Exp Brain Res (2009) 194:91–102
DOI 10.1007/s00221-008-1672-6

RESEARCH ARTICLE

Integration of auditory and visual information in the recognition of realistic objects

Clara Suied · Nicolas Bonneel · Isabelle Viaud-Delmon

Received: 8 July 2008 / Accepted: 26 November 2008 / Published online: 18 December 2008

Abstract Recognizing a natural object requires one to pool information from various sensory modalities, and to ignore information from competing objects. That the same semantic knowledge can be accessed through different modalities makes it possible to explore the retrieval of supramodal object concepts. Here, object-recognition processes were investigated by manipulating the relationships between sensory modalities, specifically, semantic content, and spatial alignment between auditory and visual information. Experiments were run under realistic virtual environment. Participants were asked to react as fast as possible to a target object presented in the visual and/or the auditory modality and to inhibit a distractor object (go/no-go task). Spatial alignment had no effect on object-recognition time. The only spatial effect observed was a stimulus–response compatibility between the auditory stimulus and the hand position. Reaction times were significantly shorter for semantically congruent bimodal stimuli than would be predicted by independent processing of information about the auditory and visual targets. Interestingly, this bimodal facilitation effect was twice as large as found in previous studies that also used information-rich stimuli. An interference effect was observed (i.e. longer reaction times to semantically incongruent stimuli than to the corresponding unimodal stimulus) only when the distractor was auditory. When the distractor was visual, the semantic incongruence did not interfere with object recognition. Our results show that immersive displays with large visual stimuli may provide large multimodal integration effects, and reveal a possible asymmetry in the attentional filtering of irrelevant auditory and visual information.

Keywords Audiovisual · Reaction time · Object recognition · Spatial disparity · Multisensory integration · Simon effect · Human

C. Suied · I. Viaud-Delmon
CNRS, UPMC UMR 7593, Hôpital de la Salpêtrière, Paris, France

C. Suied (✉) · I. Viaud-Delmon
IRCAM, CNRS UMR 9912, 1, place Igor Stravinsky, 75004 Paris, France
e-mail: clara.suied@ircam.fr

N. Bonneel
REVES, INRIA, Sophia-Antipolis, France

Introduction

Sensory cues across modalities help us to apprehend natural situations: when we are in a train station, we can see a train approaching as well as hear it. A single object concept—here, the train—is activated by different sensory components (for recent reviews, see Martin 2007; Patterson et al. 2007). How does our nervous system decide whether these different sensory components refer to a unique object and not to different objects? To perceive a single object, we must constantly bind together several cues that provide related information from different senses. They can be related among structural (e.g. time, space) or cognitive (e.g. semantic content) factors (Bedford 2001, 2004). However, structural and cognitive factors have seldom been studied in conjunction in the context of multisensory integration.

We investigated the influence of semantic content (i.e. cognitive factor) and spatial alignment (i.e. structural factor) on auditory–visual object recognition in a go/no-go task, with realistic 3D stimulation. Two meaningful objects

were presented randomly to participants. The objects were defined either by the combination of auditory and visual components or by each unimodal component presented alone. Visual and auditory components of bimodal stimuli either belonged to the same object (semantically congruent) or to different objects (semantically incongruent). Auditory objects were displayed in two possible spatial configurations: at 0° in azimuth and on the right at 40° in azimuth. Thus, bimodal stimuli could be either spatially aligned or spatially disparate. The participants' task was to react to the target object as fast as possible, irrespective of whether it was presented in the auditory or visual modality or both, and to inhibit their response to the distractor object.

Spatial alignment appeared unimportant in the case of a simple bimodal detection RT task (Hughes et al. 1994). For recognition tasks, however, conflicting results have been reported. Using an intensity recognition task, with the spatial dimension not relevant to the task, Teder-Salejarvi et al. (2005) reported no effect of spatial alignment. Gondan et al. (2005) found shorter RTs for spatially aligned stimuli than for spatially disparate stimuli. In addition, the role of the spatial alignment on semantically *incongruent* stimuli has never been studied. Does a spatial disparity help to disentangle an auditory stimulus and a visual stimulus when they are semantically incongruent? We compared RTs to spatially aligned stimuli with RTs to spatially disparate stimuli, both for semantically congruent and semantically incongruent cases.

The combined effect of auditory and visual information about the same object should lead to shorter RTs than unisensory information about this object, a phenomenon known as the redundant signal effect (RSE, Kinchla 1974). Miller (1982) has proposed a mathematical framework (the so-called race model violation) to decide whether the observed bimodal facilitation effect is the result of a separate activation or a coactivation of both sensory channels.[1] Three possible levels of coactivation were put forward in previous studies: sensory processing (Hershenson 1962; Savazzi and Marzi 2008), decision (Miller 1982; Schroger and Widmann 1998), or motor preparation (Giray and Ulrich 1993). When facilitation occurs in a situation where perceptual analysis is performed in separate modalities, coactivation is presumably taking place at a decision stage (although this is not incompatible with a coactivation taking place also at a sensory stage). In the present experiment, coactivation at a decision stage predicts a RSE for semantically congruent stimuli (redundant target). In addition, we examined whether the novel experimental setup we used (realistic objects, large screen and 3D vision) increased the size of the RSE beyond the level obtained in similar experiments with meaningful stimuli.

To investigate the possible levels of integration of the RSE (sensory level or decision level), it has been suggested to compare the redundant target conditions not only with a single target but also with a target presented with a distractor (Grice et al. 1984; Grice and Gwynne 1987; Grice and Canham 1990). With a similar approach, some recent studies focused on the role of semantic congruence in behavioral facilitation, where redundant targets were semantically congruent stimuli and non-redundant targets were semantically incongruent stimuli (Molholm et al. 2004; Laurienti et al. 2004). They observed that RTs to semantically incongruent stimuli were longer than RTs to semantically congruent stimuli. The authors thus suggested that semantic congruency between meaningful auditory and visual stimuli could influence multisensory integration. Therefore, we predicted shorter RTs in the case of semantically congruent bimodal stimuli compared to semantically incongruent stimuli. However, these findings could also be explained by the fact that semantically congruent stimuli were also redundant stimuli, whereas semantically incongruent stimuli contained only one target: redundancy and semantic congruency were intermingled. An additional experiment was performed to disentangle these two parameters, in which redundant conditions were also semantically incongruent conditions, whereas non-redundant conditions were semantically congruent.

Finally, we studied the role of a distractor on object recognition or how the semantic processing of objects could be revealed by semantic interference (see also Grice and Reed 1992 for a study on letter recognition). Longer RTs for incongruent stimuli compared to RTs for unimodal stimuli would mean that auditory and visual information belonging to different objects can interfere with the object-recognition process. In Molholm et al.'s (2004) study, semantically incongruent bimodal targets produced the same RTs as the corresponding unimodal targets. This is inconsistent with the authors' hypothesis of an interference between auditory and visual information in semantic processing. In contrast, with linguistic-type stimuli, Laurienti et al. (2004) did find a significant difference between the incongruent and unimodal conditions, with longer RTs to the incongruent conditions.

