Cognitive, Affective, & Behavioral Neuroscience
2004, 4 (2), 218-229

When does visual perceptual grouping affect multisensory integration?

DANIEL SANABRIA
University of Oxford, Oxford, England

SALVADOR SOTO-FARACO
Universitat de Barcelona, Barcelona, Spain
and University of Oxford, Oxford, England

and

JASON S. CHAN and CHARLES SPENCE
University of Oxford, Oxford, England

Several studies have shown that the direction in which a visual apparent motion stream moves can influence the perceived direction of an auditory apparent motion stream (an effect known as *cross-modal dynamic capture*). However, little is known about the role that intramodal perceptual grouping processes play in the multisensory integration of motion information. The present study was designed to investigate the time course of any modulation of the cross-modal dynamic capture effect by the nature of the perceptual grouping taking place within vision. Participants were required to judge the direction of an auditory apparent motion stream while trying to ignore visual apparent motion streams presented in a variety of different configurations. Our results demonstrate that the cross-modal dynamic capture effect was influenced more by visual perceptual grouping when the conditions for intramodal perceptual grouping were set up *prior* to the presentation of the audiovisual apparent motion stimuli. However, no such modulation occurred when the visual perceptual grouping manipulation was established at the same time as or after the presentation of the audiovisual stimuli. These results highlight the importance of the unimodal perceptual organization of sensory information to the manifestation of multisensory integration.

During the last two decades, many researchers have reported evidence demonstrating the multisensory integration of auditory and visual information presented in approximate spatial and temporal register (e.g., Calvert, Spence, & Stein, 2004; Driver & Spence, 2000; and Stein & Meredith, 1993, for reviews). Traditionally, the majority of research on multisensory integration has focused on the study of spatially *static* events (e.g., such as in the ventriloquism illusion; Bertelson & de Gelder, 2004; Howard & Templeton, 1966), seemingly neglecting the fact that our sensory systems have evolved to perceive stimuli that are frequently moving with respect to us (e.g., Soto-Faraco, Kingstone, & Spence, 2003).

More recently, a number of researchers have started to investigate the multisensory integration of visual and auditory information regarding stimulus *motion* (see Soto-Faraco & Kingstone, 2004; Soto-Faraco et al., 2003, for recent reviews). For instance, in a series of experiments reported by Soto-Faraco and colleagues (e.g., Soto-Faraco, Lyons, Gazzaniga, Spence, & Kingstone, 2002; Soto-Faraco, Spence, & Kingstone, 2004), the presentation of a visual apparent motion stream consisting of two sequentially presented light flashes was shown to influence the perception of an auditory apparent motion stream consisting of two sequentially presented tones when both apparent motion streams were presented in synchrony. Over a number of studies, participants were shown to be significantly less accurate in judging the direction of auditory motion on *conflicting* trials (i.e., ones in which the auditory and the visual stimuli moved in opposite directions) than on *congruent* trials (i.e., ones in which the auditory and the visual stimuli moved in the same direction). The existence of this *cross-modal dynamic capture* effect is consistent with a number of previous findings reported in the experimental literature (e.g., Mateeff, Hohnsbein, & Noack, 1985).

It is worth noting, however, that the paucity of research in this area over the years may in part be attributable to the contradictory findings that emerged from previous research (see Soto-Faraco & Kingstone, 2004; and Soto-Faraco et al., 2003, for reviews and discussions of this point). Whereas some researchers have argued that visual dynamic stimuli have a strong influence on the perception of auditory motion (e.g., Mateeff et al., 1985; Zapparoli & Reatto, 1969), others have come to the opposite conclusion, arguing instead that visual dynamic

This study was supported by a Network Grant from the McDonnell-Pew Centre for Cognitive Neuroscience in Oxford to S.S.-F. and C.S. Correspondence regarding this article should be directed to D. Sanabria, Department of Experimental Psychology, South Parks Road, Oxford OX1 3UD, England (e-mail: daniel.sanabria@psy.ox.ac.uk).

stimuli have little or no effect on the perception of auditory motion (e.g., Allen & Kolers, 1981; Ehrenstein & Reinhardt-Rutland, 1996; Wuerger, Hofbauer, & Meyer, 2003). However, as has been suggested by Soto-Faraco et al. (2003), it might be possible that the null results reported in certain of the early studies on cross-modal motion perception (e.g., those regarding the effects of directional congruency; Allen & Kollers, 1981; Staal & Donderi, 1983) do not reflect a genuine lack of any multisensory interaction but, instead, reflect a failure to integrate sensory information presented from different spatial locations.

Moreover, relevant to the present study are the differences in the conditions for perceptual grouping found across the different investigations, since they might also underlie some of the differences in the cross-modal effects that have been found. Both the relative number of visual and auditory stimuli used in the presentations and their relative timing and location could have led to different conditions for the manifestation of perceptual grouping and, therefore, as we argue in the present article, to differences in any cross-modal effects observed. For instance, whereas Mateeff et al. (1985, Experiment 1) presented a moving visual stimulus that started long before the presentation of the target auditory stimulus and continued long after the sound had ceased, Soto-Faraco et al. (2002) presented two visual and two auditory stimuli in synchrony and from the same spatial locations (i.e., one might argue that the conditions for grouping were much better in Soto-Faraco et al.'s, 2002, study than in Mateeff et al.'s, 1985, study). These differences in procedure could have led to the differences in cross-modal effects observed. Relevant to the present study is an investigation reported by Vroomen and de Gelder (2000). They showed that the perception of a visual target embedded in a stream of visual distractors could be enhanced by the presentation of an auditory stimulus in synchrony with the target. However, this cross-modal facilitation effect was dramatically reduced when the critical auditory stimulus was embedded within a sequence of auditory stimuli to create a melody. These results provide a dramatic example of the effect of unimodal perceptual grouping on multisensory integration.

Despite the recent growth of interest in the topic of the multisensory integration of motion information, we still know relatively little about the various factors that modulate such phenomena. For instance, intramodal perceptual grouping has been shown to have a major impact on our perceptual experience (e.g., Koffka, 1935; Kubovy & Van Valkenburg, 2001; Palmer, 2002; Wertheimer, 1950), but very little evidence has been published on the question of how such intramodal grouping principles may interact with the processes of multisensory integration, such as the ones at work in the cross-modal dynamic capture task (and which can presumably be considered as a form of *cross-modal* perceptual grouping; e.g., Sanabria, Soto-Faraco, Chan, & Spence, 2003).

We believe that the influence of intramodal perceptual grouping on the multisensory integration of motion information may be an important factor that, as was noted above, could help to explain some of the conflicting results reported in previous research (see Soto-Faraco et al., 2003, on this topic). At the same time, it is also important to gain a better understanding of the conditions under which the perceptual grouping taking place within one sensory modality can affect the nature of the grouping taking place between different sensory modalities.

