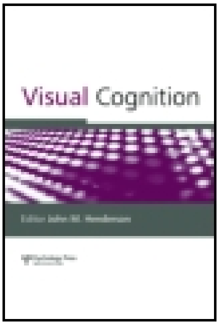


Visual Cognition
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/pvis20

Crossmodal binding in localizing objects outside the field of view

Andrea C. Piemo [a], Andrea Caria [b] & Umberto Castiello [c d]

[a] Department of Psychology, Royal Holloway, University of London, Egham, Surrey, UK

[b] Dipartimento di Scienze della Cognizione e della Formazione, Universitá di Trento, Rovereto, Italy

[c] Department of Psychology, Royal Holloway, University of London, Egham, Surrey, UK

[d] Dipartimento di Psicologia Generale, Universitá di Padova, Padova, Italy

Available online: 07 Nov 2007

To cite this article: Andrea C. Piemo, Andrea Caria & Umberto Castiello (2006): Crossmodal binding in localizing objects outside the field of view, Visual Cognition, 13:2, 223-246

To link to this article: http://dx.doi.org/10.1080/13506280544000273

VISUAL COGNITION, 2006, 13 (2), 223–246

Psychology Press
Taylor & Francis Group

Crossmodal binding in localizing objects outside the field of view

Andrea C. Pierno

Department of Psychology, Royal Holloway, University of London, Egham, Surrey, UK

Andrea Caria

Dipartimento di Scienze della Cognizione e della Formazione, Università di Trento, Rovereto, Italy

Umberto Castiello

Department of Psychology, Royal Holloway, University of London, Egham, Surrey, UK, and Dipartimento di Psicologia Generale, Università di Padova, Padova, Italy

Using virtual reality techniques we created a virtual room within which participants could orient themselves by means of a head-mounted display. Participants were required to search for a nonimmediately visually available object attached to different parts of the virtual room's walls. The search could be guided by a light and/or a sound emitted by the object. When the object was found participants engaged it with a sighting circle. The time taken by participants to initiate the search and to engage the target object was measured. Results from three experiments suggest that (1) advantages in starting the search, finding, and engaging the object were found when the object emitted both light and sound; (2) these advantages disappeared when the visual and auditory information emitted by the object was separated in time by more than 150 ms; (3) misleading visual information determined a greater level of interference than misleading auditory information (e.g., sound from one part of the room, light from the object).

Please address all correspondence to Umberto Castiello, Dipartimento di Psicologia Generale, Università di Padova, Via Venezia 8, 35131 Padova, Italy. Email: umberto.castiello@unipd.it

The authors would like to thank Andy Haswell and Philip Roberts for providing essential hardware support throughout all phases of this investigation. Angelo Maravita and Dean Lusher are also thanked for their useful comments on a previous version of this manuscript. Andrea Caria was supported by a Leverhulme Trust grant awarded to Prof. Umberto Castiello. Andrea Pierno was supported by a Thomas Holloway scholarship to Prof. Umberto Castiello.

http://www.psypress.com/viscog DOI:10.1080/13506280544000273

The notion that senses are better conceptualized as interrelated modalities rather than independent channels has been recently supported by several studies. These studies have provided evidence for neural and behavioural mechanisms in the processing of multisensory information (Bushara et al., 2003; Driver, 1996; Stein & Meredith, 1993).

The many behavioural consequences of multimodal integration have been investigated extensively with respect to the covert orienting of attention behaviour (without overt shifts such as eye or head movements), primarily concerned with the determination of stimulus location (for a review see Driver & Spence, 1998). A common finding from these studies was that better responses were elicited when stimuli presented in different modalities were located in the same (or a very close) position rather than in different positions. Covert shifts of attention in one modality tend to be accompanied by corresponding shifts in other modalities. When a target is expected on a particular side in just one modality its discrimination also improves on that side in other modalities. Further, the modality with the best spatial resolution (e.g., vision's superiority over audition) has the greatest influence on the location of the fused percept (Stekelenburg, Vroomen, & de Gelder, 2004).

Behavioural advantages of the integration of multisensory information have also been revealed using a variety of overt orienting responses (with overt shifts such as eye, head, and body movements). For example, studies of saccadic eye movements have demonstrated that saccadic latency is shorter for spatially and temporally coincident combinations of visual and auditory stimuli, compared to saccades to either single modality alone (Frens, van Opstal, & van der Willigen, 1995; Hughes, Reuter-Lorenz, Nozawa, & Fendrich, 1994; Lee, Chung, Kim, & Park, 1991). Faster responses to visual targets in the presence of spatially coincident auditory distractors have also been observed for eye–head gaze shifts (Corneil & Munoz, 1996).

Whereas the study of covert and overt crossmodal issues has allowed the establishment of a series of rules that appear to govern the optimal integration of multisensory signals, a number of issues remain unsolved. For one, it is as yet unknown how different modalities are integrated in overt searching for objects located outside the initial field of view, which we now refer to as not immediately visually available objects.

When overt responses are considered it is important to note that the transformation of sensory signals into motor output is a far from trivial problem given that the initial coding of stimulus location is very different for different sensory modalities. For example, visual information is initially topographically organized at the retina, and encoded in a retinotopic or eye-centred reference frame, whereas a tonotopic and head-centred representation is employed for auditory stimuli. Therefore, the integration of multisensory signals into one combined representation for the control of motor responses presents a considerable challenge to the nervous system. Furthermore, if a target object is located outside the field of view, the integration of multisensory information emanated by the object (light and/or sound) may be even more complex. For example, suppose that your mobile phone rings in a dimly lit room, but you have forgotten where you put it last. In this condition the search may be multimodal in nature, being guided both by the sound and by the service light. In addition, the processing of audiovisual information has to be constantly updated and maintained to the point at which the object is within the field of view and then engaged. Although the investigation of these processes poses several challenges it is of great importance because it emphasizes the multimodal nature of everyday situations in which overt responses towards objects of interest not immediately visible are often made.

One reason why this issue is still unsolved may lie in the difficulty of designing controlled stimuli and environments that allow the measurement of audiovisual crossmodal effects in such situations. Virtual reality, the simulation of real-world environments, may offer an unrivalled opportunity to explore these processes.

A number of studies have used virtual reality techniques in visual search tasks. In particular, they have emphasized the efficacy of 3-D virtual auditory displays to enhance the acquisition of visual targets (Bolia, D'Angelo, & McKinley, 1999; Nelson et al., 1998; Perrott, Cisneros, McKinley, & D'Angelo, 1996). For example, Perrott et al. (1996) demonstrated that the use of a 3-D virtual auditory cue produced a significant reduction in target acquisition time for the detection of a visual target presented at one of 264 different locations. Similarly, Nelson et al. (1998) demonstrated the beneficial effects of 3-D virtually localized auditory cues in a visual target acquisition task with respect to nonlocalized auditory cues or no auditory cues. However, while such studies tell us about the efficacy of virtual auditory cues in guiding visual search they do not take into account crossmodal issues because the combination of auditory and visual cues together to guide the search was not tested.

Other studies provide some evidence of possible multisensory integration effects comparing situations in which 3-D auditory cues were presented in isolation or accompanied by a visual cue (Bronkhorst, Veltman, & van Breda, 1996; Flanagan, McAnally, Martin, Meehan, & Oldfield, 1998). Bronkhorst et al. (1996) presented 3-D virtual auditory cues in isolation or together with a visual cue (radar display) in a flight simulation experiment in which participants had to locate and track a target aircraft as quickly as possible. Flanagan et al. (1998) administered a visual search paradigm in which participants were required to locate a target presented outside the initial field of view. Searches were guided by virtual 3-D auditory cues in isolation or together with a visual cue in the form of a 2-D arrow. The results from both studies suggested that the integration of visual and auditory spatial cues reduced search time dramatically compared to when the auditory and the visual cue were presented in isolation.

