Journal of Eye Movement Research
9(1):4, 1-10

Task-relevant spatialized auditory cues enhance attention orientation and peripheral target detection in natural scenes

Olli Rummukainen and Catarina Mendonça
Aalto University
Espoo, Finland

Concurrent auditory stimuli have been shown to enhance detection of abstract visual targets in experimental setups with little ecological validity. We presented 11 participants, wearing an eye-tracking device, with a visual detection task in an immersive audiovisual environment replicating a real-world environment. The participants were to fixate on a visual target and to press a key when they were confident of having detected the target. The visual world was accompanied by a task-relevant or task-irrelevant spatialized sound scene with different onset asynchronies. Our findings indicate task-relevant auditory cues to aid in orienting to and detecting a peripheral but not central visual target. The enhancement is amplified with an increasing amount of audio lead.

Keywords: natural scene, attention, detection, eye tracking, spatial sound

Introduction

The most important task of our selective visual attention mechanism is to shift our gaze towards spatial locations of interest in our surroundings. The fixation locations can be determined by visual saliency arising from contrast, movement, or color, for example (Itti & Koch, 2001). However, visual only cues are limited at most to the area covered by peripheral vision, leaving the majority of our surroundings unmonitored. The auditory system, on the other hand, is capable of simultaneously monitoring the whole space around us, and thus provide invaluable spatial information about events occurring outside the field of view (Blauert, 1997). Enhancement of the visual function due to auditory information has been demonstrated in numerous experiments utilizing abstract visual and sound stimuli. The study at hand presents an eye-tracking experiment showing the benefit of spatial hearing in orienting attention towards and detecting a visual target in a natural scene reproduced by a 3D loudspeaker setup and a visual screen with 226° horizontal field of view.

Mean saccadic response time (SRT) to visual target is enhanced when provided with spatially and temporally concurrent auditory signal (Colonius & Arndt, 2001). Concurrent auditory stimuli enhance the detectability of brief visual events, reflecting an increase in phenomenal visual saliency (Noesselt, Bergmann, Hake, Heinze, & Fendrich, 2008). Furthermore, in addition to detection, concurrent auditory stimuli improve the time to respond to visual target events (Ngo, Pierce, & Spence, 2012). On a broader note, McDonald, Teder-Sälejärvi, and Hillyard (2000) showed that involuntary, reflexive, orienting of attention to sound enhances early perceptual processing and saliency of visual stimuli and argued that it may be a fundamental operation for enhancing salience of natural stimuli. Similar findings of spatiotemporally aligned auditory facilitation are reported by a number of researchers (Kean & Crawford, 2008; Ngo & Spence, 2010; Li, Yang, Sun, & Wu, 2015).

Gleiss and Kayser (2013) showed that the enhancement of visual target detection by auditory facilitation depends on target eccentricity: peripheral target detection benefits more than central target detection of the auditory information. Partly contradicting results have been presented by Fiebelkorn, Foxe, Butler, and Molholm (2011) who found that target eccentricity or audiovisual spatial alignment do not play a role in the likelihood of detection, rather the co-occurring sounds improve visual target detection in a spatially non-specific manner. Similarly, Van der Burg, Olivers, Bronkhorst, and Theeuwes (2008) showed a visual target to pop out from a complex background in a spatial searching task with a synchronous non-spatial auditory cue. Moreover, the temporal preparation hypothesis suggests that reaction time to a primary stimulus can be shortened by a preparation-enhancement effect by an accessory stimulus, and potentially no multisensory integration is necessary (Nickerson, 1973). Following this theory, it has been shown that responses to a visual target are faster when the target is preceded by an auditory cue than when this cue is synchronized with the target (Los & Van der Burg, 2013).

The purpose of our study is to evaluate how spatial sound guides attention and affects detection in a complex natural scene. In addition, we examine how different sound onsets change these effects. We define *orient-*

DOI 10.16910/jemr.9.1.4
This article is licensed under a
Creative Commons Attribution 4.0 International license.
ISSN 1995-8692

ing as aligning visual attention with a source of sensory information, and *detecting* as being aware of a target stimulus, following the definitions of Posner (1980). In the study at hand we measure the latency of orienting visual attention to a target with an eye-tracking system (time to fixate on target area) and the latency of detecting a target by manual response (pressing a key).

In natural scenes task-irrelevant auditory stimuli rarely originate from the same spatial location as the given visual target, and moreover paying attention to task-irrelevant sounds in a real-world environment with multiple simultaneous sound sources is unlikely. Therefore, we constructed our experiment to contain a sound scene that was always correctly reproduced, i.e. matching the visual world both spatially and temporally, but half of the scenes contained sounds that gave the participants task-relevant information (informative scenes) and the other half did not (uninformative scenes). This enabled us to study the auditory enhancement of visual function in an ecologically valid setting. Dorr, Martinetz, Gegenfurtner, and Barth (2010) have highlighted the necessity of employing dynamic natural scenes instead of static images or professionally cut material in eye-tracking studies because they elicit unnatural viewing behavior.

Our findings indicate task-relevant auditory cues to aid in orienting to and detecting a peripheral but not central visual target. The enhancement is amplified with an increasing amount of audio lead with respect to visual scene onset. The task-irrelevant sound scene was not found as an aiding factor in either orienting or detection, and it resulted in comparable performance with the no sound condition.

