Neuron, Vol. 41, 809–823, March 4, 2004, Copyright ©2004 by Cell Press

Integration of Auditory and Visual Information about Objects in Superior Temporal Sulcus

Michael S. Beauchamp,* Kathryn E. Lee,
Brenna D. Argall, and Alex Martin
Laboratory of Brain and Cognition
National Institute of Mental Health
Bethesda, Maryland 20892

Summary

Two categories of objects in the environment—animals and man-made manipulable objects (tools)—are easily recognized by either their auditory or visual features. Although these features differ across modalities, the brain integrates them into a coherent percept. In three separate fMRI experiments, posterior superior temporal sulcus and middle temporal gyrus (pSTS/MTG) fulfilled objective criteria for an integration site. pSTS/MTG showed signal increases in response to either auditory or visual stimuli and responded more to auditory or visual objects than to meaningless (but complex) control stimuli. pSTS/MTG showed an enhanced response when auditory and visual object features were presented together, relative to presentation in a single modality. Finally, pSTS/MTG responded more to object identification than to other components of the behavioral task. We suggest that pSTS/MTG is specialized for integrating different types of information both within modalities (e.g., visual form, visual motion) and across modalities (auditory and visual).

Introduction

A central question in cognitive neuroscience is how the brain integrates information from multiple modalities. The sensations produced by the “meow” of a cat or by its photograph are completely different, yet stimuli in either modality lead to fast and efficient object identification (Stein and Meredith, 1993). Animals and manipulable man-made objects (such as a telephone) provide ideal stimulus sets for examining this integration process because these objects often have distinct visual and auditory features.

Spurred by evidence from neuropsychological testing of lesioned patients, fMRI studies of visually presented objects have shown that different categories of visual objects activate different regions of visual association cortex in occipital and temporal lobes (Beauchamp et al., 2002, 2003; Chao et al., 1999; Haxby et al., 1999; Kanwisher et al., 1997; Levy et al., 2001; Puce et al., 1996). Ventral temporal cortex responds to the form, color, and texture of objects, while lateral temporal cortex is especially responsive to the motion of objects (Beauchamp et al., 2002; for review, see Martin and Chao, 2001; Puce et al., 1998). Much less is known about cortical processing of objects presented in the auditory modality or about the integration of auditory and visual object information.

Cortical auditory processing begins in core areas of auditory cortex, located in the transverse gyrus of Heschl on the dorsal surface of the temporal lobe in the planum temporale. Anatomical and single-unit recording studies in nonhuman primates and functional neuroimaging studies in humans have shown that core areas are surrounded by belt and parabelt areas that are specialized for processing more complex aspects of auditory stimuli (Belin et al., 2000; Kaas and Hackett, 2000; Rauschecker, 1997; Rauschecker et al., 1995; Tian et al., 2001; Wessinger et al., 2001; Zatorre and Belin, 2001; Zatorre et al., 2002). We hypothesized that auditory-visual integration of complex objects might occur in midtemporal cortex, between auditory association cortex in the superior temporal gyrus (STG) and visual association cortex in posterior lateral temporal cortex. In monkeys, neurons in the superior temporal polysensory area (STP) respond to simple auditory and visual stimuli (Benevento et al., 1977), sometimes showing selectivity for the conjunction of complex auditory and visual stimuli (Bruce et al., 1981). Recent evidence from metabolic imaging studies suggests a large area of overlap between auditory and visually responsive cortex in the fundus and upper bank of the superior temporal sulcus (Poremba et al., 2003). In humans, temporal cortex is thought to be a site for heteromodal integration (Mesulam, 1998), and some human functional imaging studies of multimodal processing have reported multimodal responses in STS (reviewed in Calvert, 2001).

Functional neuroimaging of multimodal processing presents some unexpected challenges. For instance, defining the expected form of multimodal responses is not straightforward. Three general approaches have been used (Calvert, 2001). One approach is to search for areas that are responsive only to multimodal stimuli (e.g., auditory and visual together) and not to unimodal stimuli (e.g., auditory or visual alone). Across studies, this approach was not successful in identifying multimodal areas, likely because it is overly stringent: if areas responding to multimodal stimuli show some response to unimodal stimuli, they will not be identified. A second approach presents unimodal stimuli in isolation and classifies areas that respond during each modality as being multimodal (conjunction analysis). This approach succeeds in identifying potential multimodal brain regions, but may be too liberal: any region responding across conditions (not necessarily related to sensory processing) will be classified as multimodal. In a third approach, regions are classified as multimodal if they display an interaction between the response to unimodal stimulation and multimodal stimulation. For instance, if the response to combined auditory-visual stimulation is greater than the summed responses to unimodal auditory and visual stimulation, this is defined as a positive interaction effect, while if the summed responses are less than the multimodal response, this is defined as a negative interaction effect (Calvert et al., 2000). One difficulty with this expression of the interaction test is that it is not suitable for experiments in which the subject is performing a behavioral task, which is crucial for well-

*Correspondence: mbeauchamp@nih.gov

controlled imaging experiments. Regions involved in the behavioral task (such as motor cortex, if subjects make a motor response to the stimulus) are expected to be equally active during auditory, visual, and auditory-visual conditions. However, this means that they will display a negative interaction effect as defined by Calvert. Therefore, in our analysis we modify the interaction test to find those areas that show a greater response during auditory-visual stimulation than the mean response during unimodal auditory and visual stimulation.

A second hurdle to applying this approach to fMRI data is the relatively small amplitude of the interaction effect. In the face of the many thousands of multiple comparisons across the voxels in the brain volume, it is difficult to distinguish significant interaction effects from false positives. Therefore, we used an approach that has successfully detected category-related activity in visual regions (Haxby et al., 1999). We first find only those voxels showing a significant experimental effect (significant response to any experimental condition) using a high threshold ($p < 10^{-6}$) to account for the thousands of multiple comparisons across brain voxels. Then, within the much smaller pool of voxels showing an experimental effect, we use a more liberal threshold ($p < 0.05$) to search for voxels that respond positively to visual and auditory stimuli in isolation and show significantly more activity for simultaneous auditory-visual stimuli than for either modality alone.

An additional difficulty in most previous neuroimaging studies of multimodal processing is their reliance on group activation maps (Calvert et al., 1999, 2000). While averaging across subjects to create group maps increases statistical power, it may also lead to erroneous inferences. The normalization procedures (such as Talairach transformation) used for averaging across subjects align subjects based on anatomical, not functional, landmarks. This is problematic if the same anatomical location in different subjects has different functional properties, due to intersubject variability. For instance, if a particular anatomical location responds to auditory but not visual stimuli in some subjects and responds to visual but not auditory stimuli in other subjects, the region may appear to respond to both auditory and visual stimuli in an average activation map. To avoid this problem, we used an experimental design that permitted sufficient statistical power to detect effects in individual subjects. With single subject activation maps in hand, we were able to accurately locate multimodal activity in relation to sulcal and gyral anatomy by mapping activity to cortical surface models of each individual subject (Fischl et al., 1999b).

To summarize the conceptual framework of our experiments, we used criteria adapted from previous multimodal experiments to identify regions important for integrating auditory and visual information about complex objects. First, these areas should show positive responses to both auditory and visual representations of objects. Second, they should respond more to auditory or visual representations of real objects than to meaningless controls. Third, they should show an interaction effect with a stronger response to multimodal versus unimodal stimulation. Fourth, they should show a strong correlation with object identification—occurring soon after sensory stimulation—rather than with the behavioral task performed by the subject (such as a motor response). Finally, these properties should be demonstrated within individual subjects. To find brain areas meeting these criteria, we performed three imaging experiments using visual and auditory objects chosen for their characteristic auditory and visual features: animals and man-made manipulable objects (tools).

Results

Experiment 1

In the first experiment, we measured blood oxygenation level-dependent (BOLD) responses while subjects ($n = 8$) performed a one-back same/different task to blocks of stimuli. Within each block, a single type of stimulus was presented, in either the visual or auditory modality. Visual stimuli consisted of black-and-white photographs of tools, animals, or phase-scrambled photographs and auditory stimuli consisted of recordings of tools, animals, or synthesized ripple sounds (Figures 1A and 1B). Mean reaction time (RT) across stimuli was 1245 ms, with high accuracy (90%). RTs for auditory stimuli were significantly slower than for visual stimuli (1396 ms versus 1094 ms, $p < 10^{-6}$).

A number of brain regions showed a significantly greater BOLD signal during auditory or visual stimulation blocks than during fixation baseline (experimental effect, $p < 10^{-6}$). These regions were separated into three groups. Areas with greater BOLD signal during visual ($p < 0.05$) but not auditory ($p > 0.05$) blocks were located in occipital, ventral temporal, and posterior lateral temporal cortex (Figure 1C). A second set of areas was active for auditory but not visual blocks (Figure 1D). These areas were centered on Heschl's gyrus but extended anteriorly and posteriorly along the planum temporale to cover most of STG as well as into inferior frontal cortex. A third set of areas, including pSTS/MTG, dorsolateral prefrontal cortex (DLPFC), motor cortex, and ventral temporal cortex, was active during both auditory and visual blocks (Figure 1E).

Single-subject analysis confirmed that pSTS/MTG responded to both auditory and visual conditions in each individual subject and hence could not be attributed to artifacts introduced by sterotaxic normalization and group averaging. To more accurately locate the candidate multimodal region, surface models were created from three individual subjects (Figure 1F). Multimodal activation in lateral temporal cortex (white circles) was centered on the lower bank of the STS extending onto the crown of the MTG.

To examine the time course of activity, we constructed five regions of interest (ROIs) whose locations are shown in Table 1 and as white circles in Figures 1C–1E (see Experimental Procedures for details). Average MR time series from the five ROIs are shown in Figure 1G. The visual cortex ROI showed an increased BOLD signal relative to baseline during visual blocks, but decreased BOLD signal during auditory blocks. The auditory cortex ROI showed the opposite pattern, with MR signal below fixation baseline during visual blocks and large positive BOLD responses during auditory blocks. Among regions that responded to both auditory and visual blocks, differing responses to meaningful and meaningless stimuli

