Neural Correlates of Audio-Visual Object Recognition: Effects of Implicit Spatial Congruency

Tina Plank,¹ Katharina Rosengarth,¹ Wookeun Song,²
Wolfgang Ellermeier,³ and Mark W. Greenlee^{1*}

¹*Institute of Psychology, University of Regensburg, Regensburg, Germany*

²*Brüel and Kjær Sound and Vibration Measurement A/S, Nærum, Denmark*

³*Institute of Psychology, Technical University Darmstadt, Darmstadt, Germany*

Abstract: Multisensory integration assists us to identify objects by providing multiple cues with respect to object category and spatial location. We used a semantic audiovisual object matching task to determine the effect of spatial congruency on response behavior and fMRI brain activation. Fifteen subjects responded in a four-alternative response paradigm, which visual quadrant contained the object best matched to the sound presented. Realistic sounds based on head-related transfer functions were presented binaurally with the simulated sound source corresponding to one of the four quadrants. Following a random sequence, the location of the sound corresponded to the quadrant containing the semantically congruent target on half the trials, whereas on other trials the sound arose from an incongruent location. We examined the effects of spatial congruency on response latencies, hit-rates and fMRI responses. Preliminary behavioral results revealed a significant effect of spatial congruency on response latency or performance for stimuli with noise added. In the fMRI experiment, spatial congruency had a significant effect on the BOLD response. A cluster in the right middle and superior temporal gyrus was more activated when the auditory sound sources were spatially congruent with the semantically matching visual stimulus. In an exploratory post-hoc analysis, in which we correlated the BOLD signal with the subjects' ability to locate the sound sources, we found a significant cluster in the left inferior frontal cortex, where the BOLD response increased with sound-source localization performance. Thus spatial congruency appears to enhance the semantic integration of audiovisual object information in these brain regions. *Hum Brain Mapp* 33:797–811, 2012. © 2011 Wiley Periodicals, Inc.

Key words: multisensory perception; object recognition; spatial congruency; fMRI

INTRODUCTION

Everyday perception depends on multiple sources of sensory cues to determine the location and nature of objects in our environment. While waiting for a train to pass at a pedestrian crossing, we sense the location, direc-

tion and speed of the approaching train via multiple visual, auditory, and tactile cues. Normally the spatial location of these cues coincides to yield a multisensory percept of a single object. Sensory conflicts can arise when multiple objects are present yielding a mixture of sensory cues related to the nature and location of the respective

Contract grant sponsor: University of Regensburg (postdoctoral research grant out of the HWP 2004–2006 program); Contract grant sponsor: BMBF (Project “Visuospatial Cognition”); Contract grant number: 01GW0653.

*Correspondence to: Mark W. Greenlee, Institute of Psychology, University of Regensburg, Regensburg, Germany, Universitätsstraße 31, 93053 Regensburg, Germany.
E-mail: mark.greenlee@psychologie.uni-regensburg.de

Received for publication 15 June 2010; Revised 8 December 2010; Accepted 13 December 2010

DOI: 10.1002/hbm.21254

Published online 21 March 2011 in Wiley Online Library (wileyonlinelibrary.com).

objects. Nevertheless, in daily life, our brain usually has no difficulty to integrate all the sensory cues into a coherent percept. This is even more astonishing since two different networks are responsible for identifying objects (the ventral “what” pathway) and for localizing these objects in space (the dorsal “where” pathway). Evidence for two pathways has been reported for the visual modality (Courtney et al., 1998; Haxby et al., 1991; Ungerleider and Haxby, 1994; Ungerleider and Mishkin, 1982) as well as for the auditory modality (Alain et al., 2001; Maeder et al., 2001; Rauschecker and Tian, 2000; Zatorre et al., 2002; Zatorre et al., 2004). Sestieri et al. (2006) examined the “what” and “where” pathways in the audiovisual domain, using two crossmodal matching tasks. Participants had to indicate whether a picture and a simultaneously presented environmental sound matched either semantically or spatially. Pictures were presented to the left or right hemifield and sounds were presented to the left or right ear via headphones, respectively, thus operationalizing spatial congruency in a rather simplified manner. The authors found higher BOLD responses in the left superior temporal sulcus as part of the “what”-network for the object recognition task and higher BOLD responses in the right intraparietal sulcus as part of the “where”-network in the localization task. The authors also looked at congruency effects and found a cluster in the right STS region that was more responsive to spatially congruent objects than to spatially incongruent objects. In addition to these results, effects of spatial congruency have been studied using simple response tasks (Frassinetti et al., 2002; Harrington and Peck, 1998; McDonald et al., 2000) and event-related potentials (Teder-Sälejärvi et al., 2005). These findings point to the importance of spatial congruency in multisensory integration. Imaging studies so far have concentrated mainly on visuo-tactile integration (Macaluso et al., 2005; Zimmer and Macaluso, 2007), audiovisual integration for body actions and speech-related lip movements (Meyer et al., 2010) and speech processing (Macaluso et al., 2004). Meienbrock et al. (2007) conducted a study in which they used object pictures and sounds to investigate retinotopic effects during spatial audio-visual integration, presenting spatially congruent and incongruent stimuli at four different horizontal positions. The sounds were spatially rendered by using varying interaural time differences. Subjects had to perform an audio-visual matching task with regard to location. The authors found that several temporal, parietal, and frontal areas responded preferentially to spatially congruent stimuli (left ITG, right pSTG/STS, left IPS, left PreCG, left DLPFC). Earlier visual areas were more responsive to stimuli on spatially incongruent trials.

In the studies described above that addressed the influence of spatial congruency of audiovisual information on the recognition of natural objects (Meienbrock et al., 2007;

Sestieri et al., 2006), subjects had to match auditory and visual information explicitly with regard to location. Thus the spatial information from both auditory and visual channels had to be processed correctly to perform the task. The aim of this study is to explore the benefit of spatial congruency for object recognition, whereas the auditory spatial information is only implicitly varied, without the need to be processed to solve the task successfully (i.e., participants only had to perform a semantic categorization task). In a four-alternative forced-choice paradigm, they decided in which of the visual quadrants an object briefly appeared that semantically matched a simultaneously presented realistic sound (i.e., the barking of a dog). Spatial congruency was implicitly varied in the following way: Randomly on half of all trials, where objects were presented (not including baseline trials), the sounds were presented to spatially coincide with the location of the semantically matching objects, whereas on all other trials the sound origin was spatially incongruent with the visually presented and semantically matching object (i.e., the sound source was congruent to the location of the semantically nonmatching distractor). Using an event-related random effects analysis, we could compare the condition in which the audiovisual cues were spatially congruent to the condition where these cues were spatially incongruent.

Thus, we sought to determine whether there are regions in the brain that react with enhanced activation to spatially coinciding object information even though the auditory spatial information is task irrelevant. Those integration sites could be regarded as “bottom-up” regions for combining audiovisual spatial information about semantically matching objects, and thus separating them from areas only activated in explicit auditory localization tasks. The results should reveal whether those “bottom-up” integration sites are part of the “what” pathway, for example, the STG/MTG region, or the “where” pathway, for example, parietal areas. We hypothesize that the integration will be part of the “what” pathway. Areas in the right temporal lobe show enhanced activation in audiovisual discrimination tasks (Meienbrock et al., 2007; Sestieri et al., 2006) for spatially congruent object stimuli. Meienbrock et al. (2007) found additional areas with enhanced activation for spatially congruent trials in the frontal and parietal cortex, but we attribute this activation to the requirements of the auditory localization task in that study. We therefore expect higher hit-rates in trials where the location of visual target stimulus and object sound coincides in the fMRI-experiment. Depending on the amount of noise in the audiovisual stimuli, we also expect a potential effect of spatial congruency on the hit rate, because random noise added to the stimuli increases the difficulty of semantic matching. In a psychophysical pilot experiment, we therefore test three different levels of noise added to the pictures and the sounds to make the semantic matching task more difficult and to enhance the benefit of audiovisual integration on object recognition. We will then choose that noise level for the fMRI experiment that delivers moderate hit-rates in a range of about 85–90% correct.

