I can see you better if I can hear you coming: Action-consistent sounds facilitate the visual detection of human gait

James Philip Thomas
Department of Psychology, Rutgers University,
Newark, NJ, USA

Maggie Shiffrar
Department of Psychology, Rutgers University,
Newark, NJ, USA

Observers are remarkably sensitive to point-light displays of human movement. The Superior Temporal Sulcus (STS) and premotor cortex are implicated in the visual perception of point-light human actions and the integration of perceptual signals across modalities. These neurophysiological findings suggest that auditory information might impact visual sensitivity to point-light displays of human actions. Previous research has demonstrated that coincident, action-consistent sounds enhance visual sensitivity to the presence of coherent point-light displays of human movement. Here we ask whether visual detection sensitivity is modulated specifically by the meaningfulness of sounds that are coincident with observed point-light actions. To test this hypothesis, two psychophysical studies were conducted wherein participants detected the presence of coherent point-light walkers in a mask under unimodal or audiovisual conditions. Participants in audiovisual conditions heard either tones or actual footfalls coincident with the seen walkers' footsteps. Detection sensitivity increased when visual displays were paired with veridical auditory cues (footfalls), but not when paired with simple tones. The footfall advantage disappeared when the visual stimuli were inverted. These results suggest that the visual system makes use of auditory cues during the visual analysis of human action when there is a meaningful match between the auditory and visual cues.

Keywords: biological motion, point-light walkers, multisensory integration, action perception

Citation: Thomas, J. P., & Shiffrar, M. (2010). I can see you better if I can hear you coming: Action-consistent sounds facilitate the visual detection of human gait. *Journal of Vision, 10*(12):14, 1–11, http://www.journalofvision.org/content/10/12/14, doi:10.1167/10.12.14.

Introduction

Several decades ago, Johansson (1973) demonstrated that point-light displays, in which visually complex human movement is reduced to the motion of a few points of light, are sufficient to generate compelling visual percepts of human action. Since Johansson pioneered the experimental use of point-light stimuli, a sizable body of research has emerged aimed towards identifying the types of information that observers can detect in these simplified displays (see Blake & Shiffrar, 2007 for review). Consistent with the hypothesis that the typical human visual system is tuned for the detection and analysis of human movement, visual sensitivity to point-light human movement is typically more robust than sensitivity to similarly constructed point-light displays of moving objects (Kaiser, Delmolino, Tanaka, & Shiffrar, 2010) and animals (Pinto & Shiffrar, 2009).

Paramount among brain regions implicated in the perception of human movement is the posterior region of the Superior Temporal Sulcus (STSp) (e.g., Bonda, Petrides, Ostry, & Evans, 1996; Grossman & Blake, 2002; Puce & Perrett, 2003). STSp responds more during the visual perception of coherent human motion than during the perception of scrambled human motion (Grossman & Blake, 2001), coherent object motion (Pelphrey et al., 2003) and coherent "creature" motion (Pyles, Garcia, Hoffman, & Grossman, 2007). A causal relationship between visual sensitivity to point-light human motion and STSp integrity is suggested by work with brain lesioned individuals (Saygin, 2007).

Neural responsivity in the STS is not limited to visual processes. Unimodal auditory and somatosensory cortices send output to the STSp (Seltzer & Pandya, 1978) and STS responds to visual, auditory and tactile stimulation (Barraclough, Xiao, Baker, Oram, & Perrett, 2005; Bruce, Desimone, & Gross, 1981). Single-cell recordings have identified multisensory neurons in STS that respond to action cues in the visual and auditory modalities (Barraclough et al., 2005). A proportion of these cells exhibit a superadditive (i.e., greater than the summed responses to unimodal stimulation) spike response to paired visual and auditory cues when such cues relate to the same perceptual event (e.g., a hand tearing a sheet of paper).

fMRI measures have also implicated the STSp in the processing of the sounds of human footsteps in the absence

doi: 10.1167/10.12.14 Received May 6, 2010; published October 13, 2010 ISSN 1534-7362 © ARVO

of visual stimulation. Compared to no auditory stimulation and meaningless auditory noise, attending to the sounds of people walking increases activation STSp (Bidet-Caulet, Voisin, Bertrand, & Fonlupt, 2005; Saarela & Hari, 2008). Interestingly, the area within the pSTS that is recruited during the auditory perception of human walking overlaps with areas recruited during the visual perception of human walking (Bidet-Caulet et al., 2005), suggesting that the STS participates in the multimodal analysis of human actions. Some evidence suggests that this multimodal analysis depends upon on the cross-modal congruency of higher-order stimulus features. Calvert, Campbell, and Brammer (2000) located a cluster of voxels in left STS that exhibited superadditive increases in BOLD signal response to meaningfully and temporally congruent multisensory speech stimuli (i.e., heard words matched the seen speaker's facial motion), and subadditive responses to incongruent audiovisual speech pairings (i.e., heard words differed from those uttered by the seen speaker). Meaningful audio-visual congruence does not appear to impact STS activity during the perception of moving animals or tools (e.g., Beauchamp, Lee, Argall, & Martin, 2004; Hein et al., 2007).

The STS is clearly not alone in its multisensory properties. For example, the perirhinal cortex plays a key role in the crossmodal integration of meaningfully related object cues (Taylor, Moss, Stamatakis, & Tyler, 2006; Taylor, Stamatakis, & Tyler, 2009). In the analysis of human movement, cross-modal integration has also been documented in the premotor cortex, an area required for the perception of point-light displays of human motion by human observers (Saygin, Wilson, Hagler, Bates, & Sereno, 2004). In the monkey, "mirror neurons" in ventral premotor cortex fire when an action is performed, heard, and seen (Keysers et al., 2003; Kohler et al., 2002). In the human, fMRI data support the integrated analysis of heard and performed actions in the premotor, parietal, and temporal cortices (Gazzola, Aziz-Zadeh, & Keysers, 2006; Lahav, Saltzman, & Schlaug, 2007).