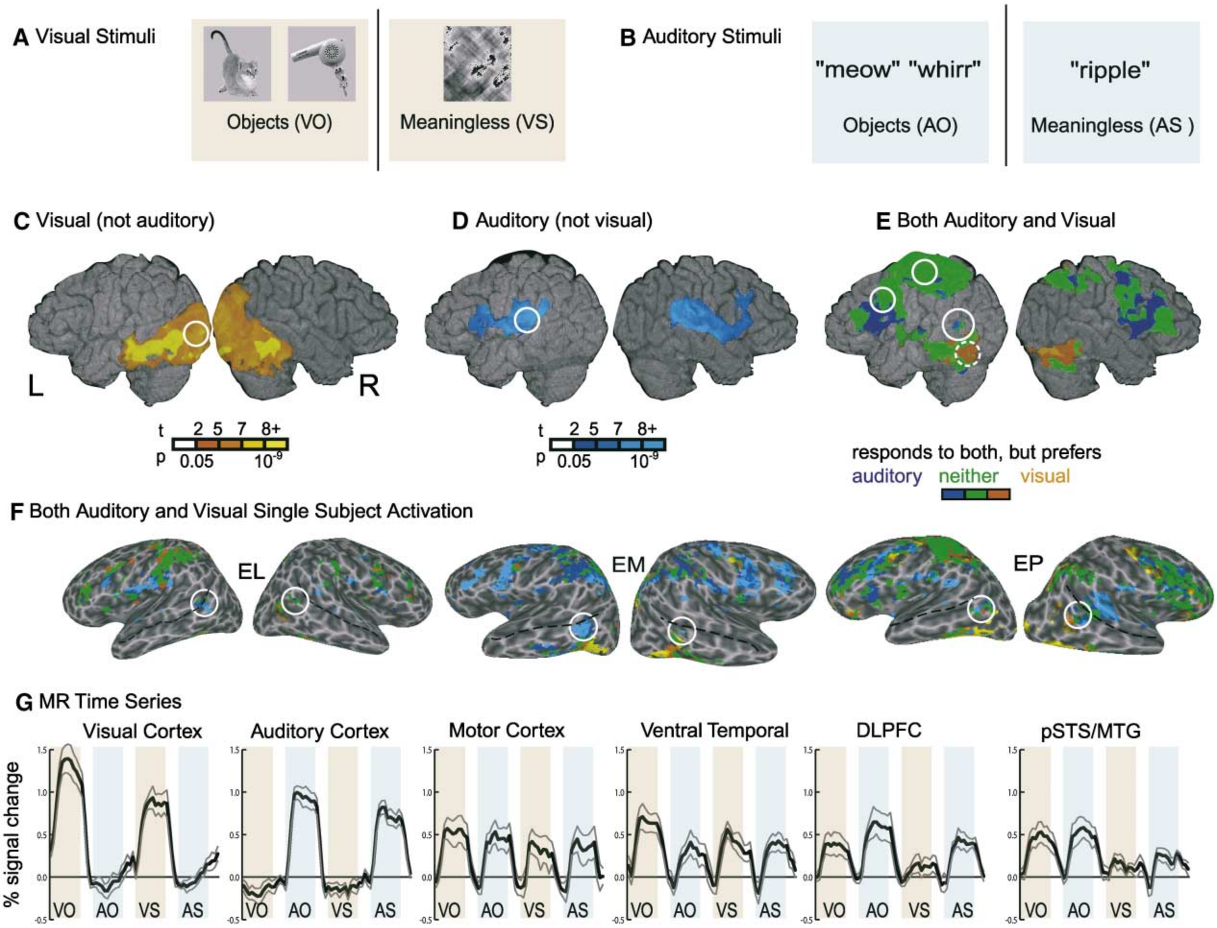


Figure 1. Stimuli and fMRI Activation from Experiment 1

(A) Visual stimuli consisted of photographs of animals and man-made manipulable objects (visual objects, VO) or meaningless scrambled photographs (visual scrambled, VS).

(B) Auditory stimuli consisted of recordings of animal and tool sounds (auditory objects, AO) or meaningless synthesized ripple sounds (auditory synthesized, AS).

(C) Brain areas active during visual but not auditory stimulation. Random-effects group map ($n = 8$) of brain regions showing a significant experimental effect ($p < 10^{-6}$) and active during visual ($p < 0.05$) but not auditory ($p > 0.05$) stimulation. Active voxels (colored) overlaid on a surface rendering of a single subject's high-resolution anatomical data set, lateral views of left (L) and right (R) hemisphere. Color scale shows significance of visual activation. White circle shows location of visual cortex ROI.

(D) Group map of brain areas active during auditory ($p < 0.05$) but not visual ($p > 0.05$) stimulation. Color scale shows significance of auditory activation. White circle shows location of auditory cortex ROI.

(E) Group map of brain areas active during both auditory and visual stimulation ($p < 0.05$). Color scale shows relative amplitude of auditory and visual activation. White circles show location (from anterior to posterior) of DLPFC, motor cortex, pSTS/MTG, and ventral temporal regions of interest (ROIs). Note that ventral temporal ROI actually sits on the ventral surface of the brain; dashed line shows position when projected onto the lateral surface.

(F) Cortical surface models of individual subject brain areas active during auditory and visual stimulation (both $p < 0.05$; same color scale as E). Dashed line shows fundus of STS; white circle shows location of pSTS/MTG multimodal region. Two-letter codes refer to experimental IDs of individual subjects.

(G) Mean time series across subjects from five brain regions (locations shown as white circles on activation maps in C–E). Central dark line shows mean MR time series, thin gray lines show ± standard error (SEM). Stimuli were presented in 21 s blocks (colored bars) followed by 9 s of fixation baseline (white interval between bars). Each block contained seven stimuli of a single type presented one at a time.

were observed. Motor cortex showed strong responses to auditory and visual blocks but was not modulated by meaning, as calculated with a repeated measures ANOVA across subjects (stimulus type as the repeated measure, subjects as replications). Ventral temporal cortex showed stronger visual compared with auditory responses ($p = 0.006$) and preferred meaningful visual objects to scrambled photographs ($p = 0.04$) but not meaningful object sounds to meaningless ripple sounds. DLPFC preferred meaningful visual stimuli ($p = 0.03$) but not meaningful sounds. pSTS/MTG was the only region that preferred real to scrambled visual stimuli ($p = 0.02$) and real to meaningless sounds ($p = 0.03$).

Experiment 2

In the second experiment, we directly tested the hypothesis that pSTS/MTG integrates auditory and visual information about complex objects by presenting auditory

Table 1. Active Regions across Experiments

	Peak Coordinates		
Anatomical Description	x	y	z
pSTS/MTG (BA 37, 19, 39)	−50	−55	7
Ventral temporal cortex (BA 37, 18, 19, 20)	−41	−44	−12
Dorsolateral prefrontal cortex (BA 6, 9, 8, 13)	−49	11	30
Motor cortex (BA 40, 3, 4, 2, 6, 7, 1, 5)	−40	−25	5
Visual cortex (BA 18, 19, 17, 37, 39)	−29	−86	0
Auditory cortex (BA 22, 13, 41, 40, 42, 43, 21, 6)	−41	−28	12

Coordinates are locations of peak significance in the group activation map, in standardized Talairach coordinates (mm). BA, Brodmann areas obtained from the San Antonio Talairach Demon (Lancaster et al., 2000). All BAs containing at least 50 active voxels are listed, ordered by number of active voxels (most-to-least).
These data are shown graphically in Figures 1–3.

and visual stimuli both in isolation (as in Experiment 1) and simultaneously. This allowed us to measure the interaction effect. Our hypothesis was that multimodal integration regions, like pSTS/MTG, should be more active when subjects are required to integrate auditory and visual information about objects than when information from a single modality is sufficient.

During separate blocks, subjects ($n = 7$) viewed line drawings of animals or man-made objects, heard the characteristic sounds of these items, or were presented with both the drawing and the sound (Figures 2A–2C). To ensure that subjects accurately identified the objects, they performed a semantic decision task. In auditory and visual blocks, subjects decided if the animal walked on four legs or not (e.g., sheep, true; bird, false) or if the tool needed electric power to operate (e.g., hair dryer, true; hammer, false). The mean RT across unimodal blocks was 1005 ms with an accuracy of 93%. Auditory RTs were significantly slower than visual RTs (1275 ms versus 735 ms, $p < 10^{-6}$). During auditory-visual blocks, subjects decided if the sound and line drawing of the object were congruent or incongruent (e.g., auditory "meow" + visual dog = incongruent). Auditory-visual RTs (mean RT, 1505 ms; accuracy, 87%) were significantly slower than auditory ($p = 0.001$) and visual ($p < 10^{-6}$) RTs, reflecting the more difficult task performed during auditory-visual blocks.

As in the first experiment, regions were classified as active based on a stringent experimental effect threshold ($p < 10^{-6}$), followed by separation into three groups based on their response to auditory or visual stimuli in isolation (threshold of $p < 0.05$). Regions responding to visual but not auditory stimulation (Figure 2D) were concentrated in occipital and temporal cortex. Auditory but not visual stimulation activated regions in and around Heschl's gyrus and inferior frontal cortex (Figure 2E). Regions that responded to both unimodal auditory and unimodal visual blocks were found in distributed frontal, parietal, and temporal regions (Figure 2F). Because auditory, visual, and auditory-visual objects were presented, we were able to construct an average activation map of regions showing an interaction effect, defined as an enhanced response to multimodal blocks (Figure 2G). This contrast revealed that pSTS/MTG, DLPFC, and ventral temporal cortex responded more strongly to auditory-visual blocks than to either auditory or visual blocks. Single-subject analysis confirmed that these regions showed an interaction effect in each individual subject (Figure 2H).

In order to calculate the amplitude and significance of the multimodal enhancement effect (defined as the response for auditory-visual blocks compared with the mean response for auditory and visual blocks), we selected regions of interest using the coordinates of peak responses to auditory and visual stimuli presented in isolation (Figures 2D–2F). This allows us to calculate the enhancement effect in an unbiased manner, since selecting voxels based on their multimodal response (Figure 2G) would bias the comparison. Time series from each ROI were averaged across subjects (shown in Figure 2I). Visual and auditory cortex showed no significant difference between multimodal stimulation and unimodal stimulation in their preferred modality. The response of motor cortex to the three conditions did not different significantly, while the response of pSTS/MTG, ventral temporal, and DLPFC to auditory-visual stimulation blocks was significantly greater than the mean response to unimodal blocks ($p = 0.04$, $p = 0.01$, $p = 0.01$). The response of pSTS/MTG to auditory-visual blocks was 39% greater than the average response to unimodal blocks.

Experiment 3

The enhanced multimodal responses observed in pSTS/MTG, DLPFC, and ventral temporal cortex in Experiment 2 might have been due to the more difficult behavioral task performed by subjects during multimodal blocks. To address this issue, in the third experiment, subjects again listened to, viewed, or simultaneously listened to and viewed objects, but performed the same behavioral task in all three conditions. In addition, an event-related design was used that allowed us to compare the amplitude of the BOLD response to object identification with the BOLD response to other elements of the behavioral task.

In the three trial types, subjects ($n = 8$) were presented with either a silent video clip of a tool moving, the sound produced by the tool, or the video and sound together (Figures 3A–3C). Then, after a 2 s delay, subjects chose the correct name of the item from a choice screen (Figures 4A–4C). Auditory RTs were significantly slower than visual RTs (1898 ms versus 1278 ms, $p = 0.01$), while multimodal RTs were intermediate (1472 ms). Subjects were least accurate for auditory stimuli (79%), more accurate for visual stimuli (92%), and most accurate for combined auditory-visual stimuli (94%).