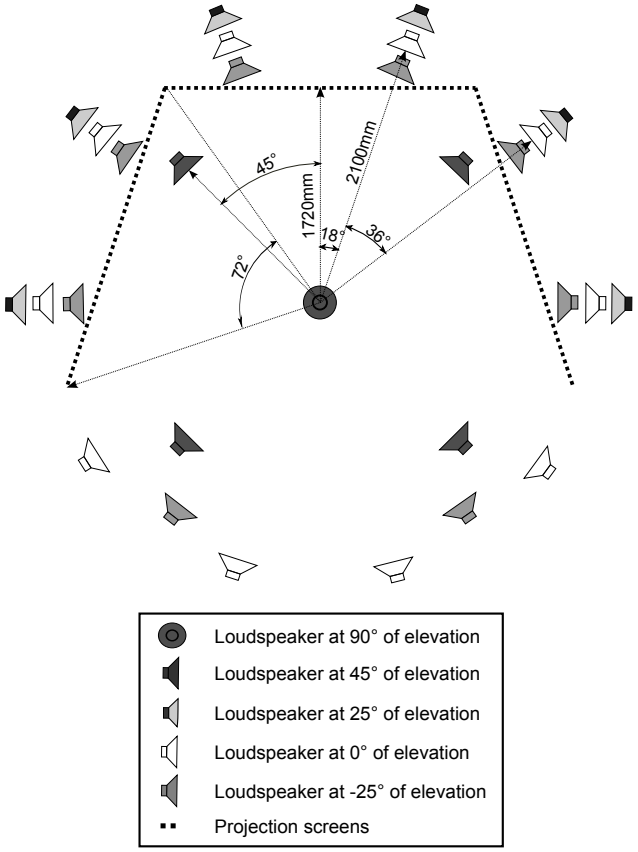

Figure 1. A schematic drawing of the loudspeaker setup and the projection screens in the immersive audiovisual environment. The observer is seated in the center of the system.

Method

Participants

A total of 11 people (2 female), naive with regard to the goal of the study, participated in the test (mean age = 33, $SD = 7.5$). A written informed consent to participate was obtained before the experimental session. All the participants reported to have normal hearing, and they were screened for visual acuity with a standard Snellen-chart at 3 m distance, to confirm normal visual function. Two participants wore glasses and one wore contact lenses.

Apparatus

The test was conducted in an immersive audiovisual environment, which consisted of three HD video projectors and 29 loudspeakers (Genelec 1029). A schematic of the system is depicted in Figure 1. The loudspeakers were installed in a spherical formation at five elevation levels around the observation position. The loudspeaker grid was made more dense behind the projection screen. The image was projected on three nearly acoustically transparent screens, 2.5 x 1.88 m each, following the shape of the base of a pentagon.

The observer was seated in the center of the system, 1.72 m from the center of each screen and 2.1 m from the loudspeaker grid. The combined visual resolution of the setup was 4320 x 1080 pixels, resulting in inter-pixel distance of 3.5 arcmin at the used viewing distance and covering horizontal and vertical field-of-views of 226° and 57°, respectively. The setup is built in an acoustically treated room. Further details of the implementation are found in (Gómez Bolaños & Pulkki, 2012). The reproduction setup has been shown to produce an accurate spatial match of auditory and visual events in reproduction of real-world environments (Rummukainen, Gómez Bolaños, & Pulkki, 2013).

Eye tracking glasses (Tobii Pro Glasses 2) were used in the experiment. The glasses enable the participants to freely explore the full area of the projection screens, while still maintaining constant tracking of the gaze with four cameras tracking the corneal reflections of the eyes. The glasses record gaze data at 50 Hz.

DOI 10.16910/jemr.9.1.4
This article is licensed under a
Creative Commons Attribution 4.0 International license.
ISSN 1995-8692

Stimuli and test cases

The stimulus videos were recorded with a camera system capable of producing spherical video (Point Grey Research: Ladybug 3), and the corresponding sound scenes were captured by a 4-capsule A-format microphone (Soundfield SPS-200). The videos were recorded and replayed at 16 frames-per-second. The loudspeaker signals were derived from the A-format microphone signals (24 bits/48 kHz) using Directional Audio Coding (DirAC) (Pulkki, 2007; Politis & Pulkki, 2011), which resulted in ecologically valid sound reproduction including the direction and distance of the sound events in the original sound scene. The spherical video was cropped to display only 226° of the full circle to make the visual world correspond to the reproduced sound scene.

Figure 2 presents a screen capture of the two scenes used as stimuli in this study. The upper panel shows the scene called *Market square*, where a view of a busy market square is shown. The soundscape consists of people chatting and noises of distant traffic. The second scene, shown in the lower panel, displays a game of floorball. In this scene the soundscape consists of the players' footsteps and shouts, and the sounds of the ball and sticks hitting the floor. The main difference between the scenes is that in the market square there is no single audiovisual event that would capture the attention, whereas in the game of floorball there is a stream of transient audiovisual events creating a storyline of the game. Van der Burg, Cass, Olivers, Theeuwes, and Alais (2010) have shown that abrupt audiovisual events are needed in order to gain benefits in visual search tasks. Moreover, the audiovisual display must not be cluttered with visual distractor events during the temporal binding window in order to avoid misbinding the auditory cue to an unwanted visual object (Van der Burg, Cass, & Alais, 2014). Examples of the stimulus scenes can be viewed online: Market square: `http://bit.ly/1ZpTzHO` and Floorball: `http://bit.ly/1OfIPTQ`. Note that the example scenes show more of the visual world than was visible in the experiment (See Figure 2).

In each trial, the two scenes were always shown concurrently, the market square scene leading the floorball. The market square scene, chosen randomly from 6 segments with similar content, was shown visually either for 9 s, 10 s, or 11 s, after which there was a 10 ms cross-fade to the game scene, which was displayed for 5 seconds. Figure 3 displays a timeline of the progress of one trial with all possible asynchronies. The sound scene was cross-faded between the scenes at 4 different lead times in relation to the visual cut. There was also a no sound condition where there was silence during the floorball segment. The audio cross-fade occurred either at 1000 ms, 500 ms, 200 ms, or 0 ms before the visual scene cut. In addition, there were 4 different segments from the game of floorball with different characteristics. There were two segments where the sound scene helped in locating the ball that was bouncing on the floor (Informative 1 & 2), and two segments with no auditory information related to the location of the ball (Uninformative 1 & 2). The different segments are summarized in Table 1 along with the target eccentricities at the beginning of the floorball scene as seen from the viewpoint of the participant. Figure 4 shows the location of the ball and the distribution of players at the visual scene cut in each segment. In addition the areas of interest (AOI), which were used to count the time to first fixation on target, are marked in the Figure.

