalities and differences in the semantic processing of words and objects. J Neurosci 33:18906–18916. CrossRef Medline

Dilks DD, Julian JB, Paunov AM, Kanwisher N (2013) The occipital place area is causally and selectively involved in scene perception. J Neurosci 33:1331–1336a. CrossRef Medline

Downar J, Crawley AP, Mikulis DJ, Davis KD (2000) A multimodal cortical network for the detection of changes in the sensory environment. Nat Neurosci 3:277–283. CrossRef Medline

Driver J, Noesselt T (2008) Multisensory interplay reveals crossmodal influences on 'sensory-specific' brain regions, neural responses, and judgments. Neuron 57:11–23. CrossRef Medline

Epstein R, Kanwisher N (1998) A cortical representation of the local visual environment. Nature 392:598–601. CrossRef Medline

Fairhall SL, Caramazza A (2013) Brain regions that represent amodal conceptual knowledge. J Neurosci 33:10552–10558. CrossRef Medline

Freedman DJ, Riesenhuber M, Poggio T, Miller EK (2001) Categorical representation of visual stimuli in the primate prefrontal cortex. Science 291:312–316. CrossRef Medline

Gaffan D, Harrison S (1991) Auditory-visual associations, hemispheric specialization and temporal-frontal interaction in the rhesus monkey. Brain 114:2133–2144. CrossRef Medline

Goel V, Tierney M, Sheesley L, Bartolo A, Vartanian O, Grafman J (2007) Hemispheric specialization in human prefrontal cortex for resolving certain and uncertain inferences. Cereb Cortex 17:2245–2250. CrossRef Medline

Grill-Spector K, Weiner KS (2014) The functional architecture of the ventral temporal cortex and its role in categorization. Nat Rev Neurosci 15: 536–548. CrossRef Medline

Hsieh PJ, Colas JT, Kanwisher N (2012) Spatial pattern of BOLD fMRI activation reveals cross-modal information in auditory cortex. J Neurophysiol 107:3428–3432. CrossRef Medline

Huth AG, de Heer WA, Griffiths TL, Theunissen FE, Gallant JL (2016) Natural speech reveals the semantic maps that tile human cerebral cortex. Nature 532:453–458. CrossRef Medline

Kastner S, De Weerd P, Desimone R, Ungerleider LG (1998) Mechanisms of directed attention in the human extrastriate cortex as revealed by functional MRI. Science 282:108–111. CrossRef Medline

Kravitz DJ, Peng CS, Baker CI (2011) Real-world scene representations in high-level visual cortex: it's the spaces more than the places. J Neurosci 31:7322–7333. CrossRef Medline

Kriegeskorte N, Goebel R, Bandettini P (2006) Information-based functional brain mapping. Proc Natl Acad Sci U S A 103:3863–3868. CrossRef Medline

MacEvoy SP, Epstein RA (2007) Position selectivity in scene-and object-responsive occipitotemporal regions. J Neurophysiol 98:2089–2098. CrossRef Medline

Mack ML, Preston AR, Love BC (2013) Decoding the brain's algorithm for categorization from its neural implementation. Curr Biol 23:2023–2027. CrossRef Medline

Malach R, Reppas JB, Benson RR, Kwong KK, Jiang H, Kennedy WA, Ledden PJ, Brady TJ, Rosen BR, Tootell RB (1995) Object-related activity revealed by functional magnetic resonance imaging in human occipital cortex. Proc Natl Acad Sci U S A 92:8135–8139. CrossRef Medline

Man K, Kaplan JT, Damasio A, Meyer K (2012) Sight and sound converge to form modality-invariant representations in temporoparietal cortex. J Neurosci 32:16629–16636. CrossRef Medline

Meddis R, Hewitt MJ, Shackleton TM (1990) Implementation details of a computation model of the inner hair-cell auditory-nerve synapse. J Acoust Soc 87:1813–1816. CrossRef

Meyers EM, Freedman DJ, Kreiman G, Miller EK, Poggio T (2008) Dynamic population coding of category information in inferior temporal and prefrontal cortex. J Neurophysiol 100:1407–1419. CrossRef Medline

Miller EK, Cohen JD (2001) An integrative theory of prefrontal cortex function. Annu Rev Neurosci 24:167–202. CrossRef Medline

Molholm S, Sehatpour P, Mehta AD, Shpaner M, Gomez-Ramirez M, Ortigue S, Dyke JP, Schwartz TH, Foxe JJ (2006) Audio-visual multisensory integration in superior parietal lobule revealed by human intracranial recordings. J Neurophysiol 96, 721–729. CrossRef Medline

Morgan LK, MacEvoy SP, Aguirre GK, Epstein RA (2011) Distances between real-world locations are represented in the human hippocampus. J Neurosci 31:1238–1245. CrossRef Medline