Sanabria et al. (2003) recently provided some of the first empirical evidence to suggest that the perceptual grouping of visual stimuli can modulate the magnitude of the multisensory integration of auditory and visual apparent motion streams. The participants in their study were required to judge the direction of an auditory apparent motion stream (elicited by the sequential presentation of two sounds from different spatial locations, one from either side of fixation) while simultaneously trying to ignore a visual apparent motion stream (elicited by the sequential presentation of a series of visual stimuli). As was reported in Soto-Faraco et al.'s (2002; Soto-Faraco et al., 2004) previous studies, the participants' judgments of the direction of motion of the auditory stimuli were less accurate on conflicting trials (in which the auditory and the visual stimuli moved in opposite directions) than on congruent trials (in which the stimuli in the two modalities moved in the same direction). Crucially, Sanabria et al.'s results also showed that the magnitude of the cross-modal dynamic capture effect was reduced when the two critical visual stimuli were embedded within an extended visual stream (consisting of four additional visual stimuli—i.e., six sequentially presented visual stimuli in total). Somewhat paradoxically, an *increase* in the strength of visual apparent motion was shown to *reduce* the magnitude of the cross-modal dynamic capture effect. Sanabria et al.'s results therefore suggest that the perceptual grouping processes taking place within vision can modulate the effect of multisensory integration of motion information taking place between different sensory modalities.

According to Sanabria et al. (2003), their results can be explained according to the laws of perceptual grouping. If we consider cross-modal grouping as an example of perceptual grouping, the same rules that apply to unimodal perceptual grouping might also govern cross-modal perceptual grouping. Therefore, when the visual and the auditory stimuli are presented at the same time and at approximately the same spatial location and are of the same number, they are more likely to be grouped than when they are different in number, as in Sanabria et al.'s spatiotemporally extended visual stream condition (six visual distractors).

In the present study, we manipulated the nature of the visual apparent motion streams in order to analyze more closely the conditions under which visual perceptual grouping modulates multisensory integration in the cross-modal dynamic capture task. One of the crucial questions concerns the role of perceptual grouping processes that occur after the critical stimuli (*late* perceptual grouping), as opposed to the role of perceptual grouping

processes that have already taken place at the time the critical cross-modal stimuli are presented.

Although the effect of *late* perceptual grouping has been widely explored for the case of unimodal sensory perception, the role played by *early* versus *late* intramodal perceptual grouping on multisensory integration remains uncharted. For instance, within the visual, auditory, and tactile modalities, it has been shown that the perceived location of a stimulus (or group of stimuli) can be shifted toward the location of a subsequent stimulus (or group of stimuli) delivered shortly afterward, from a different spatial location (as in the well-known saltation illusion; Geldard, 1976; Kilgard & Merzenich, 1995; Lockhead, Johnson, & Gold, 1980; Phillips & Hall, 2001; Shore, Hall, & Klein, 1998). Such results illustrate that the unimodal perceptual grouping of stimuli presented later in time can influence the perception of stimuli that were presented earlier in time. Of particular relevance here is the fact that although people have often argued that visual perceptual grouping takes place only at an early stage of information processing, Palmer (2002) has recently proposed that it may occur at a much later stage than had been previously thought. This then raises the possibility that any *late* perceptual grouping taking place within vision might also be capable of influencing the cross-modal grouping of earlier presented audiovisual stimuli.

In the present study, we explored the role played by the visual perceptual grouping that occurs either *prior* to or *after* the audiovisual stimuli used to elicit the cross-modal dynamic capture effect. We expected to find a modulation of the cross-modal dynamic capture effect when the conditions for intramodal perceptual grouping of the unimodal visual stimuli were available before the onset of the multisensory stimuli. The novel question addressed here was whether any visual perceptual grouping that occurred *after* the presentation of the audiovisual stimuli could also influence the cross-modal grouping of auditory and visual stimuli.

EXPERIMENT 1

The aim of Experiment 1 was to explore the role played by any perceptual grouping that might occur before or after the presentation of the critical audiovisual stimuli. The target auditory apparent motion stream consisted of two brief white noise bursts presented successively to the left and to the right of a central fixation point. The irrelevant visual stimuli were presented in various different configurations. In the two-light condition, only two lights were presented in synchrony with the sounds, moving in either the same or the opposite direction. In the four-light condition, two additional peripheral lights were presented, forming a four-light sequence together with the two central lights (that always coincided temporally with the target sounds). The multimodal event could occur either at the end of the sequence (Experiment 1A) or at the beginning (Experiment 1B). If visual perceptual grouping modulates the multisensory integration of motion information only when these grouping processes precede the multimodal event, a significant reduction in the magnitude of the cross-modal dynamic capture effect should

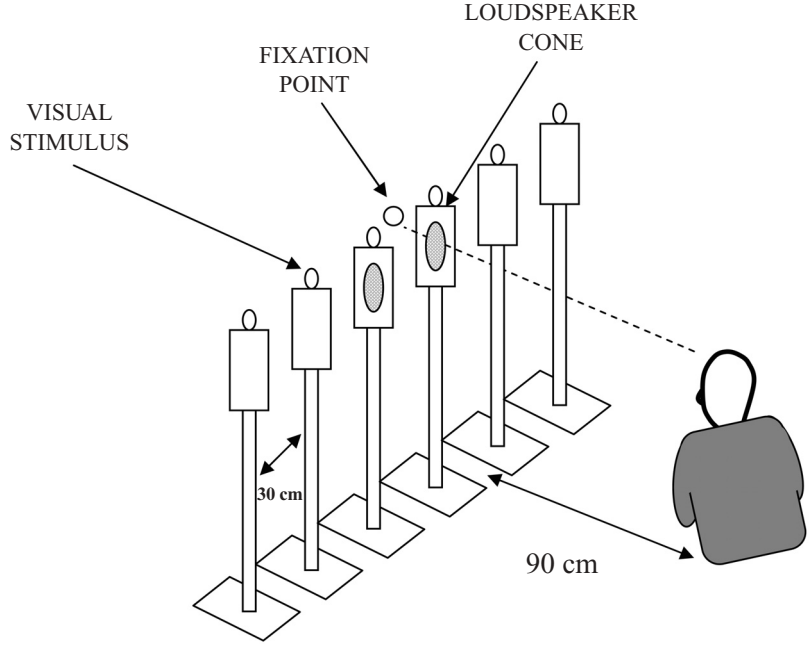

Figure 1. Schematic illustration of the setup used in the experiments reported in the present study.

be found in the four-light condition of Experiment 1A (in which conditions for perceptual grouping were established prior to the multimodal event), but not in the four-light condition of Experiment 1B (in which perceptual grouping, if any, could occur only after the presentation of the multimodal event). That is, only when the peripheral lights were presented *prior* to the central lights and the unimodal perceptual grouping started *before* the presentation of the multisensory stimuli would intramodal visual grouping modulate cross-modal grouping.

Method

Participants. Twenty-eight participants (age range, 21–30 years; mean, 23 years) from the University of Oxford took part in Experiment 1. Half were tested in Experiment 1A and the remainder in Experiment 1B. All of the participants reported normal hearing and normal or corrected-to-normal vision and received a £5 gift voucher in exchange for their participation.

Apparatus and Materials. Six orange light-emitting diodes (LEDs) were positioned in a row (30-cm separation center-to-center) at eye level in front of the seated participant. The participants sat in a comfortable chair, in complete darkness, at a distance of 90 cm from the center of the row of LEDs. A loudspeaker cone was placed just below each of the two central LEDs, and both were used to present the auditory stimuli (see Figure 1). A red LED, placed between the two central loudspeaker cones and illuminated throughout the experiment, was used as a central fixation point. The participants held the response keypad in their laps. The loudspeaker cones and LEDs were all controlled via the computer parallel port, using custom software.