Figure 2. *Market square* (upper panel) and *Floorball* (lower panel) scenes used as the free exploration and target detection stimuli, respectively. Dotted lines denote the corners of the screens.

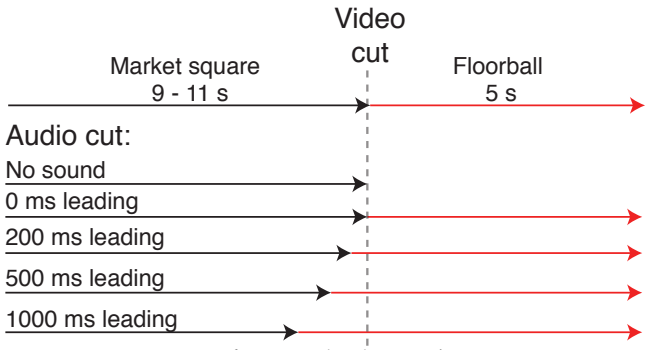

Figure 3. Structure of one trial. The *Market square* scene was visually displayed for 9, 10 or 11 seconds, after which the visual scene was cross-faded to the *Floorball* scene with a 10 ms cross-fade time. The sound scene was cross-faded similarly with a 10 ms cross-fade time with 4 different lead times in relation to the visual cross-fade. There was also a no sound condition where there was silence during the floorball segment. The task was to detect the ball in the game of floorball and press a button when the detection was confirmed.

Procedure

The participants' task was to visually detect the ball from the game of floorball. They were instructed to first freely explore the market square scene, and after the scene cut they were asked to visually find the ball and press a key on a keyboard when they were certain of having observed the ball. They were told that the

DOI 10.16910/jemr.9.1.4
This article is licensed under a
Creative Commons Attribution 4.0 International license.
ISSN 1995-8692

Table 1
Stimulus scenes and segments.

Market square	
6 segments	Ambient noise, no events
Floorball	
Informative 1	Target eccentricity 66°
Informative 2	Target eccentricity 19°
Uninformative 1	Target eccentricity -77°
Uninformative 2	Target eccentricity 51°

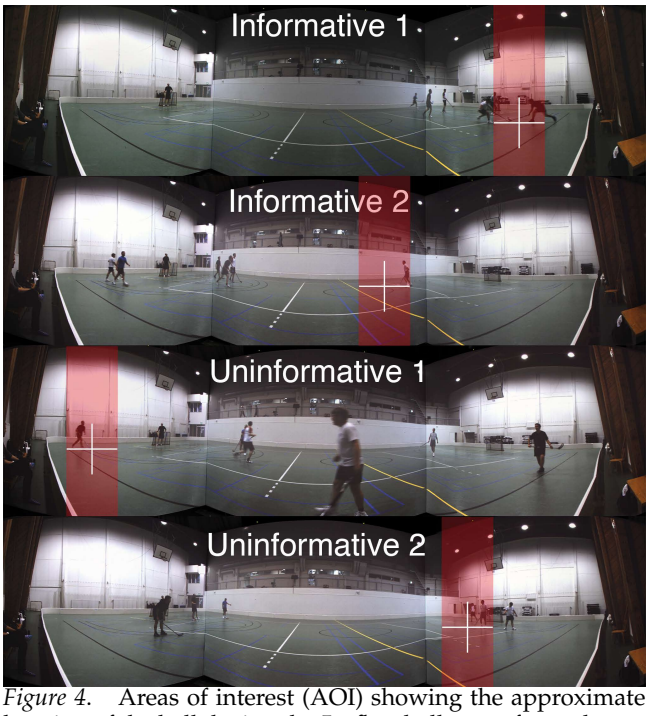

Figure 4. Areas of interest (AOI) showing the approximate location of the ball during the 5 s floorball scene for each segment. These areas were used in the analysis of the eye tracking data, as the time to first fixation was calculated according to the participants' first fixation within the respective AOI in each segment.

sound scene may change before the visual scene cut and potentially reveal information about the location of the ball. They should not, however, press the key until they had seen the ball. Between trials the participants were asked to fixate their eyes to the center of the screen where a random trial number was shown. A training phase was completed before the test to ensure the participant had understood the task and was familiar with the distinctive sound that the ball makes.

An eye-tracker (Tobii Pro Glasses 2) was used in the experiment. Before the test started, a calibration procedure took place, where the glasses' ability to track the participant's gaze was checked. The test was composed of 40 randomized trials. The trials were all the combinations of the 5 different audio lead times, and the 4 different segments of the floorball scene. Each trial was evaluated two times, resulting in 40 trials in total. The 6 different segments and the 3 different durations of the market square scene were randomly selected for each trial. The participants interacted with the experiment via a keyboard, where they were required to press the spacebar when they had detected the ball. They could take a break after each trial and continue to the next trial by pressing the arrow key. Total duration of the test, including the screening and calibration, was 25 minutes.

Results

Two-way repeated-measures ANOVAs were conducted to analyze the effect of audio lead and segment separately on the time to fixate on target and to detect it. Mauchly's test for sphericity was performed where all conditions were found to meet the criterion for sphericity. Tukey's pairwise post-hoc tests with Bonferroni correction were used to test for significant differences between conditions where a main effect was observed. The significant main effects and significant post-hoc results are summarized in Table 2.

A significant main effect of the segment was observed in both the time to fixate on target and time to press the button. In Figure 5 it can be observed that the uninformative segments were both fixated and detected with significantly longer latency than the informative segments. In the fixation latencies there is also a significant difference between the two uninformative segments where the Uninformative 1 (peripheral target) resulted in longer latency in fixation on target compared to Uninformative 2, which had a more central target. However, no difference is observed between the uninformative segments in the button press reaction time. Finally, the Informative 2 (central target) segment was fixated and detected with shorter latency than the Informative 1 (peripheral target). The average times to fixation and button press are presented in Table 3.

