PROCEEDINGS OF THE ROYAL SOCIETY B

Proc. R. Soc. B (2011) **278**, 2840–2847
doi:10.1098/rspb.2010.2757
Published online 9 February 2011

Acoustic facilitation of object movement detection during self-motion

F. J. Calabro[1], S. Soto-Faraco[2] and L. M. Vaina[1,3,4,*]

[1]*Brain and Vision Research Laboratory, Department of Biomedical Engineering, Boston, MA 02215, USA*
[2]*ICREA and Department de Tecnologies de la Informació i les Comunicacions, Universitat Pompeu Fabra, Barcelona, Spain*
[3]*Department of Neurology, and* [4]*Department of Radiology, Massachusetts General Hospital, Harvard Medical School, Boston, MA 02215, USA*

In humans, as well as most animal species, perception of object motion is critical to successful interaction with the surrounding environment. Yet, as the observer also moves, the retinal projections of the various motion components add to each other and extracting accurate object motion becomes computationally challenging. Recent psychophysical studies have demonstrated that observers use a flow-parsing mechanism to estimate and subtract self-motion from the optic flow field. We investigated whether concurrent acoustic cues for motion can facilitate visual flow parsing, thereby enhancing the detection of moving objects during simulated self-motion. Participants identified an object (the target) that moved either forward or backward within a visual scene containing nine identical textured objects simulating forward observer translation. We found that spatially co-localized, directionally congruent, moving auditory stimuli enhanced object motion detection. Interestingly, subjects who performed poorly on the visual-only task benefited more from the addition of moving auditory stimuli. When auditory stimuli were not co-localized to the visual target, improvements in detection rates were weak. Taken together, these results suggest that parsing object motion from self-motion-induced optic flow can operate on multisensory object representations.

Keywords: flow parsing; visual search; multisensory perception; visual motion; auditory motion

1. INTRODUCTION

For a stationary observer, an object moving within an otherwise still scene is uniquely identified by motion and can be effortlessly detected no matter how many elements the scene contains [1]. It has been suggested that motion-responsive mechanisms filter out the static objects, thus making detection of the unique moving elements trivial [2]. If the background objects can be grouped into a rigid surface defined by disparity [3] or common motion [4], then a single moving object will also pop out. However, from the point of view of the visual system, static or rigid backgrounds are only an exceptional case given that an observer's translation and head motion produce a complex movement pattern of the visual field, or optic flow [5,6]. This adds remarkable complexity when trying to single out object motion from the dynamic scene, given that all objects move in terms of their retinal projections. Yet perception of object motion during self-movement in humans is both accurate and critical to successful interaction with the environment. It has been proposed that object motion can be parsed out from the optic flow created by self-motion, thus allowing a moving observer to detect a moving object. Rushton and colleagues [7–9] have suggested a flow-parsing mechanism that uses the brain's sensitivity to optic flow to separate retinal motion signals into those components owing to observer movement and those owing to the movement of objects in the scene.

Previous studies addressing flow parsing have concentrated on the visual modality alone. Although vision is dominant in our perception of motion, natural environments frequently provide extra-visual cues to motion, such as the sound of a car down the street quickly approaching us (or moving away from us). The question addressed here is therefore whether extra-visual (in this case, acoustic) information can complement optic flow parsing, and hence facilitate the extraction of visual motion from dynamic visual scenes during observer movement. The benefits of congruent, cross-modal information are well known—especially the enhancement of responses to a stimulus in situations when the signal from a single modality is weak [10–12] or when processing within one sensory system is impaired by brain damage [13]. In the particular case of motion, strong synergies between different sensory modalities have been described in several recent studies. For example, in horizontal motion, directional incongruence between visual and auditory signals can lead to strong illusions regarding the perceived direction of the sound (e.g. [14–16]; see [17] for a review), whereas directional congruence can lead to improved detection performance (e.g. [18–20]; though Alais & Burr [19] suggest the improvement may be statistical, rather than owing to bimodal enhancement). Appropriately timed static sounds can drive the perceived direction of an ambiguous visual apparent-motion stimulus [21]. Speed of motion is subject to similar phenomena, given that sounds will

*Author for correspondence (vaina@bu.edu).

Electronic supplementary material is available at http://dx.doi.org/10.1098/rspb.2010.2757 or via http://rspb.royalsocietypublishing.org.

Received 21 December 2010
Accepted 20 January 2011

appear to move faster (or slower) depending on the velocity of concurrent visual stimuli [22].

Directly relevant to the present question, several previous studies have shown that cross-modal directional congruency effects can be observed in motion along the depth plane. For example, auditory looming has been shown to speed up the detection of looming visual signals [23]. In studies using motion after-effects in the depth plane [24,25], adaptation to looming (or receding) visual stimuli produced an after-effect in the reverse direction for subsequently presented sounds. When using directionally incongruent audio-visual adaptors, the after-effect is consistent with the direction of the visual adaptor. The phenomenology of these findings suggests that the interaction between visual and auditory motion signals can express at rather early levels of processing. Indeed, recent studies using fMRI have revealed that cross-modal motion congruency effects are reflected in a complex network of brain structures, including unisensory motion processing areas as well as areas of multisensory convergence [26,27]. In particular, illusory reversals of sound direction (induced by directional incongruence between auditory and visual motion) were correlated with a deactivation of auditory motion areas (the auditory motion complex, AMC) and an enhancement of activity in the cortical areas responsive to visual motion. Furthermore, in the same study it was shown that, just prior to trials leading to illusory sound motion percepts, activity in the ventral intraparietal area (VIP; an area of multi-sensory convergence that contains spatial representations) was stronger than in identical trials that resulted in veridical perception of sound direction [26].

