Behavioral/Systems/Cognitive

Sight and Sound Converge to Form Modality-Invariant Representations in Temporoparietal Cortex

Kingson Man, Jonas T. Kaplan, Antonio Damasio, and Kaspar Meyer
Brain and Creativity Institute, University of Southern California, Los Angeles, California 90089

People can identify objects in the environment with remarkable accuracy, regardless of the sensory modality they use to perceive them. This suggests that information from different sensory channels converges somewhere in the brain to form modality-invariant representations, i.e., representations that reflect an object independently of the modality through which it has been apprehended. In this functional magnetic resonance imaging study of human subjects, we first identified brain areas that responded to both visual and auditory stimuli and then used crossmodal multivariate pattern analysis to evaluate the neural representations in these regions for content specificity (i.e., do different objects evoke different representations?) and modality invariance (i.e., do the sight and the sound of the same object evoke a similar representation?). While several areas became activated in response to both auditory and visual stimulation, only the neural patterns recorded in a region around the posterior part of the superior temporal sulcus displayed both content specificity and modality invariance. This region thus appears to play an important role in our ability to recognize objects in our surroundings through multiple sensory channels and to process them at a supramodal (i.e., conceptual) level.

Introduction

Whether we see a bell swing back and forth or instead hear its distinctive dingdong, we easily recognize the object in both cases. Upon recognition, we are able to access the wide conceptual knowledge we possess about bells and we use this knowledge to generate motor behaviors and verbal reports. The fact that we are able to do so independently of the perceptual channel through which we were stimulated suggests that the information provided by different channels converges, at some stage, into modality-invariant neural representations of the object.

Neuroanatomists have long identified areas of multisensory convergence in the monkey brain, for instance in the lateral prefrontal and premotor cortices, the intraparietal sulcus, the parahippocampal gyrus, and the posterior part of the superior temporal sulcus (pSTS) (Seltzer and Pandya, 1978, 1994). Lesion and tracer studies have shown that the pSTS region not only receives projections from visual, auditory, and somatosensory association cortices, but returns projections to those cortices as well (Seltzer and Pandya, 1991; Barnes and Pandya, 1992). Also, electrophysiological studies have identified bimodal and trimodal neurons in the pSTS (Benevento et al., 1977; Bruce et al., 1981; Hikosaka et al., 1988). Recent functional neuroimaging studies in humans are in line with the anatomical and electrophysiological evidence and have located areas of multisensory integration in the lateral prefrontal cortex, premotor cortex, posterior parietal cortex, and the pSTS region (for review, see e.g. Calvert, 2001, Amedi et al., 2005; Beauchamp, 2005; Doehrmann and Naumer, 2008; Driver and Noesselt, 2008). These observations alone, however, do not address two important questions. First, are the neural patterns established in these multimodal brain regions content specific? In other words, do they reflect the identity of a sensory stimulus rather than a more general aspect of the perceptual process? Second, are the neural patterns modality-invariant? In other words, does an object evoke similar neural response patterns when it is apprehended through different modalities?

In the present study, we used multivariate pattern analysis (MVPA) of functional magnetic resonance imaging (fMRI) data to probe multimodal regions for neural representations that were both content specific and modality invariant. We first performed a univariate fMRI analysis to identify brain regions that were activated by both visual and auditory stimuli, and these regions corresponded well with those found in previous studies. Next, we tested the activity patterns in these regions for content specificity by asking whether a machine-learning algorithm could predict from a specific pattern which of several audio or video clips a subject had perceived. Finally, we tested for modality invariance by decoding the identities of objects not only within but across modalities; the algorithm was trained to distinguish neural patterns recorded during visual trials and used to classify neural patterns recorded during auditory trials. The crossmodal MVPA analysis revealed that of all the multisensory regions identified, only the pSTS region contained neural representations that were both content specific and modality invariant.

Materials and Methods

Subjects. Nine right-handed subjects were originally enrolled in the study. One subject was excluded from the analysis due to excessive head

Received May 15, 2012; revised Aug. 20, 2012; accepted Sept. 12, 2012.

Author contributions: K. Man, J.T.K., and K. Meyer designed research; K. Man, J.T.K., and K. Meyer performed research; K. Man and J.T.K. analyzed data; K. Man, J.T.K., A.D., and K. Meyer wrote the paper.

This research was supported by the Mathers Foundation and by the National Institutes of Health (NIH/NINDS 5P50NS019632). We thank Hanna Damasio for her comments on the manuscript.

The authors declare no competing financial interests.

Correspondence should be addressed to Kingson Man, Brain and Creativity Institute, University of Southern California, 3641 Watt Way, Los Angeles, CA 90089-2520. E-mail: kman@usc.edu.

DOI:10.1523/JNEUROSCI.2342-12.2012

0270-6474/12/3216629-08$15.00/0

movement during the scan. The data presented came from the remaining eight participants, five female and three male. The experiment was undertaken with the informed written consent of each subject.

Stimuli. Audiovisual clips depicting a church bell, a gong, a typewriter, a jackhammer, a pneumatic drill, and a chainsaw were downloaded from www.youtube.com. All clips were truncated to 5 s. Two additional sets of clips were generated from the original versions: an auditory set containing the audio tracks presented on a black screen, and a visual set containing the video tracks presented in silence. All audio tracks were peak leveled using The Levelator 2 software (The Conversations Network).

Offline stimulus familiarization. Before scanning, subjects watched the six original (i.e., audiovisual) stimuli on a loop for 10 min to familiarize themselves with the correspondence between the auditory and visual content of the clips.

Online stimulus presentation. Once inside the fMRI scanner, subjects were only exposed to clips from the auditory and visual sets presented in separate auditory and visual runs that alternated for a total of eight runs. Each run contained 36 stimulus presentations; each of the six clips was shown six times in randomized order with no back-to-back repeats. One clip was presented every 11 s. A sparse-sampling scanning paradigm was used to ensure that all clips were presented in the absence of scanner noise. A single whole-brain volume was acquired starting 2 s after the end of each clip, timed to coincide with the peak of the hemodynamic response to the stimulus (Meyer et al., 2010). The 2-s-long image acquisition was followed by a 2 s pause, after which the next trial began. Timing and presentation of the video clips was controlled with MATLAB 7.9.0 (The Mathworks) using the freely available Psychophysics Toolbox 3 software (Brainard, 1997). Video clips were displayed on a rear projection screen at the end of the scanner bore that subjects viewed through a mirror mounted on the head coil. All subjects had normal or corrected-to-normal vision; the sound intensity of the audio clips was adjusted to the loudest comfortable level for each subject. Participants were instructed to keep their eyes open during all runs.

