PROCEEDINGS of the
23rd International Congress on Acoustics

9 to 13 September 2019 in Aachen, Germany

Cognitive indicators for acoustic source localization and presence in a vivid 3D scene

Patrick RUEDIGER[1]; Jan SPILSKI[2]; Nûjîn KARTAL[3]; Sebastian GSUCK[3]; Nils Ove BEESE[2]; Sabine J. SCHLITTMEIER[5]; Thomas LACHMANN[2,4]; Achim EBERT[1]

[1] University of Kaiserslautern, Computer Graphics & HCI Lab, Germany

[2] University of Kaiserslautern, Center for Cognitive Science, Germany

[3] MediaApes GmbH, Germany

[4] Universidad Nebrija Madrid, Spain

[5] RWTH Aachen University, Germany

ABSTRACT

The easy access and availability of Virtual Reality (VR) technologies opens up a plethora of application fields as it is becoming increasingly easier to use from a technical point of view and more affordable from a consumers' perspective. However, current VR applications hardly or only very rudimentarily take plausible acoustics - i.e. binaural / 3D acoustics - into account, thus much of the technology's potential remains untapped. In this paper, the results of our experiments for localizing audio signals and presence using VR technologies are presented. A virtual representation of two realistic scenes (360° movie shot) was used to accurately place sound signals in the three-dimensional space using object-based audio. We compared the effect of using simple stereo sound with binaural audio which uses the overlap of the field of hearing (spatial sound) similar to the techniques used for stereoscopic depth perception in 3D movie theaters. The participants had to point out the source of the audio signals. A VR headset with eye tracker was used to measure reaction times, accuracy and eye movements. We discuss our results with respect to implications for practice, cognitive science and future VR research.

Keywords: Virtual Acoustics, Cognition, Binaural Sound Localization

1. INTRODUCTION

The two most commonly used terms describing quality of virtual reality (VR) environments are "immersion" and "presence". Presence is mainly defined as the feeling of "being there", which means that presence is a psychological state or at least the current subjective impression of being in the virtual world; whereas "immersion" describes the technology-related aspects which contribute to a particular VR environment, such as the resolution and vividness of visualization, audio quality and plausibility (1).

Presence can be measured directly or indirectly (cp. 2). A *direct method* is to ask participants to assess presented stimuli, scenarios or systems, either qualitatively-verbally or by means of questionnaires with rating scales (e.g., 3). *Indirect methods* are behavioral measures, like performance, reflexive motor acts or sensori-motor control, and physiological measures (e.g., heart rate, adrenalin). The rationale behind behavioral measures is to record them when the user is experiencing a VR environment or is working there on a specific task - and to interpret the obtained measures as indicators of the degree to which a user is "mentally immersed" (4). Although these measures are called "indirect" in the VR literature, they can be considered much more direct indicators of presence from a psychological point-of-view than a self-report by the user, which forces users to get "out of the loop" when experiencing a VR environment and to reflect on perceived presence - or to assess it retrospectively based on memory. Do note, that the aforementioned indirect measures of presence being often thought of as the psychological - i.e. perceptual, cognitive, motivational and emotional - consequences of immersion.

The technical specifications of VR environments - and thus of immersion - define, in which signal qualities the different modalities are addressed. The development of VR has been advanced mostly in the visual domain ("immersive visualization") while the other modalities – such as audio and haptic - were primarily integrated application-specifically and thus often in a more rudimentary level. Nowadays, however, the acoustics of a scene can also be integrated in binaural (3D) quality, not only as a mono or stereo signal. 3D binaural sound uses the object-oriented approach to place a sound source in the space of a scene. Knowing the position of the source and the receiver (e.g., the user) in space, and the acoustical properties of the room or other objects, one can simulate the perceived acoustic for the user at any given time and space. Generic or even individualized head-related transfer functions (HRTFs) can be used for binaural arrangements transmitted over headphones. These are used for further plausible signal processing with respect to frequency characteristics and sound pressure levels at each eardrum. However, the question arises whether 3D sound (i.e., binaural sound), really “matters” for the user in terms of presence in audiovisual VR environments.