In the time to fixation on target significant interaction effects of the segment and lead time were observed. Further inspection revealed that there is a significant effect of lead time only in the Informative 1 segment. Figure 6 presents the overall effect and Figure 7 shows the response lag information by segment and lead time. In the Informative 1 segment the *No sound* condition resulted in significantly longer latency in fixation on target than *500 ms* and *1000 ms* audio lead conditions. Furthermore, *0 ms* and *200 ms* audio leads resulted in significantly longer latency in the fixations on target compared to *1000 ms* audio lead.

In the button press task a significant main effect of the audio lead was observed. Similar significant differences between the audio lead conditions were found here as in the fixation on target task. Figure 6 shows

DOI 10.16910/jemr.9.1.4
This article is licensed under a
Creative Commons Attribution 4.0 International license.
ISSN 1995-8692

Table 2
Main effects and significant post-hoc results with $p < 0.05$.

Main effects			Post-hoc
Fixation on target:			
Segment	$F_{(3,30)} = 37.87$	$p < 0.001$	Uninformative 1 > Uninformative 2 > Informative 1 > Informative 2
Lead	$F_{(4,40)} = 1.80$	$p = 0.15$	
Segment × *Lead*	$F_{(12,120)} = 2.31$	$p = 0.01$	
$Lead_{Informative1}$	$F_{(4,40)} = 8.65$	$p < 0.001$	No sound > 500 ms & 1000 ms; 0 ms & 200 ms > 1000 ms
$Lead_{Informative2}$	$F_{(4,40)} = 0.23$	$p = 0.92$	
$Lead_{Uninformative1}$	$F_{(4,40)} = 1.70$	$p = 0.17$	
$Lead_{Uninformative2}$	$F_{(4,40)} = 0.41$	$p = 0.80$	
$Segment_{1000\,ms}$	$F_{(3,30)} = 12.71$	$p < 0.001$	Uninformative 1 & Uninformative 2 > Informative 1 & Informative 2
$Segment_{500\,ms}$	$F_{(3,30)} = 13.40$	$p < 0.001$	Uninformative 1 & Uninformative 2 > Informative 1 & Informative 2
$Segment_{200\,ms}$	$F_{(3,30)} = 13.81$	$p < 0.001$	Uninformative 1 & Uninformative 2 & Informative 1 > Informative 2
$Segment_{0\,ms}$	$F_{(3,30)} = 16.64$	$p < 0.001$	Uninformative 1 > Uninformative 2 > Informative 2; Uninformative 1 > Informative 1 > Informative 2
$Segment_{No\,sound}$	$F_{(3,30)} = 14.04$	$p < 0.001$	Uninformative 1 & Uninformative 2 & Informative 1 > Informative 2
Target detected:			
Segment	$F_{(3,30)} = 35.91$	$p < 0.001$	Uninformative 1 & Uninformative 2 > Informative 1 > Informative 2
Lead	$F_{(4,40)} = 5.46$	$p < 0.001$	No sound > 500 ms & 1000 ms 0 ms & 200 ms > 1000 ms
Segment × *Lead*	$F_{(12,120)} = 1.44$	$p = 0.14$	

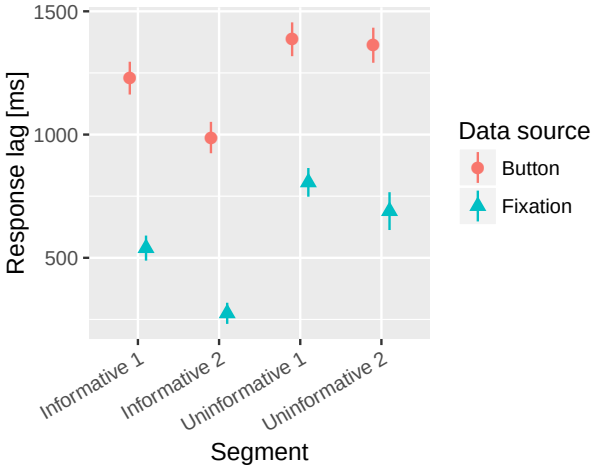

Figure 5. Combined mean times to first fixation on target and button press for the segments containing informative sounds and uninformative sounds. All audio lead times are collapsed together in this analysis. The bars represent the 95 % confidence intervals of the mean.

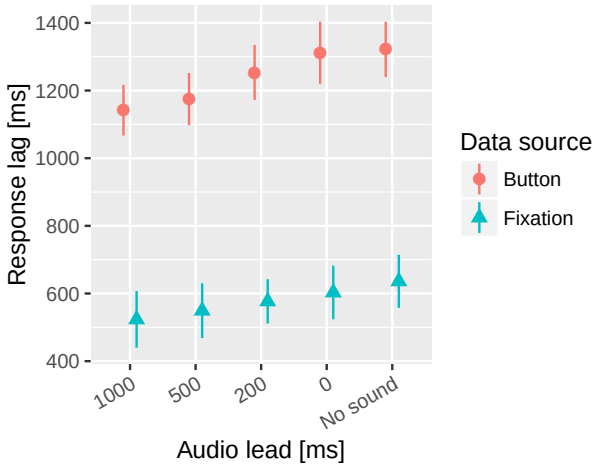

Figure 6. Combined mean times to first fixation on target and button press for the different audio lead times. All segments are collapsed together in this analysis. The bars represent the 95 % confidence intervals of the mean.

the main effect, but inspecting the detailed Figure 7 reveals the Informative 1 segment to cause most of the differences between the audio lead conditions.

No significant differences were found in the time between the first fixation on the target and the button press between scenes ($F_{(3,30)} = 1.43, p = 0.25$) or lead times ($F_{(4,40)} = 1.21, p = 0.32$). The average duration from fixation to detection was 663 ms.