The auditory apparent motion stimuli consisted of two 100-msec white noise bursts [40 dB(A) as measured from the participant's head position], one burst presented from each of the two central loudspeaker cones, and separated by an interstimulus interval (ISI) of 50 msec. The visual apparent motion stimuli consisted of a sequence of 100-msec light flashes presented from the row of LEDs, each separated by an ISI of 50 msec. In the *two-light condition*, the visual apparent motion stream consisted of the sequential activation of the two central LEDs, in time with the two sounds (as in the majority of previous studies of the cross-modal dynamic capture effect; e.g., Soto-Faraco et al., 2002; Soto-Faraco et al., 2004). In the *four-light condition*, the two central lights either followed (Experiment 1A) or preceded (Experiment 1B) the two peripheral light flashes, giving rise to a more spatiotemporally extended visual apparent motion stream consisting of the sequential illumination of four LEDs. Note that in all of the conditions, the onset of the two central lights coincided temporally with the onset of the two target sounds (see Figures 2A and 3A for a schematic illustration of the stimulus setups used in Experiments 1A and 1B).

Procedure. On each trial, the auditory and the visual apparent motion streams moved independently either to the left or to the right. The number of visual stimuli in the visual apparent motion stream (two vs. four lights) and the directional congruency between the auditory and the visual motion stimuli (conflicting vs. congruent) were combined systematically, and their presentation was randomized on a trial-by-trial basis. The participants were instructed to respond to the direction of the auditory apparent motion stream by making a keypress response with their left or right thumbs (to indicate leftward vs. rightward auditory motion, respectively). The participants were also instructed to ignore the irrelevant visual stimuli as much as possible and to try and respond to the target auditory motion as accurately as possible, regardless of the latency of their responses. The participants were also informed about the independence of the directions of the auditory and the visual apparent motion streams prior to the start of the experiment. There was a 2,000-msec interval after the participant's response was recorded before the start of the next trial. The participants completed 12 practice trials in which the sounds were presented in the absence of any visual distractors, to familiarize them with the task at hand. The experimenter made sure that the participants had understood the task, and the practice trials were repeated if the participants made more than one error. After the completion of all the practice trials, the participants completed two experimental blocks of 96 trials each. A short break was allowed between the two blocks.

Results

Experiment 1A. The accuracy data were submitted to a two-way repeated measures analysis of variance (ANOVA) with the factors of number of visual stimuli (two vs. four) and congruency (conflicting vs. congruent). The analysis revealed that the participants responded less accurately in the two-light condition ($M = 70\%$) than in the 4-light condition ($M = 78\%$), resulting in a significant main effect of the number of visual stimuli [$F(1,13) = 11.1$, $MS_e = 97.94$, $p < .01$]. The participants also responded more accurately on congruent trials ($M = 97\%$) than on conflicting trials ($M = 50\%$) overall, resulting in a significant main effect of congruency [$F(1,13) = 42.1$, $MS_e = 730.66$, $p < .01$]. Finally, the interaction between these two factors was also significant [$F(1,13) = 16.23$, $MS_e = 68.72$, $p < .01$], reflecting the fact that less cross-modal dynamic capture occurred in the four-light condition than in the two-light condition ($M = 38\%$ vs. 56%, respectively; see Figure 2B).

Experiment 1B. A similar ANOVA performed on the accuracy data from Experiment 1B revealed a significant main effect of congruency [$F(1,13) = 69.4$, $MS_e = 730.74$, $p < .01$], showing that the participants responded less accurately on conflicting trials ($M = 35\%$) than on congruent trials ($M = 95\%$) overall. Neither the main effect of the number of visual stimuli ($F < 1$) nor the interaction between these two factors ($F < 1$) approached significance (mean congruency effects of 61% and 59% in the two-light and four-light conditions, respectively; see Figure 3B).

We subsequently conducted a pooled analysis of the accuracy data from Experiments 1A and 1B with experiment as a between-subjects factor. This ANOVA revealed a significant three-way interaction between experiment, number of visual stimuli, and congruency [$F(1,26) = 10.32$, $MS_e = 47.67$, $p < .01$], confirming the fact that the addition of the two extra lights resulted in a significant reduction in the magnitude of the cross-modal dynamic capture effect in Experiment 1A (in which the multimodal stimuli occurred at the end of the visual stream), but not in Experiment 1B (in which the multimodal stimuli occurred earlier within the visual stream).

Discussion

The main conclusion to draw from the analysis of Experiment 1 is that the reduction of the cross-modal capture effect in the four-light condition occurred only when the visual apparent motion stream started *prior* to the onset of the critical audiovisual stimulus pairs, but not