As shown in Figures 4A–4C, the temporal structure of each trial allowed independent measurements of the BOLD signal triggered by object identification and the BOLD signal resulting from task components that occurred later in each trial (such as the motor response). The success of this strategy is shown in the average time series from different ROIs (Figures 4D and 4E). For example, the auditory cortex ROI responded during

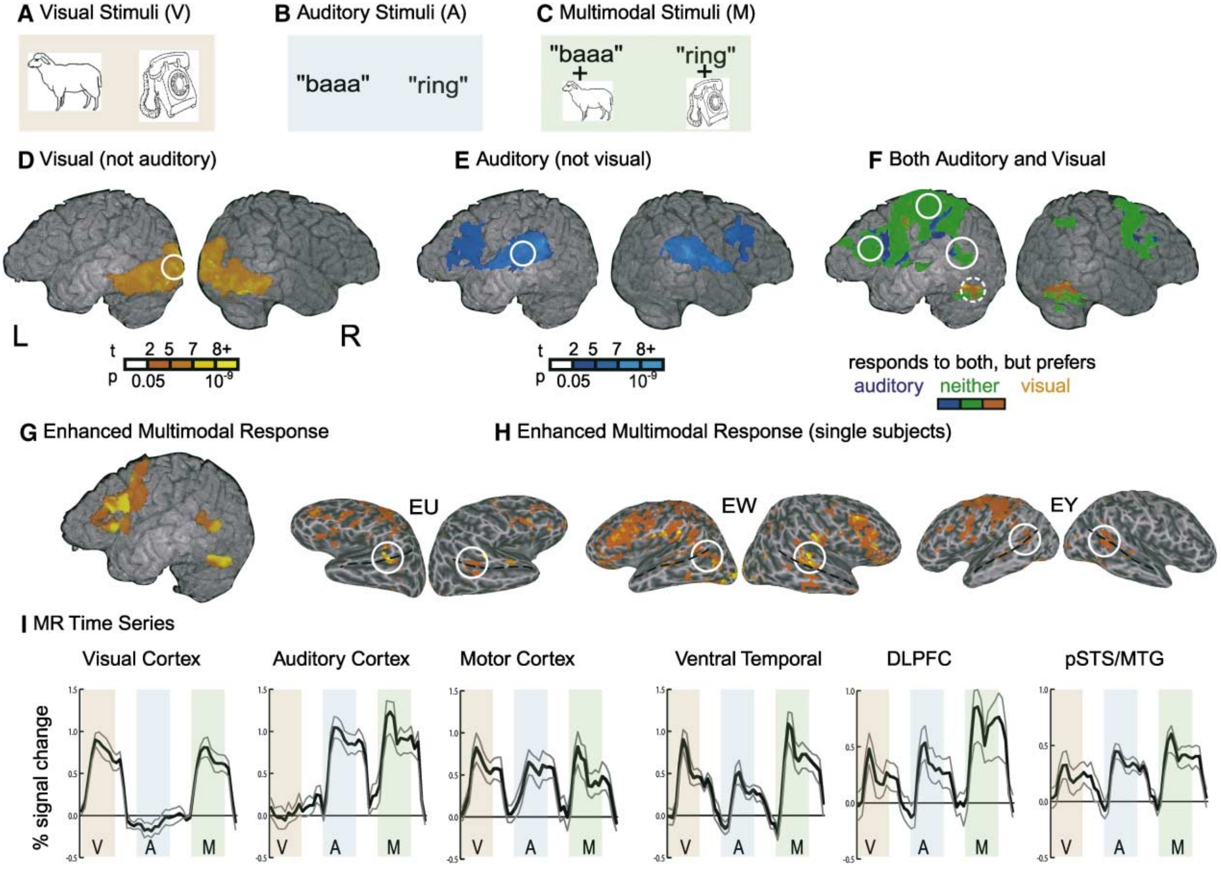


Figure 2. Stimuli and fMRI Activation from Experiment 2

(A) Visual stimuli consisted of line drawings of animals and man-made manipulable objects (tools).

(B) Auditory stimuli consisted of recordings of animal and tool sounds.

(C) Multimodal stimuli consisted of simultaneously presented line drawings and sounds from the same category (either animals or tools) that were either congruent (as shown: cat + "meow," telephone + "ring") or incongruent (not shown: e.g., cat line drawing + "woof," telephone + "bang-bang").

(D) Brain areas active during visual but not auditory object perception. Random-effects group map (n = 7) of brain regions showing a significant experimental effect ($p < 10^{-6}$) and active during visual ($p < 0.05$) but not auditory ($p > 0.05$) conditions. Active voxels (colored) overlaid on a surface rendering of a single subject's high-resolution anatomical data set, lateral views of left (L) and right (R) hemispheres. Color scale shows significance of visual activation. White circle shows location of visual cortex ROI.

(E) Areas active during auditory but not visual object presentation. Color scale shows significance of auditory activation. White circle shows location of auditory cortex ROI.

(F) Areas active during both auditory and visual object conditions (both $p < 0.05$). Color scale shows relative amplitude of auditory and visual activation. White circles show location (from anterior to posterior) of DLPFC, motor cortex, pSTS/MTG, and ventral temporal regions of interest (ROIs). Note that ventral temporal ROI actually sits on the ventral surface of the brain; dashed line shows position when projected onto the lateral surface.

(G) Areas showing an enhanced response during multimodal stimulation compared with the mean of auditory and visual stimulation ($p < 0.05$).

(H) Cortical surface models of individual subject brain areas showing an enhanced response during multimodal stimulation. Dashed line shows fundus of STS; white circle shows location of STS/MTG multimodal region. Two-letter codes refer to experimental IDs of individual subjects.

(I) Mean time series across subjects from five brain regions (locations shown as white circles on activation maps in D–F). Central dark line shows mean MR time series, thin gray lines show ± SEM. Stimuli were presented in 21 s blocks (colored bars) followed by 9 s of fixation baseline (white interval between bars). Each block contained seven objects presented in visual (yellow bar, V), auditory (blue bar, A), or simultaneous auditory-visual modalities (green bar, M).

auditory object presentation but not during the response phase of the trial, while the motor cortex ROI responded during the response phase but not during object presentation.

Figure 3 illustrates active cortical regions. As in Experiments 1 and 2, visual cortex was active during the stimulus phase of visual trials but not the stimulus phase of auditory trials, while auditory cortex showed the opposite pattern. A broad network of areas was active during both auditory and visual trials (Figure 3G), but the event-related design allowed us to functionally subdivide these areas. Motor cortex and posterior parietal cortex (Figure 3H) responded during the behavioral response phase of the trial but not during auditory or visual stimulus presentation ($p > 0.05$). Dorsolateral prefrontal cortex and parietal cortex (Figure 3I) responded to both the behavioral and stimulus phases of the trial but showed a stronger response to the behavioral phase ($p < 0.05$).

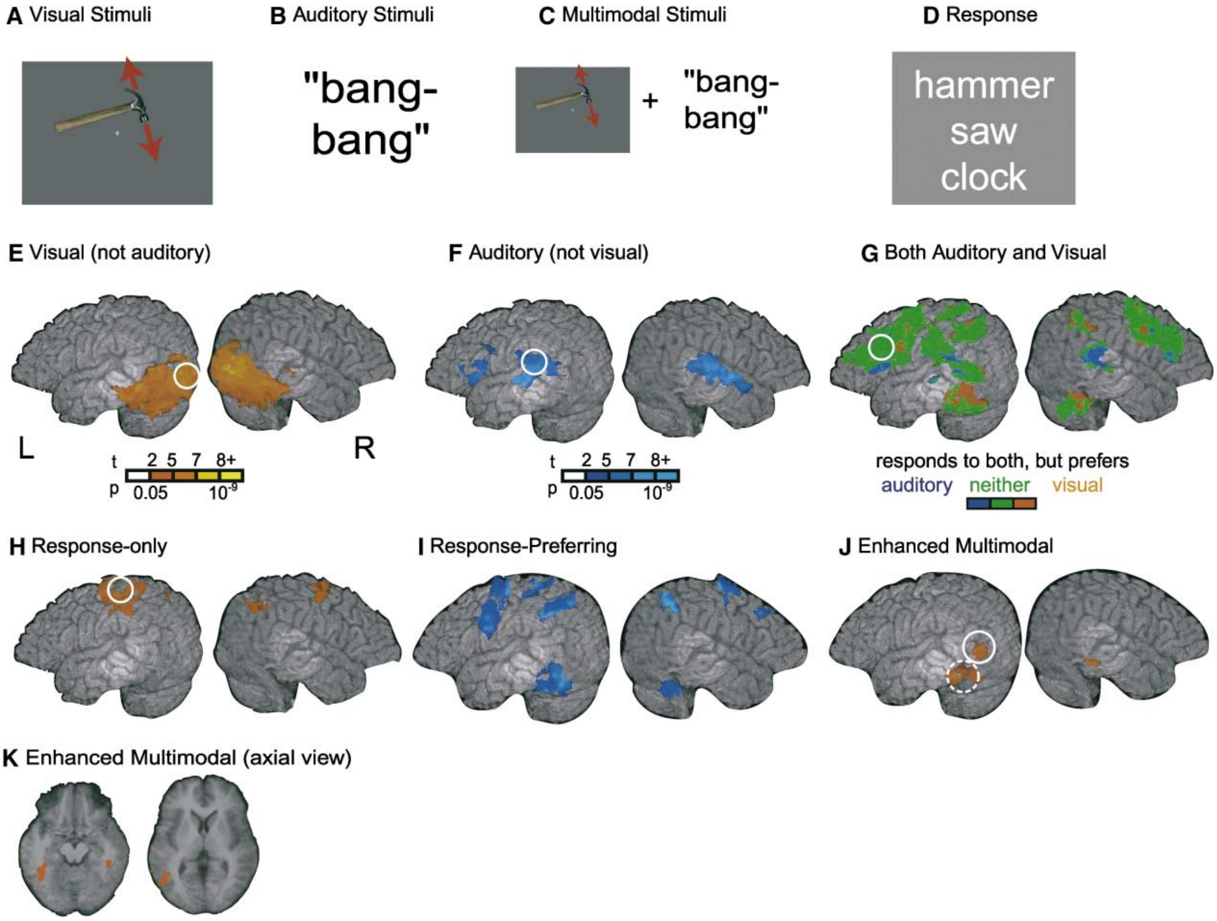


Figure 3. Stimuli and fMRI Activation Maps from Experiment 3

(A) Visual stimuli consisted of video clips of tools moving with their characteristic motion (red arrows, not present in actual display, illustrate direction of motion).

(B) Auditory stimuli consisted of recordings of tool sounds.

(C) Multimodal stimuli consisted of simultaneously presented video clip and sound from the same tool.

(D) The response screen consisted of three words presented along the horizontal meridian with a fixation square (words enlarged and displayed on multiple lines for illustration).

(E) Brain areas active during visual but not auditory object presentation. Random-effects group map ($n = 8$) of brain regions showing a significant experimental effect ($p < 10^{-6}$) and active during visual ($p < 0.05$) but not auditory ($p > 0.05$) stimulation. Active voxels (colored) overlaid on a surface rendering of a single subject's high-resolution anatomical data set, lateral views of left (L) and right (R) hemispheres. Color scale shows significance of visual activation. White circle shows location of visual cortex ROI.

(F) Brain areas active during auditory but not visual object presentation. Color scale shows significance of auditory activation. White circle shows location of auditory cortex ROI.

(G) Group map of brain areas active during auditory and visual object conditions (both $p < 0.05$). Color scale shows relative amplitude of auditory and visual activation. White circle shows location of DLPFC ROI.

(H) Brain areas active during motor responding but not auditory or visual conditions. White circle shows location of motor cortex ROI.

(I) Brain areas active during auditory and visual object conditions and during behavioral response, with greater activation during behavioral response.

(J) Brain areas active during both auditory and visual conditions, with enhanced multimodal versus unimodal response. White circle shows location of the pSTS/MTG ROI; dashed white circle shows location (projected onto lateral surface) of the ventral temporal ROI.

(K) Axial slices ($z = -12$ and $z = 7$) showing ventral temporal and pSTS/MTG activations visible in (J).

pSTS/MTG and ventral temporal cortex preferred the stimulus phase to the behavioral phase ($p < 0.05$) and showed an enhanced response to multimodal stimuli, defined as the difference between the response to combined auditory-visual stimuli and the mean response across unimodal stimuli (Figures 3J and 3K).