Another issue concerns the effect of veridical auditory localization cues on multisensory integration. In this study, the auditory stimuli were presented via headphones; the spatial information was implemented by filtering the sounds with head-related transfer functions (HRTFs) of an adequate artificial head (Minnaar et al., 2001). The resulting sounds conveyed spatial information coinciding with the location of the visual stimuli, but we assumed that our subjects might vary in their ability to correctly localize the virtual sound sources, due to individual differences in head size and characteristics of the pinnae—important factors for auditory localization (Hammershøi and Møller, 1996; Middlebrooks and Green, 1991). We expected that a correlation of the BOLD response with the post-hoc collected performance of the subjects in an auditory localization task might reveal brain areas sensitive to the matching of auditory and visual spatial cues in recognizing natural objects. Studies that examined effects of spatial congruency on audiovisual object perception so far (Meinenbrock et al., 2007; Sestieri et al., 2006) have not yet looked for the influence of individual differences in spatial hearing on the BOLD response for multisensory integration.

MATERIALS AND METHODS

Psychophysical Experiment

Subjects

A total of 10 subjects (three males, seven females; age 19–29 years, median 21.5 years) participated in the pilot behavioral experiment. They all reported normal (or corrected to normal) vision and normal hearing. All subjects gave informed consent prior to their participation for procedures approved by the Ethical Committee of the University of Regensburg. The study was conducted in accordance to the ethical guidelines of the Declaration of Helsinki.

Stimuli and experimental paradigm

Visual stimuli presented on each trial consisted of a fixation cross ($0.8^\circ \times 0.8^\circ$ of visual angle) in the centre of the screen and four squares (subtending $7.3^\circ \times 7.3^\circ$ of visual angle each) positioned in the four visual quadrants on a gray background. The four squares were filled with two juxtapositioned chromatic images of natural objects (e.g. animals, tools, vehicles, musical instruments) and two random noise pictures (see Fig. 1). One of the object images on each trial served as the target image, the other one as a distractor. The target and distractor were always chosen from different semantic categories. The clarity of the pictures was varied in three noise levels: (1) no noise added to the image, (2) Gaussian white noise with a mean of 0 and a variance of 0.25 or (3) Gaussian white noise with a mean of 0 and a variance of 0.5 added to the object images. Overall 640 different object images were pre-

Figure 1.

Example of a trial containing a target image (lower left) and a distractor (upper right) along with visual noise in all four quadrants.

sented. The photographs were collected from free internet databases or taken by the first author. Visual stimuli were shown on a Sony CRT-monitor at a viewing distance of 70 cm, whereas the subjects rested their head on a chinrest.

Three-hundred and twenty object sounds semantically corresponding to the target image on each trial were sampled from free sound effects libraries or recorded for the study, partly in the acoustics laboratory of Aalborg University. All sounds were clipped to 1 s duration, equated for root-mean-square power and, if necessary, converted from stereo to mono. Afterwards the sounds were filtered with the head-related transfer functions (HRTFs) of an artificial head (VALDEMAR) to obtain spatial information. Constructed at the Department of Acoustics (Aalborg University) as an average over 40 human heads, VALDEMAR's performance has been shown to be superior to that of other artificial heads available and yields results that are very similar to real human heads (Minnaar et al., 2001). The sounds were filtered with HRTFs placing them at directions that were spatially congruent or incongruent with the target image (angles of $\pm 10^\circ$ in azimuth and elevation, corresponding to the outer edge of each image, respectively). Spatially incongruent sounds were presented from the diagonally opposite quadrant (i.e., the quadrant containing the semantically nonmatching distractor). The clarity of the sounds was again varied in three noise levels: (1) no noise added, (2) white noise at a level of 6 dB above that of the target, or (3) white noise at a level of 12 dB above that of the target was added

diotically, that is, equally to both ears. The sounds thus processed were presented binaurally with 44.1 kHz sampling rate and 16-bit resolution, at an average level of 63 dB (A), using a Soundblaster sound card (Aureon 5.1 Sky), a digital amplifier and Beyer dynamic DT 990 headphones.

Procedure

Prior to fMRI data acquisition the 10 participants performed a 4-AFC-discrimination task in the psychophysical laboratory. On each trial, two juxta-positioned chromatic images of natural objects were shown for 500 ms. The target and distractor were always presented in the upper or lower visual quadrant on opposite sides of the vertical meridian (see Fig. 1). The other two quadrants contained only static visual noise. Simultaneously with the image onset the presentation of one object sound was started, lasting 1 s. The sound was semantically appropriate for one of the two pictures, thus defining the target image—the image that matched the sound semantically. Spatially, the virtual sound source could be congruent or incongruent with the location of the target image. Thus the auralized sound source was either spatially congruent with the semantically matching visual target (spatial congruence condition) or it was spatially congruent with the distractor (spatial incongruence condition). Visual stimuli of noise level (1) were always paired with auditory stimuli of noise level (1), the same rule applied for audiovisual stimuli of noise levels (2) and (3). Trials of all noise levels were randomly interleaved. The subjects' task was to press one of four buttons (corresponding to the four quadrants) as fast as possible indicating the quadrant that contained the target image semantically matching the sound heard. The next trial started immediately after the button press. Each participant completed 576 trials, subdivided into six blocks. The stimulus sequence was controlled by the software Presentation (Neurobehavioral Systems Inc.). Hit rate and reaction times were measured as a function of spatial congruency and the noise level.

Main Experiment (fMRI and Behavioral Sound Source Localization)

Subjects

A total of 15 subjects (4 males, 11 females; age 20–34 years, median 21) participated in the main experiment (functional MRI and behavioral sound source localization). None of the participants took part in the pilot study. All subjects had normal (or corrected to normal) vision and normal hearing as determined by audiometric screening. Two of the participants were left-handed, 13 right-handed. None of the subjects had a history of neurological or psychiatric disorders. All subjects signed an informed consent form prior to their participation in accordance to the protocols approved by the Ethical Committee of the

University of Regensburg. The study was conducted in accordance to the ethical guidelines of the Declaration of Helsinki.

Stimuli and Experimental Paradigm

The same visual stimuli as in the psychophysical pilot experiment were presented on each trial, consisting of a fixation cross ($0.8^\circ \times 0.8^\circ$ of visual angle) in the centre of the screen and four squares (subtending $7.3^\circ \times 7.3^\circ$ of visual angle each) positioned in the four visual quadrants on a gray background. The four squares were filled with two juxta-positioned chromatic images of natural objects and two random noise pictures (as depicted in Fig. 1). As in the pilot study, one of the object images on each trial served as the target image, the other one as a distractor. In the main experiment, only pictures of noise level (2) were presented. Visual stimuli were projected onto a screen placed behind the head of the subjects at the end of the scanner bore and visible for the subjects via a mirror placed within the MR head coil. Thus a constant distance of 70 cm between the subjects' eyes and the projection screen was ensured.