We used a visual search paradigm [28–32] that has been previously used to test optic flow parsing [1–3,7,33–35]. Several recent findings attest to the potential of cross-modal enhancement by sounds to improve visual selective attention in search tasks [36,37]. For example, Van der Burg *et al.* [36] showed that sounds temporally coincident with an irrelevant colour change in visual targets dramatically improved search times in a difficult search task. In fact, a difficult visual search task that led to serial search patterns in the absence of sounds reflected nearly flat search slopes when a sound was synchronized to target colour changes. Interestingly, when the sound was paired with a visual distractor colour change instead, the search became more difficult. These demonstrations, together with the strong cross-modal synergies in motion processing described above, highlight the possibility that acoustic motion could help parse out object motion from optic flow in dynamic visual displays.

Note, however, that none of the previous studies addressing cross-modal enhancement in visual search has, to our knowledge, involved dynamic scenes. Moreover, paradigms where perceptual load is high (i.e. when the matching between sound and visual events must be extracted from complex, dynamically changing events) have typically failed to demonstrate cross-modal enhancement in search tasks [38,39].

It is therefore uncertain whether the visual motion processes leading to parsing out object motion from optic flow produced by the observer's movement can benefit from cross-modal synergies. Here, we address this question empirically. We compared performance on a task of object movement detection during self-motion when paired with a static or moving auditory cue to determine whether cross-modal motion congruency enhances visual selection. Our results show that while auditory stimuli not co-localized with the visual target impart only a small benefit to detection rates, the presentation of a moving, co-localized auditory cue provides a significant gain.

2. METHODS

(a) *Subjects*

All participants ($n = 18$, eight males; age range: 19–29, mean: 22) performed the visual task, and each was tested with either the non-co-localized ($n = 10$) or co-localized ($n = 10$) auditory condition. Two of the participants, including F.J.C. (an author), performed both auditory conditions, and all except F.J.C. were naive to the purposes of the experiment.

(b) *Apparatus*

Participants viewed the visual display from a distance of 60 cm, with head position fixed by a chin and forehead rest. Stimuli were displayed on a 23″ Apple Cinema Display and were generated in MATLAB using Psychophysical Toolbox [40,41] and OpenGL libraries. Suprathreshold auditory cues were presented with Bose QC-1 QuietComfort acoustic noise cancelling headphones. We used a Minolta LS-100 for monitor luminance calibration, and a Scantek Castle GA-824 Smart Sensor SLM for acoustic calibration.

(c) *Stimulus*

Participants viewed nine textured spheres distributed within a simulated virtual environment of size $25 \times 25 \times 60$ cm (figure 1*a*), projected onto an Apple Cinema Display. Stimuli were viewed binocularly, but contained no stereo cues, such that visual motion in depth was determined only by looming motion. To avoid overlapping spheres, the viewable area was divided into nine equally sized wedges in the frontoparallel plane, and one object was placed into each wedge with a randomly chosen eccentricity. Objects were located randomly in simulated depth between 25 and 35 cm, and had a mean diameter of 1.5° (deg of visual angle) at the start of the stimulus, with a mean luminance of 28 cd m^{-2} on a background of luminance 0.3 cd m^{-2}. A red fixation mark was placed at the centre of the display and subjects were instructed to maintain fixation throughout the testing block.

Forward observer motion was simulated towards the fixation mark for 1 s. Except where noted, the observer motion was a forward translation of 3 cm s^{-1} (thus inducing a corresponding expansion of objects that were stationary within the scene). One of the nine objects (the target) was assigned an independent motion vector, moving either forward or backward at 2, 4, 6 or 8 cm s^{-1} with respect to the rest of the scene (figure 1*b*). The target's visible motion was the sum of its own motion vector and the induced motion caused by the simulated translation of the observer. At the end of the motion, the screen was cleared for 250 ms before all objects reappeared at their final locations, but projected into a single depth plane so that all were a constant size. Four objects (the target and three other randomly selected spheres) were shown with labels (marked with numerals, 1–4) and observers were asked to report which of the four was the one that had been moving within the scene (and not solely because of the observer translation). Since the labels appeared only after the end of the trial,

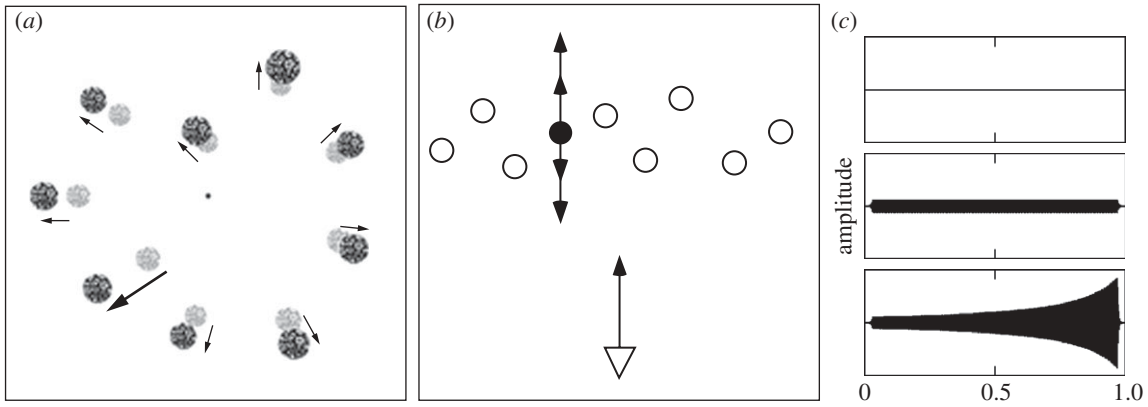

Figure 1. (*a*) Stimulus display during simulated forward translation (motion vectors indicated by arrows) with one object moving independently within the scene (indicated by a bold arrow). (*b*) Zenithal view of the stimulus layout. As the observer (triangle) moves forward (3 cm s^{-1}), the target (black circle) moves either forward or backward within the scene (white circles). (*c*) Amplitude envelope of the auditory cues in the visual-only (top), auditory-static (middle) and auditory-moving (bottom) cases.

subjects had to monitor all nine objects, although their decision was a four-alternative forced-choice task.