MR time series were created for each region, averaged across subjects, illustrating the response to the three trial types (Figures 4D and 4E). Visual cortex was deactivated during auditory stimulation (relative to fixation baseline) but responded similarly during visual stimulation and multimodal stimulation. Visual cortex also showed a moderate level of activity during the response period for all three trial types, since the response period always contained a visual display consisting of three words. In auditory cortex, the auditory and multimodal stimulus conditions evoked a large positive response, while visual stimuli produced a slight deactivation.

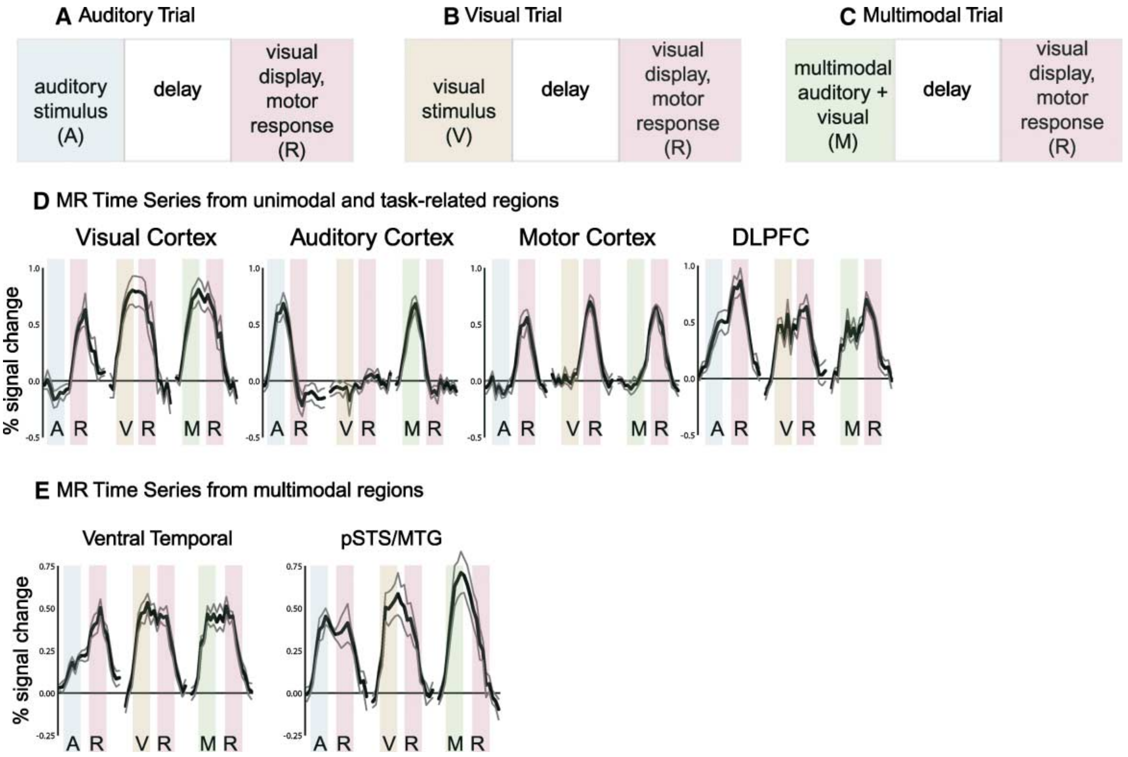


Figure 4. Details of Trial Structure and Average MR Responses from Experiment 3

(A) Each auditory trial consisted of a 2.5 s auditory stimulus (A, blue bar) followed by a 2.5 s delay followed by a 3 s response period (R, purple bar). Stimuli are illustrated in Figure 3.

(B) Visual trial consisted of a visual stimulus (yellow bar, V) followed by delay and response periods.

(C) Multimodal trials consisted of a simultaneous auditory and visual stimulus (green bar, M) followed by delay and response periods.

(D) Mean time series across subjects from four brain regions (locations shown as white circles on activation maps in Figures 3E–3H). Central dark line shows mean MR time series during each trial type, thin gray lines show ± SEM. Colored bars show approximate time of peak BOLD signal (shifted to account for the hemodynamic response lag) to the stimulus (A, V, or M) and response (R) phases of each trial type.

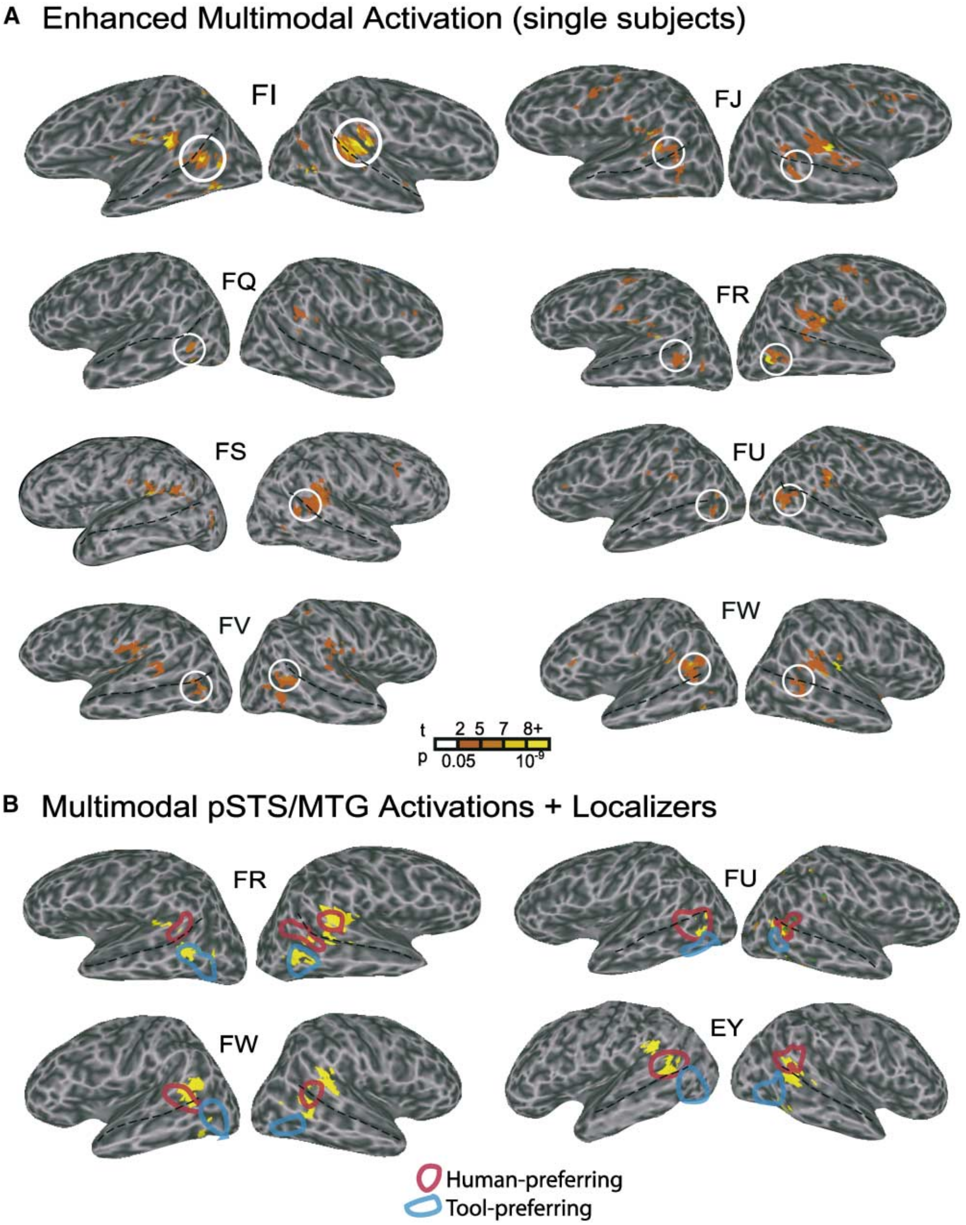
(E) Mean MR time series from two regions showing greater response during simultaneous auditory-visual object presentation compared with auditory or visual object presentation alone (locations shown as white circles on activation maps in Figure 3J).

DLPFC responded during auditory and visual stimulation but showed even greater activity during the task-response phase of the trial. DLPFC showed the greatest activity during the response phase of auditory trials, which were the most difficult trials as measured by RT and percent correct, suggesting that DLPFC was driven primarily by task demands.

In contrast to the DLPFC, pSTS/MTG responded more to the stimulus phase of the trial than to the task response phase and showed greater responses to multimodal trials than auditory trials (even though auditory trials were more difficult), suggesting that pSTS/MTG was driven primarily by object perception and auditory-visual integration. Like pSTS/MTG, ventral temporal cortex responded more to the stimulus phase of the trial than the task phase, but ventral temporal responses were predominantly visual, with only weak positive responses during auditory stimulation. Ventral temporal cortex also did not demonstrate the multimodal enhancement effect observed in pSTS/MTG. In an ANOVA with each subject as a replication, ventral temporal cortex responded similarly during visual and multimodal stimulus periods ($p = 0.22$), while pSTS/MTG responded 36% more during auditory-visual stimuli than during the average of unimodal stimuli. This response was significantly greater than either auditory stimulation alone ($p = 0.01$) or visual stimulation alone ($p = 0.02$). An additional measure of multimodal enhancement was calculated as the ratio between the response to the auditory-visual stimulus and the maximum response to unimodal stimulation (calculated for each subject and then averaged across subjects). This ratio was 1.06 for ventral temporal cortex (not significantly different from 1), while the enhancement ratio was 1.14 for pSTS/MTG ($p < 0.05$). This can be observed in Figure 4E, with the amplitude of the multimodal response in ventral temporal cortex approximately equal to the maximum unimodal (visual) response, while in pSTS/MTG the peak of the multimodal response is significantly greater than the visual response. For additional discussion of different methods of calculating multimodal enhancement, please see Supplemental Data, Section 2 at http://www.neuron.org/cgi/content/full/41/5/809/DC1.

Anatomical Relationship between Multimodal and Category-Related Activity

Figure 5A illustrates individual subject activation maps created on a model of each subject's cortical surface. While the exact anatomical location of the multimodal region varied in each subject, we consistently observed


A Enhanced Multimodal Activation (single subjects)

B Multimodal pSTS/MTG Activations + Localizers

Figure 5. Single-Subject Activations from Experiment 3 Mapped to the Cortical Surface

(A) Color scale indicates significance of multimodal enhancement (simultaneous auditory-visual versus auditory alone + visual alone) in each subject (identity shown by two-letter code). Dashed line indicates fundus of STS; white circle indicates location of pSTS/MTG multimodal region.
(B) Relationship of pSTS/MTG multimodal region to lateral temporal regions preferring moving human or tool stimuli. pSTS/MTG voxels showing enhanced multimodal response in yellow. Red line indicates boundary of human motion-preferring cortex; blue line indicates boundary of tool motion-preferring cortex.

a region of pSTS/MTG that responded to both auditory and visual stimuli and showed an enhanced multimodal response.