However, from these studies no definite conclusions in terms of crossmodal integration can be drawn. Firstly, these studies were designed to chiefly investigate the efficacy of the 3-D virtual auditory cues. Secondly, there are methodological problems concerning the nature of the cues. Whereas the auditory cues were exogenous in nature, the information carried by the visual cues (a central arrow or a radar display) needed an endogenous mode of control. Therefore, the visual and the auditory cues were based on noncomparable modes of processing, with the former requiring endogenous processing and the latter requiring an exogenous type of processing.

We have capitalized on the findings mentioned above to merge the classic reaction time approach with the ''virtual'' approach. In emphasizing the multimodal nature of everyday situations, by contrast with the nature of most previous crossmodal experiments, we seek to highlight the point that many interesting and important questions are excluded when stimuli are exclusively presented in front of the participants. These questions include some very fundamental issues about the architecture of crossmodal integration, such as: (1) Whether audiovisual binding occurs when an out-of-the-field-of-view object has to be localized on the basis of audiovisual cues that originate from it and are perceivable by the participants from the beginning of the search process; (2) whether there are limits on optimal sensory integration for this type of behaviour; (3) whether audiovisual information available to each sensory modality can be integrated for overt responses when the object from which this information comes from is not immediately visually available; (4) whether cues which are updated with respect to the movement of the participant can be retained in close temporal and spatial alignment. To shed some light on these issues we created a virtual reality room that participants could explore by means of a head-mounted display (HMD). Participants searched for an object attached to the virtual room's walls by following a light and/or a sound coming from the object's spatial location. Whereas the target object itself was not immediately visible, the auditory (sound) and/or strobe visual (light) information emanated by the target was available to vision and audition from the very beginning of the search. Thus, here we sought to examine whether and how multimodal information can be integrated and used to direct attention in order to localize a target that is itself out of the field of view. The first aim of the present study was to investigate the effect of visual and auditory sensory input relating to a single not immediately visible object. To this end participants were asked to search and engage for a nonimmediately visible object emitting both visual and auditory information as compared to conditions in which the search was guided by only one type of information (auditory or visual). Facilitation effects were found when the two cues were simultaneously presented. The second aim considered temporal proximity of different sensory inputs as a determinant for facilitation. To address this point visual and auditory information emitted by the object were presented at different intervals. Facilitation effects were found only when the two cues were presented with no delay. A final aim considers the spatial location in which the visual and the auditory information originate. To investigate this issue we contrasted the effects of spatially congruent and incongruent auditory or visual cues during the search process. Facilitation effects were evident when the visual and auditory information originated from the same spatial location that is anchored to the object.

EXPERIMENT 1

In Experiment 1 participants made a speeded overt search for a nonimmediately visually available object on the basis of auditory and/or visual information that originated from the object location. The object's location varied so as to make it unpredictable. The search could be guided by a sound, a light or the combination of sound and light. Note that both auditory and visual information was immediately available at the beginning of the search.

Method

Participants

All nine participants were volunteers and naïve as to the purpose of the experiment. Their ages ranged from 21 to 44 years with a mean of 29. All reported normal or corrected-to-normal vision and normal auditory functioning.

Apparatus and materials

A computer equipped with a Pentium III processor was used to present the visual cues and to record the time taken by participants to locate the target. Head orientation was monitored by a three-degrees-of-freedom head tracker (Intertrax^2, Intersense) that sampled head orientation at 256 Hz with the following angular range: pitch ±80°, yaw ±180°, roll ±90°. The auditory cues were presented by means of an audio module (TDT RP2.1 Real Time Processor, Tucker-Davis Technologies) designed for the delivery of three-dimensional sound. The RP2.1 audio module was equipped with a Scharc digital signal processor (DSP) running at 50 MHz able to synthesize and process wideband signals in real time (24 bit, 100 kHz bandwidth) with a 110 dB signal to noise ratio.

Each auditory cue was amplified through a precision power amplifier (TDT HB7 Headphone Buffer, Tucker-Davis Technologies) capable of delivering up to 1 W of power to headphones or other transducers. The HB7 is a stereo device with excellent channel separation, low signal distortion and a flat frequency response ranging from 10 Hz to 100 kHz. The output gain can be set between 0 and 27 dB in 3 dB steps, which allows for matching of dynamic range to the desired output level. The signal to noise ratio is 117 dB (20 Hz to 80 kHz).

The sound was delivered through a pair of earphones (ER-6 Isolator, Etymotic Research) with the following specifications: Frequency response 20 Hz to

16 kHz, impedance 48 ohms, 1 kHz sensitivity 108 dB SPL for a 4.0 volt input, maximum output 120 dB SPL, maximum continuous input 2.5 volts RMS. When properly sealed in the ears, these earphones are able to provide 15–20 dB of external noise isolation.

Visual displays and the visual cues were presented using a nonstereoscopic binocular HMD (Glasstron Sony PLM-S700E). The HMD provided a 30° horizontal and 22.5° vertical field of view (FOV) with an 800 × 600 pixel resolution (display mode VESA SVGA, vertical refresh frequency 75 Hz, horizontal refresh frequency 46.9 kHz). In order to avoid light or other visual distractors that might influence the perception of the presented visual displays, the HMD was inserted in a flexible rubber mask that was sealed to the participant's skin.

Visual virtual environment. The virtual environment presented on the HMD was a hexagonal-shaped room 6 m wide and 3.5 m in height centred on the observer's head (see Figure 1). The virtual room was uniformly illuminated by a diffuse source of light (50 cd/m^2) located above the observer's head at 3 m from the floor.

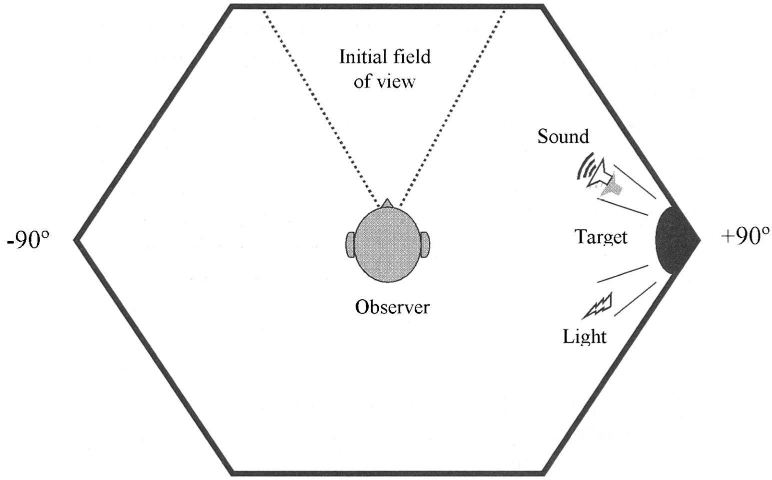
Visual display. A small circle with a radius of 1.8° of visual angle served as a sight. The sight was always presented in the centre of the observer's FOV within the HMD and followed the observer's head movements.