In summary, to explore the retrieval of supramodal object concepts, we jointly investigated the role of structural and cognitive factors on multisensory object recognition through (1) conflicting spatial cues, (2) 3D immersion, (3) semantic congruence, and (4) semantic interference.

[1] A third alternative has been proposed by Mordkoff and Yantis (1991), showing that inter-stimulus contingencies could, in some cases, entirely explain the violation of the race model, thus challenging the conclusion of an integration of the sensory channels in the presence of these contingencies.

Experimental procedures

Participants

Twenty volunteers (6 women; mean age 30 ± 6.8 years; all right-handed) participated in this experiment. All were naïve with respect to the purpose of the experiment. None of them reported having hearing problems, and all reported normal or corrected-to-normal vision. The study was carried out in accordance with the Declaration of Helsinki. All participants provided informed consent to participate in the study.

Apparatus

The experiment took place in an acoustically damped and sound proof recording studio with the light switched off. The visual scene was presented on a 300 × 225 cm^2 stereoscopic passive screen (corresponding to 90° × 74° at a viewing distance of 1.5 m) and was projected with two F2 SXGA+ ProjectionDesign projectors. Participants wore polarized glasses. Auditory stimuli were presented via two KEF loudspeakers situated at 0° and 40° in azimuth, straight ahead at a distance of 1.5 m (see Fig. 1). During the experiment, a serial response box (Cedrus Corporation, model RB-730) was used to record participants' response time and accuracy.

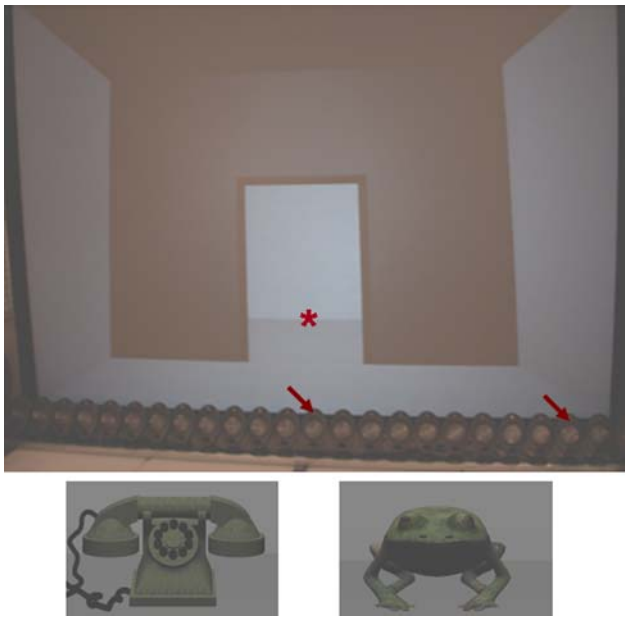

Fig. 1 Screenshots of the setup used in the experiment (*upper panel*): a large screen with the visual background (a door) and the loudspeakers. The *asterisk* indicates the location for the visual stimulus; the two *arrows* indicate the two loudspeakers used for the auditory stimuli (one at 0° in azimuth, the other at 40°). The *bottom panel* represents screenshots of the two visual stimuli, the telephone and the frog

Stimuli

Two meaningful objects were used, either visual and/or auditory; one was the target (a telephone) and the other a distractor (a frog). The duration of each stimulus was 500 ms. The two objects were situated at a virtual distance of 2.5 m.

Images

A 3D model of the frog was obtained from the CROSSMOD models database (http://www.crossmod.org). The 3D model of the telephone was obtained from a 3D library (http://www.turbosquid.com/FullPreview/Index.cfm/ID/232502). They were positioned centrally in the horizontal plane at 0° in azimuth. The two objects were adjusted to the same size in the three dimensions. Both images subtended 8° in the vertical angle and 12° in the horizontal angle. In addition, the same texture and the same illumination parameters were applied to both objects (see Fig. 1, bottom panel). The visual stimulus was embedded in a virtual environment representing a room; objects appeared behind a door situated in the center of this room (see Fig. 1, top panel).

Sounds

Auditory stimuli were complex sounds (16 bit; 44,100 Hz digitization). A frog sound was obtained from the Hollywood Edge database and a telephone sound was recorded at the INRIA lab (previously used in Moeck et al. 2007). They were modified using audio editing software (Adobe Audition version 1.5) to be 500 ms in duration. They were presented at a level of 65 dB SPL measured at the head of the listener and could be presented at two spatial locations, 0° in azimuth, i.e. straight ahead, or at 40° on the right in azimuth. The sounds used during the experiment were correctly identified by five listeners during a pilot study (with the sounds presented via loudspeakers).

Procedure

Participants were comfortably seated in a chair at 1.5 m from the screen and asked to look at the position where visual stimuli appeared. Participants were asked to give a speeded response to target stimuli (go) and to withhold any response to distractor stimuli (no-go). They were asked to press the response button with their right index if the target (the telephone) was present, either in the visual and/or auditory modality. They were invited to keep their right index finger in contact with the button between trials. Participants were explicitly instructed to ignore stimuli other than the telephone (i.e. the frog). It was further explained to them that they had to respond also to

Table 1 Each condition is defined as a function of semantic congruence and spatial alignment for both target and/or distractor stimuli

Condition	Stimuli		Name	RT (ms)	% Misses
	Auditory	Visual			
Targets					
Semantically congruent bimodal					
Aligned	Target at 0°	Target	$A+_0V+$	338 ± 8	0.1 ± 0.1
Disparate	Target at 40°	Target	$A+_{40}V+$	326 ± 7	0.3 ± 0.2
Semantically incongruent bimodal					
Aligned	Distractor at 0°	Target	$A-_0V+$	408 ± 8	0.4 ± 0.3
	Target at 0°	Distractor	$A+_0V-$	385 ± 11	0.2 ± 0.2
Disparate	Target at 40°	Distractor	$A+_{40}V-$	368 ± 12	1.2 ± 0.7
	Distractor at 40°	Target	$A-_{40}V+$	404 ± 10	0.2 ± 0.2
Unimodal					
Visual	None	Target	$V+$	392 ± 6	0.1 ± 0.1
Auditory	Target at 0°	None	$A+_0$	392 ± 10	0.4 ± 0.2
	Target at 40°	None	$A+_{40}$	374 ± 10	0.4 ± 0.2
Condition	Stimuli		Name		% FA
	Auditory	Visual			
Non-targets					
Semantically congruent bimodal					
Aligned	Distractor at 0°	Distractor	$A-_0V-$		15.4 ± 1.9
Disparate	Distractor at 40°	Distractor	$A-_{40}V-$		17.0 ± 2.1
Unimodal					
Visual	None	Distractor	$V-$		7.0 ± 2.0
Auditory	Distractor at 0°	None	$A-_0$		5.0 ± 1.7
	Distractor at 40°	None	$A-_{40}$		3.7 ± 1.0

The target stimulus was a telephone and the distractor stimulus a frog. RTs ± standard error of the mean (SEM) and percentage of misses ± SEM are detailed for each go condition, as the percentage of false alarms (FA) ± SEM for each no-go condition. RTs were first transformed to a log scale and then averaged across all participants. The log scale is converted back to ms for clarity

semantically incongruent stimuli, in which only the visual or the auditory element was the target. During the experiment, no reference was made at any time to sound localization. During each trial, participants were presented with a visual stimulus alone, an auditory stimulus alone, or a combined auditory–visual stimulus. For delivery of the bimodal stimulus conditions, the visual and auditory stimulus onsets were simultaneous. All stimulus conditions are detailed in Table 1. They were defined along the semantic and the spatial dimensions, comprising unimodal and bimodal conditions.