In a previous study, we demonstrated that videos of moving humans and tools evoked differential responses in regions of lateral temporal cortex (Beauchamp et al., 2002). STS (especially in the right hemisphere) showed stronger responses to human videos than to tool videos, while MTG (especially in the left hemisphere) showed stronger responses to tool videos. To relate our previous findings to the current study, we first examined the degree of laterality in the multimodal pSTS/MTG region. While most subjects showed multimodal pSTS/MTG activity in both hemispheres (Figure 5A), the average volume of active cortex was greater in right than left hemispheres (5881 versus 4137 mm^3, $p = 0.04$). To more directly test the relationship of human- and tool-preferring areas to multimodal cortex, we used the procedures from Beauchamp et al. (2002) to map human/tool regions in three subjects from the current study. Multimodal regions were located near the anterior portion of the human/tool regions, with the STS portion of the multimodal activity overlapping human-preferring cortex (Figure 5B). To quantify this overlap, we calculated the percentage of multimodal pSTS/MTG that responded more strongly to human or tool videos ($p < 0.05$) in three subjects. In the left multimodal region, 21% of voxels preferred human videos and 16% preferred tool videos (the remainder showed no preference). In the right hemisphere multimodal region, 35% of voxels preferred human videos and 4% preferred tool videos. Thus, the majority of the voxels (63% in left, 61% in right hemisphere) showed a multimodal response without a significant object category preference.

Discussion

Across three experiments in which subjects identified a variety of auditory, visual, and auditory-visual complex objects, pSTS/MTG matched objective criteria for a multimodal integration region. In each experiment, pSTS/MTG responded with an increased BOLD signal to both auditory and visual stimuli compared with fixation baseline, in contrast with adjacent auditory and visual cortex ROIs, in which the BOLD signal decreased below baseline to stimuli in the nonpreferred modality. In the first experiment, pSTS/MTG responded more to meaningful stimuli than to meaningless stimuli (real versus scrambled pictures; real sounds versus ripples). An enhanced response to multimodal compared with unimodal stimuli is a hallmark of regions performing sensory integration (Stein and Meredith, 1993), and in the second and third experiments, pSTS/MTG showed an interaction effect, responding more when auditory and visual object features were presented together than when they were presented in isolation. In the third experiment, pSTS/MTG (unlike other brain regions) showed a greater response during object identification than during later components of the behavioral task. These results provide strong evidence that pSTS/MTG is an important site for integrating auditory and visual information about complex objects.

Relationship to Visual and Auditory Association Areas

Consistent with previous neuroimaging studies, an area in posterior lateral occipital cortex known as area LO (Lerner et al., 2001; Malach et al., 1995) responded more to photographs of animals or tools than to scrambled stimuli (Figure 6A). Neuroimaging studies have also shown that regions of human STS show strong responses to biological stimuli, such as faces, animals, or human bodies (Allison et al., 2000; Chao et al., 1999; Haxby et al., 1999; Kanwisher et al., 1997; Puce et al., 1995) and prefer these items to tools, while regions of middle temporal gyrus (directly inferior to STS) respond more to manipulable objects than to biological stimuli (Beauchamp et al., 2002; Chao et al., 1999; Devlin et al., 2002; Martin et al., 1996). The multimodal pSTS/MTG region described in the current study lies near the boundary of these category-related visual responses. However, most of the multimodal voxels in pSTS/MTG did not show a significant category preference. Comparing their relative location (Figure 5B) suggests that regions important for integrating visual form and motion are located close to, but not overlapping, regions that integrate visual and auditory information.

The animal and tool sounds used in these experiments have complex spectral and temporal characteristics previously shown to activate multiple auditory areas along the STG (Hall et al., 2002; Rauschecker and Tian, 2000; Scott et al., 2000; Seifritz et al., 2002; Wessinger et al., 2001; Zatorre and Belin, 2001). As with previous studies of environmental sounds (Engelien et al., 1995; Maeder et al., 2001), we observed auditory activation along a significant fraction of anterior and posterior STG and STS, extending into MTG (Figure 6B). Recent research suggests that cortical auditory processing is divided into separate processing streams (Rauschecker and Tian, 2000). Posterior temporo-parietal regions, labeled the "where" or "how" stream, may be specialized for processing sound motion and location (Baumgart et al., 1999; Bushara et al., 1999; Griffiths et al., 1996; Lewis et al., 2000; Recanzone et al., 2000; Tian et al., 2001; Warren et al., 2002; Zatorre et al., 2002) Regions anterior and ventral to primary auditory cortex, labeled the "what" stream, may be specialized for processing characteristic auditory features (Alain et al., 2001; Belin et al., 2000; Scott et al., 2000; Binder et al., 2000; Tian et al., 2001). Lesioned patients with deficits in environmental sound processing have damage to STG, STS, or MTG (Clarke et al., 2000, 2002).

While we observed auditory responses in mid to anterior temporal cortex (the putative auditory "what" stream), multimodal responses were found posteriorly, in pSTS/MTG. This finding is consistent with a study of 30 aphasic patients (Saygin et al., 2003) that examined the relationship between brain lesions and the ability to process environmental sounds. Using a task in which patients made judgments about pictures of objects and their associated sounds, Saygin et al. found that the areas of maximal overlap for patients specifically impaired in this task were centered in the posterior superior temporal gyrus extending into middle temporal regions. Similarly, a PET study found that identification of animals from their characteristic sound evoked greater activity than a pitch discrimination task in ventral temporal cortex and pSTS/MTG, corresponding to the foci observed in the present study (Tranel et al., 2003). One possible explanation for these findings is that information from the auditory "what" stream is relayed both in an anterior direction and in a posterior direction, where it meets visual association regions in pSTS/MTG (Tian et al., 2001).

While the evidence suggests that pSTS/MTG plays a crucial role in integrating auditory-visual information about complex objects, this region is not likely to be important for all tasks and stimuli involving integration across modalities. Instead, the areas involved in integration will depend both on the stimulus and the behavioral task. For instance, the auditory and visual spatial processing (or "where") streams converge in parietal cortex, and enhanced activity is observed in intraparietal sulcus when subjects make fine discriminations about the relative speeds of auditory and visual moving objects (Lewis et al., 2000). In a second example, a region within the lateral occipital visual object recognition complex (LOtv) responds as strongly to tactile manipulation of objects as to visual presentation of objects (Amedi et al., 2001,

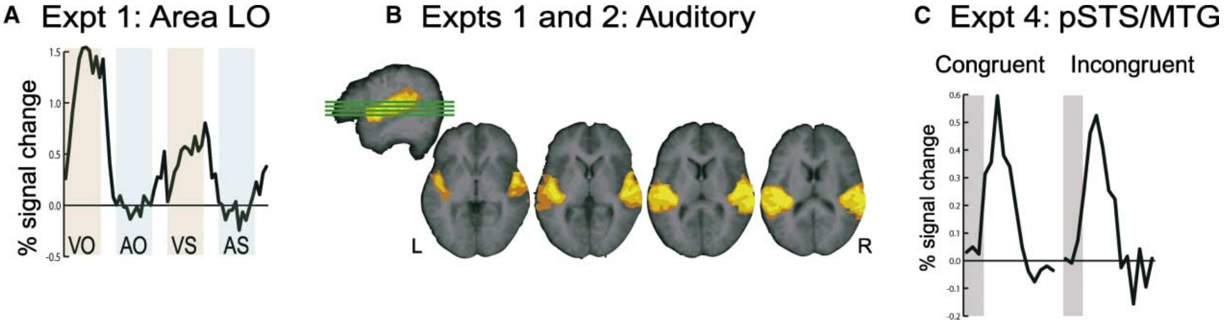


Figure 6. Additional fMRI Data

(A) Average time series from voxels in Experiment 1 showing preference for real compared with scrambled visual stimuli. Area LO in lateral occipital cortex, coordinates (−40, −88, 2). Note the enhanced response for photographs of objects (VO) compared with scrambled photographs (VS). All details as in Figure 1.

(B) Average activation map (n = 14) showing auditory-only activation from Experiments 1 and 2 (all auditory stimuli versus fixation excluding regions responding to visual stimuli versus fixation). Color scale indicates functional activity (as in Figures 1 and 2) overlaid on an average anatomical data set. Four axial slice planes (z = −5, 0, 5, 10) corresponding to green lines through left-most image (parasagittal section, x = 52 mm). Left is left in all slices.

(C) Average response from pSTS/MTG in Experiment 4 to a single presentation of a congruent auditory-visual stimulus (e.g., hammer video + "bang-bang-bang") and an incongruent stimulus (e.g., saw video + "bang-bang-bang"). Gray bar shows 2.5 s stimulus duration.

2002; James et al., 2002), suggesting that this area codes for 3-dimensional shape regardless of modality. However, the auditory modality contributes relatively little to the perception of fine details of three-dimensional object shape, and auditory stimuli do not activate this area (current study; Amedi et al., 2002). Temporal regions anterior to pSTS/MTG may be important for auditory-visual integration for stimuli other than complex objects. Belin et al. (2002) describe multiple foci of activity in response to human voices along the anterior to posterior extent of STS. Visually presented human faces, especially of familiar individuals, evoke anterior temporal responses (reviewed in Haxby et al., 2002). Therefore, we speculate that multimodal activation in anterior STS would be observed if subjects judged whether a voice matched the face of a familiar individual.

Functional Role of Auditory-Visual Integration in pSTS/MTG

Given that the brain regions important for multimodal integration depend on the nature of the stimuli and task, what precisely is the functional specialization of the pSTS/MTG multimodal region? While the present study presented complex objects, it seems unlikely that pSTS/MTG is specialized for processing only this class of stimuli. Most previous imaging studies that demonstrated multimodal activity in STS used linguistic stimuli (reviewed in Calvert, 2001). Calvert et al. used videotapes of actors speaking and recordings of voices (Calvert et al., 2000), while Raij et al. used visually presented letters and auditory phonemes (Raij et al., 2000). Given the limits of comparing locations across different neuroimaging techniques, the stereotaxic coordinates of our pSTS/MTG multimodal activation are similar to those reported in previous studies. Therefore, the pSTS/MTG region that we report is probably not specialized solely for integrating auditory and visual information about complex objects, but rather has a more general role in auditory-visual integration.

One possibility is that multimodal responses in pSTS/MTG reflect the formation of associations between auditory and visual features that represent the same object. Evidence from monkey single-unit recording experiments suggest that temporal lobe neurons rapidly form associations between paired visual stimuli, corresponding to the animal's learning of the association (Erickson and Desimone, 1999; Messinger et al., 2001; Naya et al., 2003). Because neurons in temporal cortex are both highly sensitive to stimulus differences and plastic enough to form associations between very different stimuli, they have properties suited for performing associations between the auditory and visual features of objects that generalize across low-level stimulus differences (Naya et al., 2001; Tanaka, 2003). In monkeys, the likely homolog of pSTS/MTG is known as STP (superior temporal polysensory) or TPO (temporal-parietal-occipital) and receives substantial projections from auditory and visual association cortex (Seltzer et al., 1996). In sum, the anatomical location of pSTS/MTG between high-level auditory and visual cortices (as well as the response properties of temporal neurons) renders it well situated to make links between auditory and visual object features.

pSTS/MTG may also be important for integrating different types of information within the visual modality. Visual processing takes place in anatomically distinct streams, often characterized as the ventral "what" pathway and the dorsal "where" pathway (Ungerleider and Mishkin, 1982). Just as the association between auditory and visual features corresponding to the same object must be learned, different visual features corresponding to the same object must also become associated. For instance, the brain must learn through experience the correspondence between the form of an object and its motion (for example, hammers typically move in an up and down direction while saws typically move back and forth). Single neurons in STS responded both to the form of a visual stimulus and to its direction of movement (Oram and Perrett, 1996). Evidence from neuroimaging suggests that human STS also integrates visual form and visual motion (Beauchamp et al., 2003; Puce et al., 2003). Therefore, pSTS/MTG may serve as a general-purpose association device both within and across modalities.