In separate conditions, the stimulus was presented only visually or with either a co-localized or a non-co-localized (central) auditory cue. The auditory cue was a pure tone of frequency 300 Hz that in 75 per cent of trials (*auditory-moving* trials) was simulated (via a change in amplitude) to move within the scene in the same direction as the target (forward or backward), and in the remaining 25 per cent of trials was presented at constant amplitude throughout the trial (*auditory-static* trials). The change in amplitude was modelled as a sound source at an initial distance of 4.1 m (69 dB SPL), moving towards or away from the observer at 3.5 m s^{-1} (resulting in a change of approximately 10 dB SPL; figure 1*c*). Sound attenuation as a function of distance was approximated for the testing room by measuring sound levels at various distances from a constant sound source. A least-squares fit was applied to determine the relationship of sound amplitude to distance. In both the static and moving (whether approaching or receding) auditory conditions, the auditory cue started at the same amplitude so that the initial volume did not indicate whether the auditory cue would move, nor in what direction. In all auditory conditions, the auditory stimulus was enveloped with 30 ms ramps to avoid clicks owing to a sharp onset or offset. Participants were screened to ensure they could discriminate the direction of the auditory motion.

In the non-co-localized auditory condition, the auditory cue appeared to arise directly in front of the observer (it was presented with equal amplitude to both ears). In the co-localized auditory condition, the interaural intensity difference (IID) was adjusted to match the horizontal eccentricity of the target object. For both the non-co-localized and co-localized auditory cues, we used auditory-moving and auditory-static conditions to distinguish effects owing to localizing the target, effects owing to congruent auditory motion, and effects that require both spatially co-localized and congruent-motion auditory cues.

3. RESULTS

(a) *Detection of object movement during self-motion*

Figure 2 shows the results from all 18 subjects on the visual-only condition. As expected, performance

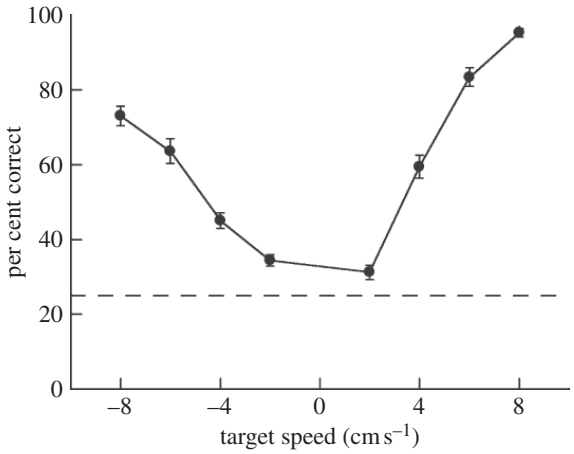

Figure 2. Performance on the visual-only condition for 18 subjects. Error bars are s.e.m across subjects. Negative speeds refer to receding targets and positive speeds to looming targets, relative to scene motion. The horizontal line indicates chance performance (25% correct).

depended on the speed of the target object, with faster speeds (6 and 8 cm s^{-1}) being detected above 80 per cent correct. Performance was above chance (chance = 25%) for all speeds. Approaching objects (those moving towards the observer within the simulated environment) were easier to detect than receding ones, as demonstrated by the increased performance for positive speeds relative to negative speeds: a two-way ANOVA showed significant effects of target speed ($F_{3,152} = 143.5$, $p < 0.001$) and direction ($F_{1,152} = 47.5$, $p < 0.001$).

(b) *How is object motion detected?*

Object motion detection in the visual search task may be accomplished by flow parsing (as suggested by Rushton and colleagues [7–9]), in which self-motion is estimated from background optic flow and parsed out from the scene. Alternatively, to resolve this task, participants may use the object's motion relative to the observer (i.e. retinal motion)—for example, detecting an object with a high perceived speed, or an object that appears nearly stationary among moving objects (as in [42]). To determine which of these mechanisms was most probably used in our experiment, 10 participants performed an

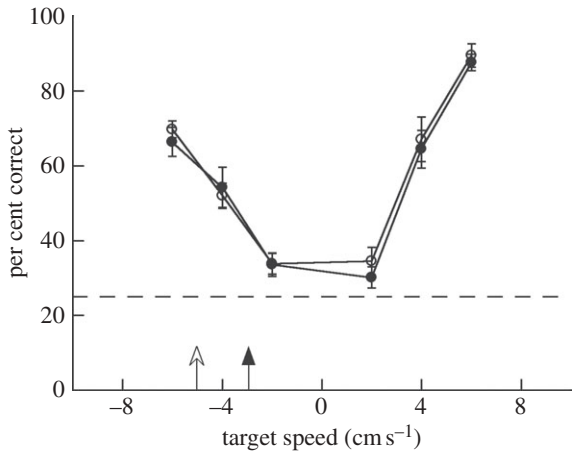

Figure 3. Performance accuracy on the visual-only condition for observer speeds of 3 cm s^{-1} (black circles) and 5 cm s^{-1} (white circles). Data from the 10 subjects who participated in both conditions are shown. Arrows indicate the speeds at which an object would appear stationary on the screen (observer velocity equal to target object velocity). Error bars are s.e.m. across subjects.

additional visual-only condition in which the speed of the simulated observer motion was increased to 5 cm s^{-1}. If observers used relative motion cues, this should have resulted in worse performance for the -6 cm s^{-1} target speed (where the target's retinal speed decreased with the faster observer motion), and better performance for objects with positive (2, 4, 6 cm s^{-1}) velocities (in which the retinal speed increased with the faster observer motion). If, on the other hand, subjects used a flow-parsing mechanism, performance levels should have been independent of the self-motion speed (which is parsed out), as long as self-motion was easily detected, as was the case in both observer speed conditions (3 and 5 cm s^{-1}).