Unimodal stimulus conditions could be visual or auditory with a sound displayed at 0° or 40° in azimuth. Bimodal conditions could be either semantically congruent (an image and a sound belonging to the same object) or incongruent (an image and a sound belonging to different objects). Incongruent conditions included both task-relevant (telephone) and task-irrelevant (frog) information. They were all target conditions. In addition, bimodal objects—congruent and incongruent—were presented in two spatial conditions: (a) spatially aligned, with no spatial disparity between the auditory and the visual stimuli, i.e. both at 0° in azimuth; (b) spatially disparate, with a spatial disparity of 40° in azimuth between the auditory and visual stimuli, i.e. the visual stimulus displayed at 0° and the auditory one at 40° in azimuth (on the right).

The three unimodal conditions were presented 65 times each in total (50 times for the telephone stimulus and 15 times for the frog). For the semantically congruent bimodal conditions, either at 0° or 40°, the telephone stimulus was presented 50 times, the frog stimulus 25 times. The semantically incongruent stimulus conditions, either at 0° or 40°, were presented 50 times each in total, 25 times with a visual target and 25 times with an auditory one. The entire experiment for each participant consisted in 445 stimuli of which 350 (i.e. 79%) were task-relevant stimuli (go responses).

These stimuli were presented on five separate blocks of trials. Participants performed practice trials until they were comfortable with the task. Breaks were encouraged between blocks to maintain high concentration and prevent fatigue. The inter-stimulus interval (ISI) was randomly varied between 1.5 and 3 s. The order of the stimuli presentation was pseudo-randomized to limit predictability. The entire experimental session lasted about 30 min.

Statistical analysis

For each participant, RTs were recorded. Responses were first analyzed to remove error trials. For the current study, errors included anticipations (RTs less than 100 ms and RTs greater than 1,000 ms). Any RT outside these limits was considered an outlier and was discarded. Percentage of false alarms was analyzed by one-way nonparametric repeated-measures analyses of variance (ANOVA; the Friedman test). $P < 0.05$ was considered to be statistically significant.

The distribution of RTs was highly skewed and the distribution of the residuals was not normal, a well-known result for RTs distribution (see Ulrich and Miller 1993; Luce 1986). Each RTs value was thus transformed to its natural logarithm (ln), before averaging ln(RT) for each condition. With such a transformation, the mean values that will be further analyzed are well representative of the mode of the distribution (another solution could be to analyze the median RTs of the non-transformed distribution).[2] To identify between-condition differences in mean ln(RTs), a repeated-measures ANOVA was conducted with the nine conditions as a within-subjects factor (V+, $A+_0$, $A+_{40}$, $A+_0V+$, $A+_{40}V+$, $A-_0V+$, $A-_{40}V+$, $A+_0V-$, $A+_{40}V-$). A Kolmogorov–Smirnov test was performed to check for the normality of the distribution of residuals of the ANOVA. For this analysis, we pooled together the results for all conditions in order to increase the power of the statistical test. Finally, to account for violations of the sphericity assumption, P values were adjusted using the Huynh-Feldt correction. $P < 0.05$ was considered to be statistically significant. To analyze these first results in more detail, two other analyses were then performed. First, to identify an overall effect of the spatial alignment, we ignored the V+ condition and performed a 2 (spatial alignment) × 4 (conditions A+, A+V+, A−V+ and A+V−) repeated-measures ANOVA. Secondly, following this result, we pooled together the spatially aligned and spatially disparate for a given stimulus condition ($A+_0$ pooled with $A+_{40}$, $A+_0V+$ pooled with $A+_{40}V+$, $A-_0V+$ pooled with $A-_{40}V+$, and $A+_0V-$ pooled with $A+_{40}V-$) and performed a repeated-measures ANOVA on the five resulting conditions (A+, V+, A+V+, A+V− and A−V+).

Concerning the bimodal facilitation effect, two models have been proposed to explain this phenomenon, a separate activation model, called the race model (Raab 1962), which is a simple probabilistic one, and a coactivation model (Miller 1982, 1986). A separate activation model assumes that for an auditory–visual stimulus, the auditory and the visual components of the bimodal stimulus are processed independently. There is a race between the auditory and the visual components, and the response to the bimodal stimulus is triggered by the winner of the race. Because the probability of getting a RT lower than a given t is higher for a bimodal stimulus than for either unimodal stimuli, the race model predicts that responses to a bimodal stimulus will be shorter than responses to either unimodal stimulus alone. In contrast, the coactivation model postulates a convergence between the two components of bimodal stimuli. As a result, a bimodal stimulus is processed faster than the fastest single stimulus composing the bimodal stimulus. Miller (1982) has developed a mathematical means to decide between the two models. Given the RT distribution to unimodal stimuli, a prediction of separate activation models is described by the race model inequality:

$$P_{AV}(t) \leq P_A(t) + P_V(t), \quad \text{for all } t,$$

where P is the cumulative probability density function (CDF) of RTs, with the subscript of P to distinguish between auditory–visual condition (AV), visual only (V) or auditory only (A) conditions. If the observed RT to bimodal stimulus is shorter than that predicted by the race model, then, the race model can be rejected in favor of a coactivation model.

Thus, to determine if the semantically congruent bimodal stimuli ($A+_0V+$ and $A+_{40}V+$) resulted in responses that were shorter than could be predicted on the basis of both unimodal stimuli ($A+_0$ or $A+_{40}$ and V+) processed independently, RTs distributions were estimated against the race model prediction (Miller 1982, 1986). We used the algorithm described by Ulrich et al. (2007) to test the race model. CDFs of RTs are first estimated for each participant in each condition. Percentile values are then computed from each side of the inequality, from the 0.025th percentile until the 0.975th percentile, in 0.05 increments (0.025, 0.075, …, 0.925, 0.975). The percentile values are further aggregated across participants. Finally, to examine whether the race model is significantly violated ($P < 0.05$), separate paired t tests are performed to compare the observed percentile values of the bimodal condition with those predicted by the race model. Here, we compared the

[2] To be able to compare our results with previous studies, all the analyses were also performed on the initial non-transformed distribution.

$A+_0V+$ condition to both V+ and $A+_0$ conditions and the $A+_{40}V+$ condition to both V+ and $A+_{40}$ ones.