Searching for an object is an everyday perceptual-cognitive task and can be extremely difficult. This holds particularly true within a cluttered visual environment, for instance when trying to find a particular person in a crowd. However, if the searched person shouts for help or blows a whistle, one would expect the task to become easier, i.e. that the search time is reduced in this audiovisual condition compared to vision-only conditions. And one might also assume that search performance - in terms of reaction time and/or accuracy - should be better if the auditory cue is available in 3D spatial audio allowing for better localization of the sound source - compared to stereo.

The auditory cue is said to accelerate the visual search when both visual and auditory signals come from the same location (e.g., 5, 6). When it is not coming from the same location, it at least grabs the attention of the listener (7). Some studies already tested whether 3D audio can enhance the search performance for visual targets in a VR-based but highly abstract image search (e.g., detecting a horizontal line in a huge amount of colored non-horizontal lines (8, see also 6 and 9-11). Until now, however, the effects of 3D audio on visual search performance has not been explored within a more realistic audiovisual scene like the aforementioned example of searching a particular person (i.e., target) in a crowd (i.e., many near-to-target distractors in low contrast conditions) via the help of spatially matched auditory cues (i.e., sound emitted “by” the visual target).

The present study tried to achieve a more plausible audiovisual stimulus and task setting by using 3D videos in VR that were either combined with binaural 3D or stereo sound. We focus on the behavioral approach by investigating the relevance of binaural (vs. stereo) audio for performance in a visual search task. The behavioral measures such as saccadic eye movement patterns are interpreted as indicators of participants’ presence in the realized audiovisual scenario and task.

We can summarize the goals of our study as follows:

1. Does a binaural audio signal increase the localization speed in a visual search task in VR compared to stereo and silence?
2. Does binaural audio have an effect on the presence of a VR scene compared to stereo and silence?

2. Method

In preparation for the main experiment, two pre-tests with different focal points were carried out. In paragraph 2.1 we briefly describe them and in paragraph 2.2 we provide the methodology of our main experiment.

2.1 Pretests

Pretest 1: To get an idea to which extent the participants’ ability to distinguish between a stereo and binaural acoustic setting, we asked them about their preference for one or the other, based on different sound samples produced in binaural and stereo. 102 participants (*N* = 102) conducted this pretest, ranging from different ages and acoustic backgrounds. We used six different audio settings, which were presented as stereo files and as binaural files. Two of these settings (stadium and organ music) were directly recorded using a 3D microphone setup, whereas for the others the spatial effect was added post-production. The pretest was conducted as an online survey. After each trial, the participants were asked which audio file they preferred (one vs. two) or if they perceived no difference in the audio files. Participants did not know in advance which version was stereo or binaural.

Pretest 2: To effectively limit visual and auditory stimulus sets to the behavioral relevant variations, we determined the maximum directional resolution the participants are able to perceive in this setup. Furthermore, we checked in the second pretest if the signal to noise ratios (SNRs) of our stimuli had an effect on the localization performance. The pretest was carried out with a total of *N*=17 participants (students and faculty staff). None of them reported an acoustic handicap. We varied 8 azimuth (horizontal plane) and 3 elevation (vertical) angles. We had a variation of 6 SNRs to determine its effects on localization performance. Out of these, we setup 144 combinations in a way so that the variations were equally distributed. We then randomly chose 20 of these combinations for each participant. A random number generator was set up to keep a near to normal distribution.

2.2 Main Experiment

Study design

The main experiment was implemented as a 2 x 3 x 6 within-subject design with the factors stadium scene (empty, full stadium), audio condition (binaural audio, stereo, no audio) and stimuli angle in azimuth plane (-135°, -90°, -45° and 45°, 90°, 135°). Each participant completed all tasks and the design was balanced over participants to avoid sequence effects. For a better understanding of the design and the factors (independent variables) an overview is given in Table 1.

Table 1 - Overview of the Experimental Design (independent variables).