Heatmaps based on the accumulated fixation counts for the Informative 1 and Uninformative 1 segments are presented in Figures 8 and 9. The amount of fixations is counted in 100 × 100 pixels squares from a scene snapshot with a resolution of 10400 × 2700 pixels, and a color is assigned to each square according to the number of fixations within the specific square. The fixations are counted during the first 1000 ms after the visual scene cut.

DOI 10.16910/jemr.9.1.4
This article is licensed under a
Creative Commons Attribution 4.0 International license.
ISSN 1995-8692

Table 3
Mean times to first fixation on target and button press.

Segment	Fixation on target	Button press
Informative 1	539 ms	1229 ms
Informative 2	275 ms	988 ms
Uninformative 1	806 ms	1387 ms
Uninformative 2	690 ms	1362 ms

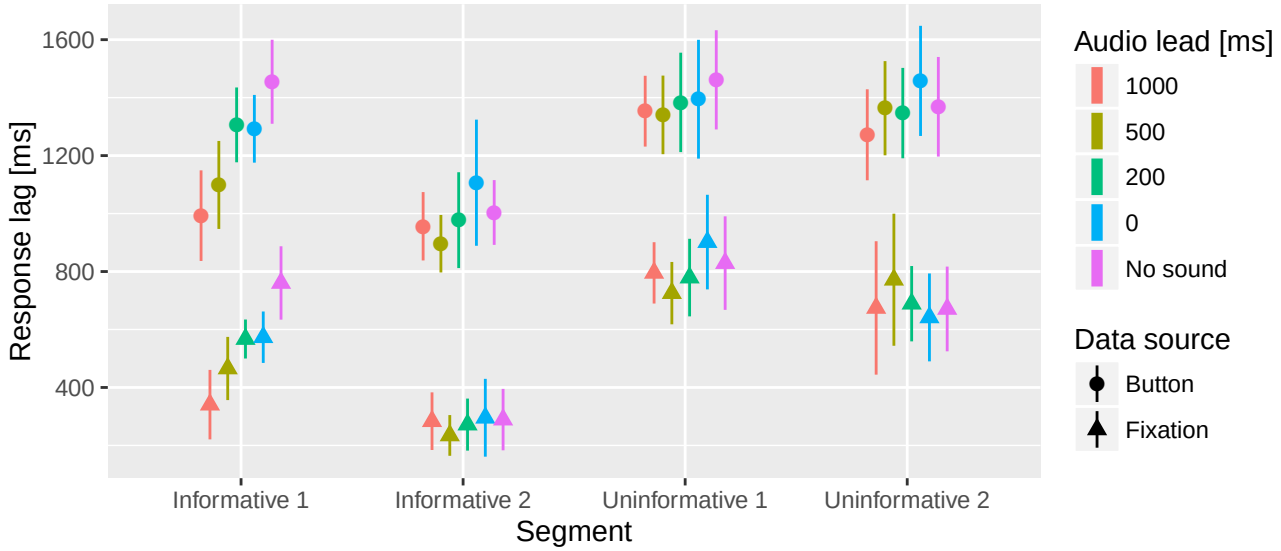

Figure 7. Mean times to first fixation on target area of interest and button press. In this analysis the data is divided according to the 4 different segments and 5 different audio offsets. The bars represent the 95 % confidence intervals of the mean.

Inspecting Figure 8, a strong clustering of gaze in the *1000 ms* condition is observed at the target location. The *200 ms* condition shows less clustering and instead more fixations at the center of the screen. In the *No sound* condition the fixations are more spread out towards the center of the screen and to the opposite side of the court. By contrast, in Figure 9, with an uninformative soundscene, there are no noticeable differences in the heatmaps between the audio conditions.

Discussion

The goal of this work was to study attention orientation and visual target detection in a real-world environment with different amounts of task-relevant or task-irrelevant auditory stimulation. The participants' task was to visually detect the ball in a game of floorball. The ability of the soundscene to guide attention was assessed by varying the sound onset time. The audio led the visual scene cut by at most 1000 ms through to no sound being played. Two measures were observed: the time to fixate on the target and the time to press a button to confirm the detection.

Task-relevant sound scenes were found to speed the time to first fixation and detection compared to uninformative ones. Between the task-relevant sound scenes the eccentricity of the target was the dominating factor in the difference of fixation latencies: the more central target was fixated on with shorter latency. A similar effect, but less pronounced, was observed between the segments whose sound scenes were task-irrelevant. Interestingly, between the uninformative segments, the latency to press the button did not differ significantly while the fixation latencies were significantly different. One possible explanation is that in both cases the overall time to fixate on the target was so long that the participants had already scanned almost all the possible locations for the ball, but lacking the confirmatory auditory cue, the decision to press the button was not made until a certain threshold in visual evidence accumulation was met. Having fixated on the Uninformative 2 target area for the first time still leaves the rightmost edge of the screen unexplored, while the Uninformative 1 is at the leftmost edge with no further areas to search. This may result in the extra time required to make the detection decision despite the faster first fixation.