In addition, to investigate more precisely a possible interference effect due to the distractors, we also computed the race model by comparing semantically congruent bimodal stimuli (redundant targets) to semantically incongruent bimodal stimuli (non-redundant targets), for both spatially aligned and spatially disparate conditions (comparison of $A+_0V+$ to $A+_0V-$ and $A-_0V+$ and comparison of $A+_{40}V+$ to $A+_{40}V-$ and $A-_{40}V+$). In this case, the percentile values were computed in 0.1 increments, because of a smaller number of RTs in the semantically incongruent conditions. Note that the experimental design included some inter-stimulus contingencies that might possibly be responsible for a violation of this race model (Mordkoff and Yantis 1991): there was a predictive relationship between the presence of the distractor and the presence of a target, and this could facilitate the processing of the target.

Results

During debriefing, all participants reported that they perceived the spatial discrepancy between the auditory and the visual stimuli.

Each single value of RTs was first transformed to a log scale and then averaged across conditions and across all participants. The log scale is converted back to ms for display purposes. Precise values of RTs, false alarms and misses (±SEM) are presented in Table 1.

Accuracy

Nonparametric repeated-measures ANOVA (Friedman's test) revealed a significant effect of condition (V−, $A-_0$, $A-_{40}$, $A-_0V-$, and $A-_{40}V-$) on percentage of false alarms ($\chi^2(4) = 41.39$; $P < 0.0001$). Percentage of false alarms was higher with a semantically congruent bimodal stimulus than with a unimodal one (see Table 1), thus revealing a small speed-accuracy trade-off (RTs were shorter for semantically congruent bimodal stimuli than for unimodal stimuli; see "Reaction times"). However, because the overall performance of participants was still very good (see also the very low percentage of misses, Table 1), these differences are not considered further.

Reaction times

No anticipations were found and only 15 RTs out of the 7,000 responses were longer than 1,000 ms and had to be discarded. A Kolmogorov–Smirnov test was performed on the distribution of the residuals of the ANOVA and revealed that this distribution was not different from a normal distribution ($d = 0.06$; $N = 180$; $P > 0.1$). This result validates the log-transformation and shows that the original distribution of RTs was indeed lognormal.

The repeated-measures ANOVA comparing ln(RTs) revealed a significant main effect between conditions ($F_{8,152} = 42.36$; $\varepsilon = 0.6$; $P < 0.0001$).[3] A first observation of these results (see Table 1) indicates that RTs were shorter for the $A+_{40}$ than for the $A+_0$ conditions (contrast: $t_1 = 3.3$; $P < 0.005$). This result can be interpreted as a spatial stimulus–response compatibility, the so-called Simon effect: reactions are performed more quickly if the response corresponds spatially to the stimulus, even when stimulus location is irrelevant to the task (Simon and Craft 1970; Simon et al. 1981; Zorzi and Umilta 1995; Lu and Proctor 1995). If all the differences observed between the spatially aligned and spatially disparate conditions (see Table 1) are solely due to a Simon effect, the difference should hold for any condition (auditory alone, semantically congruent or semantically incongruent). In contrast, if spatial alignment had a specific effect on object recognition, its effect should interact with the condition. We thus performed a 2 (spatial alignment) × 4 (conditions A+, A+V+, A−V+, and A+V−) ANOVA with both factors as within-subjects factors. It revealed a significant main effect of the spatial alignment ($F_{1,19} = 14.12$; $P < 0.005$) and a main effect of the condition ($F_{3,57} = 75.38$; $\varepsilon = 0.9$; $P < 0.0001$). Importantly, no significant interaction between spatial alignment and the four conditions was found ($F_{3,57} = 1.20$; $P = 0.32$). This means that the only influence of spatial alignment was a speeding of responses to all right-of-center target conditions, due to the Simon effect. What we observed is not an effect of the spatial relationship between the auditory and the visual stimulus, but rather an effect of the spatial compatibility between the auditory stimulus and the hand position.

To focus the rest of our analysis on other potential differences, we pooled together the spatially aligned and spatially disparate conditions and performed a new repeated-measures ANOVA on the five resulting conditions (A+, V+, A+V+, A+V−, and A−V+). These data are represented in Fig. 2. This new analysis revealed a significant effect of condition ($F_{4,76} = 58.17$; $\varepsilon = 0.7$; $P < 0.0001$). We then performed four planned comparisons to study (1) RT bimodal facilitation (A+V+ compared to A+ and V+ together, that is, coefficient 2 for A+V+, −1 for A+ and −1 for V+), (2) the comparison between both unimodal conditions (A+ compared with V+), and (3, 4) inhibition effect (comparison of V+ with A−V+ and of A+ with A+V−). Since these planned

[3] For these analyses, as for all the other ones, the ANOVA on the non-transformed distribution gave similar results.

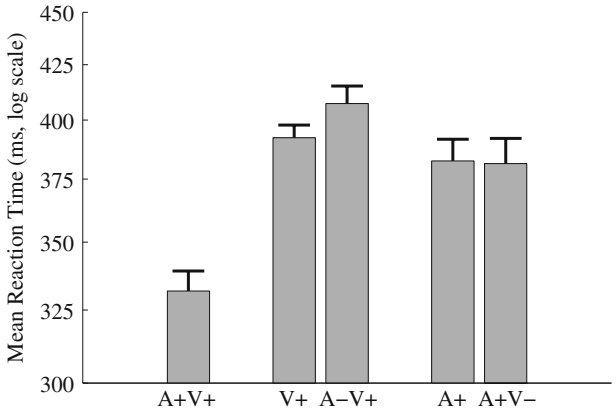

Fig. 2 RTs of the unimodal (A+ and V+), bimodal congruent (A+V+) and bimodal incongruent (A−V+ and A+V−) conditions are presented. There was no effect of the spatial alignment on object recognition; we thus pooled together the spatially aligned and disparate conditions. RTs were first transformed to a log scale and then averaged across all participants. The error bars represent one standard error of the mean. The log scale is converted back to ms for displays purposes. RTs to the A+V+ condition are significantly shorter than both unimodal conditions. RTs to the A+V+ condition are also significantly shorter than both bimodal incongruent conditions. RTs to the A−V+ condition are significantly longer than to the V+ condition, whereas RTs to the A+V− condition are similar to RTs to A+ condition

comparisons were non-orthogonal, a P value of 0.0125 was considered as statistically significant ($0.0125 = 0.05/4$, where 4 is the number of planned comparisons performed and 0.05 the alpha-level). (1) RTs were significantly shorter for the A+V+ condition than for the V+ and A+ conditions considered together ($t_4 = 16.87$, $P < 0.0001$). This is consistent with the hypothesis that bimodal stimuli are more quickly recognized than unimodal ones. (2) RTs to unimodal visual stimuli were equivalent to RTs to unimodal auditory stimuli ($t_4 = 1.7$, $P = 0.1$). (3) RTs to A−V+ were significantly slower than RTs to V+ ($t_4 = 3.8$, $P < 0.005$), thus revealing an interference effect when the target was visual and the distractor in the auditory modality. (4) Finally, RTs to A+V− were equivalent to RTs to the A+ condition ($t_4 = 0.21$, $P = 0.83$): when the target was auditory, the visual distractor did not influence RTs.