Figure 3 shows the results for an observer speed of 5 cm s^{-1} compared with results from the same 10 subjects when the observer speed was 3 cm s^{-1}. A two-way ANOVA showed a significant effect of target speed ($F_{5,113} = 84.4$, $p < 0.001$), thus reproducing the result of the main visual-only experiment, but no effect of observer speed ($F_{1,113} = 0.12$, $p > 0.7$). We further tested the two predictions of the retinal motion hypothesis separately: (1) a decrease in performance at -6 cm s^{-1} owing to lower retinal speed at the higher observer speed, and (2) an increase in performance for positive object speeds owing to the increase in retinal speed at the higher observer speed. A *t*-test considering only data from the -6 cm s^{-1} object speed (prediction 1) showed no difference with changes in observer motion ($t_9 = -0.11$, $p = 0.91$), and a two-way ANOVA restricted to positive target speeds (prediction 2) similarly showed no significant effect of observer speed ($F_{1,56} = 0.13$, $p > 0.7$). Furthermore, a two-one-sided *t*-test [43,44] for equivalence showed that performance for the two observer speed conditions across subjects and object speed was statistically equivalent at $p < 0.05$ within a tolerance of 2.5 per cent. Since a change of 2 cm s^{-1} caused on average a 21 per cent change in performance when applied to the object speed, equivalent performance within a 2.5 per cent tolerance when the 2 cm s^{-1} speed difference was applied to the observer speed indicates that the difference in retinal motion speeds cannot account for performance on the visual task. Taken together, these results suggest it was unlikely that observers solved the task by using only retinal motion cues.

(c) *Do auditory cues facilitate detection of object movement during self-motion?*

To determine whether auditory motion cues can facilitate the detection of object movement, we considered two conditions in which a moving auditory cue was presented with motion direction congruent to that of the visual target. First, we tested whether the detection of object movement during self-motion is facilitated by the presentation of a synchronous, but spatially non-co-localized, auditory cue (perceptually located at the centre of the display). Second, we tested whether facilitation depends on the spatial co-localization of the visual and auditory motions (with an IID matching the horizontal eccentricity of the visual target). In both cases (non-co-localized and co-localized), performance was compared with that of static auditory cues.

(i) *Non-co-localized auditory stimulus*

Figure 4 shows the performance of 10 subjects on the moving object detection task in the presence of a non-co-localized auditory cue (localized to the centre of the screen). A two-way ANOVA showed a small, non-significant increase in performance (3.2% mean improvement) for auditory-moving trials ($F_{1,144} = 3.39$, $p = 0.06$) as compared to auditory-static trials. There was a significant main effect of target speed ($F_{7,144} = 81.9$, $p < 0.001$), but no significant interaction between auditory condition (static versus moving) and target speed ($F_{7,144} = 0.64$, $p > 0.7$). These results suggest that the presentation of a synchronous auditory cue that is not spatially co-localized with the target produced only a very modest improvement in the detection of a moving object.

An analysis of reaction times in trials with correct responses showed that both auditory-static and auditory-moving trials resulted in faster response times than the visual-only condition in the same subjects, by 43 ($F_{1,57} = 10.12$, $p = 0.002$) and 41 ms ($F_{1,57} = 26.59$, $p < 0.001$), respectively. However, there was no significant difference between the auditory-static and auditory-motion conditions ($F_{1,59} = 0.14$, $p > 0.7$). Therefore, the use of a non-co-localized auditory motion cue contributed neither a statistically significant increase in performance nor decrease in response time.

(ii) *Co-localized auditory stimulus*

To test the effect of spatial co-localization on auditory facilitation in our task, 10 participants performed a version of the task in which the IID of the auditory cue was adjusted to match the horizontal eccentricity of the visual target. To ensure that changes in performance were not due to the spatial localization information provided by the localized auditory cue, performance between auditory-static and auditory-moving conditions was compared (see figure 4*b*; note that in both cases sounds were co-localized with the visual target). Overall, performance accuracy increased by 7.9 per cent in the presence of a moving co-localized auditory cue

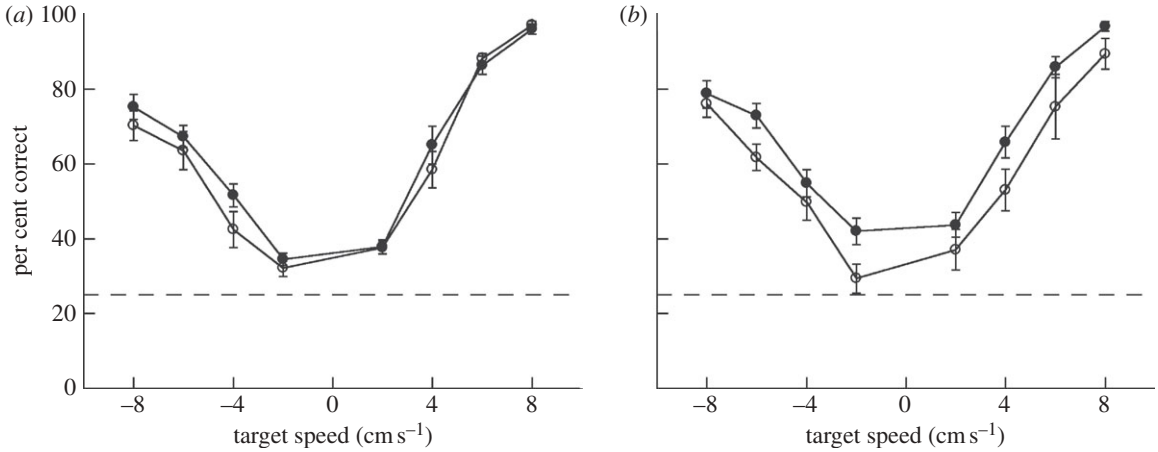

Figure 4. Performance accuracy with (*a*) a non-co-localized auditory cue and (*b*) a spatially co-localized auditory cue, each comparing moving auditory (black circles) with static auditory (white circles) conditions. Error bars are s.e.m. across subjects.

compared with the static co-localized cue. A two-way ANOVA showed significant main effects of target speed ($F_{7,144} = 42.04$, $p < 0.001$) and auditory motion ($F_{1,144} = 15.52$, $p < 0.001$), and no significant interaction between them ($F_{7,144} = 0.35$, $p > 0.9$). Thus, in contrast with non-co-localized auditory cues (where the improvement conferred by congruent motion was small and non-significant), with spatially co-localized auditory cues there was a significant improvement in visual performance.