Visual target. The visual target was a red semisphere that subtended 1.6° of visual angle. The possible target location in the virtual environment was defined by the combination of two azimuths (90° and −90°) and two elevations (30° and −30°). As a consequence, the target could only be presented at four possible locations. The 0° azimuth and 0° elevation position served as the starting position. Negative azimuths were to the left of the starting positions, positive to the right. Negative elevations were below the starting position, positive above.

Visual cue. The visual cue used to indicate target location was a 360° uniformly diffuse strobe light (250 cd/m^2) graded in contrast that flickered every 300 ms.

Auditory cue. A virtual updating sound played back at a conversion rate of 50 kHz and an intensity of 70 dB SPL served as auditory cue. The updating auditory cue consisted of a series of pulses of white noise with a rise time of 5 ms and a white noise duration of 60 ms followed by a fall time of 5 ms, separated by a silence lasting 70 ms. The updating auditory cue was designed to provide constantly updated information at a rate of 7 Hz about the target's location in relation to the head position. For each noise burst, the appropriate function simulating the spatial region in which the target was located relative to

(a)



(b)

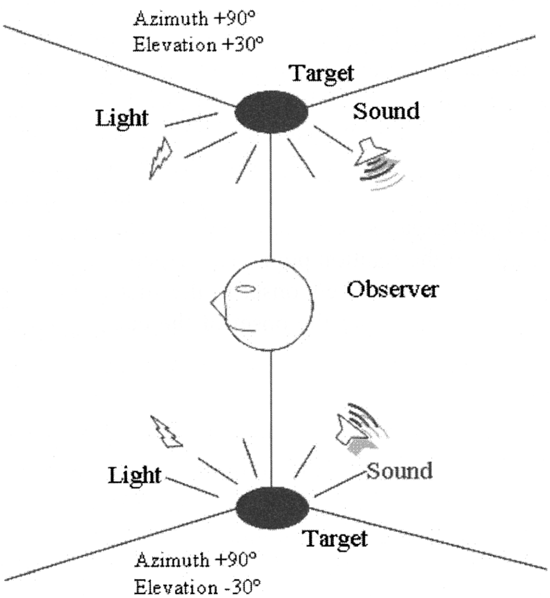


Figure 1. (a) A top view schematic representation of the virtual environment. In this example the target is located at 90°. Please note that the target could be positioned at both −30°, +30° of elevation. Both auditory and visual information originate from target location. Note that the target is not immediately visually available, but the auditory and/or visual information originating from the target is immediately perceivable by the participant. (b) A lateral view of a portion of the virtual room in which two of the four possible target locations are shown. Note that the remaining two target locations were at −90° azimuth, +30° or −30° elevation.

the instantaneous head position was selected to generate the proper spatial auditory cue from the library of the adopted Head-Related Transfer Functions (Kemar HRTFs, TDT). To generate 3-D virtual spatial auditory cues nonindividualized HRTFs were adopted.

Procedure

Participants performed the task sitting on a swivel chair wearing the HMD. The task was to locate the target in the virtual environment and overlap the sight over the target. At the beginning of each trial participants were required to overlap the sight with a fixation cross presented at 0° azimuth and 0° elevation, and hold it in this location for 2.5 s. Subsequently the fixation cross disappeared and the colour of the sighting circle turned from red to blue. This was a warning signal to the participants that the trial was initiating. Following a delay that varied randomly from 0.8 to 1.5 s, during which participants were instructed to maintain their head oriented towards the location of the fixation cross, the auditory and the visual cues were presented together or in isolation. The onset of the cue/s was the signal to start searching for the target. To perform the task participants were allowed to make head and body movements on the swivel chair. Trials in which any head movement preceded the onset of the cue/s were discarded and repeated randomly. The target was always presented outside the initial FOV of the HMD at one of the four possible locations. Once the target was found participants had to overlap the sight over the target and maintain that position for 50 ms. To indicate the end of the trial the target disappeared, the auditory and/or visual cues ceased and the HMD background turned yellow. To start a new trial participants were required to return to the starting position relocating the sight on the fixation point. The dependent measures were reaction time (RT), the time from the cue/s onset to the first head rotation over 3°, and search time (ST), the time from the onset of the cue/s to target acquisition.

Experimental design

There were three experimental conditions corresponding to the three types of cue: (1) Auditory, in which only the auditory cue was presented; (2) visual, in which only the visual cue was presented; and (3) combined, in which the auditory and the visual cue were simultaneously presented. Each participant completed a total of 108 trials divided equally into three experimental blocks.

Data analysis

Two separate analyses of variance (ANOVA) were conducted for the RT and the ST data sets with type of cue (auditory, visual, combined) as a within-subjects factor. Planned comparisons were performed using *t*-tests. Bonferroni corrections were applied when necessary.

Results and discussion

Reaction time

For the reaction time analysis the main factor type of cue was significant, $F(2, 16) = 31.395$, $p = .0001$. Reaction time was 541 ms for the auditory cue condition, 527 ms for the visual cue condition, and 451 ms for the combined cue condition (Figure 2a). Planned comparisons revealed that reaction time for the combined cue condition was significantly faster than for both the visual ($p = .0001$) and the auditory ($p = .001$) cue conditions (Figure 2a). However, no significant difference was found when reaction time obtained in the auditory cue condition was compared with reaction time obtained in the visual cue condition (Figure 2a; $p = 1$).

The reaction time results suggest that the integration of information across the senses plays an important role in the initial coding of target location. Combining auditory with visual information improves the speed of orienting towards the target location.

Search time

For the ANOVA conducted on search time the main factor type of cue was significant, $F(2, 16) = 47.834$, $p = .0001$. Search time was 1770 ms for the auditory cue condition, 1960 ms for the visual cue condition, and 1683 ms for the combined cue condition (Figure 2b). Planned comparisons revealed that search time for the combined cue condition was significantly faster than for both the visual ($p = .0001$) and the auditory ($p = .017$) cue conditions (Figure 2b). Further, in contrast with the results obtained for the reaction time analysis, search time for the visual condition was slower than search time for the auditory condition (Figure 2b; $p = .002$).

The search time results indicate that the initial facilitation found at RT level is maintained throughout the search action. However, in contrast to the reaction time results, search time varies with the type of information (visual or auditory) presented in isolation. The faster search time when only the auditory information was presented may reveal differences in the processing of visual and auditory cues during the search action. Here participants were required to perform a visual search task. In the visual condition the visual requirement of the task in addition to the coding of the visual cue may have resulted in an increase in visual load. This was not the case for the auditory condition given that the auditory cue did not put further strain on the visual modality, which could be fully utilized to perform the task. An alternative explanation, concerned with the nature of the spatial information carried by the two types of cue, may account for the faster search time obtained when only the auditory information was presented. It could be hypothesized that an auditory cue that was constantly updated at a rate of 7 Hz may have facili-

