Binaural Sound Reduces Reaction Time in a Virtual Reality Search Task

Emil R. Høeg* Aalborg University
Lynda J. Gerry† University of Copenhagen
Lui Thomsen* Aalborg University
Niels C. Nilsson‡ Aalborg University
Stefania Serafin‡ Aalborg University

ABSTRACT

Salient features in a visual search task can direct attention and increase competency on these tasks. Simple cues, such as color change in a salient feature, called the "pop-out effect" can increase task solving efficiency [6]. Previous work has shown that nonspatial auditory signals temporally synched with a pop-out effect can improve reaction time in a visual search task, called the "pip and pop effect" [14]. This paper describes a within-group study on the effect of audiospatial attention in virtual reality given a 360-degree visual search. Three cue conditions were compared (no sound, stereo, and binaural), with rising degrees of difficulty by increasing the set size. The results (n=10) indicate a statistically significant difference in reaction time between the three conditions. Overall, in spite of the small sample size, our results seem to indicate that binaural audio renders a clear advantage to the spatial visual processing.

Keywords: attention, visual search, multisensory integration, binaural, audition, vision, virtual reality

Index Terms: H.1.2 [Information Systems]: User/Machine Systems—Human factors; I.3.7 [Computer Graphics]: Three-Dimensional Graphics and Realism—Virtual Reality

1 INTRODUCTION

Bottom-up salient features in visual search tasks have been implicated in directing attention to increase competency on these tasks. For instance, Duncan and Humphreys [6] found that subjects could rapidly find a green line amongst red line segments, regardless of the number of distractors. This effect is called the "pop-out effect." Searching for an object within a visually cluttered and continuously changing environment is a much more formidable task, and more relevant for everyday wayfinding in the dynamic and complex natural and social environment. Thus, Perrot, Saberi, Brown, and Strybel [12] developed an aurally aided visual search paradigm to support finding the relevant object, quantified by a reduction in the search time for a target. Binaural sound cues have been implicated in wayfinding tasks and orienting visual attention, and binaural cues have been shown to increase visual attention to salient cues of a visual scene. The workload of processing an overloaded visual display can be aided by three-dimensional (3D) auditory cues, and moreover 3D auditory cues can enhance the detection and identification of visual targets. Previously, Van der Burg et al. [14] showed that pairing a simple auditory "pip" tone with a visual color-change effect ("pop") on a salient object during a visual search task helped subjects find the target object within a visual search task. This is called the "Pip and Pop" effect, whereby a temporal synchrony between a tone and a visual "blink" effect on the salient target (a straight line amidst several oblique line segments) caused a faster reaction time even though the sound cue did not contain any information about the location or orientation of the target.

*e-mail: {ehaeg12, lath11}@student.aau.dk
†e-mail: lynda@whatisexperience.com
‡e-mail: {ncn, sts}@create.aau.dk

One of the aims of this study is to provide a first step towards conducting visual search tasks with higher ecological validity. For the study, we re-mediate the Pip and Pop visual search task within a 360 degree Virtual Reality (VR) task space and use binaural sound integration to study visual target detection times. The Virtual Environment (VE) consisted of a set of oblique red and green line segments on a hemisphere against a black backdrop, above a fading gridline for reference orientation (see fig. 3). The subject was positioned in the middle and the VE was presented through a Head-Mounted Display (HMD). The target horizontal line appeared in random places, along the circumference of the hemisphere, with an increasing set size of oblique line segments (50, 100, 200 or 400) per trial presented.

We hypothesize that binaural tones as "pip" cues within a 360 visual search space will create a temporal and spatial synchrony for the visual "pop" that will be greater than temporal synchrony alone with a stereo tone containing no spatial information. Thus, we also hypothesize that binaural audio cues present spatial information for a salient visual cue that will contribute to a significant decrease in search times as compared to no-tone and stereo-tone conditions.