Before each auditory run, subjects were told to "imagine seeing in your mind's eye the images that go with these sound clips as vividly as possible." Analogously, each visual run was preceded by the instruction to "imagine hearing in your mind's ear the sounds that go with these video clips as vividly as possible." We included this imagery instruction based on findings from a previous study in which subjects were instructed to passively watch sound-implying (but silent) video clips (Meyer et al., 2010). Those subjects reported that they spontaneously generated auditory imagery, and the reported vividness of their imagery correlated with the classification accuracies of the evoking stimuli from neural activity in the early auditory cortices.

In addition to the eight MVPA runs, there were two runs of a functional localizer placed at the beginning and the middle of the scanning session, respectively. These runs served as an independent dataset on which we performed a conventional univariate analysis to identify regions of interest (ROIs) for MVPA. The functional localizer runs employed a slow event-related design with continuous image acquisition. Each run contained 24 stimulus presentations (two repetitions each of the six auditory clips and the six visual clips presented in randomized order with no back-to-back repeats). The duration of each trial was randomly jittered up to 1 s about the mean duration of 15 s.

Image acquisition. Images were acquired with a 3-tesla Siemens MAGNETON Trio System. Echo-planar volumes for MVPA runs were acquired with the following parameters: repetition time (TR) = 11,000 ms, acquisition time (TA) = 2000 ms, echo time (TE) = 25 ms, flip angle = 90 degrees, 64 × 64 matrix, in-plane resolution 3.0 mm × 3.0 mm, 41 transverse slices, each 2.5 mm thick, covering the whole brain. Volumes for functional localizer runs were acquired with the same parameters except in continuous acquisition with TR = 2000 ms. We also acquired a structural T1-weighted MPRAGE for each subject (TR = 2530 ms, TE = 3.09 ms, flip angle = 10 degrees, 256 × 256 matrix, 208 coronal slices, 1 mm isotropic resolution).

Univariate analysis of functional localizer runs. We performed a univariate analysis of the functional localizer runs using FSL (Smith et al., 2004). Data preprocessing involved the following steps: motion correction (Jenkinson et al., 2002), brain extraction (Smith, 2002), slice-timing correction, spatial smoothing with a 5 mm full-width at half-maximum Gaussian kernel, high-pass temporal filtering using Gaussian-weighted least-squares straight line fitting with σ (standard deviation of the Gaussian distribution) equal to 60 s, and prewhitening (Woolrich et al., 2001).

Table 1. Coordinates of the peak voxels of the two unimodal and four multimodal activity clusters, as identified in the localizer scans

		Peak voxel location (MNI space *X*, *Y*, *Z*)	
Localizer contrast	Anatomical region	Left hemisphere	Right hemisphere
Auditory versus rest	Superior temporal lobe	−60, −18, 6	42, −26, 10
Visual versus rest	Medial and lateral occipital lobe	−12, −104, 6	6, −90, −10
Audiovisual overlap	Posterior superior temporal sulcus	−54, −44, 14	64, −40, 16
	Inferior frontal cortex	−52, 34, 18	48, 16, 24
	Medial premotor cortex	−4, 12, 54	2, 8, 60
	Anterior insula	−32, 16, 0	32, 26, 2

The two stimulus types, auditory and visual, were modeled separately with two regressors derived from a convolution of the task design and a gamma function to represent the hemodynamic response function. Motion correction parameters were included in the design as additional regressors. The two functional localizer runs for each participant were combined into a second-level fixed-effects analysis, and a third-level intersubject analysis was performed using a mixed-effects design.

Registration of the functional data to the high-resolution anatomical image of each subject and to the standard Montreal Neurological Institute (MNI) brain was performed using the FSL FLIRT tool (Jenkinson and Smith, 2001). Functional images were aligned to the high-resolution anatomical image using a six-degree-of-freedom linear transformation. Anatomical images were registered to the MNI-152 brain using a 12-degree-of-freedom affine transformation.

Two activation maps were defined using the functional localizer data; the areas activated during the presentation of video clips as compared to rest, and the areas activated during the presentation of audio clips as compared to rest. The visual and auditory activation maps were thresholded with FSL's cluster thresholding algorithm using a minimum *Z*-score of 2.3 ($p < 0.01$) and a cluster size probability of $p < 0.05$. An audiovisual map was defined by the overlap of the auditory and visual activation maps. To detect the greatest number of shared voxels in the overlap, the component maps were Z-score thresholded as above but received no cluster size thresholding. Each map was normalized, overlapping voxels were summed, and the resultant map was smoothed with a 5 mm full-width at half-maximum Gaussian kernel.

Voxel selection for multivariate pattern analysis. The six ROIs for the MVPA analysis were generated based on the largest activity cluster from the group-level auditory activation map (located in the superior temporal lobe), the largest cluster from the visual activation map (located in the occipital lobe), and the four largest clusters in the audiovisual overlap map (located in an area around the posterior superior temporal sulcus, the inferior frontal gyrus and sulcus collectively referred to as the inferior frontal cortex, the medial premotor cortex, and the anterior insula, respectively). To define these ROIs in each individual subject, we centered a sphere at the peak voxel of each cluster on the group-level maps and then warped these spheres into the functional spaces of the individual subjects. All spheres had a radius of 20 mm, except the one for the anterior insula, which, due to the smaller size of the activity cluster, had a radius of 12 mm. Within the spheres located in the functional maps of each subject, we then selected the 500 voxels with the highest *Z*-scores from that subject's relevant localizer. This method allowed us to select voxels from the same anatomical regions across subjects, but with subject-specific sensitivity to activation.