We again analysed reaction times in correct trials and found that response times decreased from 930 ms in the visual-only condition to 827 ms in auditory-static trials, and decreased further to 775 ms in auditory-moving trials. Neither the difference from visual-only to auditory-static ($F_{1,61} = 3.22$, $p = 0.07$) nor from auditory-static to auditory-moving ($F_{1,63} = 2.33$, $p = 0.13$) reached statistical significance in our sample, although the difference between visual-only and auditory-moving was significant ($F_{1,61} = 20.10$, $p < 0.001$). Therefore, the accuracy differences owing to auditory motion observed in the main analyses cannot be attributed to speed–accuracy trade-offs.

We replicated the co-localized auditory stimulus experiment with a spectrally richer auditory stimulus (broadband noise filtered between 200 Hz and 12 kHz), in which auditory localization information was conveyed via both interaural level differences and interaural time differences. Although the localization information was increased in this condition, resulting in a higher baseline performance, there was still a remarkable improvement in performance for a congruently moving auditory cue compared with a static cue (electronic supplementary material). Thus, whereas better auditory stimulus localization may result in a global effect of cross-modal facilitation, our initial findings indicate that even a relatively coarse auditory motion cue is enough to provide a significant extra benefit to the detection of moving objects during observer motion.

(iii) *Auditory localization*

Since the baseline and experimental conditions in these experiments both contained an auditory stimulus presented from the same location, it is unlikely that the cross-modal benefit reported was due to a spatial-cueing effect of the sound. Nonetheless, to ensure that the increased performance in the auditory-moving trials was not due to increased localization information provided by the moving auditory cues, we constructed an auditory localization control test. Three subjects who participated in the main experiments were presented with an approaching, receding or static auditory cue (identical to those used in the co-localized auditory cue experiment) localized to one of nine locations in front of the observer, evenly spaced in 2.5° increments and with no elevation. After the sound was played, nine vertical bars matching the possible sound source locations were shown on the screen, and observers were asked to report which one corresponded to the sound origin. We measured the distribution of errors for each sound condition (electronic supplementary material, figure S2*a*). The mean absolute errors were 3.04, 3.06 and 3.04° for receding, static and approaching auditory stimuli, respectively. A Levene test of variance showed that there were no significant differences in the distributions of errors for the three cue types for these subjects ($F_{2,1733} = 0.06$, $p = 0.94$). The improved performance in the auditory-moving condition cannot, therefore, be attributed to an improved ability to localize the sound source in these trials. A similar pattern arose with the auditory stimuli containing richer localization cues (electronic supplementary material, data and figure S2*b*).

(iv) *Correlation with visual performance*

The strength of multisensory integration has been found to vary as a function of the accuracy within each modality [45,46]. We were interested in determining whether this auditory-based enhancement in visual motion was more effective in observers that performed poorly on the visual task. To test this, we performed a one-tailed Pearson correlation test for a negative correlation between the gain owing to the moving auditory cue (auditory-moving relative to auditory-static performance) and performance on the visual-only task. Note that correlations were made relative to performance on the visual-only condition, so that regression to the mean would not artificially contribute to a correlation (e.g. a noisy auditory-static data point might cause a noisy measure

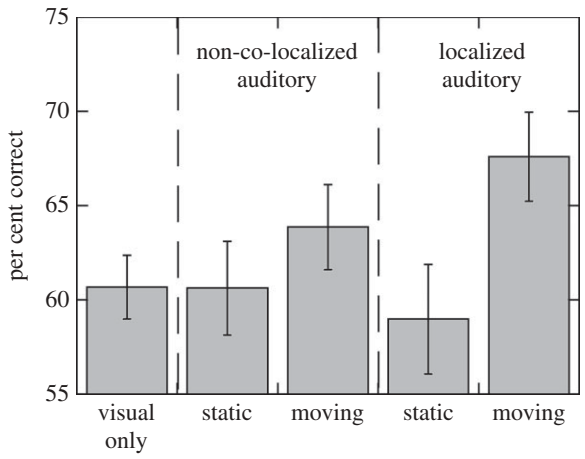

Figure 5. Interparticipant mean motion detection performance across conditions (pooled across visual target speeds) for the 3 cm s^{-1} observer speed condition. Error bars are s.e.m. across subjects.

of auditory improvement, but this would not be correlated with variations in visual-only performance). In the non-co-localized condition, where cross-modal benefit was very modest and not statistically significant, there was no significant relationship between baseline visual performance and cross-modal gain with moving auditory cues ($R^2 = -0.09$, $p = 0.2$, with R^2 sign assigned based on the r-value, and indicating a negative correlation). In contrast, with co-localized auditory cues, the correlation was considerably stronger and statistically significant ($R^2 = -0.37$, $p = 0.04$). The significant negative correlation shows that subjects who performed worse on the visual task benefited more from the auditory cue.

4. DISCUSSION

This study addressed cross-modal enhancement in the detection of visual object motion during simulated observer motion. Participants were asked to make a visual discrimination to identify a moving target sphere amid a dynamic scene simulating an observer translating forward. We first showed that the pattern of visual search results was independent of observer speed, indicating that subjects did not resort to performing the task on the basis of object motion relative to the observer. This result is consistent with the hypothesis that scene-relative object motion during simulated forward self-motion is detected by flow parsing [7–9], in which observer self-motion is estimated and subtracted from the flow field.