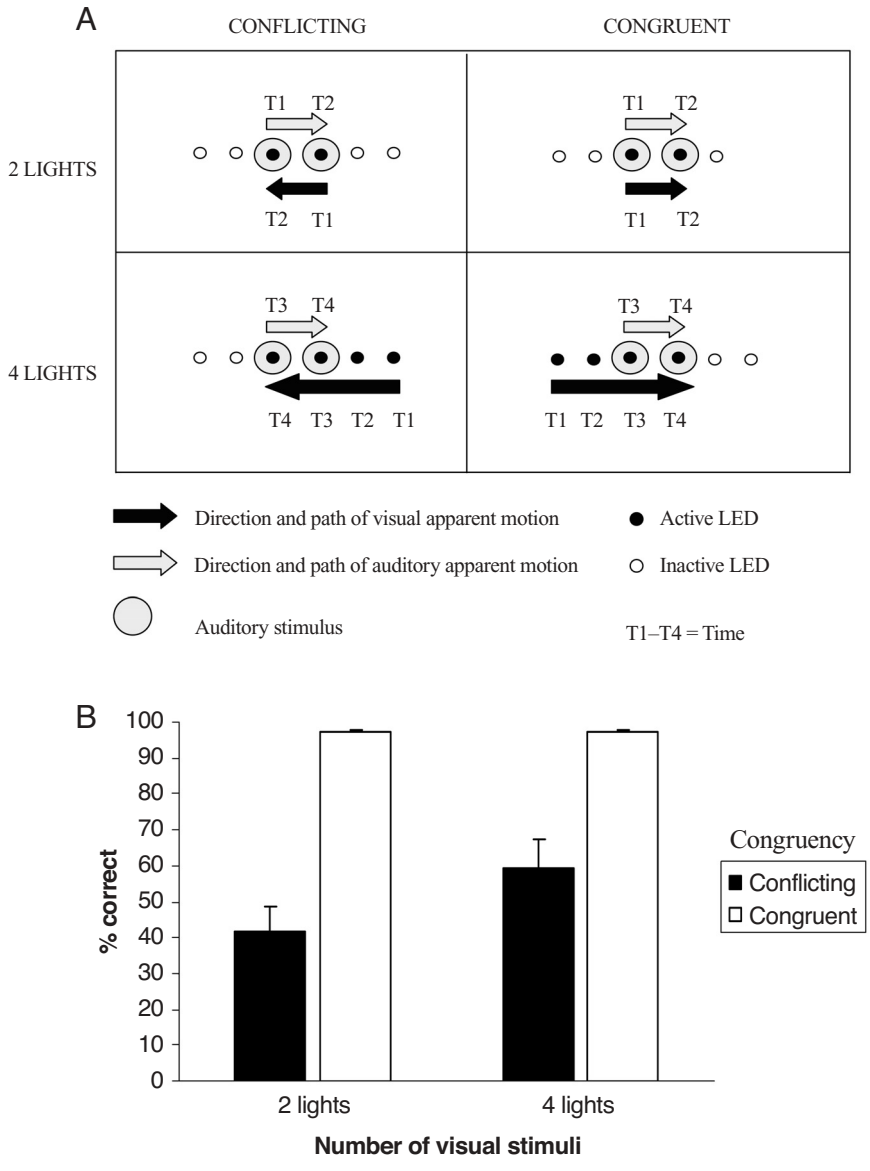

Figure 2. (A) An example of four of the trial types presented in Experiment 1A, resulting from the crossing of the number of visual stimuli (two vs. four lights) and congruency (conflicting vs. congruent) factors. There were also four more trial types (not shown) in which the stimuli moved in the opposite direction (i.e., with the auditory apparent motion moving from right to left and the visual apparent motion from left to right). T1–T4 represent the sequence of onset times for each of the stimuli presented during a particular trial. The upper labels indicate the onset times for the auditory stimuli, and the lower labels the onset times for the visual stimuli. (B) Graph showing the mean accuracy (+ *SE*) in discriminating the direction of auditory apparent motion as a function of the number of visual stimuli and congruency factors.

when the additional visual stimuli eliciting the perceptual grouping were presented *after* the audiovisual stimulus pairs. Two major results support this conclusion. First, the finding in Experiment 1A that a reduced crossmodal dynamic capture effect occurred when the central lights were embedded within a longer visual stream than when they were presented by themselves is in keeping with Sanabria et al.'s (2003, Experiment 1) previous results. Second, *no* reduction in the magnitude of the crossmodal dynamic capture effect was reported in Experiment 1B when the four-light visual apparent motion continued *after* the presentation of the audiovisual stimulus pairs. The latter result suggests that multisensory integration is unaffected by the nature of any intramodal perceptual grouping taking place within vision after the audiovisual stimulus pairs has been presented. As such, the present

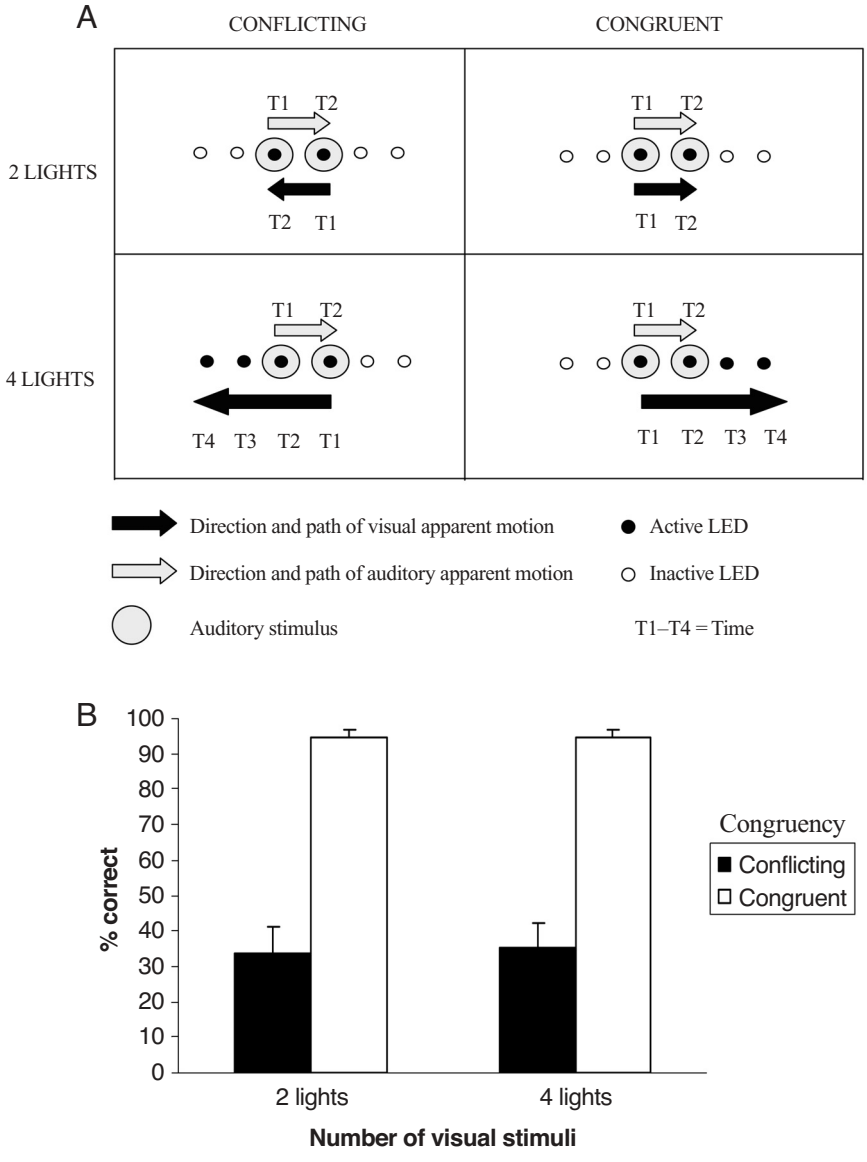

Figure 3. (A) An example of four of the trial types presented in Experiment 1B, resulting from the crossing of the number of visual stimuli (two vs. four lights) and congruency (conflicting vs. congruent) factors. There were also four more trial types (not shown) in which the stimuli moved in the opposite direction (i.e., with the auditory apparent motion moving from right to left and the visual apparent motion from left to right). T1–T4 represent the sequence of onset times for each of the stimuli (auditory and visual) presented during a particular trial. The upper labels indicate the onset times for the auditory stimuli, and the lower labels the onset times for the visual stimuli. (B) Graph showing the mean accuracy (+ *SE*) in discriminating the direction of auditory apparent motion as a function of the number of visual stimuli and congruency factors.

findings contrast with intramodal grouping phenomena, such as sensory saltation (e.g., Geldard, 1976; Kilgard & Merzenich, 1995; Lockhead et al., 1980; Shore et al., 1998), where the subsequent presentation of events in one sensory modality influences a person's perception of the spatial location of earlier events presented in that modality. For example, under the appropriate conditions, four clicks presented to the left ear can be subjectively perceived as moving toward the right ear when four additional clicks are subsequently presented to the right ear (e.g., Shore et al., 1998).