Multivariate pattern analysis. Within the six ROIs described above, we performed both intramodal and crossmodal MVPA. For both intramodal and crossmodal classification, all possible two-way discriminations among the six stimuli were carried out ($n = 15$). A classifier was trained and tested anew for each of the pairwise discriminations. For intramodal classification, we performed a leave-one-run-out cross-validation procedure; a classifier was trained on data from three of the

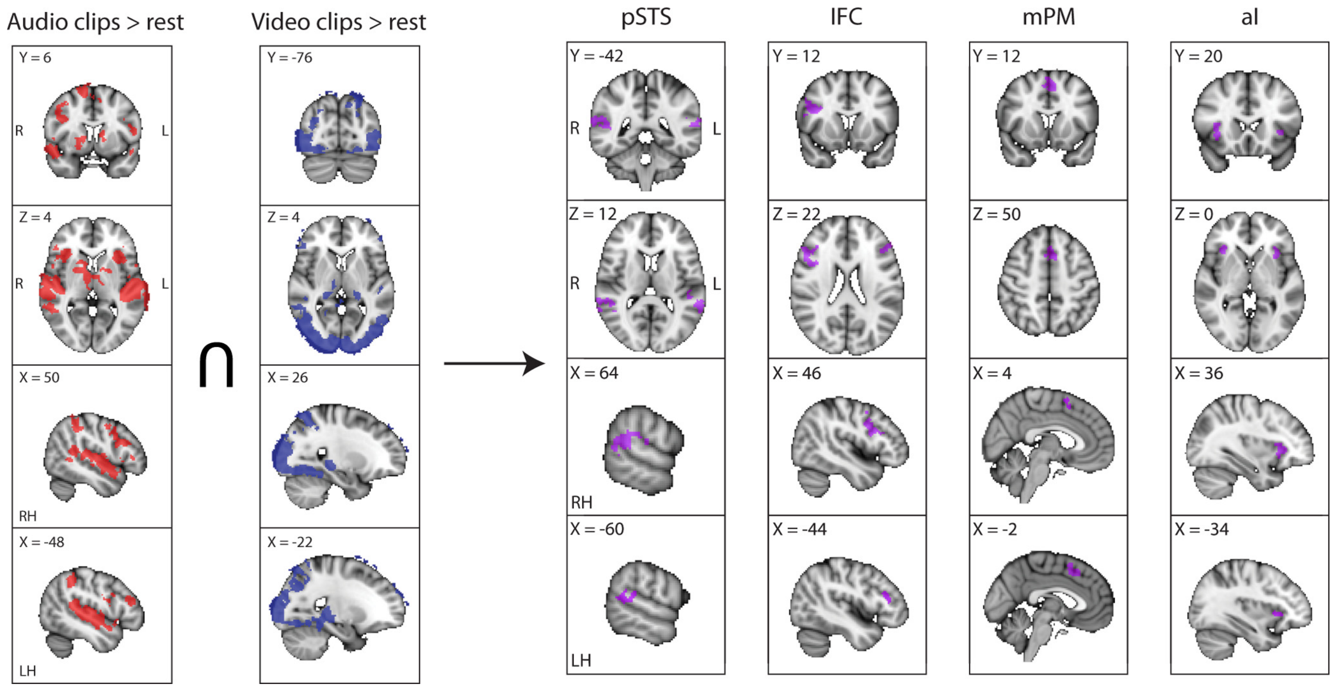

Figure 1. Brain regions activated by auditory and visual stimuli. Auditory activations (first panel on the left) most notably comprised the superior part of the temporal lobe (Heschl's gyrus, planum temporale, planum polare, and adjacent parts of the superior temporal gyrus), as well as additional regions in the parietal and frontal lobes. Visual activations (second panel from the left) included the medial and lateral portions of the occipital lobe, the posterior temporal lobe, as well as additional regions in the parietal and frontal lobes. The intersection of the auditory and visual activations was calculated and the four largest clusters on the audiovisual overlap map were subsequently used for MVPA. The four multisensory regions of interest were the posterior superior temporal sulcus (pSTS), inferior frontal cortex (IFC), medial premotor cortex (mPM), and the anterior insula (aI). Slices across different planes and regions are not to scale. R, Right; L, left; RH, right hemisphere; LH, left hemisphere.

Table 2. Average intramodal and crossmodal classification accuracies

	Classification type			
			Crossmodal	
	Intramodal		Train aud–test	Train vis–test
ROI	Auditory	Visual	vis	aud
Auditory cortices	0.838***	0.591**	0.547*	0.526*
Visual cortices	0.535 (ns)	0.843***	0.490 (ns)	0.472 (ns)
Posterior superior temporal sulcus	0.765***	0.691***	0.601**	0.588**
Inferior frontal cortex	0.563**	0.571**	0.505 (ns)	0.515 (ns)
Medial premotor	0.541*	0.555**	0.506 (ns)	0.503 (ns)
Anterior insula	0.527 (ns)	0.531*	0.497 (ns)	0.501 (ns)

Chance performance is 0.5. The *p* values were FDR-corrected to control for Type I errors given the 24 comparisons. ns, Not significant; $*p < 0.05$; $**p < 0.01$; $***p < 0.001$.

four runs of a given modality and then tested on the fourth run. This procedure was repeated four times, with the data from each run serving as the testing set once. For crossmodal classification, a classifier was either trained on all four auditory (aud) runs and tested on all four video (vis) runs (train aud–test vis) or vice versa (train vis–test aud). MVPA was performed using the PyMVPA software package (Hanke et al., 2009) in combination with implementation by LIBSVM of the linear support vector machine (Chang and Lin, 2011). Data from the eight MVPA runs of each subject were concatenated and motion corrected to the middle volume of the entire time series and then linearly detrended and converted to *Z*-scores by run. Classification was performed on the *Z*-scores from selected voxels.

Whole-brain searchlight analyses. To conduct an ROI-independent search for brain regions containing information relevant to intramodal or crossmodal classification, we performed a searchlight procedure (Kriegeskorte et al., 2006). In each subject, a six-way "all versus all" classifier (Pereira and Botvinick, 2011) was repetitively applied to small spheres ($r = 8$ mm) centered on every voxel of the brain. The classification accuracy for each sphere was mapped to its center voxel to obtain whole-brain accuracy maps for intramodal and crossmodal classification. Crossmodal searchlights were performed in both directions, train aud–test vis and train vis–test aud, and resulting accuracies were averaged across the two directions. To illustrate which brain regions showed high performance across subjects, we thresholded the individual maps at the 95th percentile, warped them into the standard space, and summed them to create an overlap map.