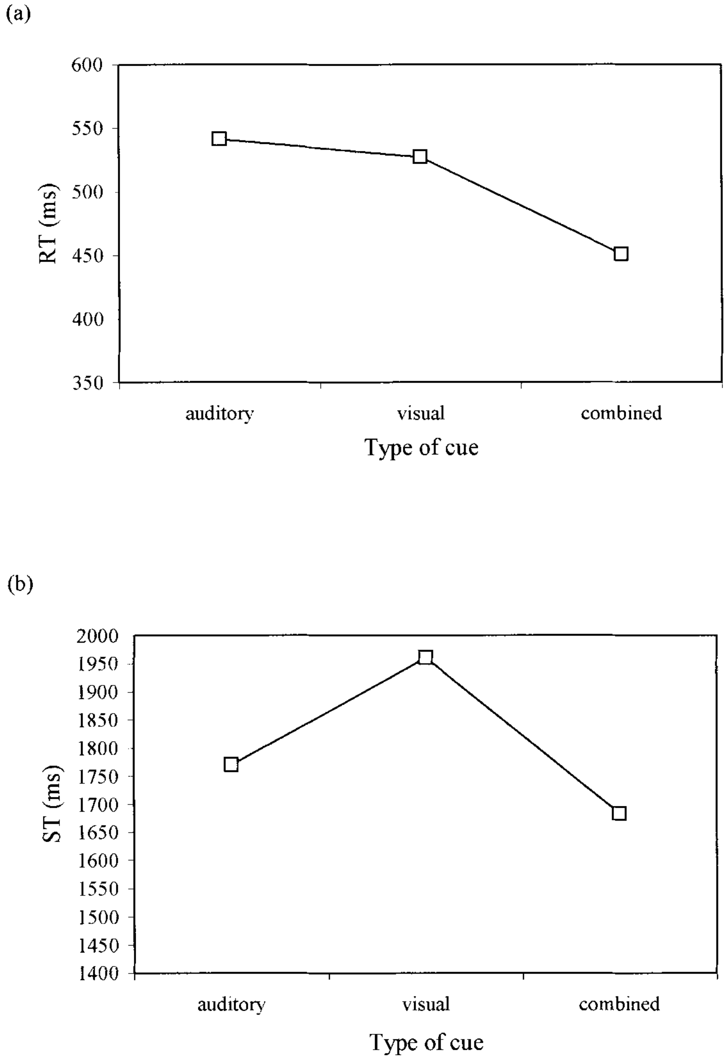


Figure 2. (a) Mean reaction time (RT) in milliseconds (ms) for Experiment 1 in the three experimental conditions: Auditory condition, visual condition, and combined condition. (b) Mean search time (ST) in milliseconds for Experiment 1 in the three experimental conditions: Auditory condition, visual condition, and combined condition.

tated the search. This is because it may have provided participants with more precise spatial information than the visual cue.

EXPERIMENT 2

Results from Experiment 1 confirm that the integration of sensory information has a beneficial effect in target localization (Perrott, Saberi, Brown, & Strybel, 1990). In Experiment 2 the nature of the determinants of such integration was investigated. In particular, we focused on the possibility that temporal delays between the presentation of the visual and the auditory information may affect the benefits of the visual-auditory binding found in Experiment 1. Here we introduced delays in the presentation of the two types of information (visual and auditory). If the simultaneous presentation of the visual and auditory information is a necessary condition to produce facilitation from crossmodal binding, then dissociating the two sources of information in time should reduce or eliminate the facilitation effects.

Method

Participants

All 10 participants were volunteers and naïve as to the purpose of the experiment. Their ages ranged from 19 to 44 years with a mean of 25. All reported normal or corrected to normal vision and normal auditory functioning.

Apparatus and materials

Apparatus and materials were the same as for Experiment 1.

Procedure

Procedure was the same as for Experiment 1 except for the following changes:

Cues onset asynchronies. Four cues onset asynchronies (COAs) were used: (1) The auditory and the visual cue were presented simultaneously (COA); (2) the interval between the presentation of the two cues was 150 ms (COA150); (3) the interval between the presentation of the two cues was 300 ms (COA300); and (4) the interval between the presentation of the two cues was 500 ms (COA500). When not simultaneously presented (COA150, COA300, and COA500) the cues could occur in two different orders: (1) Auditory cue followed by visual cue (AV order); and (2) visual cue followed by auditory cue (VA order). The start signal for the search was the onset of both cues when the COA was 0 ms and the onset of the first cue when the visual and the auditory cue were desynchronized.

Experimental design

The number of experimental conditions was defined by the combination of the four COAs (0, 150, 300, and 500 ms) and the two cues order (AV order, VA order). Given that at the shortest COA (0 ms) there could not be any difference in the cue order, only seven conditions were considered: The auditory and visual cue presented simultaneously (COA); the auditory cue was presented 150 before the visual cue (AV order, COA150); the auditory cue was presented 300 ms before the visual cue (AV order, COA300); the auditory cue was presented 500 ms before the visual cue (AV order, COA500); the visual cue was presented 150 ms before the auditory cue (VA order, COA150); the visual cue was presented 300 ms before the auditory cue (VA order, COA300); the visual cue was presented 500 ms before the auditory cue (VA order, COA500). Each participant completed a total number of 168 trials divided into three randomized experimental blocks of 56 trials each. Within each block participants performed two repetitions of all the seven possible experimental conditions for each of the four target locations. Participants were also required to perform one auditory and one visual baseline block of 24 trials each. In the auditory baseline block the auditory cue was presented in isolation. In the visual baseline block the visual cue was presented in isolation. The order of the baseline blocks was randomized with half of the participants performing the auditory baseline block before the visual baseline block and the other half performing the visual baseline block before the auditory baseline block. Further, half of the participants performed the baseline blocks before the three experimental blocks, whereas the other half performed the baseline blocks after the experimental blocks. As for Experiment 1 the dependent measures were RT and ST.

Data analysis

Two separate analyses of variance (ANOVA), one for the AV order and one for the VA order conditions with type of trial (auditory baseline, visual baseline, COA, COA150, COA300, and COA500) as within-subjects factor, were conducted on the RT and ST data set. Planned comparisons were carried out using *t*-tests. Bonferroni corrections were applied when necessary.

Results and discussion

Reaction time

AV order analysis. The main factor type of trial was significant, $F(5,45) = 22.324$, $p = .0001$. Reaction time was 502 ms for the auditory baseline, 449 ms for the visual baseline, 379 ms for the COA condition, 434 ms for the COA150 condition, 452 ms for the COA300 condition, and 451 ms for the COA500 condition (Figure 3a). Bonferroni adjusted *t*-tests revealed that