Motivated by the aforementioned findings, psychophysical researchers have recently begun to investigate audiovisual interactions in biological motion perception. Researchers have found, for example, the time needed to detect the presence of a coherent point-light walker in a mask decreases when auditory motion travels in the same direction as the walker and increases when auditory and point-light walker motions travel in opposite directions (Brooks, van der Zwan, Billard, Petreska, & Blanke, 2007). Such direction specific enhancement is not found in detection of coherent motion in random dot kinematograms or Gabor patches (Alais & Burr, 2004).

Studies of synchrony detection also inform our understanding of audio-visual integration during the perception of human motion. For example, participants can better detect when the seen footsteps of an unmasked point-light walker have the same temporal frequency as a sequence of auditory tones when the walker is coherent than when it is scrambled or inverted (Saygin, Driver, & de Sa, 2008). This benefit is lost when auditory sequences are 50% out of phase with the walker's footsteps (Saygin et al., 2008). These results not only suggest that the visual gestalt of the point-light walker facilitates audiovisual comparisons, but that this facilitation reflects temporal synchrony; that is, whether temporal evidence suggests that sounds could conceivably have been produced by the visual stimulus. Synchrony detection in more complex audio-visual actions is experience dependent. For example, expert drummers make finer grained assessments of audio-visual synchrony than novice drummers during the perception of point-light drumming actions and sounds (Petrini et al., 2009; Petrini, Russell, & Pollick, 2009). When point-light drumming stimuli are rotated away from their canonical upright orientation, audio-visual synchrony detection by novice drummers suffers while that of experts remains intact (Petrini, Holt, & Pollick, 2010).

Thus, previous researchers have determined that directionality, temporality, and experience influence the integration of auditory and visual cues in point-light displays. The question that motivates the current research is whether the *meaningful relationship* between visual and auditory information influences visual sensitivity to human movement.

According to the "unity assumption", observers are more likely to perceive inputs from two sensory streams as referencing the same event when those inputs are highly consistent along some dimension(s), such as ecological validity, perceived causality, or semantic content (e.g., Spence, 2007; Welch & Warren, 1980). Recent evidence suggests that meaning or semantic relatedness between auditory and visual information influences the perception of point-light displays. For example, gender-ambiguous point-light walkers appear more female when paired with the sounds of female footsteps (van der Zwan et al., 2009). Visual sensitivity to coherent human motion in point-light displays is also enhanced by auditory cues that are meaningful and synchronous. For example, the presentation of synchronous tap sounds increases visual sensitivity to coherent point-light displays of tapping feet presented within masks (Arrighi, Marini, & Burr, 2009). Because the audio cues in this study were always synchronous with *and* meaningfully related to the visual cues, it remains to be determined whether the meaningful relationship, per se, between auditory and visual information enhances visual sensitivity to human movement (Figure 1). The studies described below addressed this issue.

In these studies, auditory and visual events were always temporally coincident. Whether they were meaningfully related varied across conditions. Because the perception of human action is mediated by the STS and premotor cortex, and because these regions have been implicated in the integration of inputs from multiple modalities at an action-based level of representation, we hypothesized that the addition of coincident, action-appropriate auditory cues (footstep sounds) would enhance visual detection of a

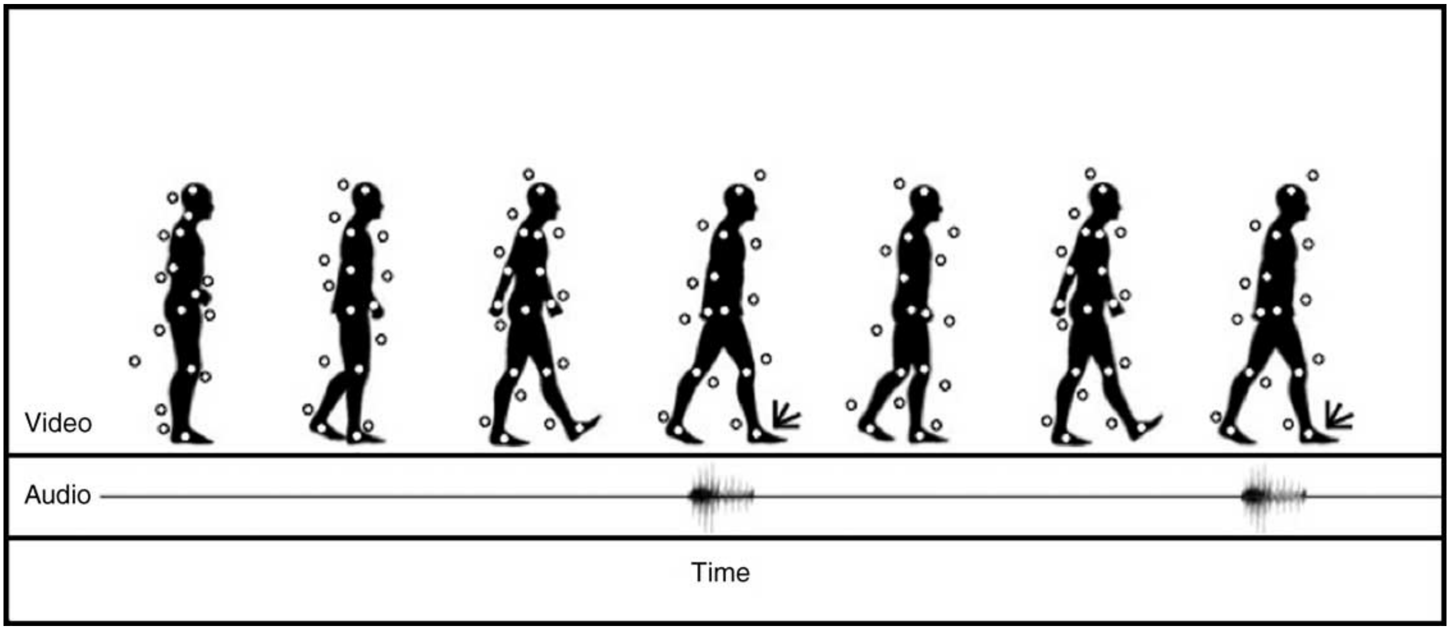

Figure 1. A diagram depicting the time course of a single gait cycle of a coherent and upright point-light walker in a mask. Across conditions, the auditory signal was either absent, synchronous but meaningfully unrelated or synchronous and meaningfully related to the repeated footfalls of the walker. The form of the human body was added here for clarity and did not appear in the experimental trials. Not all point-lights are shown.

corresponding point-light walker more than the addition of temporally coincident but meaningfully unrelated (tone sounds) auditory cues.