Statistical analyses. Given that our hypothesis was directional—classifier performance on two-way discriminations should be higher than the chance result of 0.5—we employed one-tailed *t* tests across all eight subjects to assess the statistical significance of our results. When comparing classifier performances to each other (e.g., those obtained in different ROIs), we used two-tailed, paired *t* tests across all eight subjects. The *p* values for *t* tests of classification accuracies were false discovery rate (FDR) corrected for the 24 discriminations performed: four classification types in each of six ROIs.

Results

Functional localizer scans

At the group level the presentation of auditory stimuli, as compared to rest, yielded a prominent activity cluster in the auditory cortices (Heschl's gyrus, planum temporale, planum polare, and the surrounding regions of the superior temporal gyrus), as well as smaller foci throughout the brain (Table 1, Fig. 1). The presentation of visual stimuli, as compared to rest, revealed a main cluster in the visual cortices (medial and lateral surfaces of the occipital lobe) and, again, smaller foci dispersed throughout the brain. An overlap of the two activation maps revealed several brain regions that responded to both auditory and visual stimulation; the four largest clusters were located in inferior frontal cortex, medial premotor cortex, anterior insula, and an area around the pSTS, which included parts of the superior temporal gyrus and the inferior parietal lobule (Table 1, Fig. 1).

Intramodal classification of stimuli

From the two unimodal and the four multimodal clusters described above, voxels were selected in a subject-specific manner (see Materials and Methods) to define six ROIs for MVPA: auditory cortex (AC), visual cortex (VC), inferior frontal cortex (IFC), medial premotor cortex (mPM), anterior insula (aI), and pSTS. As expected, audio clips were successfully classified within AC; averaged across all subjects and all two-way discriminations among the six stimuli, classification performance was 0.838, significantly higher than the chance level of 0.5 (FDR-corrected $p < 0.001$; one-tailed *t* test, $n = 8$; Table 2, Fig. 2*A*). Similarly, classification of video clips within VC yielded an average performance of 0.843 ($p < 0.001$). Classification of video clips from AC was less accurate but still significant (0.591; $p < 0.01$), while classification of audio clips from VC was not significant (0.535; $p > 0.05$; Table 2, Fig. 2*A*). Both audio and video clips were well classified from the pSTS region (audio clips: 0.765, $p < 0.001$; video clips: 0.691, $p < 0.001$; Fig. 2*B*). Classification performance was significantly higher in the pSTS than in any other multimodal ROI ($p < 0.001$ in all cases; two-tailed, paired *t* tests).

In each subject, we performed a whole-brain searchlight analysis to conduct an ROI-independent search for regions classifying audio or video clips, respectively. The individual subjects' searchlight maps were thresholded at the 95th percentile, warped to the standard space, and summed to create a group-level overlap map. Chance performance for six-way classification was one of six, ~0.167. The means of the individual accuracy thresholds were 0.250 (SEM = 0.004) for the auditory searchlight and 0.227 (SEM = 0.006) for the visual searchlight. Not surprisingly, highest classification accuracy for audio and video clips was found in auditory and visual cortices, respectively (Fig. 2*C*).