Yet the main finding to emerge from the present study is that the presence of a moving auditory cue facilitates parsing out relative object motion from optic flow. Figure 5 summarizes performance across five auditory conditions (visual-only, static and moving non-co-localized auditory cues, and static and moving co-localized auditory cues). The cross-modal improvement was not due to the mere presence of a sound, given that accessory static sounds did not result in any advantage, as compared with visual-only displays (static, non-co-localized auditory condition: $t_{26} = 0.18$, $p = 0.85$; static, localized auditory condition: $t_{26} = 0.67$, $p = 0.5$). Additionally, the spatial localization provided by the auditory cue did not directly improve subject performance: an ANOVA combining the non-co-localized and co-localized conditions showed a significant effect of auditory motion ($F_{1,288} = 17.8$, $p < 0.001$), and a significant interaction between auditory motion and co-localization ($F_{1,288} = 3.7$, $p = 0.05$), but no effect of co-localization alone ($F_{1,288} = 0.55$, $p > 0.4$). Thus, in our task, simply adding a temporally synchronous, static auditory stimulus did not improve subject performance by either alerting to the stimulus onset (e.g. as in [36]), or by directing the observer's attention to the region of the visual stimulus containing the target object.

However, for moving auditory cues, spatial coincidence between sound and visual object proved critical, given that congruently moving sounds significantly enhanced object motion detection only if spatially co-localized. Interestingly, the cross-modal gain seen for co-localized, moving auditory cues was negatively correlated with individual performance levels, such that participants who performed worse visually benefited more from auditory motion. This trend suggests the possibility that auditory cues may be especially useful to observers with weak visual abilities, and thus could be useful in the rehabilitation of visual deficits. This finding agrees well with previous indications that visuo-spatial deficits can be ameliorated by using co-localized accessory acoustic cues [13,47]. It also supports the idea that the gain of multisensory integration depends on prior precision levels in unisensory performance [45,46].

Our results therefore suggest that visual–auditory motion integration is more effective when both cues are presented in spatially commensurate locations within the stimulus, as has been suggested as a condition for visual–auditory motion binding [14,18]. Spatially dependent cross-modal enhancement has frequently been reported in the literature [48,49], often linked to the spatial rule of cross-modal integration derived from single-cell studies in the superior colliculus of several animal species (e.g. [50]). Yet some important exceptions to this rule have been reported recently (e.g. [51]). Indeed, the strong effect of auditory co-localization in our data is interesting, given recent reports of cross-modal improvement in visual search tasks that were obtained with spatially non-informative sounds [36]. This difference between results is, however, difficult to interpret at present, given that these previous studies did not include an auditory co-localized condition to compare with. An interesting speculation, however, is that in contrast with previous studies of cross-modal enhancement in visual search, the participants' task in our study was strongly spatial, and thus more likely to benefit from accurate information about spatial relations.

A potential mechanism underlying this spatial selectivity is that the co-localized auditory cue reduced the search space by directing the observer to the approximate location of the visual target. This could help reduce effective set size, and thus perceptual load, allowing audio-visual integration to be more effective. This explanation is indirectly supported by previous findings indicating that cross-modal integration under high-perceptual-load conditions is mediated by a serial, attentive process [38,39,52], and therefore should be more effective in conditions where there are fewer possible auditory–visual associations. Audio-visual coincidence

selection can be enabled in a variety of ways, such as using sparse visual displays (as in many multi-sensory enhancement experiments), or by the saliency and temporal informativeness of the accessory acoustic cue [36]. We hypothesize that the co-localized cues enable efficient audio-visual motion integration since they constrain the search space so that audio-visual motion integration becomes more effective.

The results presented here suggest that parsing object motion from the perceived optic flow induced by observer self-motion can be enhanced by the presentation of a spatially co-localized auditory cue of congruent motion. The use of auditory information in flow parsing suggests that flow parsing can be seen as a multisensory process, or at least it is able to operate on multisensory motion representations. A recent magneto-encephalography study of dynamic connectivity among cortical areas involved in the visual-only and auditory-motion versions of this task [53] found that the middle prefrontal cortex (MPFC) strongly and selectively modulates the middle temporal area (MT+) in the visual-only condition, while in the auditory–visual condition MPFC provides feedback to the superior temporal polysensory area (STP), to which both the auditory cortex and MT+ are functionally connected. These results suggest that in these two tasks the prefrontal cortex allocates attention to the 'target' as whole, and that the target's representation shifts from MT+ for a moving visual object when no auditory information was presented, to STP for a moving visual–auditory object. Taken together with the results we have presented here, we suggest that flow parsing, previously thought of as a purely visual process, may use multisensory object representations when detecting a moving object during observer self-motion, demonstrating that the integration of motion information across sensory modalities contributes to ecological perception that occurs at early stages of processing.

All procedures were approved by the Boston University Institutional Review Board for research with human subjects, and informed consent was obtained from each participant.

L.M.V. and F.J.C. were supported by NIH grant RO1NS064100 to L.M.V. S.S.F. was supported by grants from the Ministerio de Ciencia e Innovación (PSI2010-15426 and CSD2007-00012), by the Comissionat per a Universitats i Recerca del DIUE (SRG2009-092) and by the European Research Council (StG-2010 263145).

We thank Gerald Kidd and Chris Mason for their helpful suggestions and for generously making available to us the resources of the Sound Field Laboratory at Sargent College, Boston University, supported by grant P30 DC04663. We also thank Franco Rupcich, Benvy Caldwell and Megan Menard for helping with psychophysical data collection and subject recruitment, and Leonardo Sassi for developing and implementing a preliminary version of the object motion task.