		Factor B: Acoustic Setting		
		muted	binaural	stereo
Factor A: Visual Setting	Empty Stadium	-45° -90° -135°	-45° -90° -135°	-45° -90° -135°
		+45° +90° +135°	+45° +90° +135°	+45° +90° +135°
	Full Stadium	-45° -90° -135°	-45° -90° -135°	-45° -90° -135°
		+45° +90° +135°	+45° +90° +135°	+45° +90° +135°

Notes. Factor C (6 levels) represents the stimuli angles in azimuth plane. In the angle specifications, the plus (+) represents the right side and the minus (-) represents the left side.

VR Contents

The **visual stimuli** were produced in such a way that they were recognizable if the subject was lead in their direction, but at the same time not excessively overt. Therefore, they were kept in the same color theme as the rest of the scene, i.e. yellow color of the home team shirts, and thus blend in. For those reasons, the visual stimuli were yellow Minions out of the eponymous movie. Additionally, we used fading and light shadows to further blend the visual stimuli in sync with the audio signal. The experiment was conducted using 360° camera shots filmed in a handball stadium together with a 360° audio recording directly synced with the camera, to capture the sound spatially accurate. This allowed to produce the background noise in a spatial manner and in such a way that did not interfere with the spatial production of the acoustic stimulus. Two scenes were chosen for the main experiment: one scene with an empty stadium and one with a fully occupied stadium.

For the **acoustic stimulus**, a horn signal was used, which fitted in the stadium scene but at the same time had very unique and dominant characteristics. It was recognizable even in the presence of background noise while also being a sound typically associated with the visual target stimuli. The stimulus was then placed in the 6 angles of the azimuth plane using either the Binauralizer from Adobe Premiere Pro on the 4th order Ambisonic files for the *binaural* setting or the volume pan method for the *stereo* setting. The ambient sound was recorded via 3D microphones in a real stadium and stored in 4th order Ambisonic format. Unfortunately, most VR Players were not able to correctly render Ambisonic audio at this time, which means that the sound was not moving relatively to the head movement. As such we decided to use head-locked binaural stereo for the trials with spatial sound and all participants started from the same position to control the spatial angles from which the binaural sound came. For future trials we aim for a solution that has full support for spatial sound formats and thus enable us to have a less restricted design.

Eye tracking in VR

To determine the participants performance to localize the target stimulus, we used the saccadic eye movement patterns resulting from eye tracking measurements with the VR headset. In order to achieve reliable results, the position of the acoustic stimulus was varied while the visual scene remained unchanged except for the visual reference stimulus. The eye tracking information was collected using the Tobii Eye tracker retrofitted to an HTC Vive. The data were evaluated using the software and algorithms included in the Tobii ProLab software. The detection count was based on the participant fixating the predefined area of interest for at least 60 ms (time between fixations ≥ 75 ms) during the time interval the target was visible. The time to first fixation (TFF) was then taken accordingly as the difference in time from the appearance of the temporally and spatially synced visual and acoustic stimuli to the beginning of the first encountered fixation interval. The spatial fixation was measured as any recognized fixation in time and space. To retrieve the fixation distribution over time, the following spatial window was chosen to count the number of fixations: window speed=30 m/s, angle=0.5°.

Questionnaires

The **NASA-TLX** (12) questionnaire was used to measure subjectively perceived workload on six dimensions, namely mental demands, physical demands, temporal demands, performance, effort and frustration. In the present study, the judgments were given on a 100-point scale. The two poles of the NASA TLX are verbally indicated with high („hoch“) and low (“niedrig“) by the experimenter.

The **igroup Presence Questionnaire** (IPQ) (13) was administered as a kind of treatment check (direct indicator of perceived presence) for the whole VR experiment.

Procedure of the experiment

At first, the eye tracker was calibrated for each participant. After this, the scene started with looking at a centered blue fixation cross. The cross then disappeared, coinciding with the acoustic signal (auditory search cue) in the binaural and stereo condition, and the participants had to search for the visual target stimuli (Figure 1). They were asked to fixate the target as soon as they found it until it disappeared. The visual stimuli disappeared after 5 seconds indicated by red arrows pointing towards the blue fixation cross, leading the participant back to the starting position. After six stimuli were shown, i.e. once in every one of the six angles, the video did end and participants had to fill out the NASA-TLX. This procedure was then done with the other five videos as well, varying the sound setting and stadium setting per video. After the sixth video, the IPQ was also given to the participants.