Test of the race model

For the spatially aligned stimuli, the race model comparing $A+_0V+$ to V+ and $A+_0$ was significantly violated ($P < 0.01$) for the percentiles in the lower part of the RTs, i.e. from the 0.025th to 0.525th (see Fig. 3, left panel). For the spatially disparate stimuli, the race model comparing $A+_{40}V+$ to V+ and $A+_{40}$ was also significantly violated ($P < 0.01$) for all percentiles in the lower part of the RTs, i.e. from the 0.025th to 0.575th (see Fig. 3, right panel). Thus, in both cases, the race model cannot fully explain the pattern of RTs observed for the bimodal stimuli ($A+_0V+$ and $A+_{40}V+$).

The race model comparing $A+_0V+$ to $A+_0V-$ and $A-_0V+$ was significantly violated only for the first percentile ($P < 0.05$). However, this result is likely to be due to inter-stimulus contingencies more than to the coactivation of the two sensory channels (see "Statistical analysis"). The race model comparing $A+_{40}V+$ to

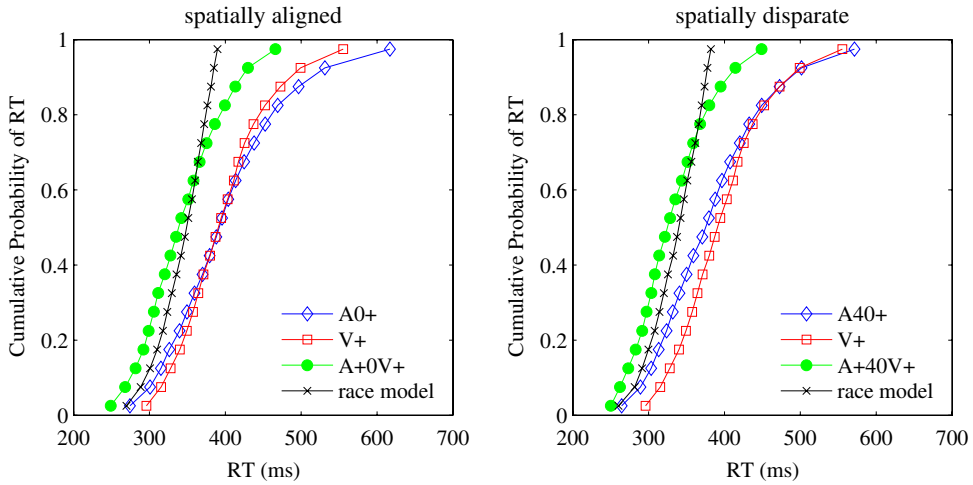

Fig. 3 Observed cumulative distribution functions (CDFs) of RTs in the two bimodal conditions ($A+_0V+$ and $A+_{40}V+$; see Table 1) and race model predictions from the respective CDFs of RTs to unimodal conditions (V+, $A+_0$ and $A+_{40}$). The *left panel* shows the spatially aligned condition; the *right panel* shows the spatially disparate condition. As can be seen, the proportions of responses to bimodal stimuli (*filled circles*) are shorter than the summed (*multiple symbol*) respective proportions for unimodal stimuli (*blank squares* and *diamonds*). This difference is significant for the percentiles in the lower part of the RTs, i.e. from the 0.025th to 0.525th in the spatially aligned condition and from the 0.025th to 0.575th in the spatially disparate condition

$A+_{40}V-$ and $A-_{40}V+$ was not significantly violated. This result confirms that the presence of a distractor had an effect on the redundancy gain. There was an interference produced by the distractor.

Size of the RSE

To quantify the auditory–visual integration more precisely, we computed the effect size (Cohen's d; see Cohen 1988) of the RSE observed in the $A+_0V+$ condition:

$$d = \frac{\min(\overline{RT_{V+}}, \overline{RT_{A+}}) - \overline{RT_{A+_0V+}}}{[\sigma(\min(RT_{V+}, RT_{A+})) + \sigma(RT_{A+_0V+})]/2}$$

where $\overline{RT}$ denotes the mean of the reaction time distribution and σ is the standard deviation (without logarithmic transformation to compare with previously reported values). Our data resulted in a value of $d = 1.79$. To compare this effect size to the size of the RSE previously observed in the literature, we also computed the d value for three studies: Giard and Peronnet (1999), Molholm et al. (2004) and Laurienti et al. (2004). These three studies were chosen because they used, like the present study, an identification paradigm with information-rich auditory–visual stimuli and they report both mean and dispersion of the data. Figure 4 illustrates the different values obtained. The effect size of the current study is clearly larger than all the other ones.

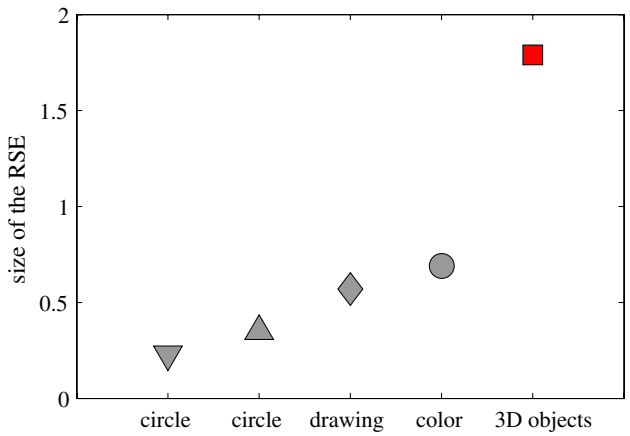

Fig. 4 The Cohen's d effect size of the RSE is shown for our data (*square*, value of 1.79) and for data reported in the literature of auditory–visual object recognition with complex and information-rich stimuli (the two *triangles* for Giard and Peronnet 1999; a *diamond* for Molholm et al. 2004; a *circle* for Laurienti et al. 2004). For the Giard and Peronnet study, we computed the effect size for the two groups of participants, AUD (auditory participants, with shorter RTs to the auditory alone condition; *downward pointing triangle*) and VIS (visual participants, with shorter RTs to the visual alone condition; *upward pointing triangle*). The effect size of the current study is clearly larger than all the other ones

Discussion and conclusions

Spatial alignment between auditory and visual information had no effect on object-recognition time. The only spatial effect observed was due to spatial compatibility between the auditory stimulus and the hand position. Responses to all right-of-center target conditions were speeded, thus revealing a Simon effect (e.g. Simon and Craft 1970). As expected, we observed a RSE. Reaction times were significantly shorter for semantically congruent bimodal stimuli than would be predicted by independent processing of information about the auditory and visual targets. The bimodal integration effect was larger than previously observed. We also found shorter RTs in the case of semantically congruent bimodal stimuli compared to semantically incongruent stimuli. Finally, we highlighted evidence of an asymmetric interference effect: only auditory distractors impaired reaction times in the case of incongruent stimuli.