Likewise, the same object sounds as in the pilot study were presented in the main experiment, on each trial semantically corresponding to the target image. As described above the sounds were filtered with the head-related transfer functions (HRTFs) of the artificial head VALDEMAR (Minnaar et al., 2001), placing them at directions that were spatially congruent or incongruent with the target image (angles of $\pm 10^\circ$ in azimuth and elevation, corresponding to the outer edge of each image, respectively). Spatially incongruent sounds were presented from the diagonally opposite quadrant (i.e., the quadrant containing the semantically nonmatching distractor). Only sounds of noise level (2) were chosen for the main experiment and presented binaurally with 44.1 kHz sampling rate and 16-bit resolution, at an average level of 63 dB (A), using a Soundblaster sound card (Aureon 5.1 Sky), an MR Confon amplifier and MRI-compatible headphones (MR Confon GmbH, Magdeburg, Germany).

In the baseline condition, only white noise was presented visually (all four quadrants filled with squares of random noise) and acoustically (white noise without any spatial information).

Procedure

fMRI experiment. An event-related design with a 4-AFC-task, comparable to that in the psychophysical pilot experiment, was performed by the participants in the scanner. On each trial, two juxta-positioned chromatic images of natural objects were shown for 500 ms. The target and distractor were always presented in the upper or lower visual quadrant on opposite sides of the vertical meridian (see Fig. 1). The other two quadrants

Figure 2.

Time course of the stimulation during fMRI-measurement. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

contained only static visual noise. Simultaneously with the image onset the presentation of one object sound was started, lasting 1 s. The sound was semantically appropriate for one of the two pictures, thus defining the target image—the image that matched the sound semantically. Spatially, the virtual sound source could be congruent or incongruent with the location of the target image. Thus the auralized sound source was either spatially congruent with the semantically matching visual target (spatial congruence condition) or it was spatially congruent with the distractor (spatial incongruence condition). The subjects' task was to press one of four buttons (corresponding to the four quadrants) as fast as possible indicating the quadrant that contained the target image semantically matching the sound heard. Thus, it was necessary to process the visual and auditory semantic information, whereas the spatial congruence between the two modalities was irrelevant to the task, and only implicitly varied. The instructions emphasized the importance of the semantic categorization, and that the subject should respond to the matching target location independently of the exact location of the sound source. Targets were randomly placed in one of the four possible quadrants with equal probability to discourage the participants from developing a spatial bias to the one or other location. In baseline trials, subjects were instructed to press a fifth button. No feedback was given. Responses were recorded using two 5-button fiber-optic response boxes (Lumitouch, Photon Control, Burnaby, BC, Canada). Thus participants used their index and middle fingers of the left hand to respond to targets in the upper and lower quadrant of the left hemifield and their index and middle fingers of the right hand to respond to targets in the upper and lower quadrant of the right hemifield, respectively. Because we used sparse temporal imaging (Hall et al., 1999) to separate brain activity elicited by target stimulus presentation and by scanner noise, the inter-trial-interval was 10 s (see Fig. 2). Two seconds before the image presentation the fixation cross changed its color to alert the subjects that the next trial was about to start. The stimulus sequence was controlled by the software Presentation (Neurobehavioral Systems) and triggered by the scanner signal at the start of each volume acquisition.

Control measurement related to sound source localization.

To determine each individual subject's ability to make use of the simulated auditory cues to spatial location, we performed a post-hoc analysis of each subject's ability to localize sound sources in the MR scanner. The 320 auditory stimuli used in the fMRI experiment were presented in the same way as under fMRI conditions. Subjects were positioned in the MR scanner with the same headphones. Their task was to identify the location of the sound source by pressing a button in a 2-alternative forced-choice procedure. As in the main experiment, two squares filled with Gaussian noise were presented on the screen in two juxtapositioned quadrants, one spatially corresponding to the auditory target stimulus. This on the one hand mimicked the procedure used during the fMRI experiment and on the other hand decreased the difficulty of the task for the subjects. They only had to perform a "left or right" decision while localizing the sounds. The more difficult "up or down" location was indicated by the relative vertical position of the simultaneously presented visual object stimuli in the main experiments. The subjects judged all 320 auditory stimuli used in the main fMRI experimental session. We split up the trials into four blocks of 80 trials each. A button press by the subject always initiated the next trial. In this post-hoc analysis only 14 of the 15 subjects who participated in the main fMRI session, took part in this control experiment. The subjects' ability to correctly localize the auditory stimuli served as a regressor in the fMRI analyses (see below).

MRI data acquisition. Scanning was performed on a Siemens Allegra 3-Tesla scanner with a single channel head-coil. Functional images were acquired using a T2* weighted gradient echo-planar imaging (EPI) sequence using a sparse imaging technique (Hall et al., 1999) (TR = 10 s, TA = 2 s, TE = 30 ms, 34 slices, FoV = 192 mm, flip angle = 90°, 3 × 3 × 3 mm voxel size) to allow audiovisual stimulation during relatively quiet periods between acquisitions. Overall 353 volumes were acquired from each subject, 320 object trials and 33 baseline trials. A high resolution T1 weighted image was also performed for each subject (TR = 2250 ms, TE = 2.6 ms, 160 slices, FoV = 256 mm, flip angle = 9°, 1 × 1 × 1 mm voxel size).

MRI data analysis. Imaging data were analyzed with the SPM5 software package (Wellcome Department of Imaging Neuroscience, London; www.fil.ion.ucl.ac.uk/spm; Friston et al., 1995a,b) running under Matlab. Two dummy scans at the beginning of the measurement were removed automatically from the dataset. Preprocessing included motion correction (realignment; images were registered to the first image in the time series), co-registration, segmentation using the standard protocol of SPM5 to produce spatial normalization parameters and after that normalization to the MNI-Standard brain ICBM 152, which is based on 152 brains and was created by the Montreal Neurological Institute (MNI) with a 2 × 2 × 2 mm resolution. In a final step

TABLE I. Choice reaction times and hit-rates for 10 participants in the psychophysical pilot experiment (upper panel) and for 15 participants in the main experiment while performing the fMRI task, matching sound and picture

		Congruent		Incongruent		Baseline	
		Mean	SE	Mean	SE	Mean	SE
Psychophysical pilot experiment							
Noise level (1)	Choice RT [ms]	815.3	42.8	809.2	44.4		
	% correct	95.3	1.7	94.7	1.1		
Noise level (2)	Choice RT [ms]	944.2	62.4	957.8	59.3		
	% correct	91.6	1.3	86.3	2.0		
Noise level (3)	Choice RT [ms]	1033.1	76.5	1091.2	78.3		
	% correct	86.8	1.3	82.2	2.3		
Main experiment: fMRI							
Noise level (2)	Choice RT [ms]	1349.5	96.2	1338.1	91.5	982.5	98.3
	% correct	85.4	1.6	85.8	1.6	95.2	1.2

Given are the mean values and standard errors (SE) respectively. Noise levels for pictures and sounds are defined as follows: (1) no noise added, (2) visually Gaussian white noise ($\mu = 0$, $\sigma^2 = 0.25$) added, auditorily White noise with a level of 6 dB above that of the target added, (3) visually Gaussian white noise ($\mu = 0$, $\sigma^2 = 0.5$) added, auditorily White noise with a level of 12 dB above that of the target added.

the images were spatially smoothed by an isotropic Gaussian Kernel with FWHM of 8 mm.