Experiment 1

Method

Participants

Sixty undergraduate students (mean age 19.8 years) from the Rutgers University-Newark campus participated in this study for partial credit towards a course requirement. All participants reported normal or corrected-to-normal visual and auditory acuity. Participants were naïve as to the hypothesis under investigation. This and the subsequent study were approved by the Rutgers University Institutional Review Board. All participants provided written informed consent.

Apparatus

Visual stimuli were constructed from motion capture data obtained from a ReActor motion capture system (Ascension Technology Corporation). Point-light movies of human walking motion were rendered with Kaydara Motion Builder 6.0 software. Footstep sounds were recorded with a Zoom H4 Handy Recorder and digitally processed on a MacBook computer with Cubase LE 4 Digital Audio Workstation software (Steinberg). Video and audio stimulus components were combined and the final movies rendered using Macintosh iMovie HD 6.0.3 video editing software.

Visual stimuli were displayed on a 22 inch ViewSonic monitor with a refresh rate set to 60 Hz and a 1280 × 1020 pixel resolution. Monitor output was controlled by a custom built AMD computer with an AMD Athlon 64 × 2 dual core processor. Auditory stimuli were presented through Bose Companion 2 speakers flanking the monitor at angles directly facing the participant. The experiment was programmed and controlled using E-Prime version 2.0 software.

Stimuli

During motion capture recording, two individuals walked with a neutral gait along a linear path within the ReActor system. Each person wore 13 sensors attached to the head, feet, knees, hips, shoulders, elbows, and wrists. As each person walked, the sensors provided kinematic information about the individual's limb, trunk and head movements. This motion was subsequently converted into point-light movies.

From each walking epoch we created two types of point-light movies: movies depicting a spatiotemporally coherent walker embedded within a scrambled walker mask ("person present"), and movies depicting a spatiotemporally scrambled walker within a scrambled walker mask ("person absent"). Following a classic psychophysical detection procedure (Bertenthal & Pinto, 1994), point-light masks were constructed from spatially scrambled duplicates of the original 13 points comprising each coherent walker. Therefore, each "person present" movie contained 13 points defining the coherent point-light walker (PLW) and 13 points defining the mask. Each "person absent" movie contained 13 points defining the scrambled walker and 13 points defining the mask (essentially, "person absent" movies depicted two spatiotemporally scrambled point-light walkers).

To construct each point-light mask, the first frame from a point-light walker movie was duplicated and the locations of each of the duplicate points was positioned within a one to five-point radius of one of the points defining the original walker. For example, for a particular "person present" movie, the head dot of the coherent PLW was duplicated and could have been relocated to a starting position near to the point defining the left foot of the coherent walker. For another "person present" movie, the head dot may have been duplicated and relocated to a starting position near the right elbow dot, and so on. For "person absent" movies, all of the points were spatially scrambled in this manner, resulting in a spatiotemporally scrambled walker embedded in a scrambled walker mask. Thus, each pair of "person present" and "person absent" movies was identical in terms of the local motion profiles but differed in regards to the presence or absence of a spatiotemporally coherent point-light walker. Since, for each movie, the mask was constructed from the original 13 points defining the walker, the points defining the mask had the same color, luminance, densities, and velocity profiles as those defining the walker itself. This masking procedure renders walker detection difficult, as the only difference between the points defining the walker and those defining the mask is the global configuration of the points defining the walker.

Once scrambling and masking were completed, each "person present" and "person absent" movie was rendered from four observer-centered orientations, such that the visual stimulus appeared to walk (or move, in the "person absent" case) towards the observer, away from the observer, from the left to the right, or vice-versa. As the point-light person (coherent or scrambled) walked, the mask moved in the same global direction; as such, the mask was tightly coupled to the target stimulus while it translated.

Point-light stimuli (walkers + masks) subtended a maximum vertical height of approximately 10 degrees of visual angle (DVA), and a maximum horizontal width of approximately 7 DVA. Each individual point subtended a maximum diameter of approximately .32 DVA. All points were white and presented on a black background. Point-light stimuli became larger as they approached observers and smaller as they moved away, with a minimum height of approximately 7 DVA. Movies in which the point-light stimuli translated from left to right or vice-versa ("rightwards" and "leftwards" movies, respectively) subtended a maximum horizontal distance of approximately 27 DVA.

The footstep sounds used in this experiment were recorded while an individual wearing flat dress shoes walked alone down an otherwise empty hallway. Each footstep was then individually extracted from the original audio recording, normalized to equate loudness, and rendered as an individual sound clip. This resulted in 20 unique footstep sounds. Each footstep sound was approximately 70 ms long in duration; the duration of the footsteps was approximated by measuring the extent of the visible waveform in time. Each clip was treated to a 10 ms fade envelope to eliminate audible pops at onset and offset.

The pure tone, 1000 Hz sound used for the control condition stimuli was downloaded from an online digital audio library (http://www.audiosparx.com). A pure tone sound was used in the control condition because pure tones are artificially generated, rarely occur in nature, and thus do not correspond to any sound that can be produced by the human body. Each tone was 100 ms long with a 10 ms fade envelope. Footstep and tone sounds were rendered so as to be approximately equal in subjective loudness. Both sounds were rendered in mono; thus, no apparent auditory motion accompanied the motion of the point-light stimuli.

The original unmasked, coherent point-light walkers were used to place the tones and footstep sounds into the correct temporal location within each movie. Movies were analyzed frame by frame, and footstep or tone sounds were inserted into the frames when each point-light foot initially reached its lowest vertical position; that is, when the foot first made contact with the ground. Sounds ended before each foot left the ground. Once the sounds were in place, the coherent, unmasked PLW was removed and the "person present" and "person absent" movies were inserted and rendered with the accompanying sounds. Thus, in "person present" movies, the footstep or tone sound was heard when the point-light foot hit the ground, and in "person absent" movies the footstep or tone was heard when a randomly located point-light with the same velocity profile as the corresponding point-light foot stopped and reversed its direction of motion. Each movie lasted 3000 ms and contained 4 to 5 discrete and non-overlapping audiovisual footstep events (see Videos 1 to 4 for sample stimuli).