Some studies of multimodal integration have found a much larger response when auditory and visual stimuli are congruent than when they are incongruent (Calvert et al., 2001). In Experiment 2 of the present manuscript, subjects viewed congruent and incongruent multimodal stimuli, but the two types were mixed together within single experimental blocks, meaning that the BOLD response to each type could not be independently estimated. In Experiment 3, an event-related design was used (allowing analysis of the response to single trials), but the stimuli were always congruent. Therefore, we performed an additional fMRI experiment (described as Experiment 4 in Experimental Procedures) in order to estimate the congruency effect for object stimuli in pSTS/MTG. Videos of tools and recordings of tools were presented simultaneously to the subject, but the stimuli were either congruent (e.g., recording of saw, video of saw) or incongruent (e.g., recording of saw, video of hammer). An event-related design was used to allow random ordering and independent estimation of the response to each stimulus type as subjects ($n = 5$) made a congruent versus incongruent decision. pSTS/MTG showed strong responses to both types of multimodal stimuli (Figure 6C) and showed a trend toward greater responses for congruent than incongruent stimuli (peak response, 0.60% MR signal increase versus 0.52%, $p = 0.07$). This shows that pSTS/MTG is sensitive to the congruency of auditory and visual object stimuli (emphasizing its involvement in multimodal processing for these stimuli). However, the relatively weak effect suggests that congruency is not the primary way in which auditory-visual stimuli are encoded in pSTS/MTG.

Other Multimodal Regions: DLPFC and Ventral Temporal Cortex

DLPFC was active during visual and auditory tasks in all three experiments, but the amplitude of its response corresponded more to the cognitive demands of the task than to the degree of sensory integration. This is entirely consistent with single-unit recording, lesion, and imaging studies that place DLPFC as the locus for the cognitive processes underlying task performance, such as working memory (Goldman-Rakic, 1999). In Experiment 1, auditory stimuli were significantly more difficult to recognize than visual stimuli and DLPFC showed the greatest response during auditory blocks. In Experiment 2, the multimodal task was more difficult than the visual or auditory tasks, and DLPFC responded most during multimodal blocks. In Experiment 3, DLPFC showed the largest response during auditory trials (the most difficult trial type), and across all trial types, responded more to the behavioral task than to object identification. These data are consistent with studies showing strong effects of task demand on DLPFC (Braver et al., 1997; Carpenter et al., 1999). In addition to task difficulty, in our experiments the retrieval of semantic information about the objects from long-term memory also likely contributed to DLPFC activity (Thompson-Schill, 2003; Wagner et al., 1999). Our focus of peak activation in inferior DLPFC was similar to that found in a previous fMRI study requiring subjects to name auditory or visually presented objects (Adams and Janata, 2002; Buckner et al., 2000). In the present study, auditory activations in frontal cortex (Figures 1D, 2E, and 2F) were concentrated in inferior regions of DLPFC, also consistent with studies in nonhuman primates that demonstrate a projection of the auditory "what" stream to inferior portions of DLPFC (Romanski et al., 1999).

Ventral temporal cortex showed a weak response to auditory objects but a strong response to visual objects, consistent with its location in the ventral visual stream. Other studies have reported responses to auditory stimuli, such as words, in similar ventral temporal sites (Buchel et al., 1998; Petersen and Fiez, 1993). One possibility is that neurons in this region respond directly to auditory and visual sensory stimuli and are important for forming the association between auditory and visual objects. Another possibility is that auditory stimuli lead to activation in this region by a less direct mechanism. Visual imagery of objects is known to activate ventral temporal regions responsive to actual visual stimuli, albeit at a weaker level (Ishai et al., 2000; O'Craven and Kanwisher, 2000). In the current experiment, presentation of an auditory stimulus might produce visual mental imagery (e.g., auditory "ring," mental image of a telephone), leading to the observed weak activity in ventral temporal regions. Because ventral temporal activity is observed in auditory naming tasks that do not require the explicit generation of mental images (Experiment 3 of the current study; Adams and Janata, 2002; Buckner et al., 2000; Tranel et al., 2003), these images may be generated automatically, perhaps to enable more rapid object identification. However, because this activity is weaker than perceptual activity, it may not be observed in all studies of auditory object identification (Amedi et al., 2002).

Other Types of Multimodal Responses

During auditory stimulus presentation, the BOLD signal in visual cortex was depressed below fixation baseline, while during visual stimulus presentation the auditory cortex BOLD signal was depressed below baseline. However, during multimodal presentation, the response in auditory and visual cortex did not differ significantly from that during stimulation in their preferred modality (instead of the expected smaller response from the linear superposition of positive preferred modality BOLD response and negative nonpreferred modality BOLD response). This suggests that even in early sensory cortices, interactions between modalities can occur (Laurienti et al., 2002).

Conclusion

Our results, along with those from previous studies, suggest that pSTS/MTG may be best viewed as an associative learning device for linking different types of information both within and across visual and auditory modalities. These associations may include naturally occurring, highly correlated features such as an animal's shape and its motion, or an animal's shape and its characteristic sound. This region may also be critical for learning arbitrary associations such as that between the

shape of a letter and its sound. The anatomical location of pSTS/MTG between areas for processing visual form and motion, and between visual and auditory association areas, makes it ideally suited for integrating these types of information. The possibility that different regions of pSTS/MTG are specialized for associating different properties within and across visual and auditory modalities remains an important avenue for future exploration.

Experimental Procedures

Human Subjects and MR Data Collection

Twenty-six subjects underwent a complete physical examination and provided informed consent (Experiment 1, n = 8; Experiment 2, n = 7; Experiment 3, n = 8; Experiment 4, n = 5; two subjects participated in Experiments 3 and 4). Subjects were compensated for participation in the study and anatomical MR scans were screened by the NIH Clinical Center Department of Radiology in accordance with the NIMH human subjects committee. MR data were collected on a General Electric 3 Tesla scanner. A high-resolution SPGR or MP-RAGE anatomical sequence (1–3 repetitions) was collected at the beginning of each scanning session. Gradient-echo echo-planar volumes were acquired with TE of 30 ms, TR of 3 s, and 3.75 mm in-plane resolution. Each volume contained 24 axial slices (slice thickness of 4.5 or 5.0 mm as necessary to cover the entire cortex) with 132 volumes per scan series and 8 to 10 scan series per subject.

Auditory and Visual Stimuli

Stimuli were presented using MATLAB (Mathworks Inc., Natick, MA) with the Psychophysics Toolbox extensions (Brainard, 1997; Pelli, 1997) running on a Macintosh G4 (Apple Computer, Cupertino, CA). The source code for the stimulus program is freely available at http://lbc.nimh.nih.gov/people/mikeb/matlab.html. Auditory stimuli were presented at approximately 80 dB SPL, using a SilentScan system from Avotec, Inc. (Stuart, FL), which attenuates gradient noise produced by the MR scanner while providing high-fidelity stimulus reproduction. Subjects reported being able to hear the stimuli in the scanner and performed the behavioral discrimination task with high accuracy (see Results). For additional details, including spectrograms of the scanner gradient sound and the auditory stimuli, please see Supplemental Data, Sections 1 and 4, and Supplemental Figures S1–S5 at http://www.neuron.org/cgi/content/full/41/5/809/DC1. Visual stimuli (which subtended between 5° and 10° of visual angle) were back-projected onto a Lucite screen using a 3-panel LCD projector (Sharp Inc., Mahwah, NJ) visible to the subject through a mirror mounted on the MR head coil. Stimulus presentation was synchronized with MR data acquisition using a DAQ board (National Instruments, Austin, TX). Subjects performed a behavioral task using an MR-compatible button device, with responses recorded using SuperLab software (Cedrus Corp., San Pedro, CA).

Experiment 1 contained 6 stimulus conditions. Three categories of visual stimuli were used (Figure 1A) consisting of stationary, black-and-white photographs of tools (man-made manipulable objects), animals, and phase-scrambled images of these same objects. Three categories of auditory stimuli (Figure 1B) were presented: sounds produced by animals, sounds produced by tools, and synthesized "ripple" sounds (Depireux et al., 2001; Klein et al., 2000). Scrambled photographs and ripple sounds were chosen as controls because of their high degree of complexity but lack of correspondence to real-world objects. Each sound clip (2.5 s duration) was sampled from commercially available sound effects libraries, converted from stereo to mono, and equated for root-mean-square power. There were 432 tool photographs, 432 animal photographs, 864 scrambled photographs, 12 object sounds, 12 animal sounds, and 8 synthesized ripple sounds. During the visual stimulation ISI and throughout auditory stimulation, subjects viewed a white fixation crosshair on a gray background (during visual stimulus conditions, no sounds were presented).

Experiment 2 also contained 6 stimulus conditions. Two categories each of visual, auditory, and simultaneously presented auditory and visual stimuli were used. Visual stimuli consisted of static line drawings of tools or animals (Figure 2A). Auditory stimuli (Figure 2B) consisted of 2.5 s clips of tools or animals sounds, either sampled from libraries or recorded de novo (sounds were processed as in Experiment 1). Auditory-visual stimuli (Figure 2C) consisted of simultaneously presented line drawings and sounds of either tools or animals. Drawing and sound either corresponded (e.g., hammer/bang, cat/meow) or did not (hammer/ring, cat/bark). There were 24 line drawings of animals and 24 of tools, and 24 sounds of animals and 24 of tools.

In Experiment 3, an event-related design was used. Each trial began with the presentation of a single stimulus (2.5 s duration) followed by a 2.5 s delay, followed by a 3 s display containing three visually presented words. Subjects pressed a button corresponding to the name of the stimulus presented (e.g., hammer/saw/telephone). Stimuli (Figure 3) consisted of either visually presented video clips of moving tools, recorded sounds of these same tools, or simultaneously presented moving video clips and sound. Eight different tools were used. Video clips of tools were presented with a central fixation square to encourage fixation; tools moved realistically without visible manipulandum (details in Beauchamp et al., 2002).

In Experiment 4, an event-related design was used. Each trial consisted of a single stimulus (2.5 s duration) followed by a 500 ms ISI. The stimulus set was the same as Experiment 3. In congruent trials, the videos and recordings represented the same tools; during incongruent trials, they represented different tools. Subjects made a 2-alternative forced choice between congruent and incongruent.

Experimental Design

Experiments 1 and 2 were conducted using a block design. Each stimulation block lasted 21 s, during which 7 stimuli from a given category were presented (2.5 s stimuli + 0.5 s ISI). Each stimulation block was followed by 9 s of a baseline condition (fixation crosshair on a gray background). Different blocks of stimuli were presented in pseudo-random order. Each MR scan series lasted 6 min and contained two blocks of each type.