Furthermore, the *No sound* condition in Informative 1 yields comparable fixation and detection times with both of the uninformative segments adding further evidence to the possible importance of an informative au-

DOI 10.16910/jemr.9.1.4
This article is licensed under a
Creative Commons Attribution 4.0 International license.
ISSN 1995-8692

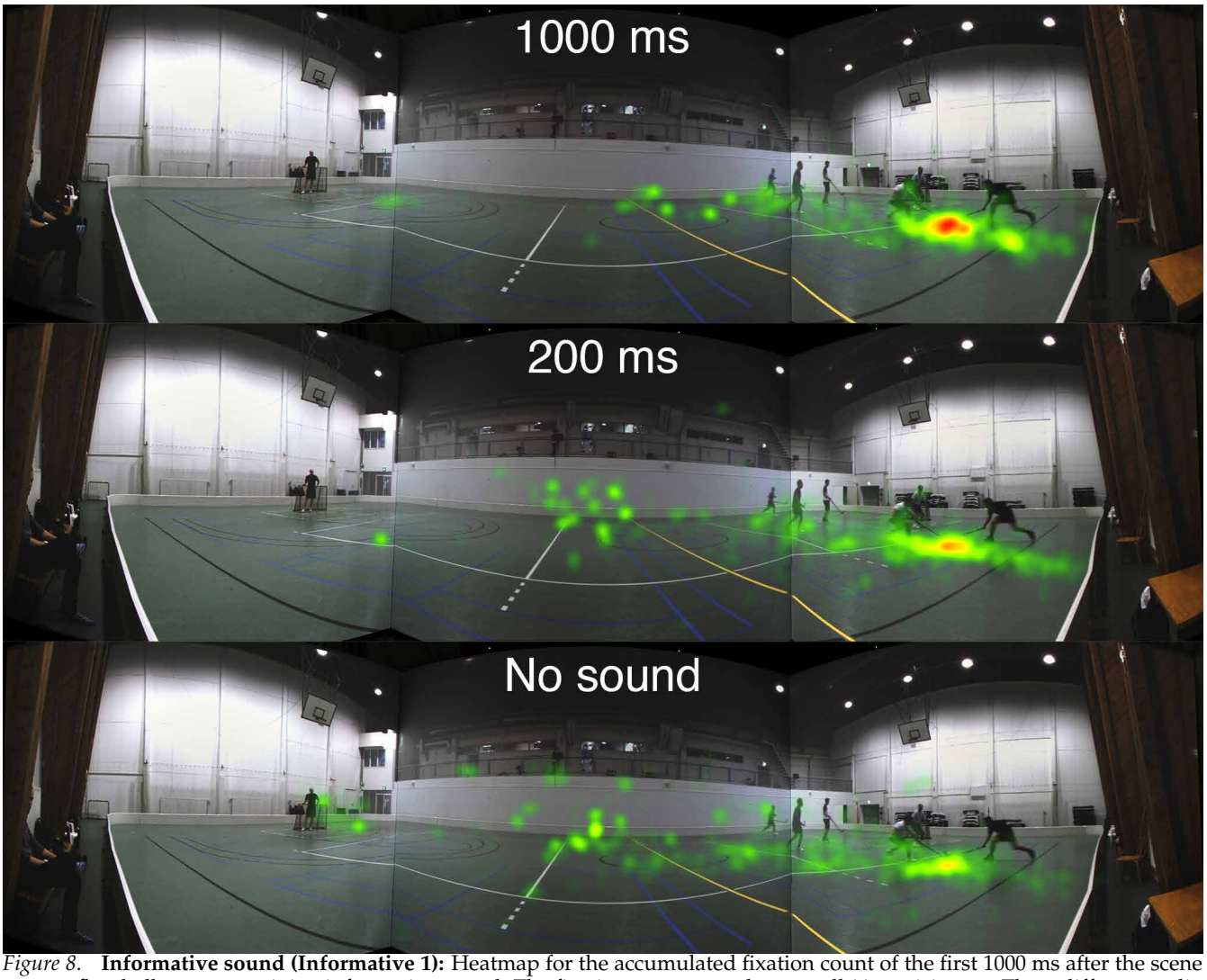

Figure 8. **Informative sound (Informative 1):** Heatmap for the accumulated fixation count of the first 1000 ms after the scene cut to a floorball scene containing informative sound. The fixations are counted across all 11 participants. Three different audio conditions are displayed: *1000 ms* lead, *200 ms* lead, and *No sound*, from top to bottom. The color codes are: red >60 fixations, yellow 30-60 fixations, and green 1-30 fixations.

ditory cue in dynamic natural scenes. This is partly in contrast with previous studies where task-irrelevant sounds have been found to enhance visual perception. However, the stimulation in previous studies has often been spatially limited to cover only central vision, and the tasks have involved detecting simple shapes or contrasts with low ecological validity, which probably explain the difference in results. The visual pop-out effect (Van der Burg et al., 2008), where a visual target is detected more easily with synchronous auditory cue, is contradicted by the Informative 2 segment. Here the visual target was centrally presented and no response enhancement was found between the *No sound* condition and synchronous audiovisual presentation. However, there is a possibility for a ceiling effect as the fixation latency was only 275 ms on average. It may be the case that, for central targets in natural scenes, already the visual cues alone are sufficient to result in low latency for fixations towards the target.

The effect of audio lead was the most prominent in the Informative 1 segment with task-relevant auditory cue and a peripheral target. In other segments the audio lead did not have a significant impact on either the time to fixation or detection. In Informative 1, at least 500 ms of audio lead was required to get significantly improved fixation and detection latency compared to a condition with no sound. In our study, latency reductions of 295 ms and 356 ms were observed in the mean times to fixation and detection, respectively. This finding is in line with previous research on saccadic response times where Colonius and Arndt (2001) have shown that accessory auditory stimulus that is lagging the visual onset increases the saccadic response time, i.e. the saccades towards the visual target are pro-

DOI 10.16910/jemr.9.1.4
This article is licensed under a
Creative Commons Attribution 4.0 International license.
ISSN 1995-8692