Spatial alignment between the auditory stimulus and the hand position had a strong influence on reaction times (Simon effect). Similar stimulus–response compatibility in RSE experiments has already been observed (e.g. Grice et al. 1984). However, object recognition was unaffected by spatial alignment between the auditory and the visual stimuli. This result is consistent with many studies of spatial disparity in detection or recognition tasks, for auditory–visual integration (Bertelson et al. 1994; Hughes et al. 1994; Stein et al. 1996; Teder-Salejarvi et al. 2005) or auditory–somatosensory integration (Murray et al. 2005; Zampini et al. 2007), but it contradicts recognition results that have shown significantly shorter RTs in response to spatially aligned stimuli than to spatially disparate stimuli (Miller 1991; Gondan et al. 2005). Although previous studies have emphasized the importance of the spatial relationship in auditory–visual integration for tasks such as saccade generation or signal detection (Stein and Meredith 1993; Hughes et al. 1994; Frens et al. 1995; Harrington and Peck 1998; Frassinetti et al. 2002), object recognition appears to be a function where spatial alignment between the auditory and visual components of a bimodal stimulus is not essential. Moreover, spatial disparity did not help to segregate two different objects: there was no effect of a spatial disparity on the ability of participants to ignore one modality when two semantically incongruent auditory and visual information were presented. As previously highlighted (Bertelson et al. 1994; Hughes et al. 1994; Stein et al. 1996; Calvert et al. 1998; Calvert and Thesen 2004; Holmes and Spence 2005), the finding that spatial alignment is not required for object recognition could reflect the fact that this function probably involves brain regions containing neurons with broad spatial receptive fields. In natural environments, because of multiple acoustic

reflections on physical obstacles, it is quite possible that auditory and visual cues appear misaligned. Bimodal processing of complex objects seems to be able to accommodate this characteristic of natural environments.

In agreement with previous research (e.g. Miller 1982; Molholm et al. 2004; Laurienti et al. 2004), we found behavioral evidence for coactivation rather than separate independent processing of the individual components of bimodal targets (see also Miller 1991 for a comparison of two class of coactivation models, the independent coactivation model and the interactive coactivation model; for an alternative to the independent race model—the interactive race model—see Mordkoff and Yantis 1991; Schwarz 1996). Shorter RTs to bimodal congruent stimuli compared to unimodal stimuli could not be simply explained on the basis of statistical facilitation, as shown by a clear race model violation. Interestingly, this effect was twice as large as previous studies with similar meaningful stimuli. The large effect size could be due to a combination of differences with previous studies: the size of the visual object, the task used (go/no-go vs. choice RT, for example; see Grice and Canham 1990), the high-degree of realism of the current stimuli (although Radeau and Bertelson 1977, 1978 found no evidence that realism increases ventriloquism), the 3D display, the immersive environment, or the particularly large display we used.

RTs to semantically congruent bimodal stimuli (A+V+) were significantly shorter than RTs to either semantically incongruent stimuli, A−V+ or A+V−. This is in line with the findings of earlier experiments with information-rich stimulus recognition tasks (Laurienti et al. 2004; Molholm et al. 2004; Yuval-Greenberg and Deouell 2007). This effect could also be explained by the fact that semantically congruent stimuli were redundant stimuli, as opposed to semantically incongruent stimuli, which contained only one target. The additional experiment was designed to unravel these two intermingled factors (semantic congruence and redundancy). No bimodal facilitation effect was found when the redundant target stimulus was also a semantically incongruent stimulus (either for the unimodal non-redundant target or for the semantically congruent non-redundant target). Although a null result is difficult to interpret, this result tends to show that this was the incongruency of the redundant stimulus that prevents any bimodal facilitation. As a consequence, we suspect that semantic congruence did have an influence on multimodal object integration. Additional support for this conclusion comes from other studies. Miller (1991) found that the response to two redundant targets depended on their congruence (defined in this case on a pseudo-spatial dimension): RTs to congruent redundant targets were significantly shorter than RTs to incongruent redundant targets. In addition, Smith et al. (2007) recently reported that semantic congruence can influence multisensory integration even when only one sensory modality was useful for object identification. Altogether, these results suggest that object-based auditory–visual interactions are sensitive to the semantic content of the stimuli. In this case, shorter RTs to bimodal congruent stimuli would be the result of an enhanced activation of a single object representation (here, a telephone). This would presumably increase the signal relative to the noise, thus improving the 'signal-to-noise ratio' (where the signal is the target object and the noise the distractor object). Incongruent stimuli would have the opposite effect: the noise level would increase and the signal-to-noise ratio would go down, thus decreasing the performance (see Lehmann and Murray 2005 for a discussion of a model of object-based multisensory interactions).

The absence of an interference effect for auditory targets in incongruent trials shows that there was no inhibition by the visual distractor (incongruent A+V− compared to unimodal A+). This means that when the distractor was visual, there was no performance cost for processing an auditory target (see also Molholm et al. 2004). In contrast, and perhaps surprisingly, it seemed impossible to ignore an auditory distractor (incongruent A−V+ compared to unimodal V+). The race model was not violated when the semantically congruent condition was compared with the semantically incongruent ones, which further specifies this interference effect. To summarize, when a *visual distractor* and an auditory target object are simultaneously presented, results are consistent with a parallel activation of each object representation and independent processing. This is not the case when an *auditory distractor* and a visual target are simultaneously presented: the co-occurrence of bimodal information belonging to different objects results in poorer performance compared to the unimodal target, consistent with an interference within the object-recognition process. Consistent with this hypothesis, Schmitt et al. (2000) observed a cueing effect in detection tasks (Posner paradigm) with visual or auditory cues and visual targets, but not with auditory targets. We extend this finding to a more complex situation involving divided attention and an identification task. This suggests that the exogenous attention system is not completely supramodal (see also Alais et al. 2006) and reveals a possible asymmetry in the attentional filtering of irrelevant auditory and visual information. The extension of these results and their link to models of the representation of semantic knowledge in the human brain (Riddoch et al. 1988; Caramazza et al. 1990; Patterson et al. 2007) seems a particularly interesting topic for future investigations.

Acknowledgments We thank Khoa-Van Nguyen, Olivier Warusfel, George Drettakis, and Grace Leslie for their help. We are grateful to Shihab Shamma, Daniel Pressnitzer, Laurence Harris and two anonymous reviewers for useful comments on a previous version of this manuscript. This research was supported by the EU IST FP6 Open FET project CROSSMOD: "Crossmodal Perceptual Interaction and Rendering" IST-04891.