2 PREVIOUS WORK

Searching for an object within a cluttered, continuously changing environment is a challenging task. Many researchers have investigated the utility of three-dimensional (3D) auditory displays for reducing the workload associated with an overloaded visual channel and for enhancing the detection and identification of visual targets [6, 7, 10, 11]. The utility advantage of spatial audio cues can be quantified by investigating its reduction on the search time for target acquisition tasks [14]. For example, recent investigations have demonstrated the utility of spatialized audio cueing for some aircraft applications [3, 9]. Van der Burg and colleagues [14] had previously shown that a pip audio cue reduces search times for a visual object that is normally hard to find. However, in these studies the pip effect did not provide any information on the location or identity of the visual object. The visual target change "pop" signal was integrated with an auditory signal through temporal synchrony such that a synchronous pip sound allowed the pop out effect of the visual target change within a complex environment. Several studies have shown that sound can guide attention towards a visual target when the visual and auditory signals come from the same location [2, 4, 5, 8, 12]. Thus, our study aimed to combine these two findings to show that binaural sound, which can provide both spatial cues about a target as well as a synchronous temporal pip and pop effect, would even more significantly reduce search times than stereo sound.

3 METHOD

3.1 Participants

In total there were 10 participants (9 males and 1 female) between the age of 24-37 (mean = 29.25, SD = 4.49). They were recruited through convenience sampling, and the test was conducted at Aalborg University in Copenhagen.

3.2 Stimuli and Apparatus

Three overall conditions were applied as repeated measures: No Sound (NS), Stereo Sound (SS), and Binaural Sound (BS). In the

NS condition, the pop alone occurred as a pop-out effect for the visual search. However, in the stereo and binaural conditions, the visual target change pop effect was paired with a short sweeping tone in temporal synchrony. The added effect of the BS was that it also gave spatial information as to the location of the target to guide the visual search to that region of the space. The binaural condition provided information about the position of the visual target, whereas the stereo condition only provided information about the moment of changing color of the visual target.

Furthermore, an independent within-subjects variable was added in the form of an increasing set size (line segments), doubling with each trial from 50, 100, 200, and 400, to make the task more complex as participants proceeded through the trials. In total, each participant would repeat the same 4 trials (increasing set size) for all three conditions for a total of 12 trials per participant. On average the participants took approximately 4.2 minutes to complete all 12 trials.

The original paper [14] used a 500 Hz sine tone, but since sine waves are hard to localize [1], we decided to adopt a different synthetic tone, i.e. a sweeping tone acquired from the notification category of the official Oculus Audio Pack 1. For ease of listening over time, the sound was processed by a high-pass filter and a slight fade in the time domain. Time and frequency domain representation are shown in figure 1 and 2 respectively.

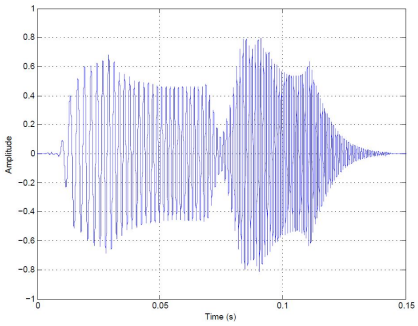


Figure 1: Time domain representation of the auditory feedback used.

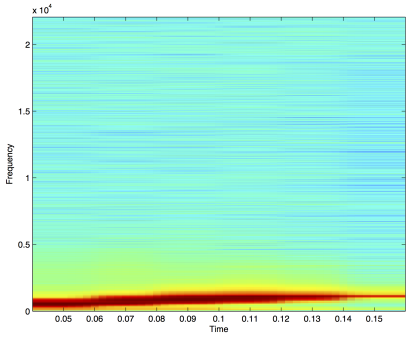


Figure 2: Frequency domain representation of the auditory feedback used.