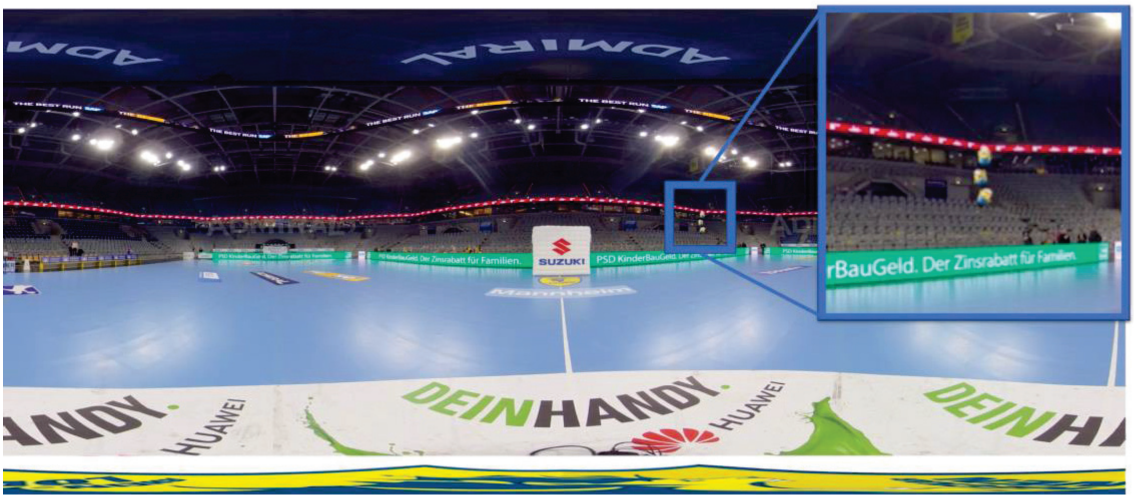

Figure 1 - Representation of the visual target, i.e. three Minions, in the VR environment

3. Results

3.1 Pretest 1

The first pretest was conducted to test whether participants could perceive a difference between binaural audio and stereo. As Figure 2 indicates, depending on sound files, 8-22% of the participants stated that they could not distinguish between the two signals. Specifically, 18% of the participants indicated not being able to differentiate binaural audio from stereo for the stadium sample, which we are using for our main experiment. In contrary to our lab experiments we were not able to control the quality of the used hardware in the online study. Assuming that the hardware used in everyday life is on average lower regarding audio quality than the one used in our lab, we can conclude that the majority of participants should be able to perceive a difference between binaural and stereo sound in our main experiment.