It is important to note, however, that there is another potentially important difference between Experiments 1A and 1B that might also account for the contrasting results between these two experiments. Given that the first two peripheral lights were presented *prior* to the onset of the

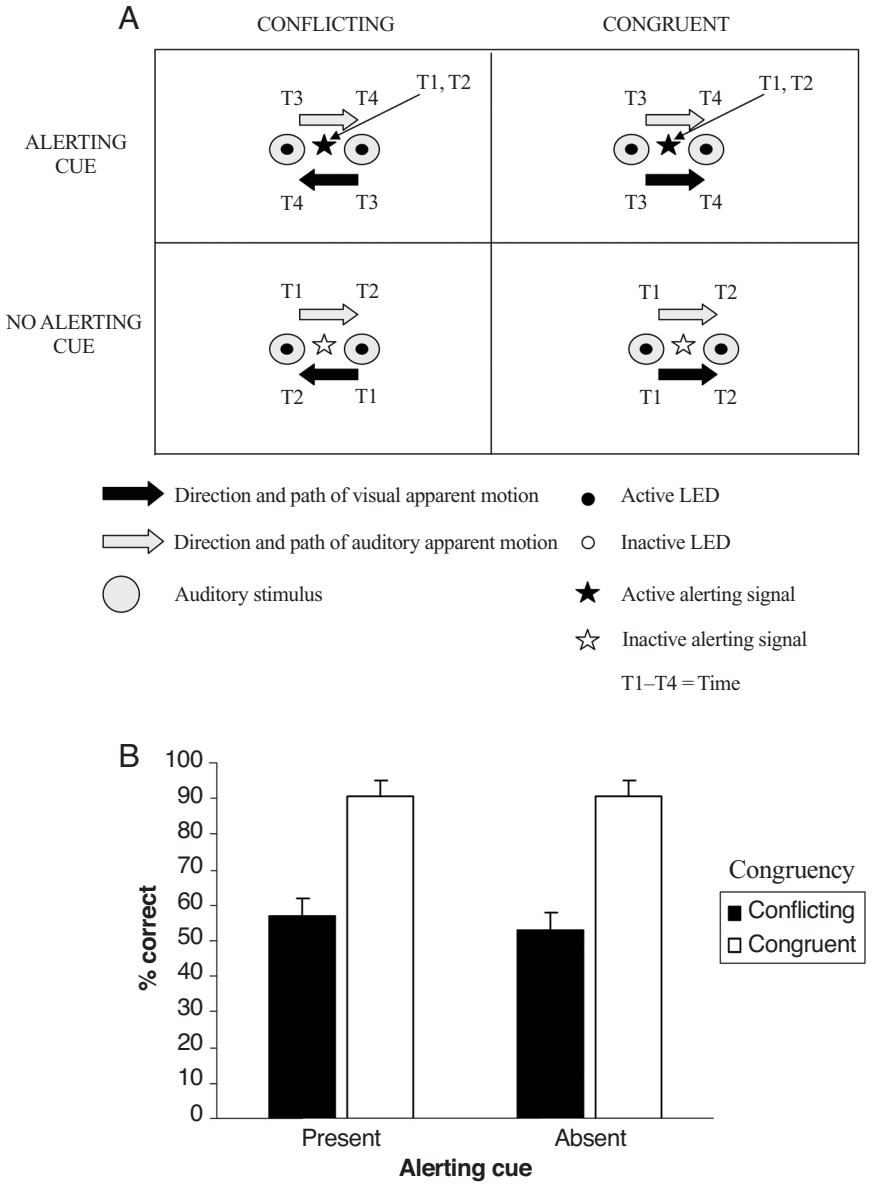

Figure 4. (A) An example of four of the trial types presented in Experiment 2, resulting from the crossing of the alerting (present vs. absent) and congruency (conflicting vs. congruent) factors. The alerting signal (two red flashes presented before the visual apparent motion stimuli) is represented by a star, to differentiate it from the other LEDs. There were also four more trial types (not shown) in which the stimuli moved in the opposite direction (i.e., with the auditory apparent motion moving from right to left and the visual apparent motion from left to right). T1–T4 represent the relative sequence of onset times for each of the stimuli (auditory and visual) presented during a particular trial. The upper labels indicate the onset times for the auditory stimuli, and the lower labels the onset times for the visual stimuli. T1 and T2 represent the onset times for the alerting cue in the two upper panels. (B) Graph showing the mean accuracy (+ *SE*) in discriminating the direction of auditory apparent motion as a function of the alerting and congruency factors.

two target sounds in Experiment 1A (but not in Experiment 1B), one could argue that the initial lights in the sequence may have provided some sort of alerting (or warning) signal to the participants with regard to the imminent onset of the target auditory stimuli in Experiment 1A (see Niemi & Näätänen, 1981; Posner, 1978; Spence & Driver, 1997). This alerting, or temporal warning, effect may have been responsible for the enhanced ability of the participants to segregate the auditory from the visual stimuli, thus perhaps making it easier for them to ignore the irrel-

evant visual stimuli. No such alerting effect would have been present in Experiment 1B, where the first light coincided with the onset of the first target sound.[1]

EXPERIMENT 2

In order to investigate whether the cross-modal dynamic capture effect is modulated by alerting, we compared participants' performance in two conditions, differing only in terms of whether or not an alerting cue was presented prior to the onset of the dynamic audiovisual stimulus pairs. In both conditions, the visual apparent motion stream consisted of the sequential presentation of the two central flashes. In one condition, these two light flashes were presented in the absence of any other visual stimuli (just as in the two-light conditions of Experiment 1). In the other experimental condition, two additional visual stimuli were presented prior to the appearance of the two central flashes, but the likelihood of perceptual grouping between these additional visual stimuli and the subsequent visual apparent motion stream was low (see the Method section for details). If alerting cues do indeed help participants to respond more accurately to auditory stimuli, one would expect to find a reduced cross-modal dynamic capture effect when the alerting signal is present, as compared with performance in those conditions in which it is absent. Instead, if the difference in perceptual grouping between the central flashes and the preceding visual stimuli was the main factor responsible for the reduction of dynamic capture in Experiment 1A, we should observe a similar cross-modal dynamic capture effect in both conditions of this new experiment, since the dynamic auditory and visual stimuli were identical in both.

Method

Participants. Sixteen participants (age range, 24–30 years; mean, 25 years) took part in Experiment 2 in exchange for a £5 gift voucher. All reported normal or corrected-to-normal vision and hearing.

Materials and Procedure. The method was the same as that in the two-light condition in Experiment 1. However, in half of the trials (alerting condition), the red central fixation LED flashed twice (for 100 msec each, with a 50-msec ISI) before the onset of the apparent motion streams. The central fixation point was used as the alerting cue to reduce the possibility of grouping effects with the other visual stimuli.[2] The ISI between the offset of the second alerting flash and the onset of the first pair of audiovisual stimuli was also set at 50 msec (i.e., the stimulus timing was exactly the same as that in the four-light condition in Experiment 1A; see Figure 4A).

Results and Discussion

The analysis of the accuracy data with an ANOVA revealed that the participants responded more accurately on congruent trials ($M = 90\%$) than on conflicting trials ($M = 55\%$) overall, resulting in a significant main effect of congruency [$F(1,15) = 35.43$, $MS_e = 562.19$, $p < .01$]. Importantly, however, the alerting signal had *no* significant effect on the accuracy of our participants' performance [$F(1,15) = 2.33$, $MS_e = 29.80$, n.s.]. Moreover, the interaction between these factors was not significant either [$F(1,15) = 2.17$, $MS_e = 28.04$, n.s.; see Figure 4B]. The results of Experiment 2 therefore strongly suggest that an alerting account cannot explain the modulation of the cross-modal dynamic capture effect reported in the four-light condition of Experiment 1A (or in the six-light condition of Sanabria et al.'s, 2003, study). Here, two light flashes with the same timing as the two initial light flashes of Experiment 1A were presented, and yet no modulation of the cross-modal dynamic capture was observed. The critical difference between the alerting condition in Experiment 2 and the four-light condition in Experiment 1A was the readiness with which the two initial light flashes were grouped into a single visual stream.