Sounds were spatialised using head-related-transfer-function (HRTF) filtering by the Oculus Audio SDK written in C. In the Oculus Audio SDK, the HRTF filter related to the incoming direction of a sound is computed in real-time and generates binaural audio by convolving the sound source with dry input audio [13].

3.3 Design and Procedure

Participants wore an Oculus Rift CV1 HMD with integrated headphones, displaying a black void with a fading aerial perspective and luminescent gridlines at the bottom of the test participant's visual field to aid orientation and establish the floor as a horizontal point of reference (see figure 3).

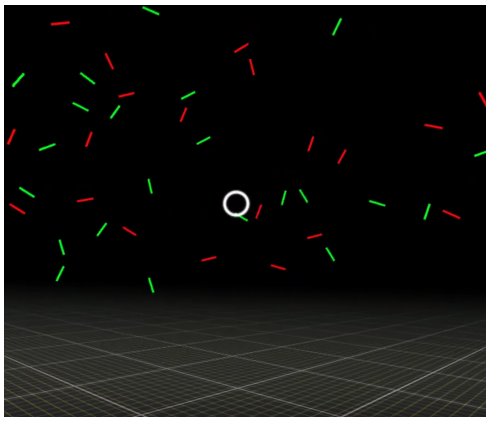


Figure 3: The view from the center of the hemisphere, where the participant would be situated inside the VE. The participant would use the white reticle to aim at the target when identified, and press the remote controller to select and hence end the search task.

Participants were instructed to search the 3D environment by rotating their head and body. Limited positional tracking was available with the Oculus' positional tracking system (The Constellation sensor) but non-essential to the experiment, since no locomotion was required. The line segments were created from primitive cylinder shapes, and procedurally generated at random locations (spawn-points) along the circumference of a hemispherical shape (see figure 4). The participant was situated in the center of the hemisphere, to have a uniform distance from the participant's position and the visual stimuli. By default, 1 unit equals 1 meter in Unity, hence the radius of the hemisphere can be said to be 37.5 meters and the length of line segments were 1.3 meters. To prevent a situation where the horizontal line may be indeterminable due to head-orientation, spawn-points were restricted by a height-limit of $0.75 * radius$. Furthermore, overlapping line segments were repositioned to prevent colliding and cluttered distractors.

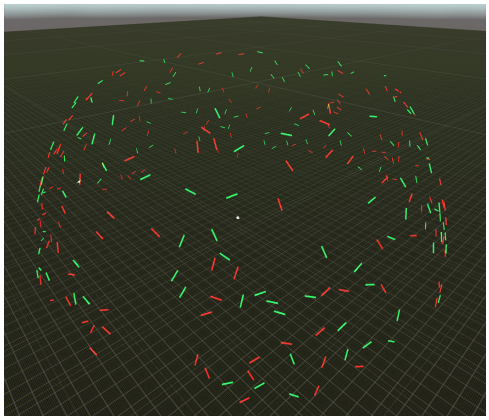


Figure 4: An aerial view of the hemisphere where line segments (distractors) are randomly generated (here with a set size of 400) and height limited to 75% of the radius.

During the task, the participant would search for, and identify, a single salient horizontal line segment (target) within the hemispherical environment with up to 400 oblique line segments at random orientations ($\pm 15^\circ - 75^\circ$) and position. The "pop-out" effect, i.e. the target color alternating between red and green, changed once every 900 ms, which is the same change-frequency used by [14]. The application was designed to prevent the target from spawning within the visual field on test-initiation to minimize the chance of immediate detection. In all between-subject conditions, the participants were instructed to respond as fast and accurately as possible to the target's position, using the white circular reticle to aim at the target to identify it, and the Oculus remote to select it by pressing the select button. Once the correct target was identified, the current task would terminate, and the reaction time (s) would be displayed on the monitor, visible only to the experimenter. The participant would keep the HMD on for the full duration of the experiment.