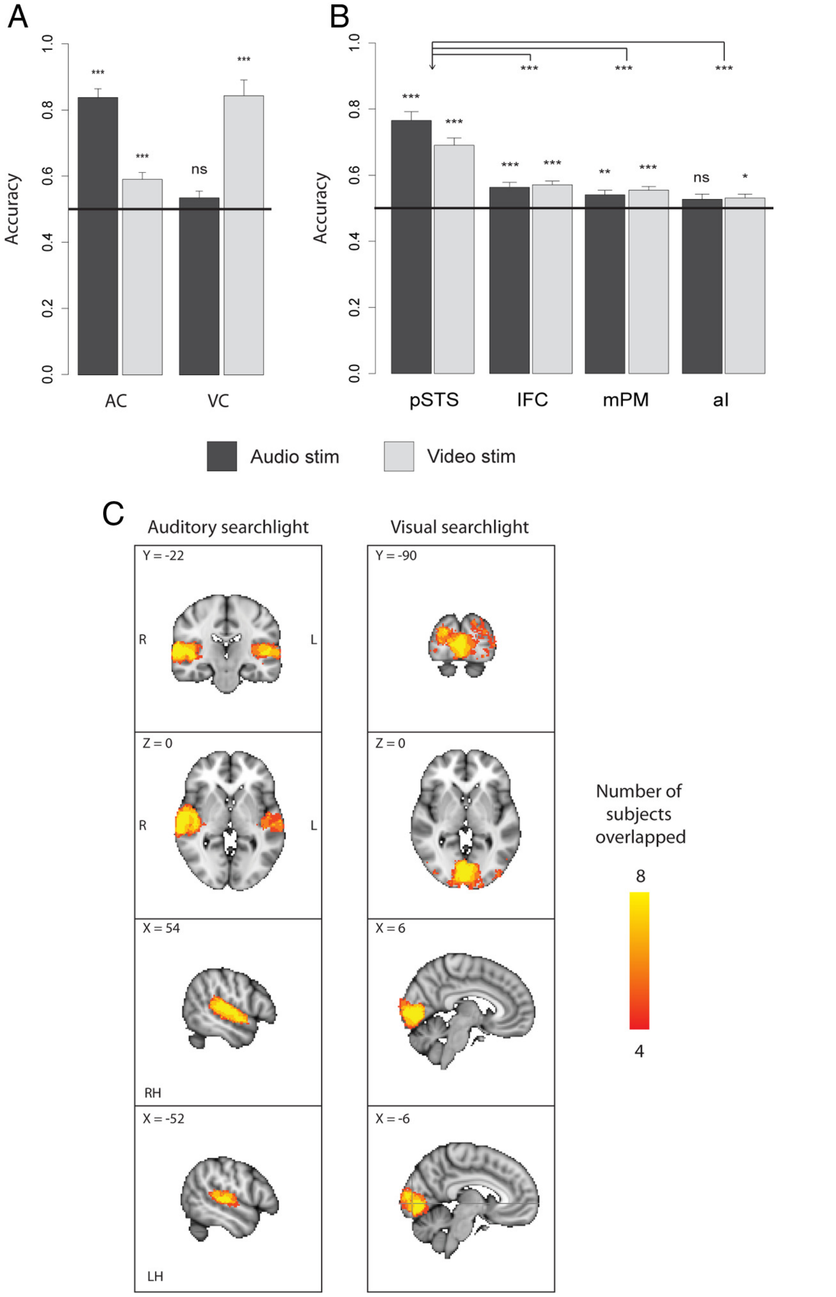

Figure 2. Intramodal classification performance. ***A***, Auditory and visual stimuli were reliably classified from their respective sensory cortices, auditory cortex (AC) and visual cortex (VC). Classification performance of unimodal stimuli in heteromodal cortices was much more modest, albeit still significant for visual stimuli classified from AC. ***B***, Audio and video clips were classified significantly more accurately in the pSTS than in the other three multimodal ROIs. ns, Not significant; $*p < 0.05$; $**p < 0.01$; $***p < 0.001$; The *p* values were FDR corrected. All error bars indicate SEM. ***C***, Intramodal searchlight analysis. Audio clips were classified most accurately from voxels in the superior temporal lobe, while video clips were best discriminated from voxels in the occipital lobe. The figure shows voxels that classified above the 95th percentile of accuracy in four or more subjects. The mean of the 95th percentile thresholds was 0.250 for the auditory searchlight and 0.227 for the visual searchlight; chance performance in the six-way searchlights was 0.167.

Crossmodal classification of stimuli

The accuracies obtained from the two directions of crossmodal classification, train aud–test vis and train vis–test aud, were not significantly different ($p > 0.05$). The accuracies indicated below for the various classifications represent the mean values derived from the results of both classification directions. Crossmodal classification performance in the pSTS region was significantly better than chance (0.595; $p < 0.01$) and significantly better than in all other multimodal ROIs, where performance was near chance level (IFC, 0.510; mPM, 0.505; aI, 0.499; $p > 0.05$ in each case; Table 2, Fig. 3*A*). Twelve of fifteen pairwise discriminations among the six items were significantly above chance in pSTS, whereas none were in IFC and mPM, and only one of fifteen was in aI (Fig. 3*B*). Crossmodal classification was modest, but also significant, in the auditory ROI (0.536, $p < 0.01$); it was at chance level in the visual ROI (0.481, $p > 0.05$).

In each subject we also performed a crossmodal searchlight analysis averaging across both directions of training and testing (Fig. 4). Again, chance performance in this six-way analysis was one in six, or 0.167. The mean of the individual 95th percentile accuracy thresholds was 0.211 (SEM = 0.003). The whole-brain crossmodal searchlight analysis confirmed that the only region consistently containing crossmodal information was located around the pSTS near the junction of the temporal and parietal lobes, almost completely lateralized to the right hemisphere (Fig. 4). This region was similar in location to the pSTS cluster identified in the audiovisual overlap map presented in Figure 1. Despite the lateralization evident in the searchlight, the classification accuracies between the right and left pSTS ROIs were not significantly different ($p > 0.05$).

Discussion

We used crossmodal MVPA to investigate the response patterns of various brain regions to corresponding audio and video clips of common objects. Successful classification of the response patterns was taken to infer content specificity. Although several regions were activated by stimuli of both modalities, both the ROI and the searchlight analyses indicated that the pSTS region may be the unique site of successful crossmodal classification. This suggests the pSTS is different from other multimodal brain areas in that it holds neural representations of common objects that are both content specific and modality invariant.

Although the searchlight analysis suggests that supramodal representations may be lateralized to the right pSTS, the ROI analysis shows no discrepancy between left and right pSTS. This may be due to the fact that our searchlight method is sensitive to voxels that consistently classify well across subjects, whereas the ROI analysis selects voxels based on individual parametric maps.

Previous studies have aimed to identify brain regions that engage in multisensory processing by using a variety of criteria: a region may simply be activated by more than one modality of stimulation; its response to a unimodal stimulus may be affected by adding a stimulus of a different modality; or, more stringently, a region may be more strongly activated by a multimodal stimulus than by the summed activations from the separate unimodal stimuli (superadditivity) (Calvert, 2001; Beauchamp, 2005; Laurienti et al., 2005). However, these approaches do not directly address the questions of modality invariance and content specificity. Modality invariance is more directly addressed by the semantic congruency effect: a brain region would display different levels of activity when responding to the sight and sound of the same object than when responding to the sight and