Appendix

An additional experiment was performed because we could not conclude from the results of our main experiment whether shorter RTs to semantically congruent stimuli than to semantically incongruent stimuli were due to semantic congruence or simply to redundancy of information. In this new experiment, the target stimuli were the sound of a frog ($A+_f$) and the image of a phone ($V+_p$). Participants had to respond to $A+_f$, $V+_p$, or when both were presented simultaneously ($A+_fV+_p$). In this case, the redundant target condition was also a semantically incongruent stimulus, whereas the non-redundant target condition was a semantically congruent stimulus. If the RSE observed in the main experiment was related to the semantic congruence between the auditory and the visual parts of the stimulus, there should be no bimodal integration for the incongruent stimuli (redundant targets) in the present control experiment. In addition, if semantically congruent trials benefited from crossmodal integration, mean RTs in semantically congruent trials (non-redundant target) should be shorter than mean RTs in semantically incongruent trials (redundant targets).

Eleven volunteers (5 women; mean age 30.9 ± 8 years; all but one right-handed) participated in the experiment. All were naïve with respect to the purpose of the experiment. None of them reported having hearing problems, and all reported normal or corrected-to-normal vision. All participants provided informed consent to participate in the study. Apparatus and stimuli were exactly the same as in the main experiment. Procedure was also highly similar, except for the definition of the go and no-go conditions. There were five go conditions: auditory frog alone ($A+_f$), visual phone alone ($V+_p$), auditory frog with a visual phone ($A+_fV+_p$), auditory frog with a visual frog ($A+_fV-_f$), and auditory phone with a visual phone ($A-_pV+_p$). The $A+_fV+_p$ condition was the only redundant target; the other four conditions were non-redundant targets. The no-go conditions were an auditory phone alone ($A-_p$), a visual frog alone ($V-_f$) and an auditory phone with a visual frog ($A-_pV-_f$). Each go condition was presented 48 times and each no-go condition was presented 20 times. In this additional experiment, there are no inter-stimulus contingencies. Thus, in the case of a potential RSE, this would not be due to the contingencies. The entire experiment for each participant consisted of 300 stimuli of which 240 (80%) were task-relevant stimuli (go responses). Statistical analyses were similar as the ones performed in the main experiment (log-transformation and ANOVA on the mean ln(RTs)), except that we did not remove RTs greater than 1,000 ms, due to the difficulty of the task that lead to RTs of the order of 650 ms on average.

Nonparametric repeated-measures ANOVA (Friedman's test) revealed a significant effect of condition (A−, V−, $A-V-_i$) on percentage of false alarms ($\chi^2(2) = 9.5$; $P < 0.01$). Percentage of false alarms was higher with a bimodal stimulus $A-_pV-_f$ (39.5 ± 6.7%) than with a unimodal one (21.4 ± 4.4% for $A-_p$ and 21.8 ± 4.9% for $V-_f$). Only 0.9 ± 0.1% of misses were observed. Overall, the larger number of false alarms here in comparison to the main experiment (around three times more) could reveal the difficulty of the task (attend to two different objects at the same time).

RTs of this additional experiment are represented in Fig. 5. The distribution of the residuals of the ANOVA was not different from a normal distribution (Kolmogorov–Smirnov test: $d = 0.09$; $N = 55$; $P > 0.2$). Overall, RTs observed in this additional experiment were much longer than those of the main experiment (more than 600 ms here, compared to around 350 ms in the main experiment). This confirms the difficulty of the task. To identify between-

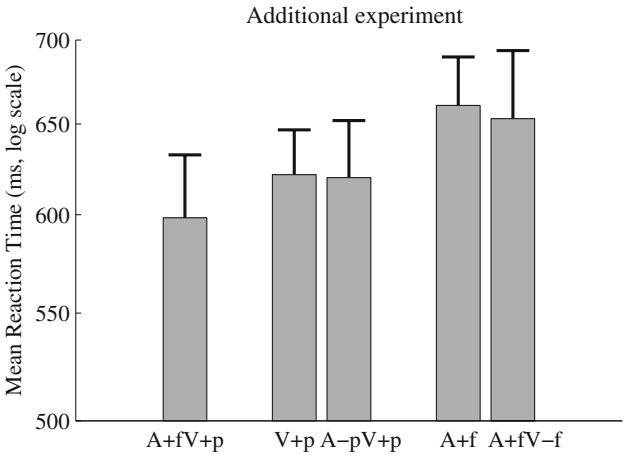

Fig. 5 RTs of the semantically incongruent bimodal (redundant target, $A+_fV+_p$), semantically congruent bimodal (non-redundant target, $A+_fV-_f$ and $A-_pV+_p$), and unimodal ($A+_f$ and $V+_p$) conditions are presented. RTs were first transformed to a log scale and then averaged across all participants. The error bars represent one standard error of the mean. The log scale is converted back to ms for displays purposes. There was no bimodal facilitation effect (RTs to the $A+_fV+_p$ condition are similar to RTs to the shortest unimodal condition, i.e. $V+_p$). RTs to the semantically congruent conditions ($A+_fV-_f$ and $A-_pV+_p$) were not shorter than RTs to the semantically incongruent condition ($A+_fV+_p$). The only significant differences observed are due to shorter RTs to the visual target alone compared to the auditory target alone

condition differences in mean ln(RTs), a repeated-measures ANOVA was conducted with the five conditions as a within-subjects factor ($A+_f$, $V+_p$, $A+_fV+_p$, $A+_fV-_f$, $A-_pV+_p$). It revealed a significant main effect of condition ($F_{4,60} = 6.14$; $\varepsilon = 0.8$; $P < 0.001$). Post hoc Tukey HSD tests revealed that this effect was due to a difference between $A+_f$ and $A+_fV+_p$ ($P < 0.001$) and a difference between $A+_fV+_p$ and $A+_fV-_f$ ($P < 0.004$). Importantly, there was no significant difference between the shortest of the unimodal conditions (here, $V+_p$) and the redundant target ($A+_fV+_p$) ($P = 0.5$). In other words, we observed no bimodal facilitation effect for redundant target (semantically incongruent stimulus). It is of course difficult to interpret unambiguously a null effect; however, it strongly suggests that semantic incongruence of redundant stimuli prevents any redundant facilitation effect.

References

Alais D, Morrone C, Burr D (2006) Separate attentional resources for vision and audition. Proc Biol Sci 273:1339–1345

Bedford FL (2001) Toward a general law of numerical/object identity. Cahiers de Psychologie Cognitive/Curr Psychol Cogn 20:113–176

Bedford F (2004) Analysis of a constraint on perception, cognition, and development: one object, one place, one time. J Exp Psychol Hum Percept Perform 30:907–912

Bertelson P, Vroomen J, Wiegeraad G, de Gelder B (1994) Exploring the relation between McGurk interference and ventriloquism. In: International conference on spoken language processing, Yokohama, Japan, pp 556–562

Calvert GA, Thesen T (2004) Multisensory integration: methodological approaches and emerging principles in the human brain. J Physiol Paris 98:191–205

Calvert GA, Brammer MJ, Iversen SD (1998) Crossmodal identification. Trends Cogn Sci 2:247–253

Caramazza A, Hillis AE, Rapp BC, Romani C (1990) The multiple semantic hypothesis: multiple confusions? Cogn Neuropsychol 7:161–189

Cohen J (1988) Statistical power analysis for the behavioral sciences, 2nd edn. Lawrence Erlbaum Associates, Hillsdale