4 Results

The conditions were evaluated simultaneously using the Friedman test, and the Wilcoxon signed-rank test for a pairwise comparison of the conditions. These tests were chosen because the assumption of normality is violated as shown by an Anderson-Darling test. To determine whether or not the test results are significant we use a significance level of $\alpha = 0.05$ for both tests.

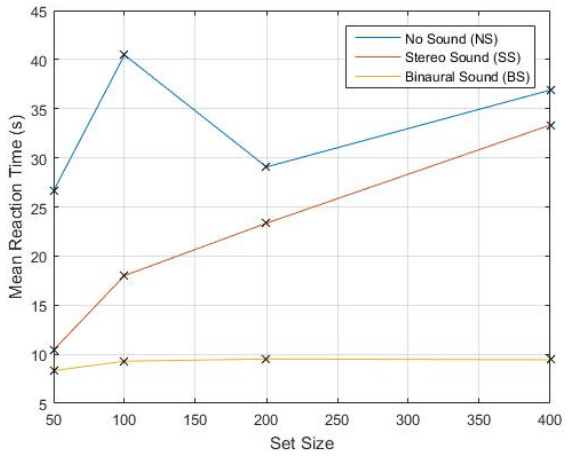


Figure 5: Mean reaction time for all three conditions (NS, SS and BS) given increasing set sizes 50, 100, 200, and 400.

Figure 5 shows the mean reaction time for all three conditions, given an increasing set size from 50 to 400. It is observable that the task completion time is rather constant for the condition with BS, while it significantly increases in the conditions with NS and SS respectively.

Table 1: Results of the Friedman-test.

Friedman	**50**	**100**	**200**	**400**
P-value	0.0608	0.0022	0.0033	0.0136

The Friedman test (see Table 1) shows an overall significant difference for all conditions, for all numbers of objects. The Wilcoxon signed-rank test (see Table 2) shows that there is a significant difference between the conditions NS and SS for a small number of objects and NS versus BS for all conditions. Also BS significantly decreases the search time when searching among a large number of objects as opposed to SS.

Table 2: Results of the Wilcoxon signed-rank test

Signed Rank	**50**	**100**	**200**	**400**
No Sound - Stereo	0.010	0.049	0.193	0.375
No Sound - Binaural	0.014	0.002	0.002	0.004
Stereo - Binaural	0.375	0.556	0.048	0.037

The Friedman test shows most the significant difference ($\rho < 0.0022$) at a set size of 100. This is due to the peak in NS (see fig. 5). Wilcoxon signed-rank highlights the fact that the declining significance in the Friedman test is due to NS and SS converging with an upward trend with increasing set size. At the same time BS remains approximately constant through all set sizes. As expected, this indicates that BS is important in a search task of increasing complexity.

5 Discussion

While the preliminary results of the experiment are encouraging, several factors can be improved in future implementation. First of all the small sample size (n=10) does not allow us to draw strong conclusions. Moreover, the trials were not randomized during the experiment, which obviously could make the improving search results, not necessarily due to the binaural sound rendering, but to learning effects. Despite these limitations, it is encouraging to observe the importance of sound rendering and binaural sound when searching for an object in a large collection.

One of the limitations of this study is, in this setup the subject's visual field is restricted, whereas the binaural audio is not. In Van der Burg et al's original study on the pip and pop effect [14], the two effects were simultaneously presented such that the pip tone was in temporal synchrony with the pop-out effect. While our study also presented the two effects simultaneously, the user may not have had the target in their field of view such as to perceive the two effects synchronously. That is, the visual synchrony effects might not have been useful to the subject if the target was not within the visual field when the tone was presented. In future experiments, we would like to account for the subject's initial head orientation in relation to the target at the moment when the auditory pip is presented. This will potentially help isolate spatialized, binaural sound as more advantageous than stereo sound in providing spatial information about the target, rather than just allowing the pop-out effect to be more noticeable (as Van der Burg and colleagues previously demonstrated). This notion seems to indicate that the decreased reaction times for binaural tones could be due to spatial attenuation cues rather than just temporal synchrony with the pop-out effect and target detection.