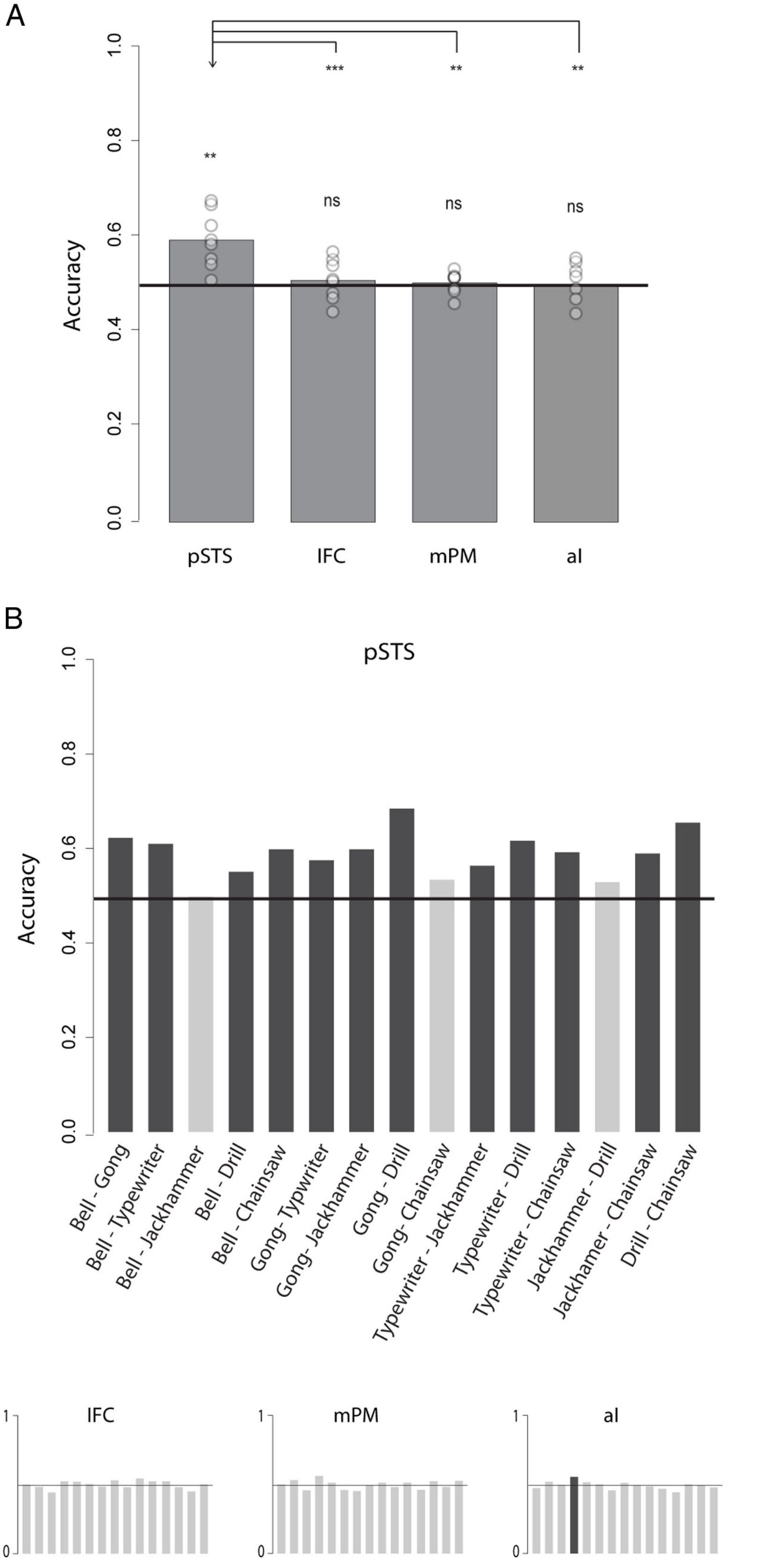

Figure 3. Crossmodal classification performance. ***A***, Mean crossmodal accuracy in the multimodal ROIs superimposed with individual subject accuracies (open circles). Classification was significantly better than chance only in the pSTS. The pSTS was significantly higher than any of the other multimodal regions. ns, Not significant; $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$. ***B***, Pairwise classification accuracies. There were 15 pairwise discriminations among the six stimuli. In the pSTS ROI, 12 of the 15 discriminations yielded classification performances significantly above chance, as indicated by darker shaded bars (top). None of the pairwise comparisons were significantly better than chance in IFC and mPM, and only one was in aI (bottom).

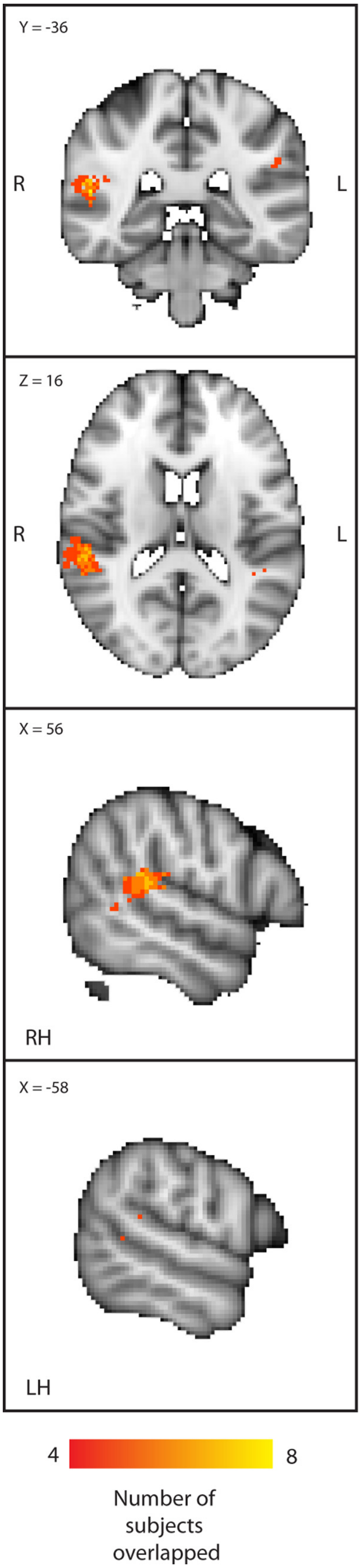

Figure 4. Crossmodal searchlight analysis. The figure shows voxels that classified above the 95th percentile of accuracy in four or more subjects. The mean of the 95 th percentile thresholds was 0.211, chance performance for the six-way searchlights being 0.167. Voxels cluster around the pSTS, close to the junction of the temporal and parietal lobes, and are almost completely lateralized to the right.

sound of different objects. Doehrmann and Naumer (2008) reviewed human fMRI studies manipulating semantic congruency and identified a general pattern of activation in lateral temporal cortex for semantically congruent stimuli and in inferior frontal cortex for semantically incongruent stimuli. Our results allow for more specific conclusions than can be drawn from the semantic congruency effect: activity in the pSTS reflects which object was presented congruently. A plausible sequence for the flow of information during semantic congruency tasks is for the (congruent or incongruent) audiovisual stimulus to first be compared against the store of congruent supramodal representations in the pSTS, generating either a match or mismatch signal that may then be broadcast to other regions, such as IFC, that display semantic congruency effects.

MVPA heretofore has been used to identify neural patterns that generalize across formats within the same sensory modality, e.g., between words and pictures (Shinkareva et al., 2011), and also across modalities, e.g., between auditory and visual displays of emotion (Peelen et al., 2010). Until now, however, there has not been an explicit test for conserved representations of object identities across different sensory modalities [although Pietrini et al. (2004) found conserved representations of object categories across the visual and tactile modalities].