REFERENCES

1 Dick, M., Ullman, S. & Sagi, D. 1987 Parallel and serial processes in motion detection. *Science* **237**, 400–402. (doi:10.1126/science.3603025)
2 McLeod, P., Driver, J. & Crisp, J. 1988 Visual search for a conjunction of movement and form is parallel. *Nature* **320**, 154–155. (doi:10.1038/332154a0)
3 Nakayama, K. & Silverman, G. H. 1986 Serial and parallel processing of visual feature conjunctions. *Nature* **320**, 264–265. (doi:10.1038/320264a0)
4 Duncan, J. 1995 Target and nontarget grouping in visual search. *Percept. Psychophys.* **57**, 117–120.
5 Gibson, J. J. 1950 *The perception of the visual world.* Boston, MA: Houghton Mifflin.
6 Vaina, L. M. 1998 Complex motion perception and its deficits. *Curr. Opin. Neurobiol.* **8**, 494–502. (doi:10.1016/S0959-4388(98)80037-8)
7 Rushton, S. K. & Warren, P. A. 2005 Moving observers, relative retinal motion and the detection of object movement. *Curr. Biol.* **15**, R542–R543. (doi:10.1016/j.cub.2005.07.020)
8 Rushton, S. K., Bradshaw, M. F. & Warren, P. A. 2007 The pop out of scene-relative object movement against retinal motion due to self-movement. *Cognition* **105**, 237–245. (doi:10.1016/j.cognition.2006.09.004)
9 Warren, P. A. & Rushton, S. K. 2007 Perception of object trajectory: parsing retinal motion into self and object movement components. *J. Vis.* **7**, 2. (doi:10.1167/7.11.2)
10 Ross, L. A., Saint-Amour, D., Leavitt, V. M., Javitt, D. C. & Foxe, J. J. 2007 Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments. *Cereb. Cortex* **17**, 1147–1153. (doi:10.1093/cercor/bhl024)
11 Molholm, S., Ritter, W., Murray, M. M., Javitt, D. C., Schroeder, C. E. & Foxe, J. J. 2002 Multisensory auditory–visual interactions during early sensory processing in humans: a high-density electrical mapping study. *Brain Res. Cogn. Brain Res.* **14**, 115–128. (doi:10.1016/S0926-6410(02)00066-6)
12 Vroomen, J. & de Gelder, B. 2000 Sound enhances visual perception: cross-modal effects of auditory organization on vision. *J. Exp. Psychol. Hum. Percept. Perform.* **26**, 1583–1590. (doi:10.1037/0096-1523.26.5.1583)
13 Ladavas, E. 2008 Multisensory-based approach to the recovery of unisensory deficit. *Ann. NY Acad. Sci.* **1124**, 98–110. (doi:10.1196/annals.1440.008)
14 Soto-Faraco, S., Lyons, J., Gazzaniga, M., Spence, C. & Kingstone, A. 2002 The ventriloquist in motion: illusory capture of dynamic information across sensory modalities. *Brain Res. Cogn. Brain Res.* **14**, 139–146. (doi:10.1016/S0926-6410(02)00068-X)
15 Soto-Faraco, S., Spence, C. & Kingstone, A. 2004 Cross-modal dynamic capture: congruency effects in the perception of motion across sensory modalities. *J. Exp. Psychol. Hum. Percept. Perform.* **30**, 330–345. (doi:10.1037/0096-1523.30.2.330)
16 Soto-Faraco, S., Spence, C., Lloyd, D. & Kingstone, A. 2004 Moving multisensory research along: motion perception across sensory modalities. *Curr. Direct. Psychol. Sci.* **13**, 29–32. (doi:10.1111/j.0963-7214.2004.01301008.x)
17 Soto-Faraco, S., Kingstone, A. & Spence, C. 2003 Multisensory contributions to the perception of motion. *Neuropsychologia* **41**, 1847–1862. (doi:10.1016/S0028-3932(03)00185-4)
18 Meyer, G. F., Wuerger, S. M., Röhrbein, F. & Zetzsche, C. 2005 Low-level integration of auditory and visual motion signals requires spatial co-localisation. *Exp. Brain Res.* **166**, 538–547. (doi:10.1007/s00221-005-2394-7)
19 Alais, D. & Burr, D. 2004 No direction-specific bimodal facilitation for audiovisual motion detection. *Brain Res. Cogn. Brain Res.* **19**, 185–194. (doi:10.1016/j.cogbrainres.2003.11.011)
20 Sanabria, D., Lupianez, J. & Spence, C. 2007 Auditory motion affects visual motion perception in a speeded discrimination task. *Exp. Brain Res.* **178**, 415–421. (doi:10.1007/s00221-007-0919-y)