Despite these limitations, it is encouraging to observe the importance of sound rendering and binaural sound when searching for an object in a large collection. The mean reaction times for each set size could have been even greater if the proper counterbalancing had been implemented. Thus, it is important to note that decreased reaction times were observed per set size, even despite the learning effects that could have been introduced by the lack of counterbalancing. Regardless, the binaural sound condition had faster reaction times in all set sizes opposing stereo and no-sound conditions, despite possible confounding bias due to learning effects.

6 Conclusion

In this experiment we investigated whether searching for an object in a complex virtual reality setting, previously known as the pip and pop effect experiment, is facilitated by the introduction of a short auditory icon temporally synchronized with the flashing object. While previous similar experiments were performed using traditional AV equipment (computer screen), we are interested in

investigating whether similar results can also be applied to a VR scenario.

As expected, a sound synchronized with the flashing visual feedback significantly helped searching for objects. However, although binaural sound did reduce reaction time, as opposed to no sound or simple stereo rendering, the small sample size and lack of counterbalancing questions the experiment's validity. Further studies are needed to better investigate to what extent searching for objects in a complex visual scenario is facilitated by the introduction of spatialized auditory feedback. While this study still uses simulated tones within a minimalistic visual search space, we believe it provides a bridge towards understanding auditory signaling effects on visual search tasks, and potentially within richer and more realistic environments.

REFERENCES

[1] D. R. Begault and L. J. Trejo. 3-d sound for virtual reality and multimedia. 2000.

[2] R. S. Bolia, W. R. D'Angelo, and R. L. McKinley. Aurally aided visual search in three-dimensional space. *Human factors*, 41(4):664–669, 1999.

[3] A. W. Bronkhorst, J. Veltman, and L. Van Breda. Application of a three-dimensional auditory display in a flight task. *Human factors*, 38(1):23–33, 1996.

[4] M. Doyle and R. Snowden. Facilitation of visual conjunctive search by auditory spatial information. In *Perception*, volume 27, pages 134–134, 1998.

[5] J. Driver and C. Spence. Cross–modal links in spatial attention. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 353(1373):1319–1331, 1998.

[6] J. Duncan and G. W. Humphreys. Visual search and stimulus similarity. *Psychological review*, 96(3):433, 1989.

[7] P. Flanagan, K. I. McAnally, R. L. Martin, J. W. Meehan, and S. R. Oldfield. Aurally and visually guided visual search in a virtual environment. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 40(3):461–468, 1998.

[8] J. J. McDonald, W. A. Teder-Sälejärvi, and S. A. Hillyard. Involuntary orienting to sound improves visual perception. *Nature*, 407(6806):906–908, 2000.

[9] R. L. McKinley and M. A. Ericson. Flight demonstration of a 3-d auditory display. 1997.

[10] W. T. Nelson, L. J. Hettinger, J. A. Cunningham, B. J. Brickman, M. W. Haas, and R. L. McKinley. Effects of localized auditory information on visual target detection performance using a helmet-mounted display. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 40(3):452–460, 1998.

[11] D. R. Perrott, J. Cisneros, R. L. Mckinley, and W. R. D'Angelo. Aurally aided visual search under virtual and free-field listening conditions. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 38(4):702–715, 1996.

[12] D. R. Perrott, T. Sadralodabai, K. Saberi, and T. Z. Strybel. Aurally aided visual search in the central visual field: Effects of visual load and visual enhancement of the target. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 33(4):389–400, 1991.

[13] C. Schissler, A. Nicholls, and R. Mehra. Efficient hrtf-based spatial audio for area and volumetric sources. *IEEE transactions on visualization and computer graphics*, 22(4):1356–1366, 2016.

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

View publication stats