Frassinetti F, Bolognini N, Ladavas E (2002) Enhancement of visual perception by crossmodal visuo-auditory interaction. Exp Brain Res 147:332–343

Frens MA, Van Opstal AJ, Van der Willigen RF (1995) Spatial and temporal factors determine auditory–visual interactions in human saccadic eye movements. Percept Psychophys 57:802–816

Giard MH, Peronnet F (1999) Auditory–visual integration during multimodal object recognition in humans: a behavioral and electrophysiological study. J Cogn Neurosci 11:473–490

Giray M, Ulrich R (1993) Motor coactivation revealed by response force in divided and focused attention. J Exp Psychol Hum Percept Perform 19:1278–1291

Gondan M, Niederhaus B, Rosler F, Roder B (2005) Multisensory processing in the redundant-target effect: a behavioral and event-related potential study. Percept Psychophys 67:713–726

Grice GR, Canham L (1990) Redundancy phenomena are affected by response requirements. Percept Psychophys 48:209–213

Grice GR, Gwynne JW (1987) Dependence of target redundancy effects on noise conditions and number of targets. Percept Psychophys 42:29–36

Grice GR, Reed JM (1992) What makes targets redundant? Percept Psychophys 51:437–442

Grice GR, Canham L, Gwynne JW (1984) Absence of a redundant-signals effect in a reaction time task with divided attention. Percept Psychophys 36:565–570

Harrington LK, Peck CK (1998) Spatial disparity affects visual–auditory interactions in human sensorimotor processing. Exp Brain Res 122:247–252

Hershenson M (1962) Reaction time as a measure of intersensory facilitation. J Exp Psychol 63:289–293

Holmes NP, Spence C (2005) Multisensory integration: space, time and superadditivity. Curr Biol 15:R762–R764

Hughes HC, Reuter-Lorenz PA, Nozawa G, Fendrich R (1994) Visual–auditory interactions in sensorimotor processing: saccades versus manual responses. J Exp Psychol Hum Percept Perform 20:131–153

Kinchla RA (1974) Detecting target elements in multielement displays: a confusability model. Percept Psychophys 15:149–158

Laurienti PJ, Kraft RA, Maldjian JA, Burdette JH, Wallace MT (2004) Semantic congruence is a critical factor in multisensory behavioral performance. Exp Brain Res 158:405–414

Lehmann S, Murray MM (2005) The role of multisensory memories in unisensory object discrimination. Brain Res Cogn Brain Res 24:326–334

Lu CH, Proctor RW (1995) The influence of irrelevant location information on performance: a review of the Simon and spatial Stroop effects. Psychon Bull Rev 2:174–207

Luce RD (1986) Response times: their role in inferring elementary mental organization. Oxford University Press, New York

Martin A (2007) The representation of object concepts in the brain. Annu Rev Psychol 58:25–45

Miller J (1982) Divided attention: evidence for coactivation with redundant signals. Cogn Psychol 14:247–279

Miller J (1986) Timecourse of coactivation in bimodal divided attention. Percept Psychophys 40:331–343

Miller J (1991) Channel interaction and the redundant-targets effect in bimodal divided attention. J Exp Psychol Hum Percept Perform 17:160–169

Moeck T, Bonneel N, Tsingos N, Drettakis G, Viaud-Delmon I, Alloza D (2007) Progressive perceptual audio rendering of complex scenes. In: ACM SIGGRAPH symposium on interactive 3D graphics and games

Molholm S, Ritter W, Javitt DC, Foxe JJ (2004) Multisensory visual–auditory object recognition in humans: a high-density electrical mapping study. Cereb Cortex 14:452–465

Mordkoff JT, Yantis S (1991) An interactive race model of divided attention. J Exp Psychol Hum Percept Perform 17:520–538

Murray MM, Molholm S, Michel CM, Heslenfeld DJ, Ritter W, Javitt DC, Schroeder CE, Foxe JJ (2005) Grabbing your ear: rapid auditory–somatosensory multisensory interactions in low-level sensory cortices are not constrained by stimulus alignment. Cereb Cortex 15:963–974

Patterson K, Nestor PJ, Rogers TT (2007) Where do you know what you know? The representation of semantic knowledge in the human brain. Nat Rev Neurosci 8:976–987

Raab DH (1962) Statistical facilitation of simple reaction times. Trans N Y Acad Sci 24:574–590

Radeau M, Bertelson P (1977) Adaptation to auditory–visual discordance and ventriloquism in semirealistic situations. Percept Psychophys 22:137–146

Radeau M, Bertelson P (1978) Cognitive factors and adaptation to auditory–visual discordance. Percept Psychophys 23:341–343

Riddoch MJ, Humphreys GW, Coltheart M, Funnell E (1988) Semantic systems or system? Neuropsychological evidence re-examined. Cogn Neuropsychol 5:3–25

Savazzi S, Marzi CA (2008) Does the redundant signal effect occur at an early visual stage? Exp Brain Res 184:275–281

Schmitt M, Postma A, de Haan E (2000) Interactions between exogenous auditory and visual spatial attention. Q J Exp Psychol A 53:105–130

Schroger E, Widmann A (1998) Speeded responses to audiovisual signal changes result from bimodal integration. Psychophysiology 35:755–759

Schwarz W (1996) Further tests of the interactive race model of divided attention: the effects of negative bias and varying stimulus-onset asynchronies. Psychol Res 58:233–245

Simon JR, Craft JL (1970) Effects of an irrelevant auditory stimulus on visual choice reaction time. J Exp Psychol 86:272–274

Simon JR, Sly PE, Vilapakkam S (1981) Effect of compatibility of SR mapping on reactions toward the stimulus source. Acta Psychol 47:63–81

Smith EL, Grabowecky M, Suzuki S (2007) Auditory–visual cross-modal integration in perception of face gender. Curr Biol 17:1680–1685

Stein BE, Meredith MA (1993) The merging of the senses. MIT, Cambridge

Stein BE, London N, Wilkinson LK, Price DD (1996) Enhancement of perceived visual intensity by auditory stimuli: a psychophysical analysis. J Cogn Neurosci 8:497–506

Teder-Salejarvi WA, Di Russo F, McDonald JJ, Hillyard SA (2005) Effects of spatial congruity on audio–visual multimodal integration. J Cogn Neurosci 17:1396–1409

Ulrich R, Miller J (1993) Information processing models generating lognormally distributed reaction times. J Math Psychol 37:513–525

Ulrich R, Miller J, Schroter H (2007) Testing the race model inequality: an algorithm and computer programs. Behav Res Methods 39:291–302

Yuval-Greenberg S, Deouell LY (2007) What you see is not (always) what you hear: induced gamma band responses reflect cross-modal interactions in familiar object recognition. J Neurosci 27:1090–1096

Zampini M, Torresan D, Spence C, Murray MM (2007) Auditory–somatosensory multisensory interactions in front and rear space. Neuropsychologia 45:1869–1877

Zorzi M, Umilta C (1995) A computational model of the Simon effect. Psychol Res 58:193–205