EXPERIMENT 3

The results of Experiment 1 demonstrated that the part of the visual stream that was presented after the critical audiovisual event was insufficient for visual perceptual grouping to modulate the multisensory integration of motion information. However, as was noted in the introduction, Sanabria et al. (2003) compared conditions in which six lights were presented with those in which only two lights were presented. Thus, it could be argued that the reduction of the cross-modal dynamic capture found in the six-light condition of their Experiment 1 was due not only to the presence of the initial two lights in the visual stream, but also to the increased distinctiveness of the auditory and the visual streams attributable to the grouping of the *whole* sequence of six lights. It is possible that any visual stimuli presented after the critical audiovisual event might still have some impact on the magnitude of any cross-modal dynamic capture effect, showing some small effect of *late* intramodal perceptual grouping on cross-modal integration.

Therefore, in Experiment 3, we attempted, once again, to investigate the possible *post facto* modulation of the cross-modal dynamic capture effect by perceptual grouping processes taking place within vision *after* the presentation of the target sounds. We compared performance in two different conditions: one in which a stream of six lights was presented (as in Sanabria et al.'s, 2003, Experiment 1) versus one in which a stream of four lights was presented (just as in Experiment 1A of the present study).

The critical question was whether the presence of two extra lights after the multisensory event in the six-light condition would modulate the magnitude of the cross-modal dynamic capture effect. Given that the visual apparent motion stream in this experiment actually started before the onset of the critical audiovisual stimuli in both conditions, we were able to measure the extra contribution (if any) of the two subsequent lights on their own. Note also that by comparing the results of Experiment 3 with those in Experiment 1B, we should be able to replicate Sanabria et al.'s (2003) original result (i.e., less cross-modal dynamic capture effect in the six-light condition than in the two-light condition).

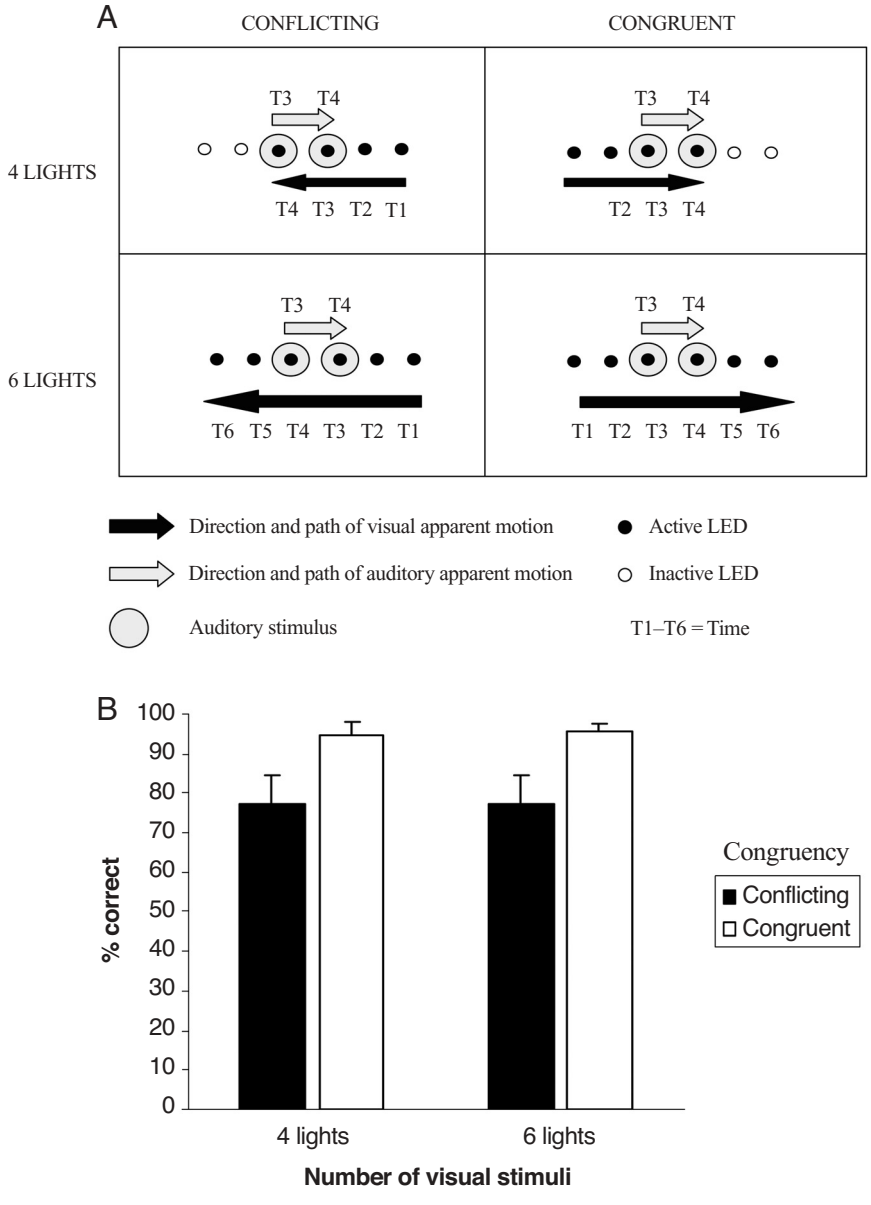

Figure 5. (A) An example of four of the trial types presented in Experiment 3, resulting from the crossing of the number of visual stimuli (four vs. six lights) and congruency (conflicting vs. congruent) factors. There were also four more trial types (not shown) in which the stimuli moved in the opposite direction (i.e., with the auditory apparent motion moving from right to left and the visual apparent motion from left to right). T1–T6 represent the sequence of onset times for each of the stimuli (auditory and visual) presented during a particular trial. The upper labels indicate the onset times for the auditory stimuli, and the lower labels the onset times for the visual stimuli. (B) Graph showing the mean accuracy (+ *SE*) in discriminating the direction of auditory apparent motion as a function of the number of visual stimuli and congruency factors.

Method

Participants. Ten participants (age range, 24–28 years; mean, 23 years) took part in this experiment in exchange for a £5 gift voucher. All the participants reported normal hearing and normal or corrected-to-normal vision.

Materials and Procedure. The method was the same as that in Experiment 1A, with the following exception: Two different types of visual stream were presented with the auditory stimuli, one containing a sequence of six lights, and the other consisting of a sequence of four lights (i.e., the same as in the six-light condition, but with the last two lights missing; see Figure 5A).

Results and Discussion

Analysis of the accuracy data revealed that the participants' performance was significantly more accurate on

congruent trials ($M = 95\%$) than on conflicting trials ($M = 77\%$), overall [$F(1,9) = 10.04$, $MS_e = 312.25$, $p < .02$], as in the previous experiments. However, the number of visual stimuli had no significant influence on performance ($F < 1$), and the interaction between these two factors also failed to reach significance ($F < 1$), with an equivalent cross-modal dynamic capture effect being reported in both the four-light and the six-light conditions ($M = 18\%$ in both conditions; see Figure 5B), as was predicted.[3]

Once again, the results of Experiment 3 point to the fact that the nature of any intramodal visual perceptual grouping taking place *after* the presentation of the auditory target stimuli has no influence on the cross-modal dynamic capture effect. Varying the presence/absence of visual stimuli after the offset of the sounds had no measurable effect on the magnitude of any cross-modal dynamic capture effects reported (just as in Experiment 1B), despite the fact that the priority for visual grouping had already been established by the initial presentation of the two leading lights in the present experiment.