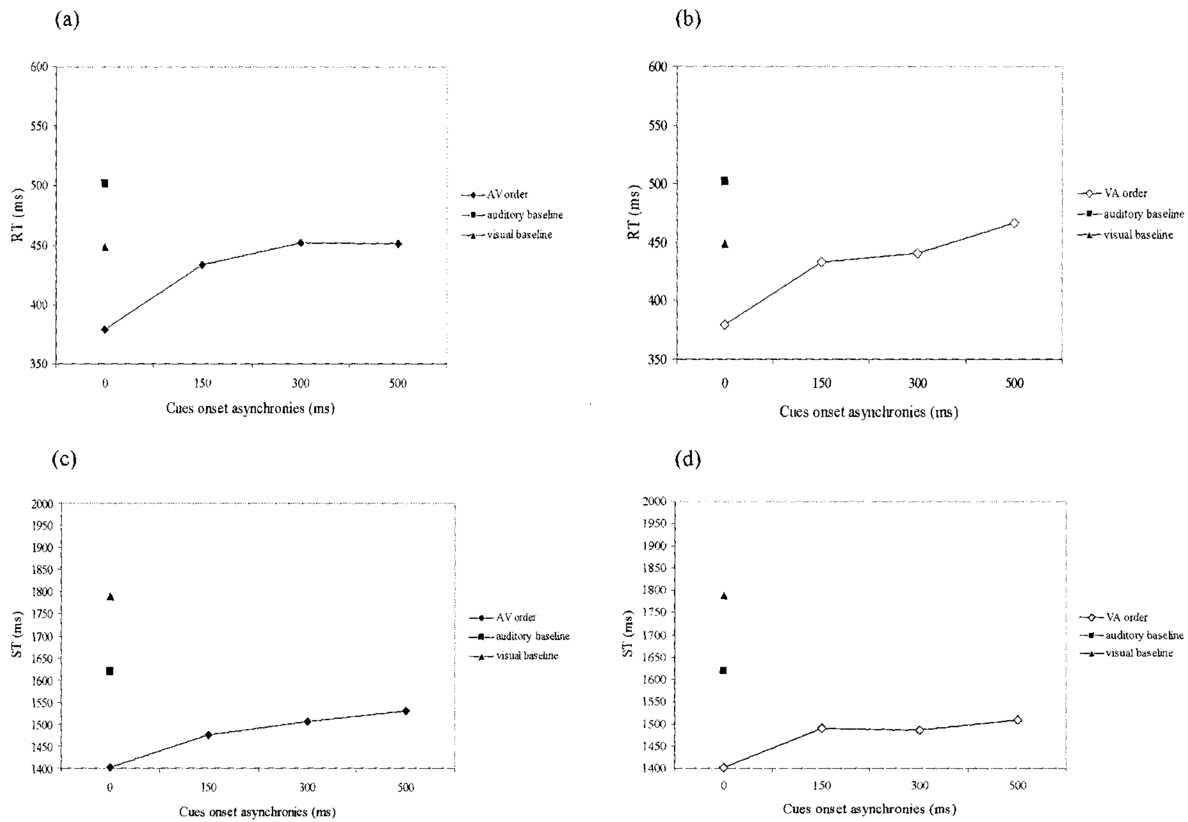


Figure 3. Mean reaction time (RT) and search time (ST) in milliseconds (ms) for Experiment 2. (a) and (b) Reaction time for the AV and VA order respectively relative to the cues onset asynchronies and baseline trials. (c) and (d) Search time for the AV and VA order respectively relative to the cues onset asynchronies and baseline trials.

reaction time for the auditory baseline was not significantly different from reaction time for the visual baseline (p = .07; Figure 3a). Reaction time for the COA condition was faster than reaction time for both the auditory (p = .001) and visual (p = .001) baselines (Figure 3a). Reaction time for the auditory baseline was slower than reaction time for the COA150 (p = .011), COA300 (p = .012), and COA500 (p = .021) conditions (Figure 3a). In contrast, reaction time for the visual baseline was not different from reaction time obtained in the COA150 (p = 1), COA300 (p = 1), and COA500 (p = 1) conditions (Figure 3a) Further, reaction time for the COA condition was faster than reaction time for the COA150 (p = .002), COA300 (p = .002), and COA500 (p = .005) conditions (Figure 3a). No differences were found when comparing reaction time obtained in the COA150 with reaction time obtained in the COA300 (p = .4) and in the COA 500 (p = .181) conditions (Figure 3a). Finally there was no difference between reaction time for the COA 300 and the COA 500 conditions (p = 1; Figure 3a).

VA order analysis. The main factor type of trial was significant, $F(5, 45)$ = 15.971, p = .0001. Reaction time was 502 ms for the auditory baseline, 449 ms for the visual baseline, 379 ms for the COA condition, 433 ms for the COA150 condition, 441 ms for the COA300 condition, and 467 ms for the COA500 condition (Figure 3b). Bonferroni adjusted t-tests revealed that reaction time for the auditory baseline was not significantly different from reaction time for the visual baseline (Figure 3b; p = .07). Reaction time for the COA condition was faster than reaction time for both the auditory (p = .001) and visual (p = .001) baselines (Figure 3b). Reaction time for the auditory baseline did not differ from reaction time for the COA150 (p = .103), COA300 (p = .334), and COA500 (p = 1) conditions (Figure 3b). Similarly, reaction time for the visual baseline did not differ from reaction time for the COA150 (p = 1), COA300 (p = 1), and COA500 (p = 1) conditions (Figure 3b). Reaction time for the COA condition was faster than reaction time for the COA150 (p = .002), COA300 (p = .001), and COA 500 (p = .0001) conditions. No differences were found when comparing reaction time obtained in the COA150 with reaction time obtained in both the COA300 (p = 1) and the COA 500 (p = .093) conditions (Figure 3b). Finally reaction time for the COA 300 condition was faster than reaction time for the COA 500 condition (p = .016; Figure 3b).

The RT results suggest that the simultaneous presentation of the auditory and visual information appear to be a necessary condition to elicit audiovisual integration.

First, presenting both cues simultaneously proved to generate a faster reaction time than both the auditory and visual baselines. Second, desynchronizing visual and auditory information by more than 150 ms precluded the facilitation-integration effect observed when both cues were simultaneously presented.

Search time

AV order analysis. The main factor type of trial was significant, $F(5, 45) = 7.222$, $p = .0001$. Search time was 1619 ms for the auditory baseline, 1790 ms for the visual baseline, 1420 ms for the COA condition, 1491 ms for the COA150 condition, 1487 ms for the COA300 condition, and 1510 ms for the COA500 condition. Bonferroni adjusted *t*-tests revealed that search time for the auditory baseline was faster than search time for the visual baseline, $p = .03$ (Figure 3c). Other *t*-tests did not reveal any significant difference between the conditions.

VA order analysis. The main factor type of trial was significant, $F(5, 45) = 7.850$, $p = .0001$. Search time was 1619 ms for the auditory baseline, 1790 ms for the visual baseline, 1420 ms for the COA condition, 1475 ms for the COA150 condition, 1507 ms for the COA300 condition, and 1531 ms for the COA500 condition. Bonferroni adjusted *t*-tests revealed that search time for the auditory baseline was faster than search time for the visual baseline, $p = .03$ (Figure 3c). Further, search time for the COA condition was faster than search time for both the COA300 ($p = .019$) and COA500 ($p = .006$) conditions (Figure 3c). Other *t*-tests did not reveal any significant difference between the conditions.

The ST results suggest that the facilitation found at the RT level when both cues were simultaneously presented is only evident at the very beginning of the search task. However, as shown in Figure 3c and 3d, a trend similar to the RT pattern can be noticed (Figure 3a and 3b). This may indicate that the integration of audiovisual information that remains available throughout the whole search might be sustained for an extended period of time. In other words, the audiovisual binding elicited at an early level continues to influence performance across a prolonged search period. A point worth mentioning is that, in contrast to the RT baseline results, the auditory baseline was faster than the visual baseline. This pattern of results is in line with findings for Experiment 1. We are inclined to explain such inversion of pattern in terms of task demands. Searching for a visual target on the basis of solely a visual cue may have brought to a higher level of visual load than when the search was aurally guided. In the latter case visual resources could be allocated at a greater extent to the visual task without the demand of coding the information carried by the visual cue.

EXPERIMENT 3

Results from Experiment 2 suggested that a slight temporal displacement in the presentation of the visual and auditory cues prevents the emergence of facilitation effects during the overt search for a target object. However, apart from temporal factors spatial factors also play a role as major determinants for facilitation (Stein, Hunneycutt, & Meredith, 1988). In Experiment 3 spatially congruent and incongruent conditions were administered. Participants were instructed to follow one specific modality (primary modality) in the presence of information from another modality (secondary modality). For the congruent condition, the information given through the primary and the secondary modality was always congruent with respect to the location of the target object. For the incongruent condition participants were instructed to follow one specific (primary) modality, but the secondary modality always provided incongruent information with respect to the location of the target object. We predict that if the target object is strongly cued on a particular side by both modalities (audition and vision), then facilitation effects should be revealed with respect to when spatially incongruent information is provided. In contrast, if the target object is cued on a particular side by just one modality (e.g., vision) whereas the secondary modality provides incongruent spatial information regarding the location of the target object, then this conflict should result in a decrease or disappearance of the facilitation effects.