In Experiment 3, each event-related trial contained stimulation and response epochs, separated in time to allow separation of their neural substrates. Each trial lasted 8 s and was separated from the next trial by 0–6 s of fixation baseline. Different trial types were randomly ordered for optimal experimental efficiency (Dale, 1999) using the optseq program written by Doug Greve (http://surfer.nmr.mgh.harvard.edu/optseq/). The combination of 3 s stimuli with 2 s time for brain acquisition allowed for an effective TR of 1 s, allowing estimation of the hemodynamic response to a single stimulus of each type with 1 s resolution (see below). Experiment 4 used the same rapid event related method as Experiment 3, except that each trial lasted 3 s and did not contain separate epochs.

fMRI Data Analysis

MR data were analyzed within the framework of the general linear model in AFNI 2.50 (Cox, 1996). The first two volumes in each scan series, collected before equilibrium magnetization was reached, were discarded. Then, all volumes were registered to the volume collected nearest in time to the high-resolution anatomy. Next, a spatial filter with a root-mean-square width of 4 mm was applied to each echo-planar volume. The response to each stimulus category compared with the fixation baseline was calculated using multiple regression. All areas that showed a response to any stimulus type were included in the analysis.

For the first and second experiments (block design), multiple regression was performed using 32 regressors of no interest (mean, linear trend, and second-order polynomial within each scan series to account for slow changes in the MR signal; 8 outputs from volume registration to account for residual variance from subject motion not corrected by registration); and 6 regressors of interest, one for each stimulus type. Each regressor of interest consisted of a square wave for each stimulation block of that stimulus type, convolved with a γ-variate function to account for the slow hemodynamic response (Cohen, 1997). In the third experiment (event-related), a separate regressor was used to model the response in each 1 s period in a 20 s window following each stimulus onset. With three stimulus types, this resulted in 60 regressors of interest (each consisting of a series of delta functions), resulting in an estimate of the response to a single stimulus of each type with no assumptions about the shape of the hemodynamic response (along with 32 regressors of no interest, described above) (Miezin et al., 2000). This resulted in a 20 s time series for each stimulus category in each voxel. This time series contained BOLD responses to both the stimulus (presented at $t = 0$ s) and the motor response (occurring at $t = 6$ s). Because the hemodynamic signal peaks 4–6 s after neural activity, the amplitude of the response to the stimulus was estimated by summing the β weights of the regressors representing the 5th through the 8th s of the response, while the amplitude of the MR signal to the motor response estimated by summing the 11th through the 14th s of the response.

Individual subject activation maps were created by using the overall experimental effect (all regressors of interest) to find voxels showing a response to any type of stimulus at a threshold of $p < 10^{-6}$ to correct for the multiple comparisons produced by 20,000–25,000 intracranial functional voxels. Following stringent thresholding by the experimental-effect contrast, voxels were categorized by their response to different stimulus types using a more liberal threshold of $p < 0.05$, described below. Functional data were interpolated to 1 mm^3 resolution using cubic interpolation and overlaid on single subject anatomical data.

To create group maps, a random-effects model was used. For each subject, the regression model provided a single estimate of the response to each stimulus type in each voxel (either from the amplitude of the single regressor representing that stimulus in the block design experiments, or from the estimated amplitude of the event-related response as described above). After stereotactic normalization to Talairach space (Talairach and Tournoux, 1988), a two-way mixed-effect ANOVA was performed on each voxel in standard space. Planned contrasts on stimulus type were undertaken (fixed effect), with each individual subject serving as the repeated measure (random effect).

Unimodal and Multimodal Regions

The criteria for a voxel to be considered "active" was a stringent statistical threshold of $p < 10^{-6}$. For individual subject maps, this threshold was applied to the experimental effect or F-statistic of the general linear model (ratio of full model to baseline model). For group maps, this threshold was applied to the treatment factor of the ANOVA (mean across conditions across subjects significantly different than 0).

In order to create maps of auditory, visual, and multimodal regions (Figures 1–3) voxels were categorized with a separate statistical test. Unimodal auditory regions were defined as those showing responses less than $t < 2$ ($p > 0.05$) to visual stimuli, while unimodal visual regions showed responses of $t < 2$ ($p > 0.05$) to auditory stimuli. If voxels responded at $t > 2$ to both auditory and visual stimuli, they were classified as responding to both auditory and visual stimulus conditions. Early auditory and visual cortex showed significant decreases in the BOLD signal (below fixation baseline) to stimulation in the nonpreferred modality (e.g., Figure 1). These deactivations were not considered in the classification.

In Experiment 3, the time resolution of the event-related design allowed us to categorize additional sets of active regions. In Experiment 3, regions were classified as response related (Figure 3D) if they responded during the response epoch ($t > 2$, $p < 0.05$) but not the stimulus epoch ($t < 2$, $p > 0.05$). Task-related activations (Figure 3E) were defined as those that responded during both auditory ($t > 2$) and visual ($t > 2$) stimulus epochs but showed as great or greater responses during the response epoch (response versus stimulus contrast, $t > 0$). Multimodal activations (Figure 3F) were classified as those that that responded during auditory and visual stimulus epochs but showed greater responses during combined auditory and visual stimulation than unimodal stimulation ($t > 0$). In Experiment 4, the pSTS/MTG region was identified as in Experiment 3, and the average time series was calculated across subjects.

Regions of Interest

For each subject, the average response to each stimulus category within five regions of interest (ROI) (visual cortex, auditory cortex, DLPFC, motor cortex, STS) was calculated. Then, the MR time series from each ROI was averaged across subjects to create group MR time series (Figures 1, 2, and 4). For additional details on ROI construction, please see Supplemental Data, Section 3 at http://www.neuron.org/cgi/content/full/41/5/809/DC1.

Surface Modeling

Three-dimensional models of the cortical surfaces were constructed using FreeSurfer software (Cortechs, Inc., http://www.cortechs.net). From one to five high-resolution MP-RAGE scans for each subject were collected and averaged. An automated segmentation routine then extracted the gray-white boundary and constructed a surface model, which was then inflated to allow inspection of active areas buried deep in cortical sulci (Fischl et al., 1999a). The overall model significance was thresholded and blurred with a spatial Gaussian filter of root-mean-square width 8 mm before painting to the cortical surface. Only voxels intersecting surface nodes were mapped to the cortical surface. Surfaces were visualized using SUMA software (http://afni.nimh.nih.gov/afni/SUMA).

Acknowledgments

The authors would like to thank David Poeppel for providing the auditory ripple stimuli; Ziad Saad for developing the SUMA surface analysis package and Bob Cox for continued development of AFNI; Doug Greve for providing RSFGEN to generate random stimulus sequences; and Jill Weisberg and Adam Messinger for helpful comments on the manuscript.

Received: September 17, 2003
Revised: November 25, 2003
Accepted: January 20, 2004
Published: March 3, 2004

References

Adams, R.B., and Janata, P. (2002). A comparison of neural circuits underlying auditory and visual object categorization. Neuroimage *16*, 361–377.

Alain, C., Arnott, S.R., Hevenor, S., Graham, S., and Grady, C.L. (2001). "What" and "where" in the human auditory system. Proc. Natl. Acad. Sci. USA *98*, 12301–12306.

Allison, T., Puce, A., and McCarthy, G. (2000). Social perception from visual cues: role of the STS region. Trends Cogn. Sci. *4*, 267–278.

Amedi, A., Malach, R., Hendler, T., Peled, S., and Zohary, E. (2001). Visuo-haptic object-related activation in the ventral visual pathway. Nat. Neurosci. *4*, 324–330.

Amedi, A., Jacobson, G., Hendler, T., Malach, R., and Zohary, E. (2002). Convergence of visual and tactile shape processing in the human lateral occipital complex. Cereb. Cortex *12*, 1202–1212.

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

Beauchamp, M.S., Lee, K.E., Haxby, J.V., and Martin, A. (2002). Parallel visual motion processing streams for manipulable objects and human movements. Neuron *34*, 149–159.

Beauchamp, M.S., Lee, K.E., Haxby, J.V., and Martin, A. (2003). fMRI responses to video and point-light displays of moving humans and manipulable objects. J. Cogn. Neurosci. *15*, 991–1001.

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

Belin, P., Zatorre, R.J., and Ahad, P. (2002). Human temporal-lobe response to vocal sounds. Brain Res. Cogn. Brain Res. *13*, 17–26.

Benevento, L.A., Fallon, J., Davis, B.J., and 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.

Binder, J.R., Frost, J.A., Hammeke, T.A., Bellgowan, P.S., Springer, J.A., Kaufman, J.N., and Possing, E.T. (2000). Human temporal lobe activation by speech and nonspeech sounds. Cereb. Cortex *10*, 512–528.

Brainard, D.H. (1997). The psychophysics toolbox. Spat. Vis. *10*, 433–436.

Braver, T.S., Cohen, J.D., Nystrom, L.E., Jonides, J., Smith, E.E., and Noll, D.C. (1997). A parametric study of prefrontal cortex involvement in human working memory. Neuroimage *5*, 49–62.

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

Buchel, C., Price, C., and Friston, K. (1998). A multimodal language region in the ventral visual pathway. Nature *394*, 274–277.

Buckner, R.L., Koutstaal, W., Schacter, D.L., and Rosen, B.R. (2000). Functional MRI evidence for a role of frontal and inferior temporal cortex in amodal components of priming. Brain *123*, 620–640.

Bushara, K.O., Weeks, R.A., Ishii, K., Catalan, M.J., Tian, B., Rauschecker, J.P., and Hallett, M. (1999). Modality-specific frontal and parietal areas for auditory and visual spatial localization in humans. Nat. Neurosci. *2*, 759–766.

Calvert, G.A. (2001). Crossmodal processing in the human brain: insights from functional neuroimaging studies. Cereb. Cortex *11*, 1110–1123.

Calvert, G.A., Brammer, M.J., Bullmore, E.T., Campbell, R., Iversen, S.D., and David, A.S. (1999). Response amplification in sensory-specific cortices during crossmodal binding. Neuroreport *10*, 2619–2623.

Calvert, G.A., Campbell, R., and Brammer, M.J. (2000). Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. Curr. Biol. *10*, 649–657.

Calvert, G.A., Hansen, P.C., Iversen, S.D., and Brammer, M.J. (2001). Detection of audio-visual integration sites in humans by application of electrophysiological criteria to the BOLD effect. Neuroimage *14*, 427–438.

Carpenter, P.A., Just, M.A., Keller, T.A., Eddy, W., and Thulborn, K. (1999). Graded functional activation in the visuospatial system with the amount of task demand. J. Cogn. Neurosci. *11*, 9–24.

Chao, L.L., Haxby, J.V., and Martin, A. (1999). Attribute-based neural substrates in temporal cortex for perceiving and knowing about objects. Nat. Neurosci. *2*, 913–919.

Clarke, S., Bellmann, A., Meuli, R.A., Assal, G., and Steck, A.J. (2000). Auditory agnosia and auditory spatial deficits following left hemispheric lesions: evidence for distinct processing pathways. Neuropsychologia *38*, 797–807.

Clarke, S., Bellmann Thiran, A., Maeder, P., Adriani, M., Vernet, O., Regli, L., Cuisenaire, O., and Thiran, J.P. (2002). What and where in human audition: selective deficits following focal hemispheric lesions. Exp. Brain Res. *147*, 8–15.

Cohen, M.S. (1997). Parametric analysis of fMRI data using linear systems methods. Neuroimage *6*, 93–103.

Cox, R.W. (1996). AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. Comput. Biomed. Res. *29*, 162–173.

Dale, A.M. (1999). Optimal experimental design for event-related fMRI. Hum. Brain Mapp. *8*, 109–114.

Depireux, D.A., Simon, J.Z., Klein, D.J., and Shamma, S.A. (2001). Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex. J. Neurophysiol. *85*, 1220–1234.