21 Freeman, E. & Driver, J. 2008 Direction of visual apparent motion driven solely by timing of a static sound. *Curr. Biol.* **18**, 1262–1266. (doi:10.1016/j.cub.2008.07.066)
22 Lopez-Moliner, J. & Soto-Faraco, S. 2007 Vision affects how fast we hear sounds move. *J. Vis.* **7**, 1–7. (doi:10.1167/7.12.6)
23 Cappe, C., Thut, G., Romei, V. & Murray, M. M. 2009 Selective integration of auditory–visual looming cues by humans. *Neuropsychologia* **47**, 1045–1052. (doi:10.1016/j.neuropsychologia.2008.11.003)
24 Kitagawa, N. & Ichihara, S. 2002 Hearing visual motion in depth. *Nature* **416**, 172–174. (doi:10.1038/416172a)
25 Valjamae, A. & Soto-Faraco, S. 2008 Filling in visual motion with sounds. *Acta Psychol. (Amst.)* **129**, 249–254. (doi:10.1016/j.actpsy.2008.08.004)
26 Alink, A., Singer, W. & Muckli, L. 2008 Capture of auditory motion by vision is represented by an activation shift from auditory to visual motion cortex. *J. Neurosci.* **28**, 2690–2697. (doi:10.1523/JNEUROSCI.2980-07.2008)
27 Baumann, O. & Greenlee, M. W. 2007 Neural correlates of coherent audiovisual motion perception. *Cereb. Cortex* **17**, 1433–1443.
28 Wolfe, J. M. 1998 What can 1 million trials tell us about visual search? *Am. Psychol. Soc.* **9**, 33–39.
29 Verghese, P. & Pelli, D. G. 1992 The information capacity of visual attention. *Vis. Res.* **32**, 983–995. (doi:10.1016/0042-6989(92)90040-P)
30 Duncan, J. & Humphreys, G. 1989 Visual search and stimulus similarity. *Psychol. Rev.* **96**, 433–458. (doi:10.1037/0033-295X.96.3.433)
31 Bravo, M. & Blake, R. 1990 Preattentive vision and perceptual groups. *Perception* **19**, 515–522. (doi:10.1068/p190515)
32 Cavanagh, P., Arguin, M. & Treisman, A. 1990 Effect of surface medium on visual search for orientation and size features. *J. Exp. Psychol. Hum. Percept. Perform.* **16**, 479–491. (doi:10.1037/0096-1523.16.3.479)
33 Royden, C. S., Wolfe, J. M. & Klempen, N. 2001 Visual search asymmetries in motion and optic flow fields. *Percept. Psychophys.* **63**, 436–444. (doi:10.3758/BF03194410)
34 Rushton, S. K. & Bradshaw, M. F. 2000 Visual search and motion—is it all relative? *Spat. Vis.* **14**, 85–86.
35 Rushton, S. K., Bradshaw, M. F. & Warren, P. A. 2006 The pop out of scene-relative object movement against retinal motion due to self-movement. *Cognition* **105**, 237–245.
36 Van der Burg, E., Olivers, C. N., Bronkhorst, A. W. & Theeuwes, J. 2008 Pip and pop: nonspatial auditory signals improve spatial visual search. *J. Exp. Psychol. Hum. Percept. Perform.* **34**, 1053–1065. (doi:10.1037/0096-1523.34.5.1053)
37 Iordanescu, L., Guzman-Martinez, E., Grabowecky, M. & Suzuki, S. 2008 Characteristic sounds facilitate visual search. *Psychon. Bull. Rev.* **15**, 548–554. (doi:10.3758/PBR.15.3.548)
38 Fujisaki, W., Koene, A., Arnold, D., Johnston, A. & Nishida, S. 2006 Visual search for a target changing in synchrony with an auditory signal. *Proc. R. Soc. B* **273**, 865–874. (doi:10.1098/rspb.2005.3327)
39 Alsius, A., Navarra, J., Campbell, R. & Soto-Faraco, S. 2005 Audiovisual integration of speech falters under high attention demands. *Curr. Biol.* **15**, 839–843. (doi:10.1016/j.cub.2005.03.046)
40 Brainard, D. H. 1997 The psychophysics toolbox. *Spat. Vis.* **10**, 433–436. (doi:10.1163/156856897X00357)
41 Pelli, D. G. 1997 The VideoToolbox software for visual psychophysics: transforming numbers into movies. *Spat. Vis.* **10**, 437–442. (doi:10.1163/156856897X00366)
42 McLeod, P., Driver, J., Dienes, Z. & Crisp, J. 1991 Filtering by movement in visual search. *J. Exp. Psychol. Hum. Percept. Perform.* **17**, 55–64. (doi:10.1037/0096-1523.17.1.55)
43 Barker, L. E., Luman, E. T., McCauley, M. M. & Chu, S. Y. 2002 Assessing equivalence: an alternative to the use of difference tests for measuring disparities in vaccination coverage. *Am. J. Epidemiol.* **156**, 1056–1061. (doi:10.1093/aje/kwf149)
44 Barker, L., Rolka, H., Rolka, D. & Brown, C. 2001 Equivalence testing for binomial random variables: which test to use? *Am. Stat.* **55**, 279–287. (doi:10.1198/000313001753272213)
45 Alais, D. & Burr, D. 2004 The ventriloquist effect results from near-optimal bimodal integration. *Curr. Biol.* **14**, 257–262.
46 Ernst, M. O. & Banks, M. S. 2002 Humans integrate visual and haptic information in a statistically optimal fashion. *Nature* **415**, 429–433. (doi:10.1038/415429a)
47 Frassinetti, F., Bolognini, N., Bottari, D., Bonora, A. & Ladavas, E. 2005 Audiovisual integration in patients with visual deficit. *J. Cogn. Neurosci.* **17**, 1442–1452. (doi:10.1162/0898929054985446)
48 Frassinetti, F., Bolognini, N. & Ladavas, E. 2002 Enhancement of visual perception by crossmodal visuo-auditory interaction. *Exp. Brain Res. Exp. Hirnforsch. Exp. Cereb.* **147**, 332–343. (doi:10.1007/s00221-002-1262-y)
49 Bolognini, N., Frassinetti, F., Serino, A. & Ladavas, E. 2005 'Acoustical vision' of below threshold stimuli: interaction among spatially converging audiovisual inputs. *Exp. Brain Res.* **160**, 273–282. (doi:10.1007/s00221-004-2005-z)
50 Stein, B. & Meredith, M. 1993 *The merging of the senses.* Cambridge, MA: MIT Press.
51 Murray, M. M., Molholm, S., Michel, C. M., Heslenfeld, D. J., Ritter, W., Javitt, D. C., Schroeder, C. E. & Foxe, J. J. 2005 Grabbing your ear: rapid auditory-somatosensory multisensory interactions in low-level sensory cortices are not constrained by stimulus alignment. *Cereb. Cortex* **15**, 963–974. (doi:10.1093/cercor/bhh197)
52 Talsma, D., Senkowski, D., Soto-Faraco, S. & Woldorff, M. G. 2010 The multifaceted interplay between attention and multisensory integration. *Trends Cogn. Sci.* **14**, 400–410. (doi:10.1016/j.tics.2010.06.008)
53 Vaina, L., Calabro, F., Lin, F. & Hamalainen, M. 2010 Long-range coupling of prefrontal cortex and visual (MT) or polysensory (STP) cortical areas in motion perception. *17th International Conference on Biomagnetism Advances in Biomagnetism, Biomag 2010, IFMBE Proceedings*, vol. 28, pp. 197–201. Berlin, Germany: Springer Verlag.