This stimulus construction process yielded three sets of movies, one for each experimental condition (silent, footsteps, and tones). Each movie set was identical in regards to the visual stimulus. Each contained a total of 56 individual movies ((7 scrambled + 7 coherent) × (4 motion directions)). Half of the movies (28) depicted a coherent point-light walker embedded in a scrambled walker mask and the other half of the movies depicted a scrambled walker in a scrambled walker mask. Of these 56 movies, 8 were presented during practice and 48 were presented during the experimental trials.

Procedure

This study utilized a between subjects design. This design was chosen so as to minimize the likelihood that participants became aware of the experimental hypothesis and to avoid possible order effects. Participants were randomly assigned to one of three conditions (silent, footstep, or tone displays). The experiment took place in a quiet 213 cm by 203 cm room. During the instruction phase, participants were informed that they would see a series of short movies that might or might not contain a point-light person walking within a mask of visually similar point-lights. They were instructed to watch each movie and report whether they saw a person in each display by means of a button press. No mention was made of the possible presence of auditory cues during the instruction phase. Participants sat in a chair and placed their heads in a chinrest 90 cm from the display while completing the experiment.

Participants first completed a short block of 8 practice trials. During practice, participants observed 4 "person present" and 4 "person absent" movies. Participants watched each movie in its entirety, after which a response screen appeared and participants provided their answers via a key press. No feedback was provided.

After practice, participants completed two experimental blocks during which the remaining 48 movies were presented. Each of the remaining 48 movies was presented twice; once during block 1 and again during block 2. During each block, participants observed 24 "person present" and 24 "person absent" movies (6 of each rendered from one of the four observer-centered orientations described above). Movie presentation within the experimental blocks was completely randomized. Again, no feedback was provided.

Results

D-Prime was the primary dependent variable of interest. D-Prime, an unbiased measure of sensitivity, was estimated by subtracting the normalized rate of false alarms from the normalized rate of hits (McMillan & Creelman, 1991). A one-way ANOVA revealed a significant effect of audio condition on D-Prime, $F(2,57) = 7.107$, $p = .002$. Subsequent Post hoc *t*-tests (Bonferroni corrected) revealed enhanced visual sensitivity in the footstep condition, such that performance in the footstep condition (M = 1.905) was significantly better than performance in the silent (M = 1.296), $p = .002$ and tone (M = 1.481) conditions, $p = .039$. Detection performance did not significantly differ between the silent and tone conditions, $p = .806$. To assess any potential sensitivity differences as a function of walker direction, two mixed measures ANOVAs were conducted with audio condition as the between-subjects factor and walker direction as the within-subjects factor. A 3 (audio condition) × 4 (direction: towards, away, leftwards, and rightwards) mixed measures ANOVA revealed no main effect of direction, $F(3,171) = 1.32$, $p = .271$ and no audio condition × direction interaction, $F(6,171) = .743$, $p = .615$ For the second analysis, the four walker directions were combined to create two global directions: horizontal (leftwards and rightwards) and depth (towards and away). This analysis also revealed no main effect of direction, $F(1,57) = .139$, $p = .771$, and no condition × direction interaction, $F(2,57) = .773$, $p = .466$.

Because the results of the aforementioned analyses proved significant and supported the *a-priori* hypotheses detailed above, Criterion C was also analyzed. Criterion C describes the strategy observers adopt regarding their willingness to respond "yes" or "no" given the available information, and thus functions as an average measure of response bias (McMillan & Creelman, 1991). A one-way ANOVA on Criterion C revealed no significant differences between the three audio conditions, $F = (2,57) = .945$, $p = .395$. Thus, it is not the case that participants were more likely to report the presence of a person simply because footstep sounds were present.

Discussion

The results from Experiment 1 revealed better detection sensitivity to coherent point-light walkers when they were paired with veridical auditory cues (footsteps) than when they were paired with simple tones. This result suggests that temporally coincident auditory cues are not always sufficient to improve visual sensitivity to actions with which they do not naturally co-occur. Such selective improvement in detection performance is consistent with results from single-cell recording studies of the STS in which meaningfully related auditory cues have been found to modulate the activity of cells responsive to visual actions (Barraclough et al., 2005), as well as with neuroimaging work implicating the STS in the multimodal perception of observed speech (Calvert et al., 2000). Additionally, the current results are consonant with evidence for audiovisual mirror neurons in premotor cortex, a proportion of which exhibit the strongest response when visual and auditory action cues are presented together (Keysers et al., 2003). These results also replicate and extend those of Arrighi et al. (2009), who found that temporally coincident tap dancing sounds facilitate the visual detection of masked point-light tapping feet.

Experiment 2

It is well known that biological motion perception is orientation-specific (e.g. Bertenthal, 1993; Dittrich, 1993; Pavlova & Sokolov, 2000, 2003; Sumi, 1984). Like the visual perception of faces (e.g., Yin, 1969) and static bodies (e.g., Reed, Stone, Bozova, & Tanaka, 2003), biological motion perception suffers from inversion effects, such that point-light movies of human motion are less recognizable when presented upside-down. Behavioral evidence of deficits in perceiving inverted biological motion dovetails with neuroimaging evidence demonstrating diminished STS response to inverted, relative to upright, point-light displays of human motion (e.g. Grossman & Blake, 2001).

Inversion has a strong negative impact on global processing (Bertenthal & Pinto, 1994). If the benefit for footstep sounds observed in Experiment 1 was simply due to detection of local coincidence between heard footsteps and visual footfalls (directional changes of foot dots relative to sounds), then a similar benefit should be observed with inverted displays. Conversely, if the benefit observed for footstep sounds was specific to perceiving a global human form in motion, then one would expect to observe no benefit in detection sensitivity for inverted displays paired with meaningfully related sound cues. Experiment 2 sought to test this hypothesis.