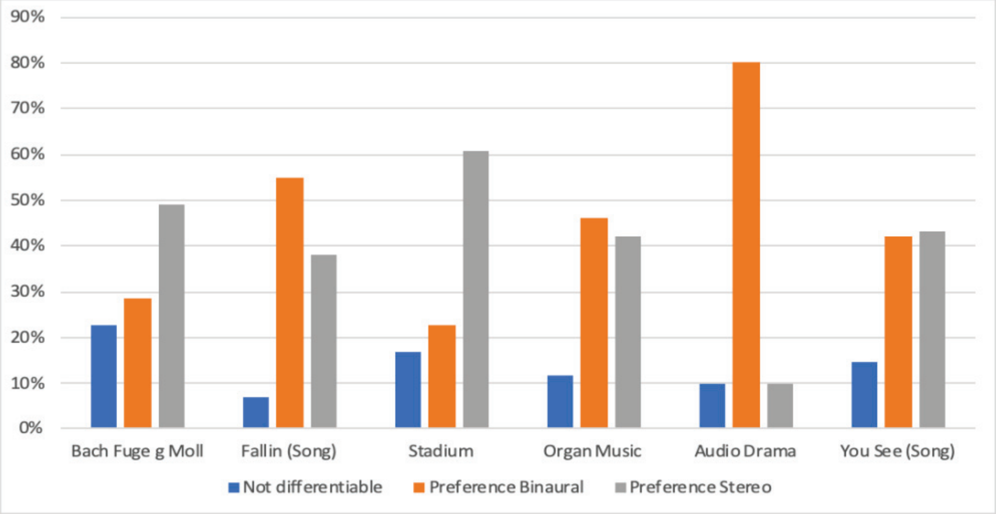

Figure 2 – % of participants indicating a preference for binaural audio, stereo audio or not hearing a difference between the two ($N = 102$).

3.2 Pretest 2

In the second pretest we aimed to find out how many spatial angle variations of an acoustic stimulus can be reliably identified by participants. This was done to effectively limit visual and auditory stimulus sets to the behavioral relevant variations and thus the number of trials for the subsequent main experiment. Localization errors and variation of localization performance decreased the closer the sound sources were located to the axis of the ears. For the signal-to-noise ratio no significant effects were found. The following means were derived: For elevation perception (horizontal plane) $M = 40.28°$ ($n = 340$, $SD = 34.06°$), for azimuth variations (vertical plane) $M = -8.86°$ ($n = 340$, $SD = 91.50°$). The highest errors and variations were observed for 0° and 180° (cp. Figure 3). Opting out these values as our true negatives, leads to a mean of $M = 3.3°$ ($n = 270$, $SD = 60.61°$) elevation perception (horizontal plane).

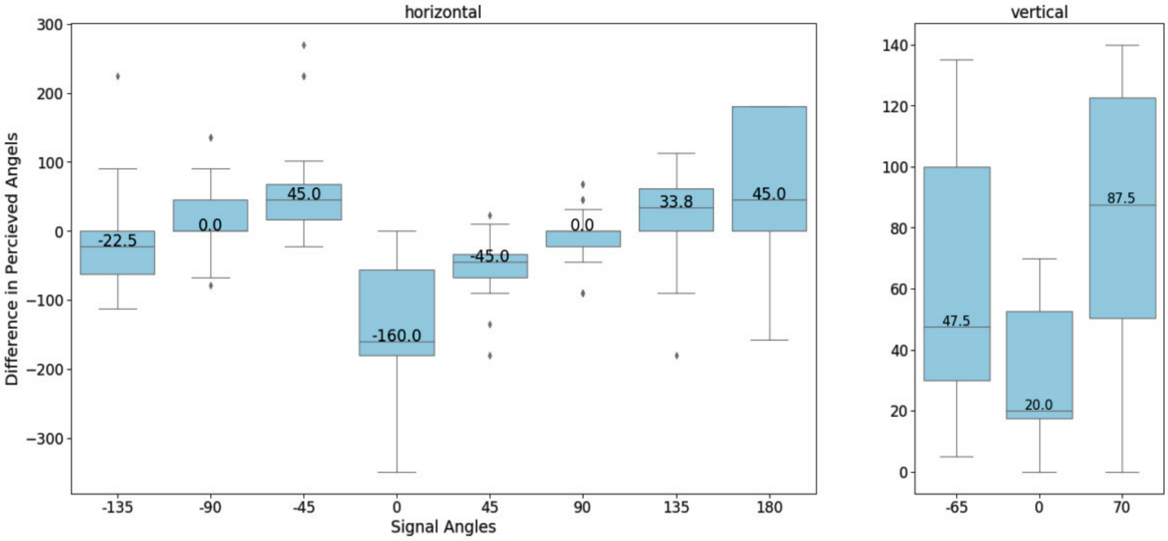

Figure 3 - Boxplot of the differences between marked angles and actual spatial angles of the sound source (8 angles); *N* = 17

3.3 Preliminary Results from the main experiment

Results for localization performance: The data collection was not yet fully completed at the time of reporting, so that preliminary results are reported below for *n* = 16 participants. Each participant solved 36 search tasks in the different settings, namely 2 stadium scenes (empty vs. full stadium) x 3 audio conditions (binaural audio, stereo, no audio) x 6 trials (-135°, -90°, -45° and 45°, 90°, 135°). The previous VR experience of the participants was on average *M* = 2.83 on a scale from 1 to 5, 1 "being the first time in VR" and 5 "living in the VR" (e.g., daily using VR applications). For the previous acoustic experience, the average was *M* = 3.5 (1 being "judging sound as negligible" and 5 "being an audio enthusiast"). The time to first fixation (TFF) was *M* = 2.12 seconds (*SD* = 701).