Degrees of modality invariance

Our claim that neural patterns in the pSTS display modality invariance may appear tempered by the fact that crossmodal classification accuracies are not perfect. While many MVPA studies attempt to predict perceptual stimuli from neural activity in early sensory cortices, the current investigation was concerned with activity patterns in association cortices. Conceivably, the topographical organization of the early cortices would be particularly conducive to successful pattern classification, and it is well known that this organization is gradually lost at higher levels of the sensory hierarchies (Felleman and Van Essen, 1991). This is reflected in our intramodal classification accuracies; they were high for the early sensory cortices (around 0.84 for both audio and video clips) but declined in the pSTS (0.69 for video clips and 0.77 for audio clips). It stands to reason that crossmodal classification would be unlikely to yield accuracies exceeding the lower of the two intramodal classification accuracies. The pSTS crossmodal accuracy of 0.595 thus should be appreciated from the perspective of 0.69 as a "soft ceiling."

The functional organization of the pSTS may provide additional insight into the comparison between crossmodal and intramodal classification accuracies. Superior temporal cortex has been found to contain intermixed millimeter-scale patches that respond preferentially to auditory (A), visual (V), or audiovisual stimuli (AV) (Beauchamp et al., 2004; Dahl et al., 2009). Accordingly, auditory stimuli would activate A and AV patches, and visual stimuli would activate V and AV patches. Crossmodal classification, however, could rely solely on the AV patches. Consequently, crossmodal classification accuracy would be expected to be lower than intramodal classification accuracy because: (1) it has access to fewer information-bearing voxels; and (2) the A and V patches, from the classifier's perspective, add noise to the analysis. In light of these arguments, we wish to make clear that we use "invariance" as a relative term. What our analysis shows is that the neural activity patterns induced by corresponding auditory and visual stimuli are more similar than the patterns induced by noncorresponding stimuli.

Supramodal representations as convergence–divergence zones

According to a neuroarchitectural framework proposed by Damasio (Damasio, 1989; Meyer and Damasio, 2009), neuron ensembles in higher-order association cortices constitute convergence–divergence zones (CDZs), which register associations among perceptual representations from multiple sensory modalities. Due to the convergent bottom-up projections they receive from early sensory cortices, CDZs can be activated, for instance, both by the sight and the sound of a specific object. Once activated, the CDZs can re-instantiate the associated representations in different sensory cortices by means of divergent top-down projections. This re-instantiation of activity patterns in early sensory cortices distinguishes the CDZ framework from other models that posit nonspecific, modulatory feedback mechanisms. In brief, according to the CDZ framework, the bottom-up processing of sensory stimuli would be continuously accompanied by the top-down reconstruction of associated patterns in different modalities.

In keeping with this prediction, previous MVPA studies have shown that visual stimuli implying sound or touch lead to content-specific representations in early auditory and somatosensory cortices, respectively (Meyer et al., 2010, 2011). Conversely, auditory stimuli induce content-specific neural patterns in early visual cortices (Vetter et al., 2011). The current results provide an additional piece of support for the framework, suggesting that the pertinent audiovisual CDZs may be located around the pSTS.

In the context of the CDZ framework, we also considered whether modality invariance would extend from multisensory association cortices toward early sensory cortices. If watching a silent video clip of a jackhammer results in the reconstruction of a neural activity pattern in the early auditory cortices, it is conceivable that the reconstructed pattern would resemble the one established when the jackhammer was actually heard. This question is the reason why our subjects received an explicit instruction to imagine the sensory counterpart (auditory or visual) of the stimuli (visual or auditory) they perceived. We did not find any indication of modality invariance in the early visual cortices, and while crossmodal classification performance was indeed significant in the auditory ROI, the searchlight analysis suggests this may have been due to the partial overlap of the AC and pSTS ROIs. Thus, the present study does not provide evidence for modality invariance at the level of early sensory cortices.

Is it possible that crossmodal classification was successful only because of mental imagery? In other words, have we demonstrated classification between perception and imagery within a single modality? We do not believe this is the case for two reasons. First, the pSTS ROI was defined based on activations from the localizer runs in which subjects did not receive any imagery instruction. Second, previous studies performing classification between visual perception and visual imagery found above-chance performance in ventral visual stream regions such as lateral occipital cortex and ventral temporal cortex (Stokes et al., 2009; Reddy et al., 2010; Lee et al., 2012). Our crossmodal analysis did not identify any of these regions and, conversely, the mentioned studies did not find above-chance classification in the pSTS. As for the auditory modality, to the best of our knowledge classification between perception and imagery heretofore has not been reported. Auditory perception and auditory imagery do activate overlapping areas in planum temporale and the posterior superior temporal gyrus (Bunzeck et al., 2005; Zatorre and Halpern, 2005), admitting of the theoretical possibility for auditory imagery–perception classification in our experiment. However, such an interpretation would be pressed to explain the asymmetry between (successful) auditory and (unsuccessful) visual classification. Given the large amount of evidence that implicates the pSTS in multisensory processing, it appears most parsimonious to interpret the neural representations in pSTS as generalizing across the auditory and visual modalities.

To summarize, we have shown that a region around the posterior superior temporal sulcus is activated in content-specific fashion by both visual and auditory stimuli, and that the activity patterns induced by the same object presented in the two modalities exhibit a certain degree of similarity. Such modality-invariant representations of the objects of perception may be the initial stage of the neural process that allows us to recognize and react to sensory stimuli independently of the modality through which we perceive them.