For statistical analysis on the basis of the General Linear Model individual design matrices were generated. A matrix consisted of one session with three regressors for the experimental conditions “spatially congruent,” “spatially incongruent,” and “noise only” as the baseline condition. A boxcar model function of every condition was convolved with the canonical hemodynamic response function (hrf). The signal time course was high-pass filtered (1000 s) and globally normalized with proportional scaling. Due to the long TR of 10 s no correction for autocorrelation was included in the model. T-tests were carried out for every voxel to compare the BOLD activity respectively for the β -weights of the three different conditions in every paradigm. At a single-subject level, four different contrasts were calculated: two contrasts against the modelled baseline (“spatially congruent > baseline” and “spatially incongruent > baseline”) and two differential contrasts between the modeled experimental conditions (“spatially congruent > spatially incongruent” and “spatially incongruent > spatially congruent”).

These calculated contrasts were entered into the second-level random-effects analysis. The baseline contrasts were thresholded at P_{corr} (FWE) < 0.05 (corrected for multiple comparisons) on cluster level. Additional analyzes were performed, where we exclusively masked the contrast “spatially congruent > baseline” with the contrast “spatially incongruent > baseline” and vice versa. Thus we could directly compare the significant clusters in the two experimental conditions relative to baseline. These contrasts were also thresholded at P_{corr} (FWE) < 0.05 (corrected for multiple comparisons) on cluster level. For the differential contrasts, we used a threshold of $P_{\text{uncor}} = 0.001$ on voxel level and a P -value of $P_{\text{corr}} < 0.05$ on clus-

ter level. In addition, a regression analysis was conducted, correlating the performance (% correct) of each participant in the behavioral auditory localization task to the BOLD response for the contrasts “spatially congruent > spatially incongruent” and “spatially incongruent > spatially congruent”. Here again, we used a threshold of $P_{\text{uncor}} = 0.001$ at voxel level and a P -value of $P_{\text{corr}} < 0.05$ on cluster level. Region of interest (ROI) analyzes were performed using the SPM5 extension Marsbar (Brett et al., 2002) to obtain percent signal change for the conditions “spatially congruent,” “spatially incongruent,” and “baseline” in significant clusters.

RESULTS

Psychophysical Experiment

The results for reaction times and percent correct for the 10 participants, who performed the discrimination task in the psychophysical laboratory, are presented in Table I. A repeated-measures ANOVA on choice reaction times with the factor levels congruent/incongruent and noise level revealed significant main effects for the factors spatial congruency/incongruency [$F(1,9) = 6.4$; $P = 0.032$] and noise level [$F(1,9) = 40.5$; $P < 0.001$; Greenhouse-Geisser corrected], as well as a significant interaction between the two factors [$F(2,18) = 6.1$; $P = 0.01$]. The effect of spatial congruency was only significant to the highest noise level [$t(9) = 3.47$; $P = 0.021$, Bonferroni corrected for multiple comparisons]. Hit-rates decreased with increasing noise level [$F(1,9) = 70.96$; $P < 0.001$; Greenhouse-Geisser corrected], but also the factor spatial congruency/incongruency had a significant effect on performance [$F(1,9) = 21.5$; $P = 0.001$]. The interaction between the two factors returned no

significant result [$F(2,18) = 3.4$; $P = 0.056$]. Here, the effect of spatial congruency was only significant to the middle noise level [$t(9) = 4.35$; $P = 0.006$, Bonferroni corrected for multiple comparisons]. These findings suggest that the implicit variation of spatial congruency can have an implicit effect on performance in the semantic discrimination task that seems to be especially pronounced in the condition with noise level (2). This noise level, that also delivered moderate hit-rates in a range of 86–92% correct, was then chosen for the main experiment under fMRI conditions.

Main Experiment

Behavioral performance during fMRI experiment

Results for reaction times and percentage correct for 15 participants during the fMRI paradigm are shown in Table I. A one-factor repeated-measures ANOVA on choice reaction times with the factor levels congruent, incongruent and baseline revealed a significant main effect [$F(2,28) = 29.3$; $P < 0.001$], basically due to the significantly faster responses in the baseline condition [$P < 0.001$ for pairwise comparisons using Bonferroni correction]. For the performance data (number of hits), a χ^2 -test for frequency distribution was chosen to analyze the influence of the factors congruence and target quadrant. The results show that hit-rates are independent of target quadrant and congruence of auditory and visual information [$\chi^2(3) = .39$; $P > 0.25$]. Thus neither subjects' response times nor their hit-rates differed significantly between the two conditions spatially congruent and spatially incongruent, suggesting that the two tasks were equally difficult.

Performance in sound source localization

The hit rates (percent correct) in the post-hoc experiment on virtual sound-source localization revealed each subject's ability to locate the sound source in the MR scanner: mean = 79.94%; SE = 3.72; minimum = 57.50; maximum = 95.94. Results show that the subjects varied considerably in how well they could process the spatial information contained in the auditory stimuli (see Fig. 3 for a histogram of the hit rates). No systematic bias for the left or right side was detected in the tested subjects. A Kolmogorov-Smirnov test revealed no significant deviation from an equal distribution ($P = 0.131$). There was no significant correlation between the subjects' performance on the sound-source localization control task and their performance in the main fMRI experiment (Spearman correlation coefficient $\rho = 0.26$; $P = 0.37$).

fMRI results

The results of the random-effects analysis of the fMRI findings are presented in Fig. 4 for the contrasts against

Figure 3.

Histogram of the hit rates (% correct) in sound source localization, performed by 14 subjects in a post-hoc control experiment to determine differences in sound-source localization ability.

baseline. The corresponding locations of these activated clusters are described in detail in Table II. Regions in the temporal and occipital lobes exhibited more activation for the spatially incongruent stimuli. In addition, the right precentral and postcentral gyri were activated. Only small clusters in the left fusiform gyrus and the right STG showed more activation for the spatially congruent stimuli relative to baseline. Figure 5 and Table III show those clusters that were exclusively activated in the conditions with spatial congruence or spatial incongruence, respectively, each of them relative to baseline. These clusters were obtained by exclusively masking the contrast “spatially congruent > baseline” with the contrast “spatially incongruent > baseline” and vice versa. A small region in the right middle and superior temporal gyrus (BA 21/22; Fig. 5) was more activated when the auditory sound sources were spatially congruent with the semantically matching visual stimulus (MNI coordinates of peak voxel $(x,y,z) = 60, -16, -4$; z -value = 4.73; cluster size = 98 voxels, $P_{\text{corr}} < 0.05$ on cluster level; differential contrast “spatially congruent > incongruent”). This region appears to preferentially respond to audiovisual stimuli that share the same spatial location. ROI analysis was conducted in this cluster to reveal the mean percent BOLD signal change for the conditions “spatially congruent”, “spatially incongruent” and “baseline”. A repeated-measures ANOVA revealed a significant main effect [$F(1,14) = 17.9$; $P = 0.001$; Greenhouse-Geisser corrected] of % signal change. Pairwise comparisons using Bonferroni correction showed a significantly higher BOLD-response for “spatially congruent” trials than for “spatially incongruent” trials [$P < 0.001$]. Both conditions also differed significantly from baseline [$P = 0.001$ for “spatially congruent” in comparison to baseline; $P = 0.006$ for “spatially incongruent” in comparison to baseline] (see also Fig. 6, right panel). The opposite contrast “spatially incongruent > congruent” yielded no significant cluster.

Figure 4.