In order to explore further the role of *post facto* perceptual grouping in the cross-modal dynamic capture effect, we conducted a pooled ANOVA of the data from Experiments 1B and 3, with experiment as the between-subjects factor (note that the four-light condition was identical in both experiments, whereas what differed was the two-light vs. six-light conditions). This analysis revealed a significant main effect of experiment [$F(1,22) = 13.8$, $MS_e = 790.08$, $p < .01$], reflecting the difference in terms of the overall accuracy of performance between the two experiments, with a lower level of accuracy being reported in Experiment 1B than in Experiment 3. The interaction between experiment and congruency was also significant [$F(1,22) = 19$, $MS_e = 562.81$, $p < .01$], supporting the claim that a stronger cross-modal dynamic capture effect was demonstrated in Experiment 1B than in Experiment 3 ($M = 60\%$ vs. 18%, respectively). Importantly, there was no interaction between number of visual stimuli, congruency, and experiment ($F < 1$). Therefore, the between-experiments analysis of Experiments 1B and 3 points to the fact that the nature of any intramodal perceptual grouping taking place after the presentation of the critical audiovisual stimuli (i.e., even if it had been initiated prior to the onset of the audiovisual stimuli) had no significant effect on multisensory integration. Moreover, a *t* test pairwise comparison of the six-light condition of Experiment 3 with the two-light condition of Experiment 1B revealed a significant difference between the cross-modal capture effects found in these two experimental conditions [17% vs. 61%, respectively; $t(22) = -4.23$, $p < .01$], further supporting the original results reported by Sanabria et al. (2003).

GENERAL DISCUSSION

The main conclusion to be drawn from the experiments reported here is that the nature of any visual perceptual grouping taking place *after* the presentation of the audiovisual stimuli has no effect on the multisensory integration of motion information (as measured by the cross-modal dynamic capture effect). The results of Experiment 1 concur with those recently reported by Sanabria et al. (2003) but go further by demonstrating that the crucial manipulation modulating audiovisual multisensory integration via unimodal perceptual grouping lies in the relative temporal onset of the visual apparent motion stream. When the initial portion of the visual apparent motion stream is presented first, it presumably allows for the establishment of the perceptual grouping of the visual stimuli prior to the presentation of the audiovisual stimuli, hence enabling the visual events to be more easily segregated from the auditory events.

Experiment 2 demonstrated that this modulation was not due simply to an alerting, or warning signal, effect (e.g., Niemi & Näätänen, 1981; Posner, 1978; Spence & Driver, 1997). Experiment 3 corroborated the results of Experiment 1B whereby no modulation of the cross-modal dynamic capture effect was demonstrated as a function of the perceptual grouping that took place *after* the critical audiovisual stimuli eliciting the cross-modal dynamic capture effect. In particular, Experiment 3 demonstrated that the crucial part of the visual stream modulating the cross-modal dynamic capture effect consisted of the stimuli that preceded (or co-occurred with) the multisensory stimuli.

Contrary to the situation for intramodal perceptual grouping phenomena, such as sensory saltation (e.g., Geldard, 1976; Kilgard & Merzenich, 1995; Lockhead et al., 1980; Shore et al., 1998), the *late* perceptual grouping processes occurring within one sensory modality (vision, in this case) appear to have no significant effect on the cross-modal grouping occurring between earlier presented auditory and visual apparent motion streams.

A better understanding of the way in which intramodal and cross-modal perceptual groupings interact may also help to resolve certain conflicting findings present in the literature on the topic of the multisensory integration of motion signals. For instance, Mateeff et al. (1985) reported a study in which the presentation of a visual motion stimulus was shown to modulate the perceived motion of a sound source. Mateeff et al. (Experiment 1) found that while participants were tracking a moving visual target with their eyes, a static sound presented for about 1 sec appeared to move with the visual stimulus. Note, however, that in this experiment, the visual stimulus started to move *before* the onset of the target auditory motion stimulus and continued to move *after* it had finished, just as in the six-light condition of Experiment 3. Consequently, the conditions for cross-modal perceptual grouping were presumably not optimized in their study. By contrast, the onset and offset of the auditory and visual stimuli in Zapparoli and Reatto's (1969) study (and in the majority of previous studies of the cross-modal dynamic capture effect; e.g., Soto-Faraco et al., 2002) occurred synchronously, presumably enhancing any cross-modal perceptual grouping that might have taken place.

Taken together with the results of a number of other recent studies (e.g., Sanabria et al., 2003; Vroomen & de Gelder, 2000; Watanabe & Shimojo, 2001), it seems likely that intramodal perceptual grouping principles may come to play an increasingly important role in helping researchers to explain the presence versus absence of multisensory integration (or grouping) effects across different modalities.

The data reported here suggest that multisensory integration depends on how the environment is carved up into units or perceptual objects. In the present study, the auditory and visual stimuli were perceptually grouped under particular conditions giving rise to two perceptual objects (one visual and the other auditory; see Kubovy & Van Valkenburg, 2001). Interestingly, when both perceptual objects were simultaneous in time and presented from the same spatial locations (i.e., spread over the same medium, using Kubovy & Van Valkenburg's, 2001, terminology) and were of the same number (i.e., two lights and two sounds), cross-modal dynamic capture took place more readily. As a consequence, our participants were more likely to perceive a single multisensory object moving coherently through space and time. When a greater number of visual stimuli (i.e., a more extended visual stream) were presented, a different visual object "emerged." Thus, because of the difference in the number of components of the more extended visual stream, relative to the auditory stream, multisensory integration was *less likely* to occur, presumably because the participants were *more likely* to segregate the auditory and the visual stimuli into two separate unimodal perceptual objects.

Although widely investigated, the levels of *processing* at which visual perceptual grouping and audiovisual cross-modal integration occur remain an issue of some debate in the literature. For instance, whereas it has traditionally been accepted that visual perceptual grouping occurs at a low level of perceptual processing (e.g., Francis & Grossberg, 1996), Palmer and his collaborators (Beck & Palmer, 2002; Palmer, 2002; Palmer, Brooks, & Nelson, 2003) have recently challenged this view. These authors have proposed that "higher" levels of perceptual processing—described as the processes that occur after the creation of an initial set of discrete elements (giving rise to perceptual constancy; e.g., Palmer, 2002)—may also be responsible for the manifestation of perceptual grouping.

With regard to the neural basis of such processes as perceptual grouping and the multisensory integration of motion information, a range of brain areas have been highlighted in previous studies. Relevant to the present study is Francis and Grossberg's (1996) suggestion that interactions between primary visual areas V1, V2, and hMT may provide the locus of perceptual grouping by apparent motion. A number of researchers have also started to highlight certain brain areas that appear to play a critical role in the multisensory integration of motion information (e.g., the ventral premotor cortex, the ventral intraparietal area, the lateral parietal cortex, the lateral frontal cortex, and the hMT), most of which are considered "higher level" association areas (see Soto-Faraco et al., 2003, for a review). Moreover, "high level" sensory interactions would depend on the perceptual organization of unimodal sensory information, processed at *earlier* stages (e.g., such as in areas V1 and V2).