References

Amedi A, von Kriegstein K, van Atteveldt NM, Beauchamp MS, Naumer MJ (2005) Functional imaging of human crossmodal identification and object recognition. Exp Brain Res 166:559 –571. CrossRef Medline

Barnes CL, Pandya DN (1992) Efferent cortical connections of multimodal cortex of the superior temporal sulcus in the rhesus monkey. J Comp Neurol 318:222–244. Medline

Beauchamp MS (2005) Statistical criteria in FMRI studies of multisensory integration. Neuroinformatics 3:93–113. CrossRef Medline

Beauchamp MS, Argall BD, Bodurka J, Duyn JH, Martin A (2004) Unraveling multisensory integration: patchy organization within human STS multisensory cortex. Nat Neurosci 7:1190 –1192. CrossRef Medline

Benevento LA, Fallon J, Davis BJ, Rezak M (1977) Auditory-visual interaction in single cells in the cortex of the superior temporal sulcus and the orbital frontal cortex of the macaque monkey. Exp Neurol 57:849 – 872. Medline

Brainard DH (1997) The psychophysics toolbox. Spat Vis 10:433– 436. CrossRef Medline

Bruce C, Desimone R, Gross CG (1981) Visual properties of neurons in a polysensory area in superior temporal sulcus of the macaque. J Neurophysiol 46:369 –384. Medline

Bunzeck N, Wuestenberg T, Lutz K, Heinze HJ, Jancke L (2005) Scanning silence: mental imagery of complex sounds. Neuroimage 26:1119 –1127. CrossRef Medline

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

Chang CC, Lin CJ (2011) LIBSVM: a library for support vector machines. ACM Trans Intell Syst Technol 2:27:1–27:27. CrossRef

Dahl CD, Logothetis NK, Kayser C (2009) Spatial organization of multisensory responses in temporal association cortex. J Neurosci 29:11924 – 11932. CrossRef Medline

Damasio AR (1989) Time-locked multiregional retroactivation: a systems-level proposal for the neural substrates of recall and recognition. Cognition 33:25– 62. CrossRef Medline

Doehrmann O, Naumer MJ (2008) Semantics and the multisensory brain: how meaning modulates processes of audio-visual integration. Brain Res 1242:136 –150. 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

Felleman DJ, Van Essen DC (1991) Distributed hierarchical processing in the primate cerebral cortex. Cereb Cortex 1:1– 47. CrossRef Medline

Hanke M, Halchenko YO, Sederberg PB, Hanson SJ, Haxby JV, Pollmann S (2009) PyMVPA: A python toolbox for multivariate pattern analysis of fMRI data. Neuroinformatics 7:37–53. CrossRef Medline

Hikosaka K, Iwai E, Saito H, Tanaka K (1988) Polysensory properties of neurons in the anterior bank of the caudal superior temporal sulcus of the macaque monkey. J Neurophysiol 60:1615–1637. Medline

Jenkinson M, Smith S (2001) A global optimisation method for robust affine registration of brain images. Med Image Anal 5:143–156. CrossRef Medline

Jenkinson M, Bannister P, Brady M, Smith S (2002) Improved optimization for the robust and accurate linear registration and motion correction of brain images. Neuroimage 17:825– 841. 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

Laurienti PJ, Perrault TJ, Stanford TR, Wallace MT, Stein BE (2005) On the use of superadditivity as a metric for characterizing multisensory integration in functional neuroimaging studies. Exp Brain Res 166:289 –297. CrossRef Medline

Lee SH, Kravitz DJ, Baker CI (2012) Disentangling visual imagery and perception of real-world objects. Neuroimage 59:4064 – 4073. CrossRef Medline

Meyer K, Damasio A (2009) Convergence and divergence in a neural architecture for recognition and memory. Trends Neurosci 32:376 –382. CrossRef Medline

Meyer K, Kaplan JT, Essex R, Webber C, Damasio H, Damasio A (2010) Predicting visual stimuli on the basis of activity in auditory cortices. Nat Neurosci 13:667– 668. CrossRef Medline

Meyer K, Kaplan JT, Essex R, Damasio H, Damasio A (2011) Seeing touch is correlated with content-specific activity in primary somatosensory cortex. Cereb Cortex 21:2113–2121. CrossRef Medline

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, Botvinick M (2011) Information mapping with pattern classifiers: A comparative study. Neuroimage 56:476 – 496. CrossRef Medline

Pietrini P, Furey ML, Ricciardi E, Gobbini MI, Wu WH, Cohen L, Guazzelli M, Haxby JV (2004) Beyond sensory images: Object-based representation in the human ventral pathway. Proc Natl Acad Sci U S A 101:5658 – 5663. CrossRef Medline

Reddy L, Tsuchiya N, Serre T (2010) Reading the mind's eye: decoding category information during mental imagery. Neuroimage 50:818 – 825. CrossRef Medline

Seltzer B, Pandya DN (1978) Afferent cortical connections and architectonics of the superior temporal sulcus and surrounding cortex in the rhesus monkey. Brain Res 149:1–24. CrossRef Medline

Seltzer B, Pandya DN (1991) Post-rolandic cortical projections of the superior temporal sulcus in the rhesus monkey. J Comp Neurol 312:625– 640. Medline

Seltzer B, Pandya DN (1994) Parietal, temporal, and occipital projections to cortex of the superior temporal sulcus in the rhesus monkey: a retrograde tracer study. J Comp Neurol 343:445– 463. Medline

Shinkareva SV, Malave VL, Mason RA, Mitchell TM, Just MA (2011) Commonality of neural representations of words and pictures. Neuroimage 54:2418 –2425. CrossRef Medline

Smith SM (2002) Fast robust automated brain extraction. Hum Brain Mapp 17:143–155. CrossRef Medline

Smith SM, Jenkinson M, Woolrich MW, Beckmann CF, Behrens TEJ, Johansen-Berg H, Bannister PR, De Luca M, Drobnjak I, Flitney DE, Niazy RK, Saunders J, Vickers J, Zhang Y, De Stefano N, Brady JM, Matthews PM (2004) Advances in functional and structural MR image analysis and implementation as FSL. Neuroimage 23 [Suppl 1]:S208 –S219. CrossRef Medline

Stokes M, Thompson R, Cusack R, Duncan J (2009) Top-down activation of shape-specific population codes in visual cortex during mental imagery. J Neurosci 29:1565–1572. CrossRef Medline

Vetter P, Smith FW, Muckli L (2011) Decoding natural sounds in early visual cortex. J Vis 11:779. CrossRef

Woolrich MW, Ripley BD, Brady M, Smith SM (2001) Temporal autocorrelation in univariate linear modeling of FMRI data. Neuroimage 14: 1370 –1386. CrossRef Medline

Zatorre RJ, Halpern AR (2005) Mental concerts: musical imagery and auditory cortex. Neuron 47:9 –12. CrossRef Medline