Method

Participants

All eight participants were volunteers and naïve as to the purpose of the experiment. Their age ranged from 25 to 44 years with a mean of 31. All reported normal or corrected-to-normal vision and normal auditory functioning.

Apparatus and materials

Apparatus and materials were the same as for Experiment 1.

Procedure

Procedure was the same as for Experiment 1 except for the following changes:

Type of trial. Two different types of trial were administered to the participants: (1) Congruent trials in which the auditory and the visual cue were simultaneously presented and both cued the target azimuth; and (2) incongruent trials, in which the auditory and the visual cue were presented simultaneously but one cue indicated the target azimuth while the other indicated one of the two possible locations contralateral to the target.

Type of instruction. Participants completed four blocks of trials under different instructions. In the auditory blocks of trials, participants were instructed to follow the auditory cue ignoring the visual cue when present. In the visual blocks of trials, participants were instructed to follow the visual cue ignoring the auditory cue.

Regardless of the type of trial and the type of instruction, once the target was found participants had to overlap the sight over the target and maintain that position for 50 ms. In addition to RT and ST there was a third dependent measure, the number of times that participants started to search for the target moving their head contralaterally to the target position. We defined these as errors. This dependent measure has been introduced to test whether irrelevant incongruent information relative to the type of instruction could be easily ignored by participants.

Experimental design

There were four experimental conditions derived from the combination of the two types of trial (congruent, incongruent) and the two types of block (auditory, visual): (1) Auditory congruent, where within an auditory block, the auditory and the visual cue originated from the target location; (2) auditory incongruent, where, within an auditory block, the auditory cue originated from the target location, whereas the visual cue originated from one of the two locations contralateral to the target; (3) visual congruent, where, within a visual block, the auditory and the visual cue originated from the target location; and (4) visual incongruent, where, within a visual block, the visual cue originated from the target location whereas the auditory cue originated from the one of the two locations contralateral to the target. Each participant completed a total of 144 trials equally divided into four experimental blocks, two auditory and two visual. The order of blocks was counterbalanced across participants.

Data analysis

Three separate analyses of variance (ANOVA) with type of trial (congruent, incongruent) and type of block (auditory, visual) as within-subject factors were conducted for the RT, ST, and the error data sets respectively. In all analyses planned comparisons were performed using t-tests. Bonferroni adjustments were carried out when necessary.

Results and discussion

Reaction time

For the reaction time analysis the main factor type of trial was significant, $F(1,7) = 8.601$, $p = .022$. Regardless of the type of instruction, reaction time for the congruent conditions was faster than reaction time for the incongruent conditions (469 vs. 504 ms). The main factor type of block was also significant, $F(1,7) = 31.107$, $p = .001$. Reaction time was 543 ms for the auditory block conditions and 431 ms for the visual block conditions. The interaction between type of trial and type of block was significant, $F(1,7) = 38.189$, $p = .0001$. Planned comparisons revealed that reaction time for the auditory congruent condition was faster than for the auditory incongruent condition, $p = .004$ (Figure 4a). No difference was found between the visual congruent and visual incongruent conditions, $p = .392$. Reaction time for the auditory congruent condition was slower than for the visual congruent condition, $p = .002$. Finally, reaction time for the auditory incongruent condition was also slower than for the visual incongruent condition (Figure 4 a; $p = .001$), suggesting that subjects were more reactive when performing a visual block than when performing an auditory block.

Search time

For the search time analysis the main factor type of trial was significant, $F(1, 7) = 9.699$, $p = .017$. Search time was 1529 ms for the congruent condition and 1625 ms for the incongruent conditions. The main factor type of block was also significant, $F(1, 7) = 32.143$, $p = .001$. Search time was 1648 ms for the auditory block and 1506 ms for the visual block conditions (Figure 4b). The interaction between type of trial and type of block was not significant, $F(1, 7) = 2.179$, $p = .183$.

Errors

Errors represented 5.2% of the total number of trials and were therefore analysed. For the error analysis the main factor type of trial was significant, $F(1, 7) = 18.617$, $p = .004$. Error rate was 1.3% for the congruent condition and 14.3% for the incongruent conditions. The main factor type of block was also significant, $F(1, 7) = 11.225$, $p = .012$. Error rate was 1.6% for the visual block and 8.8% for the auditory block conditions. The interaction between type of trial and type of block was significant, $F(1, 7) = 6.443$, $p = .039$ (Figure 4c). Planned comparisons revealed that error rate for the auditory congruent condition was lower than for the auditory incongruent condition, $p = .011$ (Figure 4c). Error rates for the visual congruent condition were lower than for the visual incongruent conditions, $p = .04$ (Figure 4c). Comparisons were also made between the auditory congruent and the visual congruent conditions and between the auditory incongruent and the visual incongruent conditions. In particular error rate for the auditory congruent condition was not significantly different from error rate obtained for the visual congruent condition, $p = .197$ (Figure 4c). However, error rates for the auditory incongruent condition were significantly higher than for the visual incongruent condition ($p = .021$; Figure 4c).

Results from both reaction time and search time confirm that when both the visual and the auditory cue signal the position in which the target is located, a facilitation effect occurs. In addition, they reveal that during the incongruent conditions it is more challenging to inhibit irrelevant visual information than irrelevant auditory information. The potency of the visual information is also reflected by the result that reaction time was not affected by the presence of