Method

Participants

Forty-five undergraduate students (mean age 21.2 years) from the Rutgers University-Newark campus participated in the study for partial course credit. All participants reported normal or corrected-to-normal vision and hearing, and all were naïve as to the hypothesis under investigation. None had participated in the previous study.

Stimuli and procedure

The three sets of point-light movies used in Experiment 1 were inverted (flipped along the horizontal axis) and re-rendered. Thus, all movies were presented in an upside-down orientation relative to the observer. These stimuli were otherwise identical to those used in the previous experiments. The experimental procedure was identical to that followed in Experiment 1. During instruction, participants were informed that the point-light movies would be presented in an upside-down orientation, as if walking on an invisible ceiling.

Results and discussion

Performance in this experiment was generally poor, which is typical for experiments utilizing inverted displays. A one-way ANOVA on D-Prime revealed no difference in performance accuracy across audio conditions, $F(2,42)$ = .047, p = .954. As well, no differences in Criterion C were observed, $F(2,42)$ = .817, P = .449. Nineteen of the 45 participants did not perform the task above chance (7 in the silent, 5 in the tone, and 7 in the footstep condition). A Chi-Square test of independence revealed no significant difference in the proportion of chance performers as a function of audio condition, X^2 (2, N = 45) = .616, p = .735. Thus, the sound specific enhancement of visual sensitivity to coherent point-light gaits does not generalize to the perception of inverted point-light walkers. Since inversion disrupts global percepts of human motion (e.g., Bertenthal & Pinto, 1994), the current results suggest that the results of Experiment 1 cannot be attributed to coincidence of heard footsteps and the seen motions of individual points depicting footfalls. Importantly, the results of this study indicate that auditory footsteps do not enhance visual sensitivity to all point-light stimuli (Figure 2).

General discussion

In the real world, we typically both hear and see the people around us. Human actions produce characteristic sounds, and these sounds can inform us about the actions being performed. Two psychophysical studies were conducted to test whether visual sensitivity to point-light depictions of human gait reflects the action specific co-occurrence of visual and auditory cues typically produced by walking people. Visual sensitivity to coherent human gait was greatest in the presence of temporally coincident and action-consistent sounds (footsteps). Visual sensitivity to human gait with coincident sounds that were not action-consistent (tones) was significantly lower and did not significantly differ from visual sensitivity to gaits presented without sound. There was no evidence of a criterion shift as a function of audio condition. Thus, the current results cannot be attributed to increases in the likelihood of reporting the presence of human motion whenever footsteps were heard. These results are difficult to interpret in terms of low-level detection of coincident multimodal stimuli. Coherent point-light walkers differed from scrambled point-light walkers and from point-light masks only in the spatial arrangement of the points. Because auditory footfalls and visual footfalls were coincident in both the walker present and walker absent trials, low-level coincidence detectors are insufficient to account for the current findings. Were a low-level coincidence detection mechanism solely responsible for the observed effects, one would expect to see comparable detection sensitivity increases in the tone and footstep conditions as stimuli in both conditions contained identical instances of audiovisual coincidence. Such an effect was not observed. Additionally, the fact that no differences were observed as a function of audio condition with inverted stimuli also speaks against a low-level audiovisual coincidence account of the current data.

Previous research has shown that hearing tap dancing sounds increases visual sensitivity to masked point-light tapping dancing feet (Arrighi et al., 2009). Specifically, it was found that while desynchronized tap sounds produced modest increases in sensitivity, a greater benefit was conferred when the tapping was synchronized with the

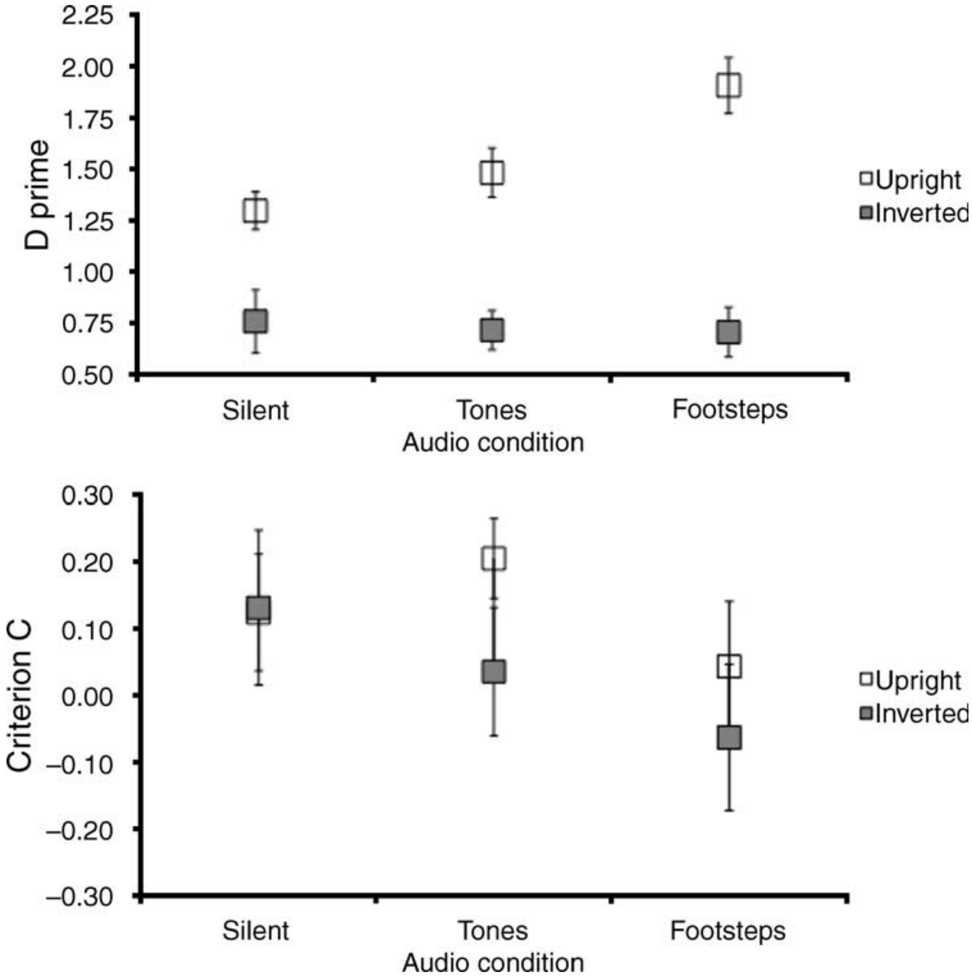

Figure 2. Experiments 1 (white boxes) and 2 (gray boxes): Detection sensitivity as assessed by D-Prime, and response bias as assessed by Criterion C. Error bars indicate standard errors.