Devlin, J.T., Moore, C.J., Mummery, C.J., Gorno-Tempini, M.L., Phillips, J.A., Noppeney, U., Frackowiak, R.S., Friston, K.J., and Price, C.J. (2002). Anatomic constraints on cognitive theories of category specificity. Neuroimage *15*, 675–685.

Engelien, A., Silbersweig, D., Stern, E., Huber, W., Doring, W., Frith, C., and Frackowiak, R.S. (1995). The functional anatomy of recovery from auditory agnosia. A PET study of sound categorization in a neurological patient and normal controls. Brain *118*, 1395–1409.

Erickson, C.A., and Desimone, R. (1999). Responses of macaque perirhinal neurons during and after visual stimulus association learning. J. Neurosci. *19*, 10404–10416.

Fischl, B., Sereno, M.I., and Dale, A.M. (1999a). Cortical surface-based analysis. II: Inflation, flattening, and a surface-based coordinate system. Neuroimage *9*, 195–207.

Fischl, B., Sereno, M.I., Tootell, R.B., and Dale, A.M. (1999b). High-resolution intersubject averaging and a coordinate system for the cortical surface. Hum. Brain Mapp. *8*, 272–284.

Goldman-Rakic, P.S. (1999). The physiological approach: functional architecture of working memory and disordered cognition in schizophrenia. Biol. Psychiatry *46*, 650–661.

Griffiths, T.D., Rees, A., Witton, C., Shakir, R.A., Henning, G.B., and Green, G.G. (1996). Evidence for a sound movement area in the human cerebral cortex. Nature *383*, 425–427.

Hall, D.A., Johnsrude, I.S., Haggard, M.P., Palmer, A.R., Akeroyd, M.A., and Summerfield, A.Q. (2002). Spectral and temporal processing in human auditory cortex. Cereb. Cortex *12*, 140–149.

Haxby, J.V., Ungerleider, L.G., Clark, V.P., Schouten, J.L., Hoffman, E.A., and Martin, A. (1999). The effect of face inversion on activity in human neural systems for face and object perception. Neuron *22*, 189–199.

Haxby, J.V., Hoffman, E.A., and Gobbini, M.I. (2002). Human neural systems for face recognition and social communication. Biol. Psychiatry *51*, 59–67.

Ishai, A., Ungerleider, L.G., and Haxby, J.V. (2000). Distributed neural systems for the generation of visual images. Neuron *28*, 979–990.

James, T.W., Humphrey, G.K., Gati, J.S., Servos, P., Menon, R.S., and Goodale, M.A. (2002). Haptic study of three-dimensional objects activates extrastriate visual areas. Neuropsychologia *40*, 1706–1714.

Kaas, J.H., and Hackett, T.A. (2000). Subdivisions of auditory cortex and processing streams in primates. Proc. Natl. Acad. Sci. USA *97*, 11793–11799.

Kanwisher, N., McDermott, J., and Chun, M.M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. J. Neurosci. *17*, 4302–4311.

Klein, D.J., Depireux, D.A., Simon, J.Z., and Shamma, S.A. (2000). Robust spectrotemporal reverse correlation for the auditory system: optimizing stimulus design. J. Comput. Neurosci. *9*, 85–111.

Lancaster, J.L., Woldorff, M.G., Parsons, L.M., Liotti, M., Freitas, C.S., Rainey, L., Kochunov, P.V., Nickerson, D., Mikiten, S.A., and Fox, P.T. (2000). Automated Talairach atlas labels for functional brain mapping. Hum. Brain Mapp. *10*, 120–131.

Laurienti, P.J., Burdette, J.H., Wallace, M.T., Yen, Y.F., Field, A.S., and Stein, B.E. (2002). Deactivation of sensory-specific cortex by cross-modal stimuli. J. Cogn. Neurosci. *14*, 420–429.

Lerner, Y., Hendler, T., Ben-Bashat, D., Harel, M., and Malach, R. (2001). A hierarchical axis of object processing stages in the human visual cortex. Cereb. Cortex *11*, 287–297.

Levy, I., Hasson, U., Avidan, G., Hendler, T., and Malach, R. (2001). Center-periphery organization of human object areas. Nat. Neurosci. *4*, 533–539.

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

Maeder, P.P., Meuli, R.A., Adriani, M., Bellmann, A., Fornari, E., Thiran, J.P., Pittet, A., and Clarke, S. (2001). Distinct pathways involved in sound recognition and localization: a human fMRI study. Neuroimage *14*, 802–816.

Malach, R., Reppas, J.B., Benson, R.R., Kwong, K.K., Jiang, H., Kennedy, W.A., Ledden, P.J., Brady, T.J., Rosen, B.R., and Tootell, R.B. (1995). Object-related activity revealed by functional magnetic resonance imaging in human occipital cortex. Proc. Natl. Acad. Sci. USA *92*, 8135–8139.

Martin, A., and Chao, L.L. (2001). Semantic memory and the brain: structure and processes. Curr. Opin. Neurobiol. *11*, 194–201.

Martin, A., Wiggs, C.L., Ungerleider, L.G., and Haxby, J.V. (1996). Neural correlates of category-specific knowledge. Nature *379*, 649–652.

Messinger, A., Squire, L.R., Zola, S.M., and Albright, T.D. (2001). Neuronal representations of stimulus associations develop in the temporal lobe during learning. Proc. Natl. Acad. Sci. USA *98*, 12239–12244.

Mesulam, M.M. (1998). From sensation to cognition. Brain *121*, 1013–1052.

Miezin, F.M., Maccotta, L., Ollinger, J.M., Petersen, S.E., and Buckner, R.L. (2000). Characterizing the hemodynamic response: effects of presentation rate, sampling procedure, and the possibility of ordering brain activity based on relative timing. Neuroimage *11*, 735–759.

Naya, Y., Yoshida, M., and Miyashita, Y. (2001). Backward spreading of memory-retrieval signal in the primate temporal cortex. Science *291*, 661–664.

Naya, Y., Yoshida, M., and Miyashita, Y. (2003). Forward processing of long-term associative memory in monkey inferotemporal cortex. J. Neurosci. *23*, 2861–2871.

O'Craven, K.M., and Kanwisher, N. (2000). Mental imagery of faces and places activates corresponding stiimulus-specific brain regions. J. Cogn. Neurosci. *12*, 1013–1023.

Oram, M.W., and Perrett, D.I. (1996). Integration of form and motion in the anterior superior temporal polysensory area (STPa) of the macaque monkey. J. Neurophysiol. *76*, 109–129.

Pelli, D.G. (1997). The VideoToolbox software for visual psychophysics: transforming numbers into movies. Spat. Vis. *10*, 437–442.

Petersen, S.E., and Fiez, J.A. (1993). The processing of single words studied with positron emission tomography. Annu. Rev. Neurosci. *16*, 509–530.

Poremba, A., Saunders, R.C., Crane, A.M., Cook, M., Sokoloff, L., and Mishkin, M. (2003). Functional mapping of the primate auditory system. Science *299*, 568–572.

Puce, A., Allison, T., Gore, J.C., and McCarthy, G. (1995). Face-sensitive regions in human extrastriate cortex studied by functional MRI. J. Neurophysiol. *74*, 1192–1199.

Puce, A., Allison, T., Asgari, M., Gore, J.C., and McCarthy, G. (1996). Differential sensitivity of human visual cortex to faces, letterstrings, and textures: a functional magnetic resonance imaging study. J. Neurosci. *16*, 5205–5215.

Puce, A., Allison, T., Bentin, S., Gore, J.C., and McCarthy, G. (1998). Temporal cortex activation in humans viewing eye and mouth movements. J. Neurosci. *18*, 2188–2199.

Puce, A., Syngeniotis, A., Thompson, J.C., Abbott, D.F., Wheaton, K.J., and Castiello, U. (2003). The human temporal lobe integrates facial form and motion: evidence from fMRI and ERP studies. Neuroimage *19*, 861–869.

Raij, T., Uutela, K., and Hari, R. (2000). Audiovisual integration of letters in the human brain. Neuron *28*, 617–625.

Rauschecker, J.P. (1997). Processing of complex sounds in the auditory cortex of cat, monkey, and man. Acta Otolaryngol. Suppl. *532*, 34–38.

Rauschecker, J.P., and Tian, B. (2000). Mechanisms and streams for processing of "what" and "where" in auditory cortex. Proc. Natl. Acad. Sci. USA *97*, 11800–11806.

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

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

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

Saygin, A.P., Dick, F., Wilson, S.M., Dronkers, N.F., and Bates, E. (2003). Neural resources for processing language and environmental sounds: evidence from aphasia. Brain *126*, 928–945.

Scott, S.K., Blank, C.C., Rosen, S., and Wise, R.J. (2000). Identification of a pathway for intelligible speech in the left temporal lobe. Brain *123*, 2400–2406.

Seifritz, E., Esposito, F., Hennel, F., Mustovic, H., Neuhoff, J.G., Bilecen, D., Tedeschi, G., Scheffler, K., and Di Salle, F. (2002). Spatio-temporal pattern of neural processing in the human auditory cortex. Science *297*, 1706–1708.

Seltzer, B., Cola, M.G., Gutierrez, C., Massee, M., Weldon, C., and Cusick, C.G. (1996). Overlapping and nonoverlapping cortical projections to cortex of the superior temporal sulcus in the rhesus monkey: double anterograde tracer studies. J. Comp. Neurol. *370*, 173–190.

Stein, B.E., and Meredith, M.A. (1993). The Merging of the Senses (Cambridge, MA: MIT Press).

Talairach, J., and Tournoux, P. (1988). Co-Planar Stereotaxic Atlas of the Human Brain (New York: Thieme Medical Publishers).

Tanaka, K. (2003). Columns for complex visual object features in the inferotemporal cortex: clustering of cells with similar but slightly different stimulus selectivities. Cereb. Cortex *13*, 90–99.

Thompson-Schill, S.L. (2003). Neuroimaging studies of semantic memory: inferring "how" from "where." Neuropsychologia *41*, 280–292.

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

Tranel, D., Damasio, H., Eichhorn, G.R., Grabowski, T., Ponto, L.L., and Hichwa, R.D. (2003). Neural correlates of naming animals from their characteristic sounds. Neuropsychologia *41*, 847–854.

Ungerleider, L.G., and Mishkin, M. (1982). Two cortical visual systems. In Analysis of Visual Behavior, D.J. Ingle, M.A. Goodale, and R.J.W. Mansfield, eds. (Cambridge, MA: MIT Press), pp. 549–586.

Wagner, A.D., Koutstaal, W., and Schacter, D.L. (1999). When encoding yields remembering: insights from event-related neuroimaging. Philos. Trans. R. Soc. Lond. B Biol. Sci. *354*, 1307–1324.

Warren, J.D., Zielinski, B.A., Green, G.G., Rauschecker, J.P., and Griffiths, T.D. (2002). Perception of sound-source motion by the human brain. Neuron *34*, 139–148.

Wessinger, C.M., VanMeter, J., Tian, B., Van Lare, J., Pekar, J., and Rauschecker, J.P. (2001). Hierarchical organization of the human auditory cortex revealed by functional magnetic resonance imaging. J. Cogn. Neurosci. *13*, 1–7.

Zatorre, R.J., and Belin, P. (2001). Spectral and temporal processing in human auditory cortex. Cereb. Cortex *11*, 946–953.

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