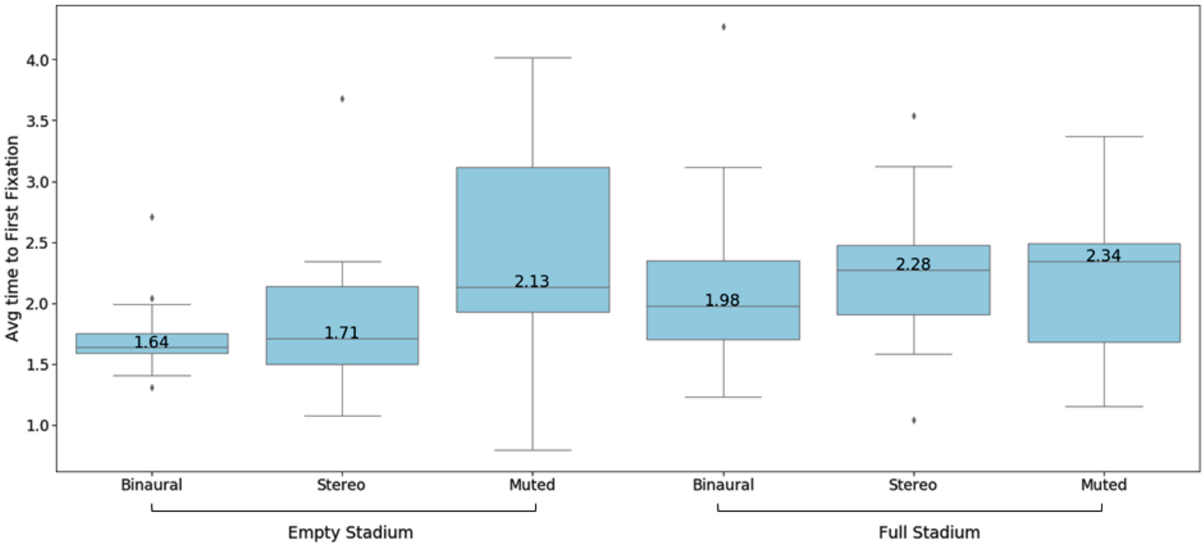

Figure 4 - Boxplot for the average time to first fixation for all trials and participants (in seconds after the stimulus). The results are grouped by the visual and acoustic setting.

The results are summarized in Figure 4. If no acoustic cues were given (cp. Empty Stadium Muted; Full Stadium Muted), the participants varied very strongly in their localization performance. This applies to the empty stadium condition as well as to the full stadium condition (Empty Stadium Muted: *SD* = 0.96; Full Stadium Muted: *SD* = 0.65). Differences in scattering between the binaural and stereo condition are not pronounced (see figure 3). However, in the full-stage condition binaural cues are likely to lead to faster localization compared to stereo sound cues (*MD Binaural* = 1.98 vs. *MD* Stereo = 2.28), whereas in the empty-stage condition the differences are nominally smaller. In view of the incomplete study and the small number of participants (*n* = 16) so far, no inference statistics were calculated and the eye tracking results are still under analysis.