visual displays. While these results suggest an important role for temporal synchrony, they do not indicate whether the meaningful relationship between the stimuli impacts their perception because the auditory and visual stimuli used were always meaningfully related. The current research builds upon the work of Arrighi et al. (2009) in three ways. First, it utilizes whole-body, point-light walker stimuli, and in doing so, shows that the greater detection sensitivity observed for human foot motion extends to whole-body motion. Second, it extends the results of Arrighi et al. (2009) to the perception of a category of actions with which all observers have extensive visual and motor experience. Finally, the current work directly tests and supports the hypothesis that the meaningful relationship between visual and auditory action cues impacts visual sensitivity to human motion. While the experiments reported herein do not establish the degree to which temporal audiovisual coincidence is *necessary* for this effect, coincidence alone is not *sufficient* to explain these results. Instead, the current results specifically implicate processes related to the natural associations or meaningful relationship between high-level auditory and visual cues.

While visual processes have traditionally been studied in isolation, it is becoming increasingly clear that multisensory interactions are the rule rather than the exception (Shimojo & Shams, 2001). In humans, enhancements in visual sensitivity as a function of auditory cues have been found for low-level stimuli (e.g., Bolognini, Frassinetti, Serino, & Làdavas, 2005; Driver & Spence, 1998; Frassinetti, Bolognini, & Làdavas, 2002; Vroomen & de Gelder, 2000), and these effects often depend upon some degree of spatial and/or temporal coincidence between stimuli. Here we find evidence for an enhancement in visual sensitivity as a function of meaningful auditory cues. Previous evidence exists for the role of meaningful stimulus correspondence in audiovisual detection tasks. For example, semantically congruent multimodal color stimuli (colored circles and color word vocalizations) are detected faster than either unimodal or incongruent multimodal stimuli (Laurienti, Kraft, Maldjian, Burdette, & Wallace, 2004). Reaction times for congruent multimodal stimuli are faster than those predicted by probability summation, suggesting multimodal integration. Recent evidence points towards audiovisual interactions at higher levels of complexity as well. For instance, semantically congruent sounds speed the visual detection of related pictures (Iordanescu, Guzman-Martinez, Grabowecky, & Suzuki, 2008; Molholm, Ritter, Javitt, & Foxe, 2004). A similar detection benefit is not found when sounds are paired with related *words,* which suggests that the effect occurs at the level of visual object processing rather than at some perceptually removed, amodal semantic or decision level (Iordanescu et al., 2008). Additionally, semantically congruent sounds increase participants' visual sensitivity to briefly presented, backwards masked pictures of animals and tools, while incongruent sounds tend to result in a decrease (Chen & Spence, 2010). Such examples of meaning-based crossmodal enhancement suggest that perceptual representations of objects and events may be fundamentally multimodal. The current results suggest the same for representations of human actions. However, as evidenced in Experiment 2, such representations may be constrained by factors such as stimulus orientation. In order for such high-level stimuli to be integrated, it is likely that they must match the internal model of the observed actions.

Evidence for the role of meaningful congruence in the integration of visual and auditory cues to human action depicted in full-light displays has been mixed. Vatakis and Spence (2007a) found evidence for perceptual integration based on gender cues in audiovisual speech using a temporal order judgment task. On the other hand, similar studies by these authors have failed to find evidence of integration based on meaningful congruence in displays depicting human actions with objects (Vatakis & Spence, 2008a). Furthermore, inversion of audiovisual speech, monkey vocalizations, and musical stimuli did not decrease participants' sensitivity to the temporal order of the visual and auditory streams (Vatakis & Spence, 2008b). However, inversion did impact the point of subjective simultaneity for speech stimuli, such that when stimuli were inverted, the visual stream had to lead the auditory stream by a greater interval in order for simultaneity to be perceived (Vatakis & Spence, 2007b, 2008b). This suggests that the impact of inversion may be related to a loss of configural processing of the inverted face, resulting in additional processing time required for the visual component of the stimulus.

Schutz and Kubovy (2009) did find evidence for the unity assumption in the binding of audiovisual displays depicting musical performance. Utilizing a novel audiovisual illusion, they showed that visual cues can affect the perceived duration of auditory cues, but only when the visual stimuli could have potentially caused the sound. The current results as well support a role for meaningful congruence and/or plausible causality in fulfilling the "unity assumption", whereby cues perceived to relate to the same event are perceptually bound. In sum, the current results suggest that the multimodal binding of heard and seen action cues enhances visual sensitivity to human movement.

While the current studies cannot speak to the neural mechanisms behind the observed effect, some speculation is warranted. As discussed above, potential mechanisms include the STS, in which multimodal neurons with overlapping sensitivity to human action information in multiple modalities have been found. The STS, in concert with unimodal visual and auditory areas and perhaps other multimodal cortical and subcortical regions, may play a role in a functional circuit subserving action detection and perception. Additionally, it is possible that the motor system, in particular multisensory premotor neurons, may also play a role (Gazzola et al., 2006; Keysers et al., 2003; Kohler et al., 2002; Lahav et al., 2007).