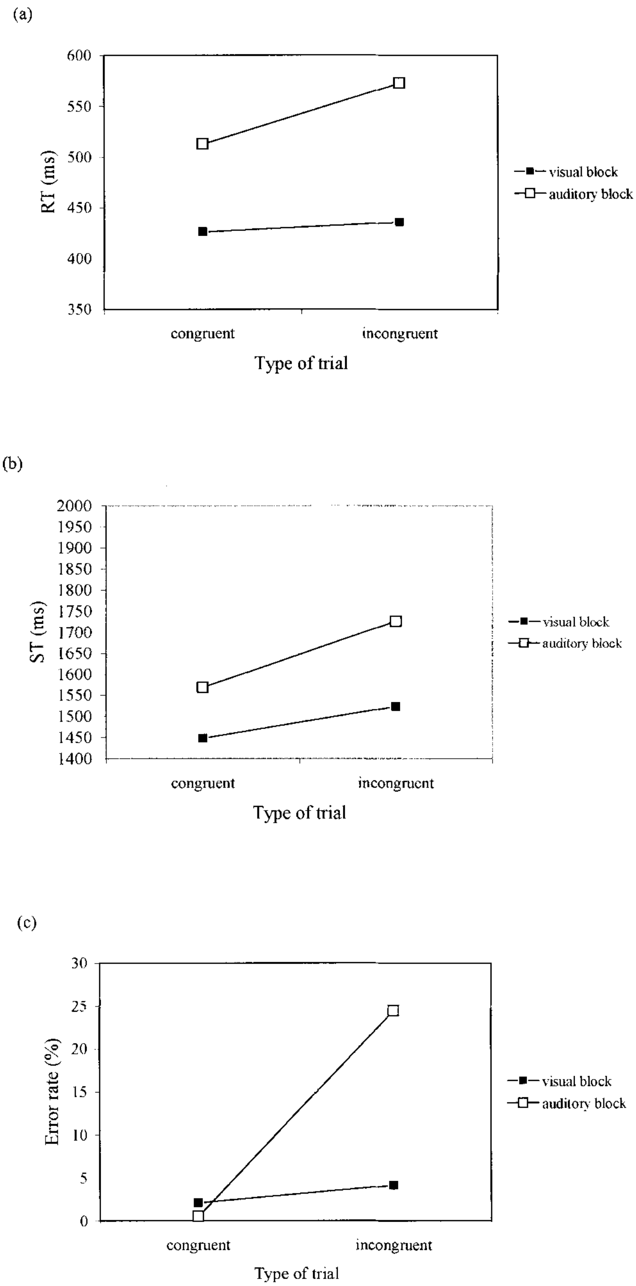


Figure 4. (a) Mean reaction time (RT) in milliseconds (ms) for Experiment 3 in the congruent and incongruent trials for both the auditory and the visual blocks. (b) Mean search time (ST) in milliseconds (ms) for Experiment 3 in the congruent and incongruent trials for both the auditory and the visual blocks. (c) Mean error rate (%) for Experiment 3 in the congruent and incongruent trials for both the auditory and the visual blocks.

irrelevant auditory information when the visual cue was congruent. This latter result is also confirmed by the error analyses. When participants were required to ignore an incongruent visual cue they made more errors than when they had to ignore an incongruent auditory cue.

Although participants were given clear instructions on which type of cue they had to attend, we could not exclude that they followed a different strategy to perform the task. For example, it may be argued that on incongruent trials the irrelevant and invalid information served as a 100% valid countercue to orient participants towards target location. However, given that we tested four possible different locations the invalid cue could be used as countercue only at the level of azimuth but not elevation. Further and more important, the error data suggest that within the auditory blocks a spatially incongruent visual cue elicited a higher percentage of initial head rotation towards locations contralateral to the target. If we hypothesize that our participants were using the incongruent cue as a 100% valid countercue, then it may be suggested that they were not able to operationalize this strategy, at least in the incongruent trials of the auditory blocks.

GENERAL DISCUSSION

We set out to investigate multimodal integration during an overt search for a nonimmediately visible object. We investigated this issue by creating a virtual reality room that participants could explore by means of a head-mounted display. Participants searched for an object attached to the virtual room's walls following visual and auditory cues originating from the spatial region containing the object. When the object was found participants engaged it by aligning it with a sight attached to the observer's head. Visual and auditory information relating to the object's location were provided in different combinations. In Experiment 1 they were presented either alone or simultaneously. In Experiment 2 they were presented either simultaneously or separated by various time intervals. In Experiment 3 they originated from either the same or different spatial locations.

The present results reveal that temporal and spatial coincidence is an important determinant of facilitation effects. When the administration of auditory and visual cues was delayed in time or presented at different locations no facilitation took place. Our results also go some way towards answering questions about how multisensory integration might operate in forming appropriate motor behaviours during an overt search task that resembles a search task performed in real-life situations and how they might best be measured. The results of Experiment 1 showed that reaction time and search time were faster when spatially and temporally coincident combinations of visual and auditory cues were presented, compared to when a visual or an auditory cue was presented alone. These results parallel and extend those obtained in previous work using overt responses such as eye movements in which saccade or gaze latency is faster for spatially and temporally coincident combinations of visual and auditory stimuli compared to saccades to each single modality alone (Frens et al., 1995; Goldring, Dorris, Corneil, Ballantyne, & Munoz, 1996; Hughes et al., 1994). However, in those studies stimuli were always presented in front of the subjects in visually available positions. Here, although the updating auditory and visual cues were immediately perceivable, the target object was not immediately visible. This allows for new insights about how the multisensory representation can be maintained and refreshed along time and space dimensions.

In contrast to the amount of published work on the mechanisms involved in the multisensory integration of spatial information, comparatively little is known about the nature of the integration of sensory inputs for the purpose of searching for a nonimmediately visible object. In such a task at least two different phases can be delineated. First, a phase in which the information emitted by the object has to be interpreted in the spatial register. This phase would determine the initial orientation of the body towards the object. Second, a search/identification phase in which the object is searched, recognized, and engaged. The interesting issue is to understand the underlying rules for multisensory integration at these two different phases. As revealed by the reaction time results, which reflect the time to start orienting gaze and body towards an object's location, it is clear that the benefits of being initially guided by both auditory and visual exogenous cues are evident. This is in line with the notion that extensive multisensory integration has been reported for overt and covert shifts of exogenous spatial attention (Sokolov, 1963; Spence & Driver, 1997). Of interest, however, is that the present data provide some indication that the initial advantage is maintained throughout the search process as reflected by the trend in the search time results. With a certain degree of caution, this may suggest that the integration of sensory cues can be sustained for quite a long period of time given that the object is not visually available at the start of the search action. We suspect that this integration occurs because of the updating nature of the cues. Recall that the information provided by the cues is constantly reiterated during the search process. This constant updating may therefore allow for a continuous refreshing of a multisensory binding created at the start of the search action.

In Experiment 1 the long lasting effect (e.g., search time) of the multisensory integration could be attributed to the updating nature of the utilized cues that allowed the formation of a continuously refreshed and maintained multisensory representation. An interesting question is whether this representation can still be created and maintained even when the visual and the auditory information originate from the object at different times. A possible answer can be found in Experiment 2, in which using the same procedures as in Experiment 1 we varied the delay between the presentation of the auditory and the visual information cues. Benefits related to the presentation of the two cues were tied to the condition in which the two cues were presented simultaneously. When a temporal gap between the presentation of the visual and the auditory cues at the start of the search action was introduced, facilitation effects diminished dramatically. These results demonstrate that the integration of the auditory and visual information at the beginning of the search action is fundamental to bringing the formation of a multisensory representation and to reveal facilitation. That is, if the two cues are asynchronous at the beginning of the search, they don't integrate, despite the very fact that they remain available during the entire search action. This result is further supported by the advantage of the combined cue conditions with respect to the baseline conditions at RT level.

Whereas these results are in line with most studies that have found that simultaneous presentation of bimodal stimuli will produce a reduction in saccadic and gaze latency, they also fit with the notion of a brief temporal window for neural integration (Wallace, Wilkinson, & Stein, 1996). Taken together, findings from Experiments 1 and 2 suggest that the simultaneous presentation of multimodal information is a crucial factor for the formation of a multisensory representation that can be refreshed and maintained to optimally guide the search for the object. Displacing the multimodal information in time not only revealed a lack of integration at the reaction time level as previously demonstrated by Stein and Meredith (1993), but also at the search time level. This may imply that a multisensory representation can only be established at the very beginning of the search because even when information from another modality becomes available it does not seem to be integrated.

So far we have assumed that the beneficial effects found in Experiments 1 and 2 are due to the integration of the auditory and visual cues. However, a possibility is that an independent cue model might also account for the benefit in those studies. In this respect it is important to mention the distinction between "sensory combination" and "sensory integration" proposed by Ernst and Bulthoff (2004). They suggest that sensory combination of cues describes interactions between sensory signals that are not redundant, whereas sensory integration describes interactions between redundant signals. In these terms our results speak clearly in favour of an integration process given the high level of redundancy of the adopted cues. Furthermore, our paradigm meets the necessary conditions for the two cues to be integrated (Ernst & Bulthoff, 2004). First, the cues derived exactly from the same objects. Second, integration occurred only when the two cues were presented with no spatial or temporal discrepancy.