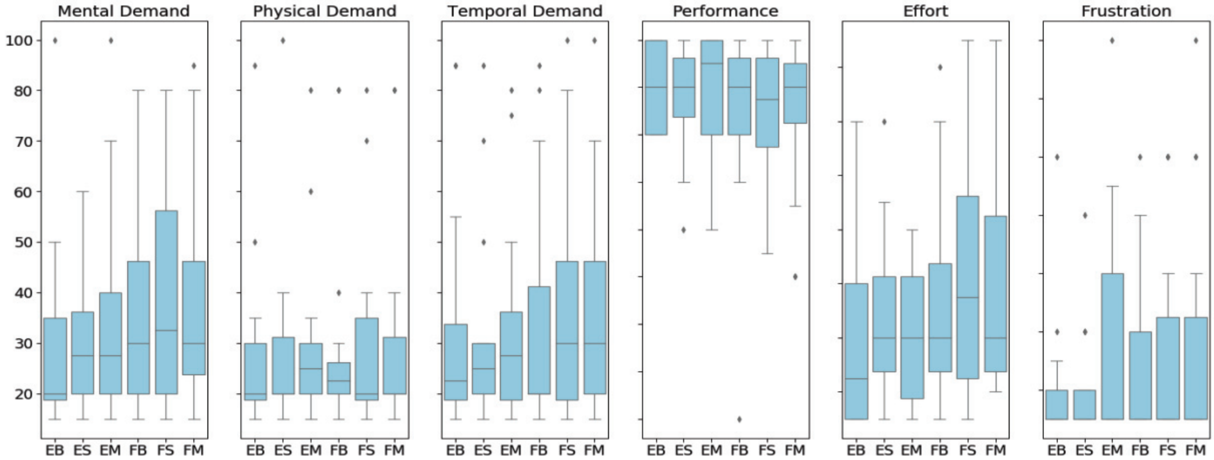

Figure 5 - Boxplots of the results of the NASA TLX questionnaire for each of the six conditions (EB= empty stadium binaural, ES= empty stadium stereo, EM= empty stadium muted, FB= full stadium binaural, FS= Full stadium stereo, FM= full stadium muted).

Results NASA-TLX: After participants had worked on the search trials corresponding to one of the 6 experimental conditions (stadium scene x sound condition), the NASA-TLX were given to assess subjectively perceived mental workload. As Figure 5 visualizes, there is not much variation between experimental conditions on a specific dimension of the NASA-TLX. Statistical testing is reasonable when data from a larger participant sample is available.

Results IPQ: The IPQ consists of 14 items with a seven-point Likert scale (0 to 6). The 14 items load on the four factors Spatial Presence (SP), General Presence (GP), Involvement (INV) and Experienced Realism (REAL). SP was rather high (*M* = 4.2, *SD* = 1.0), the single item factor GP (*M* = 3.7, *SD* = 1.5), INV (*M* = 3.3, *SD* = 1.4) and REAL (*M* = 2.6, *SD* = 1.1) were expectedly lower.

4. DISCUSSION AND CONCLUSIONS

The present study aimed to explore whether binaural 3D sound "matters" for user performance in audiovisual VR environments. This was tested exemplarily with respect to the effects of auditory cues on visual search performance in a vivid, audiovisual 3D VR scene. Reaction time in a visual search task and eye movements were measured as indicators for presence. We compared visual search performance during three conditions: [1] without auditory cue, i.e. vision-only condition (baseline condition), [2] presentation of the auditory search cue in binaural quality, and [3] in stereo quality.

The preliminary results of our main experiment indicate that an auditory search cue increases speed and accuracy of localizing the corresponding visual target when the auditory cue is presented binaurally instead of stereo. However, the auditory cue presented in stereo quality also helped localization performance. In fact, the only trials, in which participants failed to locate the visual target were trials in which visual search had to be performed without auditory cues. Do note, however, that no differences between the three experimental conditions (binaural auditory cues, stereo auditory cues, no auditory cues) could be statistically supported with respect to eye tracking data analyzed via heat maps as well as regarding mental workload measured by the questionnaire NASA-TLX.

As for the current state of the art, there was no appropriate VR Player that supports the ambisonic sound format, which is needed to have a truly object-based audio experience. In future experiments we would like to investigate how the participants perform if they are not forced to look at a given position at the start of each search trial. We expect that in such a setup the differences between stereo and binaural will increase even further as the search task is not as guided as it was in this study. For future research, we want to further investigate the effects of different audio settings, e.g. binaural, stereo and Ambisonic sound, and take hardware and software parameters into account as well.