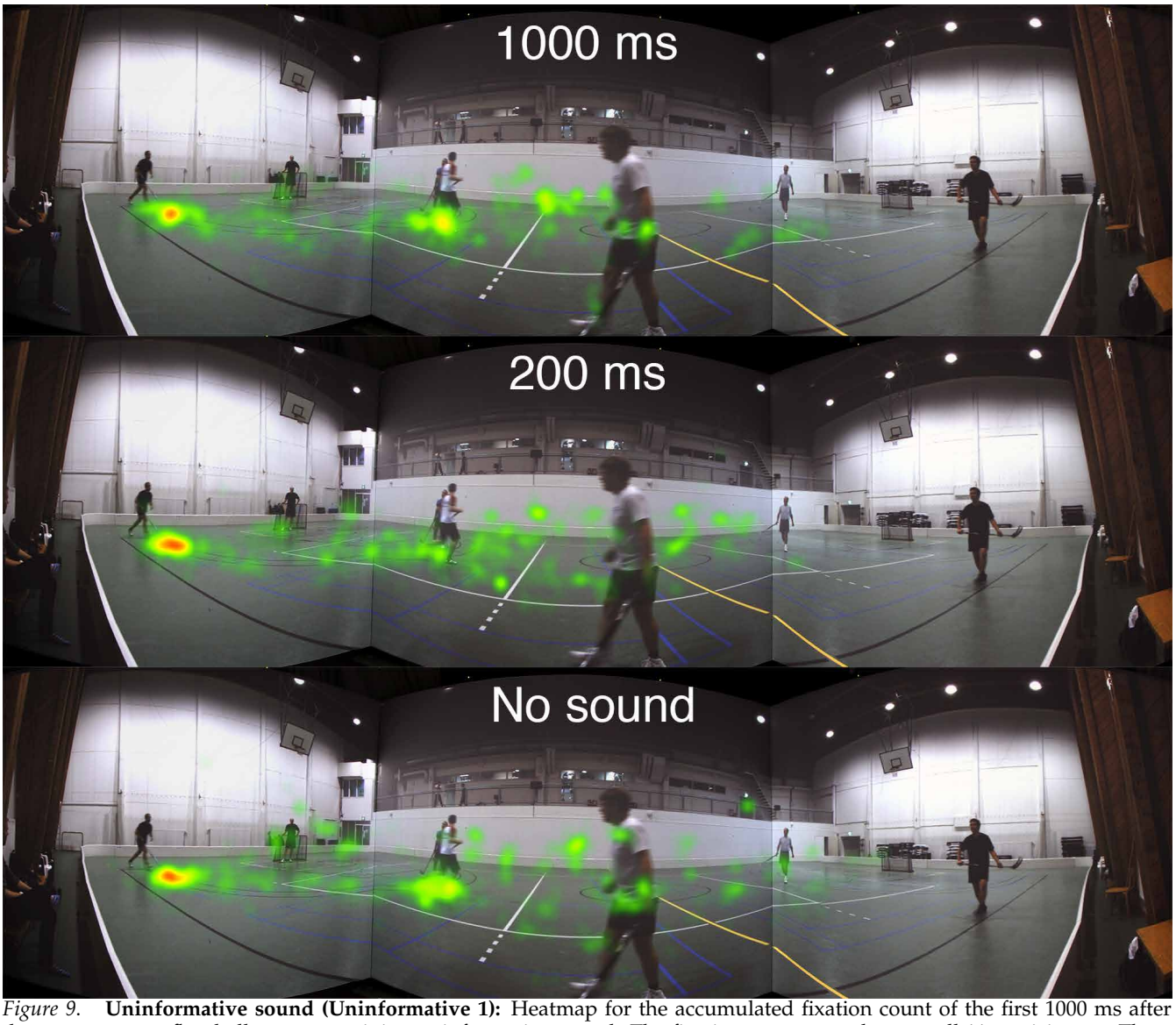

Figure 9. **Uninformative sound (Uninformative 1):** Heatmap for the accumulated fixation count of the first 1000 ms after the scene cut to a floorball scene containing uninformative sound. The fixations are counted across all 11 participants. Three different audio conditions are displayed: *1000 ms* lead, *200 ms* lead, and *No sound*, from top to bottom. The color codes are: red >60 fixations, yellow 30-60 fixations, and green 1-30 fixations.

grammed more slowly compared to synchronous onset.

Similarly, the temporal preparation theory (Nickerson, 1973; Los & Van der Burg, 2013) supports the observations of the effect of audio lead in Informative 1. The theory suggests that the enhancement of reaction time to the visual target is partly or completely due to the preceding alarming effect of the accessory auditory cue. However, inspecting the three other segments, no temporal preparation enhancement nor enhancement due to multimodal integration is observed in either of the uninformative segments or the Informative 2 segment, where the visual target was centrally presented. Therefore, based on our results, it appears that the accessory auditory cue needs to be task-relevant and the audiovisual target must not be in the central visual field for the temporal preparation and multimodal integration to have an enhancing effect.

When comparing to a synchronized presentation of audio and video, an audio lead of 1000 ms was needed to get a significant fixation and detection latency improvement. In this case the resulting improvements of the mean latencies were 232 ms for fixation and 300 ms for detection. On average, in both comparisons (No sound vs. 500 ms lead and synchronous vs. 1000 ms lead), the detection process benefitted more than the attention orientation process from the earlier auditory

DOI 10.16910/jemr.9.1.4
This article is licensed under a
Creative Commons Attribution 4.0 International license.
ISSN 1995-8692

cue. This fact would imply that the detection decision is made as a combination of the unimodal sensory evidence instead of one modality capturing the decision process.

Observing these improvements in the peripheral target (66°) case and not in the central target case (19°) is in line with previous research presented by Gleiss and Kayser (2013) who found the detection of peripheral but not central visual targets to be enhanced with spatialized simultaneous auditory cues. In our study significant differences were found only with 1000 ms audio lead while Gleiss and Kayser (2013) found significant effects already with simultaneous onset of the auditory and visual stimuli. The difference stems probably from the experimental setup: natural scenes versus abstract stimuli and the visual angles to the peripheral target (66° versus 14°).