Experiment 3, in which the spatial congruency of visual and auditory information was manipulated, revealed asymmetries in the ability of dissociating simultaneously presented audiovisual information. Searching for an object on the basis of auditory information is significantly affected by irrelevant and spatially incongruent visual information. In contrast, searching for an object on the basis of visual information does not seem to be influenced by irrelevant and spatially incongruent auditory information. All in all, results from this experiment suggest that irrelevant visual information has more power to interfere than irrelevant auditory information. This finding is also supported by the error results in which participants tended to be tuned to the visual modality even though it is irrelevant or distracting. These results confirm the notion that in a localization task visual inputs tend to dominate other modalities not only when perceptual speeded responses are required under covert situations (e.g., Colavita, 1982; Colavita & Weisberg, 1979; Posner, Nissen, & Klein, 1976; for a review see also Shimojo & Shams, 2001), but also for an overt search task such as that utilized here. Furthermore, these results are consistent with those of Experiment 1 and are in line with studies of bimodal stimulus presentation in which saccades to spatially aligned auditory and visual stimuli have a reduced response latency (Zahn, Abel, & Dell'Oso, 1978, 1979). However, in these studies participants were not instructed to respond to a specific modality in bimodal conditions. Therefore it is not clear whether the facilitation of response was a product of a saccade being generated by the fastest modality or, alternately, was due to the convergence or integration of the two stimuli in the nervous system (Raab, 1962). Here we instructed participants to follow one specific modality in the presence of an irrelevant cue from another modality. Facilitation effects with bimodal congruent rather than incongruent cues were evident when participants were instructed to follow only one modality at any one time.

In sum, we have found several factors responsible for mediating the responses to the type of bimodal stimulation presented here. The characterization of these factors allows for the determination of new conditions through which multisensory processing occurs in an ecologically valid experimental condition. Future studies with more complex virtual environments may offer a way forward for future research, for example when participants are trained to search for objects in the presence of distracting information. Finally, future models of these sensorimotor transformations should gain from the information provided here between the initial topographical encoding of visual and auditory information, and the complex transformations necessary to retain close alignment between the spatial and the temporal register (these two dynamic sensory systems) during movement.

REFERENCES

Bolia, R. S., D'Angelo, W. R., & McKinley, R. L. (1999). Aurally aided visual search in three-dimensional space. *Human Factors*, *41*, 664–669.

Bronkhorst, A. W., Veltman, J. A., & van Breda, L. (1996). Application of a three- dimensional auditory display in a flight task. *Human Factors*, *38*, 23–33.

Bushara, K. O., Hanakawa, T., Immisch, I., Toma, K., Kansaku, K., & Hallett, M. (2003). Neural correlates of cross-modal binding. *Nature Neuroscience*, *6*, 190–195.

Colavita, F. B (1982). Visual dominance and attention in space. *Bulletin of the Psychonomic Society*, *19*, 261–262.

Colavita, F. B., & Weisberg, D. (1979). A further investigation of visual dominance. *Perception and Psychophysics*, *25*(4), 345–347.

Corneil, B. D., & Munoz, D. P. (1996). The influence of auditory and visual distractors on human orienting gaze shifts. *Journal of Neuroscience*, *16*(24), 8193–8207.

Driver, J. (1996). Enhancement of selective listening by illusory mislocation of speech sounds due to lip-reading. *Nature*, *381*, 66–68.

Driver, J., & Spence, C. (1998). Crossmodal attention. *Current Opinion in Neurobiology*, *8*, 245–253.

Ernst, M. O., & Bulthoff, H. H. (2004). Merging the senses into a robust percept. *Trends in Cognitive Sciences*, *8*(4), 162–169.

Flanagan, P., McAnally, K. I., Martin, R. L., Meehan, J. W., & Oldfield, S. R. (1998). Aurally and visually guided visual search in a virtual environment. *Human Factors*, *40*, 461–468.

Frens, M. A., van Opstal, A. J., & van der Willigen, R. F. (1995). Spatial and temporal factors determine auditory-visual interactions in human saccadic eye movements. *Perception*, *57*(6), 802–816.

Goldring, J. E., Dorris, M. C., Corneil, B. D., Ballantyne, P. A., & Munoz, D. P. (1996). Combined eye-head gaze shifts to visual and auditory targets in humans. *Experimental Brain Research*, *111*, 68–78.

Hughes, H. C., Reuter-Lorenz, P. A., Nozawa, G., & Fendrich, R. (1994). Visual-auditory interactions in sensorimotor processing: Saccades versus manual responses. *Journal of Experimental Psychology: Human Perception and Performance*, *20*(1), 131–153.

Lee, C., Chung, S., Kim, J., & Park, J. (1991). Auditory facilitation of visually guided saccades. *Society for Neuroscience Abstracts*, *17*, 862.

Nelson, W. T., Hettinger L. J., Cunningham, J. A., Brickman, B. J., Haas, M. W., & McKinley, R. L. (1998). Effects of localized auditory information on visual target detection performance using a helmet-mounted display. *Human Factors*, *40*, 452–460.

Perrott, D. R., Cisneros, J., McKinley, R. L., & D'Angelo, W. R. (1996). Aurally aided visual search under virtual and free-field listening conditions. *Human Factors*, *38*, 702–715.

Perrott, D. R., Saberi, K., Brown, K., & Strybel, T. Z. (1990). Auditory psychomotor coordination and visual search performance. *Perception and Psychophysics*, *48*, 214–226.

Posner, M. I., Nissen, M. J., & Klein, R. M. (1976). Visual dominance: An information processing account of its origins and significance. *Psychological Review*, *83*, 157–171.

Raab, D. H. (1962). Statistical facilitation of simple reaction times. *Transactions of the New York Academy of Sciences*, *24*, 574–590.

Shimojo, S., & Shams, L. (2001). Sensory modalities are not separate modalities: Plasticity and interactions. *Current Opinion in Neurobiology*, *11*, 505–509.

Sokolov, E. N. (1963). *Perception and the conditioned reflex*. New York: Macmillan.

Spence, C., & Driver, J. (1997). Audiovisual links in exogenous covert spatial attention. *Perception and Psychophysics*, *59*(1), 1–22.

Stein, B. E., Hunneycutt, W. S., & Meredith, M. A. (1988). Neurons and behaviour: The same rules of multisensory integration apply. *Brain Research*, *448*, 335–358.

Stein, B. E., & Meredith, M. A. (1993). *The merging of the senses*. London: MIT Press.

Stekelenburg, J. J., Vroomen, J., & de Gelder, B. (2004). Illusory sound shifts induced by the ventriloquist illusion evoke the mismatch negativity. *Neuroscience Letters*, *357*, 163–166.

Wallace, M. T., Wilkinson, L. K., & Stein, B. E. (1996). Representation and integration of multiple sensory inputs in primate superior colliculus. *Journal of Neurophysiology*, *76*(2), 1246–1266.

Zahn, J. R., Abel, L. A., & Dell'Oso, L. F. (1978). The audio-ocular response characteristics. *Sensory Processes*, *2*, 32–37.

Zahn, J. R., Abel, L. A., & Dell'Oso, L. F. (1979). The audio-ocular response: Intersensory delay. *Sensory Processes*, *3*, 60–65.

Manuscript received April 2004
Manuscript accepted March 2005