Results of the random-effects event-related analysis for the contrasts “spatially congruent > baseline” (depicted in red) and “spatially incongruent > baseline” (depicted in green), $n = 15$ subjects. Yellow voxels depict regions of common activation for the two conditions [P_{corr} (FWE) < 0.05 on cluster level]. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Correlation Between Psychophysical Performance and Brain Activation. The participants’ ability to correctly localize the auditory stimuli varied. In a post-hoc analysis we correlated their ability ($n = 14$) to localize the auditory stimuli in the two-alternative forced choice test with the brain activations reported above. Each subject was ranked according to his/her ability to correctly locate the simulated sound source in the scanner. A regression analysis of the second-level contrast “spatial congruence > spatial incongruence” yielded a significant correlation (Spearman’s rank order correlation $\rho = 0.88$, $P < 0.0001$) between the amplitude of the BOLD response at the peak voxel and the performance level of the subject in the auditory localization task. A significant cluster of activation was determined mainly in the

left inferior frontal gyrus (BA 10/11/47; MNI coordinates of peak voxel $(x,y,z) = -42, 46, -12$; z -value = 4.16; cluster size = 94 voxels, $P_{\text{corr}} < 0.05$ on cluster level; see Fig. 7). The higher the performance on the auditory localization task the higher the response in this region. This is also demonstrated by the difference of % BOLD signal change between spatially congruent and incongruent trials that correlates with performance in the auditory localization task in this cluster. (Spearman’s rank order correlation $\rho = .877$, $P < 0.001$; Fig. 7 lower panel) These results point to the importance of the prefrontal cortex in the analysis of implicit audiovisual stimulus congruency during a semantic matching task.

TABLE II. Significant clusters for the contrasts “spatially congruent > baseline” and “spatially incongruent > baseline”; P_{corr} (FWE) < 0.05 on cluster level

Hemisphere and Region			Congruent > Baseline						Incongruent > Baseline					
			MNI coordinates			Cluster			MNI coordinates			Cluster		
			x	y	z	z-value	Voxels		x	y	z	z-value	Voxels	
L	SOG,MOG,IOG, LinG,FG,MTG, ParaHippG, Cer	18/19/20/37	-44	-58	-20	6.52	241		-42	-58	-20	6.17	340	
			-28	-54	-12	6.06	50		-36	-88	22	5.69	40	
R	MOG, IOG, FG	18/19/37	46	-74	-12	5.71	16		44	-74	-14	6.33	108	
			40	-48	-20	6.09	63							
R	SOG,AngG, MTG	19/39							34	-76	26	5.56	21	
L	STG, MTG	21/22	-56	-8	-6	6.05	61		-56	-8	-6	6.35	92	
R	STG	22/42	68	-32	10	5.60	11		68	-32	10	5.90	21	
L	STG,TransTG, Insula, PostCG	22/29/40/41/42	-50	-24	8	5.82	188		-46	-28	12	5.80	244	
R	STG, MTG, TransTG, Insula, PostCG	13/21/22/29/40/41/42/43	58	-12	6	6.74	445		56	-12	0	6.39	516	
R	PreCG, PostCG	3/4/6	36	-28	58	5.74	109		34	-20	58	6.32	199	
L	MFG, IFG	9/46	-38	14	26	5.65	21		-42	16	26	5.51	20	

p_{corr} (FWE) < .05 on cluster level, extent threshold > 10 voxels.

MOG, middle occipital gyrus; IOG, inferior occipital gyrus; FG, fusiform gyrus; STG, superior temporal gyrus; MTG, middle temporal gyrus; MFG, middle frontal gyrus; IFG, inferior frontal gyrus; LingG, Lingual gyrus; PreCG, Precentral gyrus; PostCG, Postcentral gyrus; ParaHippG, parahippocampal gyrus; Cer, cerebellum; TransTG, Transverse temporal gyrus; AngG, Angular gyrus.

DISCUSSION

In this fMRI study, subjects performed a semantic audio-visual matching task, while the spatial congruency of auditory and visual object information was implicitly varied. Participants had to pick one of two pictures that semantically matched a simultaneously presented object sound. Thus the semantic information of both modalities had to be processed to solve the task, whereas the spatial congruency of the audiovisual information was varied implicitly.

When we look at the clusters activated in comparison to the baseline condition, where only white noise was presented both visually and acoustically, we find regions in the occipital and temporal lobe of both hemispheres typically reported to be involved in auditory and visual object recognition and audiovisual integration (Amedi et al., 2005; Beauchamp et al., 2004; Belardinelli et al., 2004; Calvert, 2001; Grill-Spector, 2003; Hein et al., 2007; Hocking and Price, 2008; Lewis et al., 2004; Macaluso and Driver, 2005; Naghavi et al., 2007; Stevenson and James, 2009; Werner and Noppeney, 2009). Activation is also found in a cluster of the left DLPFC/IFG, probably related to semantic retrieval (Wagner et al., 2001) and to task-related attention processes, inhibition or working memory (Beauchamp et al., 2004). In a recent paper, Noppeney et al. (2010) also found the left inferior frontal sulcus functioning as an accumulator of audiovisual information in preparation of forming perceptual decisions. In addition, the right pre- and postcentral gyri are activated, probably related to the planning of the motor response. The results

indicate that auditory and visual object information is integrated by the subjects in favor of localizing the visual object matching the sound semantically, regardless of conflicting spatial cues from the auditory modality.

Despite differences in the behavioral results collected in a separate subject sample prior to the fMRI experiment, we found no significant difference in hit rates and also no significant difference in mean reaction times between spatially congruent and incongruent trials. Participants in the psychophysical pilot experiment responded on average more than 300 ms faster than participants in the fMRI experiment. This could have contributed to the equalization of the hit-rates for spatially congruent and incongruent trials that can be observed in the fMRI experiment. Despite this apparent lack of difference in behavioral results, spatial congruency between auditory and visual stimuli led to a more pronounced BOLD activation in the right STG (Fig. 6).

The present findings point to a—albeit modest—role of spatial congruency between visual and auditory stimuli in a semantic matching task. A single area in the superior temporal gyrus (Fig. 6) was more active when the auditory stimulus shared the same location as the semantically matching visual stimulus. This region in the right temporal cortex, as part of the ventral “what” pathway in auditory processing, appears to be sensitive to the auditory localization cues, perhaps to provide spatial information to facilitate performance in the semantic matching task. This finding is in line with the results of Sestieri et al. (2006) and Meinenbrock et al. (2007), who also found enhanced

Figure 5.

Results of the random-effects event-related analysis for the contrasts “spatially congruent > baseline,” exclusively masked with the contrast “spatially incongruent > baseline,” and for the contrast “spatially incongruent > baseline,” exclusively masked with the contrast “spatially congruent > baseline” ($n = 15$ subjects). Clusters that show significant activations

exclusively in the spatial congruence condition are depicted in red, clusters showing significant activations exclusively in the spatial incongruence condition are depicted in green [P_{corr} (FWE) < 0.05 on cluster level]. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

activation in the right STG/MTG region for spatially congruent audiovisual stimuli in comparison to spatially incongruent stimuli. The presently measured effects seem to be less pronounced than those of the studies cited, but this might be due to the fact that in these studies, participants had to perform an explicit location-matching task, while in the present experiment auditory spatial information was merely implicitly varied and not relevant for the semantic matching task. Our result also supports a possible left and right hemispheric specialization of the superior temporal sulcus for processing “what” and “where” information as suggested by Sestieri et al. (2006). According to this approach, left STS would be especially involved in the integration of semantic information, while the right

STS is not as restricted and combines more general aspects of cross-modal stimuli. The differential role of left and right STS in spatial congruency effects for audiovisual processing needs further independent confirmation in further studies.