Müller VI, Cieslik EC, Turetsky BI, Eickhoff SB (2012) Crossmodal interactions in audiovisual emotion processing. Neuroimage 60:553–561. CrossRef Medline

Nishimoto S, Vu AT, Naselaris T, Benjamini Y, Yu B, Gallant JL (2011) Reconstructing visual experiences from brain activity evoked by natural movies. Curr Biol 21:1641–1646. CrossRef Medline

Norman-Haignere S, Kanwisher N, McDermott JH (2013) Cortical pitch regions in humans respond primarily to resolved harmonics and are located in specific tonotopic regions of anterior auditory cortex. J Neurosci 33:19451–19469. CrossRef Medline

Oosterhof NN, Connolly AC, Haxby JV (2016) CoSMoMVPA: multi-modal multivariate pattern analysis of neuroimaging data in Matlab/GNU octave. Front Neuroinform 10:27. CrossRef Medline

Park JY, Gu BM, Kang DH, Shin YW, Choi CH, Lee JM, Kwon JS (2010) Integration of cross-modal emotional information in the human brain: an fMRI study. Cortex 46:161–169. CrossRef Medline

Park S, Brady TF, Greene MR, Oliva A (2011) Disentangling scene content from spatial boundary: complementary roles for the parahippocampal place area and lateral occipital complex in representing real-world scenes. J Neurosci 31:1333–1340. CrossRef Medline

Paton A, Petro L, Muckli L (2016) An investigation of sound content in early visual areas. J Vis 16:153. CrossRef

Peelen MV, Atkinson AP, Vuilleumier P (2010) Supramodal representations of perceived emotions in the human brain. J Neurosci 30:10127–10134. CrossRef Medline

Pereira F, Mitchell T, Botvinick M (2009) Machine learning classifiers and fMRI: a tutorial overview. Neuroimage 45:S199–S209. CrossRef Medline

Portilla J, Simoncelli EP (2000) A parametric texture model based on joint statistics of complex wavelet coefficients. Int J Comput Vis 40:49–70. CrossRef

Romanski LM (2007) Representation and integration of auditory and visual stimuli in the primate ventral lateral prefrontal cortex. Cereb Cortex 17 [Suppl. 1]:i61–i69. CrossRef Medline

Sereno MI, Huang RS (2006) A human parietal face area contains aligned head-centered visual and tactile maps. Nat Neurosci 9:1337–1343. CrossRef Medline

Slotnick SD, Moo LR (2006) Prefrontal cortex hemispheric specialization for categorical and coordinate visual spatial memory. Neuropsychologia 44:1560–1568. CrossRef Medline

Smith FW, Goodale MA (2015) Decoding visual object categories in early somatosensory cortex. Cereb Cortex 25:1020–1031. CrossRef Medline

Sugihara T, Diltz MD, Averbeck BB, Romanski LM (2006) Integration of auditory and visual communication information in the primate ventrolateral prefrontal cortex. J Neurosci 26:11138–11147. CrossRef Medline

Teng S, Sommer VR, Pantazis D, Oliva A (2017) Hearing scenes: a neuromagnetic signature of auditory source and reverberant space separation. eNeuro 4:ENEURO.0007–17.2017. CrossRef Medline

Torralbo A, Walther DB, Chai B, Caddigan E, Fei-Fei L, Beck DM (2013) Good exemplars of natural scene categories elicit clearer patterns than bad exemplars but not greater BOLD activity. PLoS One 8:e58594. CrossRef Medline

Vetter P, Smith FW, Muckli L (2014) Decoding sound and imagery content in early visual cortex. Curr Biol 24:1256–1262. CrossRef Medline

Walther DB, Caddigan E, Fei-Fei L, Beck DM (2009) Natural scene categories revealed in distributed patterns of activity in the human brain. J Neurosci 29:10573–10581. CrossRef Medline

Walther DB, Chai B, Caddigan E, Beck DM, Fei-Fei L (2011) Simple line drawings suffice for functional MRI decoding of natural scene categories. Proc Natl Acad Sci U S A 108:9661–9666. CrossRef Medline

Walther DB, Beck DM, Fei-Fei L (2012) To err is human: correlating fMRI decoding and behavioral errors to probe the neural representation of natural scene categories. In: Understanding visual population codes–Toward a common multivariate framework for cell recording and functional imaging (Kriegeskorte N, Kreiman G, eds), pp 391–416. Cambridge, MA: MIT Press.

Wang D, Brown GJ (2006) Computational auditory scene analysis: principles, algorithms, and applications. New York, NY: Wiley.

Westfall PH, Young SS (1993) Resampling-based multiple testing: examples and methods for *p*-value adjustment, Vol 279. New York: Wiley.

Wood JN, Grafman J (2003) Human prefrontal cortex: processing and representational perspectives. Nat Rev Neurosci 4:139–147. CrossRef Medline