The current work represents one set of a small but growing number of studies that have investigated the effects of auditory information on the visual perception of human movement. Meaningfully related auditory cues impact visual sensitivity to human action, and that this effect is not a bias but rather an increase in visual sensitivity when auditory cues match the action presented. While the observed effect was found under artificial laboratory conditions, with passive observers viewing simplified kinematic depictions of human motion, one can speculate as to the perceptual and ultimately the behavioral advantages of the processes supporting this effect. If hearing action-related sounds enhances visual processing of related human actions, this may aid in person detection and action recognition under natural circumstances in which visual cues are degraded or ambiguous. Furthermore, while spatial coincidence is a strong cue to belongingness, one may ask what happens in real-world situations in which cues provide conflicting source information, as when sound localization is difficult due to acoustic reflection. The meaningful association between visual and auditory action cues may serve to increase the observer's visual sensitivity in situations in which the spatial or temporal relationship between cues is ambiguous. If one finds oneself in the unfortunate situation of walking alone down a dark alley at night, and one hears the footsteps of an approaching stranger, the benefit conferred from an increase in visual detection sensitivity to human motion is plain to see.

Acknowledgments

The authors thank Ashley Blanchard, Adam Doerrfeld, John Franchak Sr., Steve Ivory, Christina Joseph, Martha Kaiser, and Songjoo Oh, two anonymous reviewers their for helpful feedback, and the NSF (grant EXP-SA 0730985) and the Simons Foundation (grant 94915) for funding.

Commercial relationships: none.

Corresponding author: James Thomas.
Email: jpthomas@psychology.rutgers.edu.
Address: Department of Psychology, 301 Smith Hall, Rutgers University, Newark, NJ 07102, USA.

References

Alais, D., & Burr, D. (2004). No direction-specific bimodal facilitation for audiovisual motion detection. *Cognitive Brain Research, 19,* 185–194.

Arrighi, R., Marini, F., & Burr, D. (2009). Meaningful auditory information enhances perception of visual biological motion. *Journal of Vision, 9*(4):25, 1–7, http://www.journalofvision.org/content/9/4/25, doi:10.1167/9.4.25. [PubMed] [Article]

Barraclough, N. E., Xiao, D., Baker, C. I., Oram, M. W., & Perrett, D. I. (2005). Integration of visual and auditory information by superior temporal sulcus neurons responsive to the sight of actions. *Journal of Cognitive Neuroscience, 17,* 377–391.

Beauchamp, M. S., Lee, K. E., Argall, B. D., & Martin, A. (2004). Integration of auditory and visual information about objects in superior temporal sulcus. *Neuron, 41,* 809–823.

Bertenthal, B. I. (1993). Perception of biomechanical motions by infants: Intrinsic image and knowledge-based constraints. In C. Granrud (Ed.), *Carnegie symposium on cognition: Visual perception and cognition in infancy* (pp. 175–214). Hillsdale, NJ: Erlbaum.

Bertenthal, B. I., & Pinto, J. (1994). Global processing of biological motion. *Psychological Science, 5,* 221–225.

Bidet-Caulet, A., Voisin, J., Bertrand, O., & Fonlupt, P. (2005). Listening to a walking human activates the temporal biological motion area. *Neuroimage, 28,* 132–139.

Blake, R., & Shiffrar, M. (2007). Perception of human motion. *Annual Review of Psychology, 58,* 47–74.

Bolognini, N., Frassinetti, F., Serino, A., & Làdavas, E. (2005). "Acoustical vision" of below threshold stimuli: Interaction among spatially converging audiovisual inputs. *Experimental Brain Research, 160,* 273–282.

Bonda, E., Petrides, M., Ostry, D., & Evans, A. (1996). Specific involvement of human parietal systems and the amygdala in the perception of biological motion. *Journal of Neuroscience, 16,* 3737–3744.

Brooks, A., van der Zwan, R., Billard, A., Petreska, B., & Blanke, O. (2007). Auditory motion affects visual biological motion processing. *Neuropsychologia, 45,* 523–530.

Bruce, C., Desimone, R., & Gross, C. G. (1981). Visual properties of neurons in a polysensory area in superior temporal sulcus of the macaque. *Journal of Neurophysiology, 46,* 369–384.

Calvert, G. A., Campbell, R., & Brammer, M. J. (2000). Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. *Current Biology, 10,* 649–657.

Chen, Y.-C., & Spence, C. (2010). When hearing the bark helps to identify the dog: Semantically-congruent sounds modulate the identification of masked pictures. *Cognition, 114,* 389–404.

Dittrich, W. H. (1993). Action categories and the perception of biological motion. *Perception, 22,* 15–22.

Driver, J., & Spence, C. (1998). Attention and the crossmodal construction of space. *Trends in Cognitive Sciences, 2,* 254–262.

Frassinetti, F., Bolognini, N., & Làdavas, E. (2002). Enhancement of visual perception by crossmodal visuo-auditory interaction. *Experimental Brain Research, 147,* 332–323.

Gazzola, V., Aziz-Zadeh, L., & Keysers, C. (2006). Empathy and the somatotopic auditory mirror system in humans. *Current Biology, 16,* 1824–1829.

Grossman, E. D., & Blake, R. (2001). Brain activity evoked by inverted and imagined biological motion. *Vision Research, 40,* 1475–1482.

Grossman, E. D., & Blake, R. (2002). Brain areas active during visual perception of biological motion. *Neuron, 35,* 1167–1175.

Hein, G., Doerhrmann, O., Muller, N. G., Kaiser, J., Muckli, L., & Naumer, M. J. (2007). Object familiarity and semantic congruency modulate responses in cortical audiovisual integration areas. *Journal of Neuroscience, 27,* 7881–7887.

Iordanescu, L., Guzman-Martinez, E., Grabowecky, M., & Suzuki, S. (2008). Characteristic sounds facilitate visual search. *Psychonomic Bulletin & Review, 15,* 548–554.

Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Perception & Psychophysics, 14,* 201–211.

Kaiser, M. D., Delmolino, L., Tanaka, J. W., & Shiffrar, M. (2010). Comparison of visual sensitivity to human and object motion in Autism Spectrum Disorder. *Autism Research, 3,* 191–195.

Keysers, C., Kohler, E., Umilta, M. A., Nanetti, L., Fogassi, L., & Gallese, V. (2003). Audiovisual mirror neurons and action recognition. *Experimental Brain Research, 153,* 628–636.