A post-hoc significant correlation between the subjects’ ability to locate the sound sources and the BOLD response in a small region in the left inferior frontal cortex was determined for 14 subjects. Inferior and middle frontal regions have been identified as being related to error processing and conflict monitoring (e.g., Botvinick et al., 2004; see Doehrmann and Naumer, 2008, for a review covering semantic aspects of multisensory processing; Hein et al., 2007; Noppeney et al., 2007; Weissman et al., 2004). With

TABLE III. Significant clusters for the contrast “spatially congruent > baseline” exclusively masked with the contrast “spatially incongruent > baseline” and for the contrast “spatially incongruent > baseline” exclusively masked with the contrast “spatially congruent > baseline”; P_{corr} (FWE) < 0.05 on cluster level

Hemisphere and Region	BA	Congruent > baseline exclusively masked with incongruent > baseline						Incongruent > baseline exclusively masked with congruent > baseline					
		MNI coordinates			Cluster			MNI coordinates			Cluster		
		x	y	z	z-value	Voxels		x	y	z	z-value	Voxels	
L SOG,MOG,IOG, LinG,FG,MTG, ParaHippG, Cer	19/20/37	-42	-52	-18	5.35	13		-34	-58	-10	5.68	109	
		-34	-40	-18	5.45	10		-36	-88	22	5.69	40	
		-30	-50	-12	5.44	14							
		-42	-70	-12	5.50	18							
R MOG, IOG, FG	18/19/37	38	-58	-16	5.61	14		36	-44	-20	5.32	15	
								42	-76	-12	6.05	35	
R SOG, AngG, MTG	19/39							34	-76	26	5.56	21	
L STG, MTG	21/22							-52	-10	-6	5.46	32	
R STG, MTG	21/22							58	-8	-2	5.28	20	
L STG, TransTG, PostCG	22/29/40/41/42							-58	-28	12	5.65	75	
R STG, TransTG, Insula, PostCG	13/22/29/40/41/42/43	40	-20	12	5.64	22		56	-28	6	5.61	72	
R PreCG, PostCG	3/4/6							34	-20	58	6.32	88	

P_{corr} (FWE) < 0.05 on cluster level, extent threshold > 10 voxels.

MOG, middle occipital gyrus; IOG, inferior occipital gyrus; FG, fusiform gyrus; STG, superior temporal gyrus; MTG, middle temporal gyrus; LingG, Lingual gyrus; PreCG, Precentral gyrus; PostCG, Postcentral gyrus; ParaHippG, parahippocampal gyrus; Cer, cerebellum; TransTG, Transverse temporal gyrus; AngG, Angular gyrus

respect to these earlier studies, our cluster appears to be located a bit more anterior and inferior in the left ventrolateral and orbitofrontal cortex. Thus we suppose that this activation is not related to conflict monitoring processes, especially since we found that this cluster was activated in trials with spatially congruent audiovisual information and the BOLD-response positively correlated with the ability of the subjects to process the provided auditory spatial cues.

Rather this region in the left ventrolateral and orbitofrontal cortex seems to play a role in combining and integrating coherent audiovisual information from the “what” and “where” pathways.

The orbitofrontal cortex is known as a region that receives input from different modalities including visual and auditory object information (e.g., Barbas, 1988; see Kringelbach and Rolls, 2004 for a review; Morecraft et al., 1992; Rolls and Deco, 2002; Romanski et al., 1999; Thorpe et al., 1983). Other work showed that orbitofrontal areas are active when subjects have to make intuitive judgments about the coherence of visual (Volz and von Cramon, 2006) and auditory (Volz et al., 2008) object information. In the work of Volz et al., the BOLD-response in this region rises, when subjects make an initial experience of coherence. In our experiment, the BOLD-response in this region was increased in subjects who could accurately judge the spatial

Figure 6.

Results of the random-effects event-related fMRI analysis with the second-level contrast spatial congruence > spatial incongruence ($n = 15$ subjects; $P_{\text{corr}} < 0.05$ on cluster level). The right panel shows mean % signal change in the significant cluster (right STG/MTG region) for the conditions “spatially congruent,” “spatially incongruent,” and “baseline.” [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Figure 7.

Results of the regression between the BOLD response (difference of % signal change between spatially congruent and incongruent trials) and the performance level of each subject in the auditory spatial localization task (lower panel) and location of the activated cluster [upper panel; IFG, MFG (BA 10, 11, 47); peak voxel: $x, y, z = -42, 46, -12$] ($n = 14$; $P_{\text{corr}} < 0.05$ on cluster level). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

location of the sound source, when audiovisual information was presented spatially congruent. As such it seems to be ideally capable of detecting perceived coherences in semantic and spatial information that it receives from both the visual and auditory modalities.

The role of the PFC in combining “what” and “where” information has been postulated ever since several studies have shown that spatial information from both, the auditory and the visual modalities, is projected to the dorsolateral part of the PFC, while information about the identity of an object is projected to the ventrolateral part of the PFC (e.g., Petrides and Pandya, 1984; Rauschecker and Tian, 2000; Ungerleider et al., 1989). Rao et al. (1997) revealed the role of the prefrontal cortex of macaques in combining visual “what” and “where” information. The authors recorded the activity of 195 neurons of the lateral PFC and found that 52% of these neurons responded to object and spatial working memory related task phases as well. These types of cells were equally distributed between the dorsolateral and the ventrolateral part of the prefrontal cortex. But there are a number of conflicting findings concerning the role of ventrolateral PFC in processing spatial information. Cohen et al. (2009) for example found that neurons in the ventrolateral prefrontal cortex of nonhuman primates were only responsive to nonspatial auditory stimuli. Also Kennerley and Wallis (2009a) found no evidence that orbitofrontal areas were engaged in spatial attention processes in maca-

ques, although in another study (Kennerley and Wallis, 2009b) they found the ventrolateral PFC to encode spatial information in working memory as well. Postle et al. (2000) did not find a dissociation of processing object and spatial information in visual working memory in prefrontal areas. In contrast, Sala et al. (2003) showed that middle and inferior prefrontal cortices were more active in humans during a visual working memory task related to object identity rather than object location, although the authors suggest that the representation of objects in working memory may be distributed across dorsal and ventral parts of the PFC. The psychophysical study by Sala and Courtney (2009) showed that spatial and nonspatial information in working memory only seems to be integrated in cases where the spatial information is task-relevant. Russ et al. (2008) confirmed the role of the ventrolateral PFC in auditory object identification as part of the ventral processing stream by single-cell recordings in macaques. Rämä et al. (2004) showed in humans that the ventrolateral part of the PFC is more specialized in processing auditory object information. However, the magnitude of this dissociation between “what” and “where” processing in prefrontal cortex during working memory tasks also seems to be dependent on the cognitive requirements in the auditory domain (see Rämä, 2008, for a review).

In this study we found that increased activation in the left ventrolateral PFC for spatially congruent trials correlated with subjects’ ability to localize the auditory stimuli. From these results we propose that those participants, who were able to perceive spatial congruency between auditory and visual object information, could form an initial experience of coherent object information (Volz et al., 2008) that is reflected in the found higher brain activation in the orbitofrontal cortex. It could stem from an integration of information from the audiovisual ventral “what” pathway and the dorsal “where” pathway in the PFC. But it could also result from an integration of audiovisual spatial information with audiovisual object information, reaching the orbitofrontal cortex from ventral pathways (e.g. Kringelbach and Rolls, 2004), to form an early and rough percept of the audiovisual environment. Another explanation for the exhibited activation could be that the integration of audiovisual “what” and “where” information from ventral and dorsal pathways in the PFC enhances the retrieval of conceptual representations. According to Badre and Wagner (2007), the anterior part of the left ventrolateral PFC plays a role in controlled access to stored conceptual representations. From the present data it cannot be decided which of these possibilities caused the orbitofrontal activation. Methods that offer better temporal resolution to follow the spreading of information along the cortical pathways (e.g., ERPs) should help to gain insight on this issue. A further limitation of our study is that we did not measure BOLD responses during the auditory localization task. Thus it remains uncertain whether the activation in orbitofrontal cortex would also be exhibited for explicit auditory localization tasks.