In principle, our data regarding the modulation of the multisensory integration of motion information by the perceptual grouping taking place within vision seem to fit quite well with the results of these neurophysiological and neuroimaging studies. However, it is worth mentioning that some researchers (e.g., Hagen et al., 2002) have suggested the hMT area ("low-level" perceptual area) may also be implicated in the multisensory integration of motion information. Therefore, an important issue for future research will be to uncover exactly how the modulation of multisensory integration by unimodal perceptual grouping established in the present study is implemented in neural terms.

REFERENCES

Allen, P., & Kolers, P. (1981). Sensory specificity of apparent motion. *Journal of Experimental Psychology*, **7**, 1318-1326.

Beck, D. M., & Palmer, S. E. (2002). Top-down influences on perceptual grouping. *Journal of Experimental Psychology: Human Perception & Performance*, **28**, 1071-1084.

Bertelson, P., & de Gelder, B. (2004). The psychology of multimodal perception. In C. Spence & J. Driver (Eds.), *Crossmodal space and crossmodal attention* (pp. 141-179). Oxford: Oxford University Press.

Calvert, G. A., Spence, C., & Stein, B. E. (Eds.), (2004). *The handbook of multisensory processes*. Cambridge, MA: MIT Press.

Driver, J., & Spence, C. (2000). Multisensory perception: Beyond modularity and convergence. *Current Biology*, **10**, 311-331.

Ehrenstein, W. H., & Reinhardt-Rutland, A. H. (1996). A cross-modal aftereffect: Auditory displacement following adaptation to visual motion. *Perceptual & Motor Skills*, **82**, 23-26.

Francis, G., & Grossberg, S. (1996). Cortical dynamics of form and motion integration: Persistence, apparent motion, and illusory contours. *Vision Research*, **36**, 149-173.

Geldard, F. A. (1976). The saltatory effect in vision. *Sensory Processes*, **1**, 77-86.

Hagen, M. C., Franzen, O., McGlone, F., Essick, G., Dancer, C., & Pardo, J. V. (2002). Tactile motion activates the human middle temporal (MT/V5) complex. *European Journal of Neuroscience*, **16**, 957-964.

Howard, L. P., & Templeton, W. B. (1966). *Human spatial orientation*. New York: Wiley.

Kilgard, M. P., & Merzenich, M. M. (1995). Anticipated stimuli across skin. *Nature*, **373**, 663.

Koffka, K. (1935). *Principles of Gestalt psychology*. New York: Harcourt Brace.

Kubovy, M., & Van Valkenburg, D. (2001). Auditory and visual objects. *Cognition*, **80**, 97-126.

Lockhead, G. R., Johnson, R. C., & Gold, F. M. (1980). Saltation through the blind spot. *Perception & Psychophysics*, **27**, 545-549.

Mateeff, S., Hohnsbein, J., & Noack, T. (1985). Dynamic visual capture: Apparent auditory motion induced by a moving visual target. *Perception*, **14**, 721-727.

Niemi, P., & Näätänen, R. (1981). Foreperiod and simple reaction time. *Psychological Bulletin*, **89**, 133-162.

Palmer, S. T. (2002). Perceptual grouping: It's later than you think. *Current Directions in Psychological Science*, **11**, 101-106.

Palmer, S. T., Brooks, J. L., & Nelson, R. (2003). When does grouping happen? *Acta Psychologica*, **114**, 311-330.

Phillips, D. P., & Hall, S. E. (2001). Spatial and temporal factors in auditory saltation. *Journal of the Acoustical Society of America*, **110**, 1539-1547.

Posner, M. I. (1978). *Chronometric explorations of mind*. Hillsdale, NJ: Erlbaum.

Sanabria, D., Soto-Faraco, S, Chan, J., & Spence, C. (2003). *Intramodal perceptual grouping modulates multisensory integration: Evidence from the crossmodal dynamic capture task*. Manuscript submitted for publication.

Shore, D. I., Hall, S. E., & Klein, R. M. (1998). Auditory saltation: A new measure for an old illusion. *Journal of the Acoustical Society of America*, **103**, 3730-3733.

Soto-Faraco, S., & Kingstone, A. (2004). Multisensory integration of dynamic information. In G. Calvert, C. Spence & B. E. Stein (Eds.), *The handbook of multisensory processes* (pp. 49-69). Cambridge, MA: MIT Press.

Soto-Faraco, S., Kingstone, A., & Spence, C. (2003). Multisensory contributions to the perception of motion. *Neuropsychologia*, **41**, 1847-1862.

Soto-Faraco, S., Lyons, J., Gazzaniga, M., Spence, C., & Kingstone, A. (2002). The ventriloquist in motion: Illusory capture of dynamic information across sensory modalities. *Cognitive Brain Research*, **14**, 139-146.

Soto-Faraco, S., Spence, C., & Kingstone, A. (2004). Cross-modal dynamic capture: Congruency effects of motion perception across sensory modalities. *Journal of Experimental Psychology: Human Perception & Performance*, **30**, 330-345.

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

Staal, H. E., & Donderi, D. C. (1983). The effect of sound on visual apparent movement. *American Journal of Psychology*, **96**, 95-105.

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

Vroomen. J., & de Gelder, B. (2000). Sound enhances visual perception: Cross-modal effects of auditory organization on vision. *Journal of Experimental Psychology: Human Perception & Performance*, **26**, 1583-1590.

Watanabe, K., & Shimojo, S. (2001). When sound affects vision: Effects of auditory grouping on visual perception. *Psychological Science*, **12**, 109-116.

Wertheimer, M. (1950). Laws of organization in perceptual forms. In W. D. Ellis (Ed.), *A sourcebook of Gestalt psychology* (pp. 71-81). New York: Humanities Press. [Original work published 1923]

Wuerger, S. M., Hofbauer, M., & Meyer, G. F. (2003). The integration of auditory and visual motion signals at threshold. *Perception & Psychophysics*, **65**, 1188-1196.

Zapparoli, G. C., & Reatto, L. L. (1969). The apparent movement between visual and acoustic stimulus and the problem of intermodal relations. *Acta Psychologica*, **29**, 256-267.

NOTES

1. Note that one could also argue that instead of improving performance, any alerting effect might actually have interfered with the perception of auditory apparent motion, given that alerting effects are normally characterized by a speed–accuracy tradeoff in human performance (e.g., Posner, 1978).

2. In several pilot experiments, we tried to use the most peripheral lights (one or two flashes from the most distant LED on either side of fixation) as the alerting signal, but a continuous visual stream was perceived by most participants, presumably caused by a saltation-like effect produced by the rapid sequential presentation of the visual stimuli in the periphery (e.g., Geldard, 1976; Lockhead et al., 1980).

3. As the conclusions from Experiments 2 and 3 were partially based on null results, we conducted statistical power analyses in both cases. The results of these analyses revealed a lack of power (less than 30% for the two-way interactions in both experiments), and therefore, our conclusions should be qualified by this fact. However, it should also be noted that this reduced power derives not only from the small sample size used, but also, critically, from the very small numerical differences between the conditions compared in both Experiment 2 and Experiment 3, which were used as the predicted effect size to conduct the power calculations (4% difference in Experiment 2 and less than 1% difference in Experiment 3).

(Manuscript received August 15, 2003;
revision accepted for publication April 22, 2004.)