Regarding the presence measurement via the IPQ, we found out that the participants struggled applying the questions to the scene and the different acoustic settings they perceived. In fact, the IPQ questionnaire is mostly focused on the visual quality of a virtual scene and does not specifically ask for the perception of sound. In order to measure the effect of different audio settings on presence, additional items and measures would need to be added in future experiments.

With the chosen experimental VR setup, there were tendencies of positive effects of spatially plausible acoustic cues in a visual search task and thus of cognitive performance in a visual task being supported by plausible co-visual acoustic cues. Furthermore, we could show that there is a difference between binaural and stereo, with binaural leading to slightly better effects. Using a controlled VR environment with only minor artificial changes for the targets, we were able to achieve a quite plausible setup, while at the same time having full control over our variables and measurements.

This experiment showed that the auditory aspect of presence should not be overlooked, as we did find slightly better performance in binaural scenes than in stereo scenes, even with our small sample size. In general, achieving a high level of presence can be a crucial aspect for the usability, user experience (UX) and acceptance of VR technologies. The "perfect" level of presence cannot be specified in principle, as it highly depends on the application and the intended use. Applications like VR-based emergency trainings, training in flight simulators or psychotherapies require high presence and plausibility for the desired training effects to have the potential to generalize into real life performance.

REFERENCES

1. Slater M, Wilbur S. A framework for immersive virtual environments (FIVE): Speculations on the role of presence in virtual environments. Presence: Teleoperators & Virtual Environments. 1997;6(6):603–616.
2. Möller S, Raake A. Quality of experience: advanced concepts, applications and methods. Springer; 2014.
3. Schubert T, Friedmann F, Regenbrecht H. The experience of presence: Factor analytic insights. Presence: Teleoperators & Virtual Environments. 2001;10(3):266–281.
4. Sherman WR, Craig AB. Understanding virtual reality: Interface, application, and design. Morgan Kaufmann; 2018.
5. Doyle MC, Snowden RJ. Facilitation of visual conjunctive search by auditory spatial information. In: Perception. Pion LTD 207 Brondesbury Park, London NW2 5JN, England; 1998. p. 134–134.
6. Perrott DR, Sadralodabai T, Saberi K, Strybel TZ. Aurally aided visual search in the central visual field: Effects of visual load and visual enhancement of the target. Human Factors. 1991; 33(4):389–400.
7. Van der Burg E, Olivers CN, Bronkhorst AW, Theeuwes J. Audiovisual events capture attention: Evidence from temporal order judgments. Journal of vision. 2008;8(5):2–2.
8. Hoeg ER, Gerry LJ, Thomsen L, Nilsson NC, Serafin S. Binaural sound reduces reaction time in a virtual reality search task. In: 2017 IEEE 3rd VR workshop on sonic interactions for virtual environments (SIVE). IEEE; 2017. p. 1–4.
9. Flanagan P, McAnally KI, Martin RL, Meehan JW, Oldfield SR. Aurally and visually guided visual search in a virtual environment. Human factors. 1998;40(3):461–468.
10. Nelson WT, Hettinger LJ, Cunningham JA, Brickman BJ, Haas MW, McKinley RL. Effects of localized auditory information on visual target detection performance using a helmet-mounted display. Human factors. 1998;40(3):452–460.
11. Perrott DR, Cisneros J, McKinley RL, D'Angelo WR. Aurally aided visual search under virtual and free-field listening conditions. Human factors. 1996;38(4):702–715.
12. Hart SG. NASA-task load index (NASA-TLX); 20 years later. In: Proceedings of the human factors and ergonomics society annual meeting. Sage publications Sage CA: Los Angeles, CA; 2006. p. 904–908.
13. Witmer BG, Jerome CJ, Singer MJ. The Factor Structure of the Presence Questionnaire. Presence: Teleoperators and Virtual Environments. 2005 Jun;14(3):298–312.