Finally, it is of importance to note that the natural scenes used in this study do not provide as easily generalizable results than the abstract stimulation often employed in previous research. A further difficulty in our stimuli is quantifying the impact of visual cues: The players are centered around the ball in Informative 1, and in all scenes the players are looking and moving to the direction of the ball. These effects remain unchanged within a segment, but comparisons between segments may be affected by differing visual cues in addition to auditory cues. In future studies it may be beneficial to experiment with mirroring the scene laterally around the center axis to get balanced conditions, and remove any possibility for employing a search strategy. In our experiment only one of the four targets was located on the left side of the screen, potentially biasing the search locations towards the right. However, there is no evidence of such bias in Figures 8 and 9 in the No sound condition, nor was there a mention of a search strategy in informal discussions with the participants after the experiment.

We stress the importance of studying human behavior in a natural setting. Our study provides the first insights into realistic latencies of attention orienting and target detection in the real world, and future work is required to build theoretical models of perception based on these results.

Conclusions

We employed an immersive audiovisual environment to study visual target detection with varying amounts of spatialized auditory information. We evaluated the benefit of accumulating auditory information about the location of the target object on orienting to and detecting the object in a complex natural scene. Data from eye-tracking and a manual response indicated task-relevant auditory cues to aid in orienting to and detecting a peripheral but not central visual target. The enhancement was amplified with an increasing amount of audio lead with respect to the visual onset. The task-irrelevant sound scene was not found as an aiding factor in either orienting or detection, and it resulted in comparable performance with no sound condition.

Acknowledgments

The research leading to these results has received funding from the Academy of Finland (decision n^o [266239]) and from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant n^o [659114].

References

Blauert, J. (1997). *Spatial Hearing - The Psychophysics of Human Sound Localization*. Cambridge (MA): The MIT Press.

Colonius, H., & Arndt, P. (2001). A two-stage model for visual-auditory interaction in saccadic latencies. *Perception & psychophysics*, *63*(1), 126–147.

Dorr, M., Martinetz, T., Gegenfurtner, K., & Barth, E. (2010). Variability of eye movements when viewing dynamic natural scenes. *Journal of vision*, *10*(10), 1–17.

Fiebelkorn, I. C., Foxe, J. J., Butler, J. S., & Molholm, S. (2011). Auditory facilitation of visual-target detection persists regardless of retinal eccentricity and despite wide audiovisual misalignments. *Experimental Brain Research*, *213*, 167–174.

Gleiss, S., & Kayser, C. (2013). Eccentricity dependent auditory enhancement of visual stimulus detection but not discrimination. *Frontiers in Integrative Neuroscience*, *7*, 1–8.

Gómez Bolaños, J., & Pulkki, V. (2012). Immersive audiovisual environment with 3D audio playback. In *Audio engineering society 132nd convention* (pp. 1–9). Budapest, Hungary.

Itti, L., & Koch, C. (2001). Computational modelling of visual attention. *Nature reviews neuroscience*, *2*(February), 1–11.

Kean, M., & Crawford, T. (2008). Cueing Visual Attention to Spatial Locations With Auditory Cues. *Journal of Eye Movement Research*, *2*(3), 1–13.

Li, Q., Yang, H., Sun, F., & Wu, J. (2015). Spatiotemporal relationships among audiovisual stimuli modulate auditory facilitation of visual target discrimination. *Perception*, 1–11.

Los, S. a., & Van der Burg, E. (2013). Sound speeds vision through preparation, not integration. *Journal of Experimental Psychology: Human Perception and Performance*, *39*(6), 1–13.

McDonald, J. J., Teder-Sälejärvi, W. A., & Hillyard, S. A. (2000). Involuntary orienting to sound improves visual perception. *Nature*, *407*, 906–908.

Ngo, M. K., Pierce, R. S., & Spence, C. (2012). Using Multisensory Cues to Facilitate Air Traffic Management. *The Journal of the Human Factors and Ergonomics Society*, *54*(6), 1093–1103.

Ngo, M. K., & Spence, C. (2010). Crossmodal facilitation of masked visual target discrimination by informative auditory cuing. *Neuroscience Letters*, *479*(2), 102–106.

Nickerson, R. (1973). Intersensory facilitation of reaction time: energy summation or preparation enhancement? *Psychological Review*, *80*(6), 489–509.

DOI 10.16910/jemr.9.1.4
This article is licensed under a
Creative Commons Attribution 4.0 International license.
ISSN 1995-8692

Noesselt, T., Bergmann, D., Hake, M., Heinze, H. J., & Fendrich, R. (2008). Sound increases the saliency of visual events. *Brain Research*, *1220*, 157–163.

Politis, A., & Pulkki, V. (2011). Broadband analysis and synthesis for DirAC using A-format. In *Audio engineering society 131st convention* (pp. 1–11). New York, (NY).

Posner, M. I. (1980). Orienting of attention. *The Quarterly Journal of Experimental Psychology*, *32*, 3–25.

Pulkki, V. (2007). Spatial sound reproduction with directional audio coding. *Journal of the Audio Engineering Society*, *55*(6), 503–516.

Rummukainen, O., Gómez Bolaños, J., & Pulkki, V. (2013). Horizontal localization of auditory and visual events with Directional Audio Coding and 2D video. In *Quality of multimedia experience (qomex)* (pp. 94–99). Klagenfurt am Wörthersee, Austria.

Van der Burg, E., Cass, J., & Alais, D. (2014). Window of audio-visual simultaneity is unaffected by spatiotemporal visual clutter. *Scientific Reports*, *4*, 1–7.

Van der Burg, E., Cass, J., Olivers, C. N. L., Theeuwes, J., & Alais, D. (2010). Efficient Visual Search from Synchronized Auditory Signals Requires Transient Audiovisual Events. *PLoS ONE*, *5*(5), e10664.

Van der Burg, E., Olivers, C. N. L., Bronkhorst, A. W., & Theeuwes, J. (2008). Pip and pop: Nonspatial auditory signals improve spatial visual search. *Journal of Experimental Psychology: Human Perception and Performance*, *34*(5), 1053–1065.

DOI 10.16910/jemr.9.1.4

This article is licensed under a
Creative Commons Attribution 4.0 International license.

ISSN 1995-8692