Kohler, E., Keysers, C., Umiltà, M. A., Fogassi, L., Gallese, V., & Rizzolatti, G. (2002). Hearing sounds, understanding actions: Action representation in mirror neurons. *Science, 297,* 846–848.

Lahav, A., Saltzman, E., & Schlaug, G. (2007). Action representation of sound: Audiomotor recognition network while listening to newly acquired actions. *Journal of Neuroscience, 27,* 308–314.

Laurienti, P. J., Kraft, R. A., Maldjian, J. A., Burdette, J. H., & Wallace, M. T. (2004). Semantic congruence is a critical factor in multisensory behavioral performance. *Experimental Brain Research, 158,* 405–414.

McMillan, N., & Creelman, C. (1991). *Detection theory: A user's guide*. Cambridge, UK: Cambridge University Press.

Molholm, S., Ritter, W., Javitt, D. C., & Foxe, J. J. (2004). Multisensory visual-auditory object recognition in humans: A high-density electrical mapping study. *Cerebral Cortex, 14,* 452–465.

Pavlova, M., & Sokolov, A. (2000). Orientation specificity in biological motion perception. *Perception & Psychophysics, 6,* 889–899.

Pavlova, M., & Sokolov, A. (2003). Prior knowledge about display inversion in biological motion perception. *Perception, 32,* 937–946.

Pelphrey, K. A., Mitchell, T. V., McKeown, M. J., Goldstein, J., Allison, T., & McCarthy, G. (2003). Brain activity evoked by the perception of human walking: Controlling for meaningful coherent motion. *Journal of Neuroscience, 23,* 6819–6825.

Petrini, K., Dahl, S., Rocchesso, D., Waadeland, C. H., Avanzini, F., Puce, A., et al. (2009). Multisensory integration of drumming actions. Musical experience affects perceived audiovisual asynchrony. *Experimental Brain Research, 198,* 339–352.

Petrini, K., Holt, S. P., & Pollick, F. (2010). Expertise with multisensory events eliminates the effect of biological motion rotation on audiovisual synchrony perception. *Journal of Vision, 10*(5):2, 1–14, http://www.journalofvision.org/content/10/5/2, doi:10.1167/10.5.2. [PubMed] [Article]

Petrini, K., Russell, M., & Pollick, F. (2009). When knowing can replace seeing in audiovisual integration of actions. *Cognition, 111,* 432–439.

Pinto, J., & Shiffrar, M. (2009). The visual perception of human and animal motion in point-light displays. *Social Neuroscience, 4,* 332–246.

Puce, A., & Perrett, D. (2003). Electrophysiology and brain imaging of biological motion. *Philosophical Transactions of the Royal Society of London, 358,* 435–445.

Pyles, J. A., Garcia, J. O., Hoffman, D. D., & Grossman, E. D. (2007). Visual perception and neural correlates of novel "biological motion". *Vision Research, 47,* 2278–2797.

Reed, C. L., Stone, V. E., Bozova, S., & Tanaka, J. (2003). The body-inversion effect. *Psychological Research, 14,* 302–308.

Saarela, M. V., & Hari, R. (2008). Listening to humans walking together activates the social brain circuitry. *Social Neuroscience, 3,* 401–409.

Saygin, A. P. (2007). Superior temporal and premotor brain areas necessary for biological motion perception. *Brain, 130,* 2452–2461.

Saygin, A. P., Driver, J., & de Sa, V. R. (2008). In the footsteps of biological motion and multisensory perception: Judgments of audiovisual temporal relations are enhanced for upright walkers. *Psychological Science, 19,* 469–475.

Saygin, A. P., Wilson, S. M., Hagler, D. J., Bates, E., & Sereno, M. I. (2004). Point-light biological motion perception activates human premotor cortex. *Journal of Neuroscience, 24,* 6181–6188.

Schutz, M., & Kubovy, M. (2009). Causality and cross-modal integration. *Journal of Experimental Psychology: Human Perception and Performance, 35,* 1791–1810.

Seltzer, B., & Pandya, D. N. (1978). Afferent cortical connections and architectonics of the superior temporal sulcus and surrounding cortex in the rhesus monkey. *Brain Research, 149,* 1–24.

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

Spence, C. (2007). Audiovisual multisensory integration. *Acoustical Science and Technology, 28,* 61–70.

Sumi, S. (1984). Upside-down presentation of the Johansson moving light-spot pattern. *Perception, 13,* 283–286.

Taylor, K. I., Moss, H. E., Stamatakis, E. A., & Tyler, L. K. (2006). Binding crossmodal object features in perirhinal cortex. *Proceedings of the National Academy of Sciences, 21,* 8239–8244.

Taylor, K. I., Stamatakis, E. A., & Tyler, L. K. (2009). Crossmodal integration of object features: Voxel-based correlations in brain-damaged patients. *Brain, 132,* 671–683.

van der Zwan, R., MacHatch, C., Kozlowski, D., Troje, N. F., Blanke, O., & Brooks, O. (2009). Gender bending: Auditory cues affect visual judgments of gender in biological motion displays. *Experimental Brain Research, 198,* 373–382.

Vatakis, A., & Spence, C. (2007a). Crossmodal binding: Evaluating the "unity assumption" using audiovisual speech stimuli. *Perception & Psychophysics, 69,* 744–756.

Vatakis, A., & Spence, C. (2007b). How special is the human face? Evidence from an audiovisual temporal order judgment task. *Neuroreport, 18,* 1807–1811.

Vatakis, A., & Spence, C. (2008a). Evaluating the influence of the "unity assumption" on the temporal perception of realistic audiovisual stimuli. *Acta Psychologica, 127,* 12–23.

Vatakis, A., & Spence, C. (2008b). Investigating the effects of inversion on configural processing with an audiovisual temporal-order judgment task. *Perception, 37,* 143–160.

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

Welch, R. B., & Warren, D. H. (1980). Immediate perceptual response to intersensory discrepancy. *Psychological Bulletin, 88,* 638–667.

Yin, R. K. (1969). Looking at upside-down faces. *Journal of Experimental Psychology, 81,* 141–145.