Our results, however, emphasize the importance of using veridical auditory spatial stimuli for examining audiovisual spatial congruency effects with fMRI. Auditory spatial perception is highly individual and depends on several idiosyncratic cues like head size, and the shape, cavities and contours of the pinnae (Hammershøi and Møller, 1996; King et al., 2001; Middlebrooks and Green, 1991; Mrsic-Flogel et al., 2001). In an fMRI experiment, it is therefore difficult to present realistic auditory spatial information to every subject, since the stimuli can only be presented via headphones or custom-built earplugs. If the considerable additional measurement effort for determining individual head-related transfer functions (HRTFs) cannot be afforded, it is crucial to use adequate (generic) HRTFs for creating auditory cues that render the spatial impression optimally for most subjects. In this study we used the HRTFs of a carefully designed head-and-torso simulator (VALDEMAR, Aalborg University) that was shown to yield filter functions that produce a more veridical result than other artificial heads (Minnaar et al., 2001). Nevertheless, as shown in our post-hoc auditory localization task, the spatial impression was not optimal for all of our subjects and they differed to a great extent in their ability to localize the stimuli, from near guessing rates to almost 100% correct (Fig. 3). For examining the integration of audiovisual spatial information it therefore is crucial to improve the auralization of the stimuli. The best solution would be to use individually measured HRTFs. This seems to be especially important, since the BOLD signal appears to vary with the quality of the auditory spatial cues. However, caution is necessary here, since the correlation between sound-source localization performance and activation in the inferior frontal gyrus was established post hoc, whereas the main task performed during MRI recordings did not require explicit judgments about the location of the sound sources.

CONCLUSION

Our results show that spatially congruent presented audiovisual object information leads to enhanced BOLD signal in a cluster in the right temporal lobe even though the auditory spatial information was only implicitly varied without the necessity of being processed to do the task. We also found that the activation in a cluster of the left inferior frontal cortex correlated with the subjects' ability to locate the sound source. These points to improved integration of audiovisual object information when the perceived location of both modalities coincides.

ACKNOWLEDGMENTS

The authors would like to thank Alfred Zeitler and Mark Dirst for recording some of the sounds that were used in this study.

REFERENCES

- Alain C, Arnott SR, Hevenor S, Graham S, Grady CL (2001): "What" and "where" in the human auditory system. *PNAS* 98:12301–12306.
- 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.
- Badre D, Wagner AD (2007): Left ventrolateral prefrontal cortex and the cognitive control of memory. *Neuropsychologia* 45:2883–2901.
- Barbas H (1988): Anatomic organization of basoventral and medio-dorsal visual recipient prefrontal regions in the rhesus monkey. *J Comp Neurol* 276:313–342.
- Beauchamp MS, Lee KE, Argall BD, Martin A (2004): Integration of auditory and visual information about objects in superior temporal sulcus. *Neuron* 41:809–823.
- Belardinelli MO, Sestieri C, Di Matteo R, Delogu F, Del Gratta C, Ferretti A, Caulo M, Tartaro A, Romani GL (2004): Audio-visual crossmodal interactions in environmental perception: An fMRI investigation. *Cogn Process* 5:167–174.
- Botvinick MM, Cohen JD, Carter CS (2004): Conflict monitoring and anterior cingulate cortex: An update. *Trends Cogn Sci* 8:539.
- Brett M, Anton JL, Valabregue R, Poline JB (2002): Region of interest analysis using an SPM toolbox [abstract] Presented at the 8th International Conference on Functional Mapping of the Human Brain, June 2–6, Sendai, Japan. Available on CD-ROM in *NeuroImage* 16(2).
- Calvert GA (2001): Crossmodal Processing in the Human Brain: Insights from Functional Neuroimaging Studies. *Cereb Cortex* 11:1110–1123.
- Cohen YE, Russ BE, Davis SJ, Baker AE, Ackelson AL, Nitecki R. (2009): A functional role for the ventrolateral prefrontal cortex in non-spatial auditory cognition. *PNAS* 106:20045–20050.
- Courtney SM, Petit L, Maisog JM, Ungerleider LG, Haxby JV (1998): An Area Specialized for Spatial Working Memory in Human Frontal Cortex. *Science* 279:1347–1351.
- Doehrmann O, Naumer MJ (2008): Semantics and the multisensory brain: How meaning modulates processes of audio-visual integration. *Brain Res* 1242:136–150.
- Frassinetti F, Bolognini N, Làdavas E (2002): Enhancement of visual perception by crossmodal visuo-auditory interaction. *Exp Brain Res* 147:332–343.
- Friston KJ, Ashburner J, Poline JB, Frith CD, Heather JD, Frackowiak RSJ (1995a) Spatial registration and normalization of images. *Hum Brain Mapp* 2:165–189.
- Friston KJ, Holmes AP, Worsley KJ, Poline JP, Frith CD, Frackowiak RSJ (1995b) Statistical parametric maps in functional imaging: a general linear approach. *Hum Brain Mapp* 2:189–210.
- Grill-Spector K (2003): The neural basis of object perception. *Curr Opin Neurobiol* 13:159–166.
- Hall DA, Haggard MP, Akeroyd MA, Palmer AR, Summerfield AQ, Elliott MR, Gurney EM, Bowtell RW (1999): "Sparse" temporal sampling in auditory fMRI. *Hum Brain Mapp* 7:213–223.
- Hammershøi D, Møller H (1996): Sound transmission to and within the human ear canal. *J Acoust Soc Am* 100:408–427.
- Harrington LK, Peck CK (1998): Spatial disparity affects visual-auditory interactions in human sensorimotor processing. *Exp Brain Res* 122:247–252.

- Haxby JV, Grady CL, Horwitz B, Ungerleider LG, Mishkin M, Carson RE, Herscovitch P, Schapiro MB, Rapoport SI (1991): Dissociation of Object and Spatial Visual Processing Pathways in Human Extrastriate Cortex. *Proc Natl Acad Sci U S A* 88:1621–1625.
- Hein G, Doehrmann O, Muller NG, Kaiser J, Muckli L, Naumer MJ (2007): Object familiarity and semantic congruency modulate responses in cortical audiovisual integration areas. *J Neurosci* 27:7881–7887.
- Hocking J, Price CJ (2008): The Role of the Posterior Superior Temporal Sulcus in Audiovisual Processing. *Cereb Cortex* 18:2439–2449.
- Kennerley SW, Wallis JD (2009a): Encoding of Reward and Space During a Working Memory Task in the Orbitofrontal Cortex and Anterior Cingulate Sulcus. *J Neurophysiol* 102:3352–3364.
- Kennerley SW, Wallis JD (2009b): Reward-Dependent Modulation of Working Memory in Lateral Prefrontal Cortex. *J Neurosci* 29:3259–3270.
- King AJ, Schnupp JWH, Doubell TP (2001): The shape of ears to come: dynamic coding of auditory space. *Trends Cogn Sci* 5:261–270.
- Kringelbach ML, Rolls ET (2004): The functional neuroanatomy of the human orbitofrontal cortex: evidence from neuroimaging and neuropsychology. *Prog Neurobiol* 72:341–372.
- Lewis JW, Wightman FL, Brefczynski JA, Phinney RE, Binder JR, DeYoe EA (2004): Human Brain Regions Involved In Recognizing Environmental Sounds. *Cereb Cortex* 14:1008–1021.
- Macaluso E, Driver J (2005): Multisensory spatial interactions: a window onto functional integration in the human brain. *TINS* 28:264–271.
- Macaluso E, George N, Dolan R, Spence C, Driver J (2004): Spatial and temporal factors during processing of audiovisual speech: a PET study. *Neuroimage* 21:725–732.
- Macaluso E, Frith CD, Driver J (2005): Multisensory stimulation with or without saccades: fMRI evidence for crossmodal effects on sensory-specific cortices that reflect multisensory location-congruence rather than task-relevance. *Neuroimage* 26:414–425.
- Maeder PP, Meuli RA, Adriani M, Bellmann A, Fornari E, Thiran JP, Pittet A, Clarke S (2001): Distinct pathways involved in sound recognition and localization: a human fMRI study. *Neuroimage* 14:802–816.
- McDonald JJ, Teder-Sälejärvi WA, Hillyard SA (2000): Involuntary orienting to sound improves visual perception. *Nature* 407:906–908.
- Meinenbrock A, Naumer MJ, Doehrmann O, Singer W, Muckli L (2007): Retinotopic effects during spatial audio-visual integration. *Neuropsychologia* 45:531–539.
- Meyer GF, Wuerger S, Greenlee MW: Interactions between auditory and visual semantic stimulus classes: Evidence for common processing networks for speech and body actions. *J Cogn Neurosci* (in press).
- Middlebrooks JC, Green DM (1991): Sound localization by human listeners. *Annu Rev Psychol* 42:135–159.
- Minnaar P, Olesen SK, Christensen F, Møller H (2001): Localization with binaural recordings from human and artificial heads. *J Audio Eng Soc* 49:323–336.
- Morecraft RJ, Geula C, Mesulam M-M (1992): Cytoarchitecture and neural afferents of orbitofrontal cortex in the brain of the monkey. *J Comp Neurol* 232:341–358.
- Mrsic-Flogel TD, King AJ, Jenison RL, Schnupp JWH (2001): Listening Through Different Ears Alters Spatial Response Fields in Ferret Primary Auditory Cortex. *J Neurophysiol* 86:1043–1046.
- Naghavi HR, Eriksson J, Larsson A, Nyberg L (2007): The claustrum/insula region integrates conceptually related sounds and pictures. *Neurosci Lett* 422:77–80.
- Noppeney U, Josephs O, Hocking J, Price CJ, Friston KJ (2007): The effect of prior visual information on recognition of speech and sounds. *Cereb Cortex* 18:598–609.
- Noppeney U, Ostwald D, Werner S (2010): Perceptual Decisions Formed by Accumulation of Audiovisual Evidence in Prefrontal Cortex. *J Neurosci* 30:7434–7446.
- Petrides M, Pandya DN (1984): Projections to the Frontal Cortex From The Posterior Parietal Region in the Rhesus Monkey. *J Comp Neurol* 228:105–116.
- Postle BR, Stern CE, Rosen BR, Corkin S (2000): An fMRI Investigation of Cortical Contributions to Spatial and Nonspatial Visual Working Memory. *Neuroimage* 11:409–423.
- Rämä P (2008): Domain-dependent activation during spatial and nonspatial auditory working memory. *Cogn Process* 9:29–34.
- Rämä P, Poremba A, Sala JB, Yee L, Malloy M, Mishkin M, Courtney SM (2004): Dissociable Functional Cortical Topographies for Working Memory Maintenance of Voice Identity and Location. *Cereb Cortex* 14:768–780.
- Rao SC, Rainer G, Miller EK (1997): Integration of What and Where in the Primate Prefrontal Cortex. *Science* 276:821–824.
- Rauschecker JP, Tian B (2000): Mechanisms and streams for processing of “what” and “where” in auditory cortex. *Proc Natl Acad Sci U S A* 97:11800–11806.
- Rolls ET, Deco G (2002): *Computational Neuroscience of Vision*. Oxford: Oxford University Press, pp 569.
- Romanski LM, Tian B, Fritz J, Mishkin M, Goldman-Rakic PS, Rauschecker JP (1999): Dual streams of auditory afferents target multiple domains in the primate prefrontal cortex. *Nat Neurosci* 2:1131–1136.
- Russ BE, Ackelson AL, Baker AE, Cohen YE (2008): Coding of Auditory-Stimulus Identity in the Auditory Non-Spatial Processing Stream. *J Neurophysiol* 99:87–95.
- Sala JB, Courtney SM (2009): Flexible working memory representation of the relationship between an object and its location as revealed by interactions with attention. *Atten Percept Psychophys* 71:1525–1533.
- Sala JB, Rämä P, Courtney SM (2003): Functional topography of a distributed neural system for spatial and nonspatial information maintenance in working memory. *Neuropsychologia* 41:341–356.
- Sestieri C, Di Matteo R, Ferretti A, Del Gratta C, Caulo M, Tartaro A, Olivetti Belardinelli M, Romani GL (2006): “What” versus “Where” in the audiovisual domain: An fMRI study. *Neuroimage* 33:672–680.
- Stevenson RA, James TW (2009): Audiovisual integration in human superior temporal sulcus: Inverse effectiveness and the neural processing of speech and object recognition. *Neuroimage* 44:1210–1223.
- Teder-Sälejärvi WA, Di Russo F, McDonald JJ, Hillyard SA (2005): Effects of spatial congruity on audio-visual multimodal integration. *J Cognitive Neurosci* 9:1396–409.
- Thorpe SJ, Rolls ET, Maddison S (1983): Neuronal activity in the orbitofrontal cortex of the behaving monkey. *Exp Brain Res* 49:93–115.
- Ungerleider LG, Mishkin M (1982): Two cortical visual systems. In: Ingle DJ, Goodale MA, Mansfield RJW, editors. *Analysis of Visual Behavior*, Cambridge, MA: MIT press. pp 549–586.

- Ungerleider LG, Haxby JV (1994): “What” and “where” in the human brain. *Curr Opin Neurobiol* 4:157–165.
- Volz KG, von Cramon DY (2006): What neuroscience can tell about intuitive processes in the context of perceptual discovery. *J Cog Neurosci* 18:2077–2087.
- Ungerleider LG, Gaffan D, Pelak VS (1989): Projections from inferior temporal cortex to prefrontal cortex via the uncinate fascicle in rhesus monkeys. *Exp Brain Res* 76:473–84.
- Volz KG, Rübsamen R, von Cramon DY (2008): Cortical regions activated by the subjective sense of perceptual coherence of environmental sounds: A proposal for a neuroscience of intuition. *Cogn Affect Behav Neurosci* 8:318–328.
- Wagner AD, Paré-Blagoev EJ, Clark J, Poldrack RA (2001): Recovering Meaning: Left Prefrontal Cortex Guides Controlled Semantic Retrieval. *Neuron* 31:329–338.
- Weissman DH, Warner LM, Woldorff MG (2004): The neural mechanisms for minimizing cross-modal distraction. *J Neurosci* 24:10941–10949.
- Werner S, Noppeney U (2010): Superadditive Responses in Superior Temporal Sulcus Predict Audiovisual Benefits in Object Categorization. *Cereb Cortex* 20(8):1829–1842.
- Zatorre RJ, Bouffard M, Ahad P, Belin P (2002): Where is where in the human auditory cortex? *Nat Neurosci* 5:905–909.
- Zatorre RJ, Bouffard M, Belin P (2004): Sensitivity to Auditory Object Features in Human Temporal Neocortex. *J Neurosci* 24:3637–3642.
- Zimmer U, Macaluso E (2007): Processing of multisensory spatial congruency can be dissociated from working memory and visuo-spatial attention. *Eur J Neurosci* 26:1681–1691.