九州大学学術情報リポジトリ
Kyushu University Institutional Repository

The Impact of Sound Systems on the Perception of Cinematic Content in Immersive Audiovisual Productions

Korshunova, Victoria
Faculty of Sound in New Media, Department of Media, Aalto University

Remijn, Gerard Bastiaan
Faculty of Design, Department of Human Science, Kyushu University

Elischka, Synes
Department of Film, Television and Scenography, Aalto University

Mendonça, Catarina
Department of Signal Processing and Acoustics, Aalto University

https://hdl.handle.net/2324/7449713

出版情報：pp.1-5, 2019-05-09. IEEE
バージョン：
権利関係：© 2019 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

KYUSHU UNIVERSITY

The Impact of Sound Systems on the Perception of Cinematic Content in Immersive Audiovisual Productions

1st Victoria Korshunova
Faculty of Sound in New Media, Department of Media
Aalto University
Espoo, Finland
viktoriya.korshunova@gmail.com

2nd Gerard B. Remijn
Faculty of Design, Department of Human Science
Kyushu University
Fukuoka, Japan
remijn@design.kyushu-u.ac.jp

3rd Synes Elischka
Department of Film, Television and Scenography
Aalto University
Espoo, Finland
synes.elischka@aalto.fi

4th Catarina Mendonça
Department of Signal Processing and Acoustics
Aalto University
Espoo, Finland
mendonca.catarina@gmail.com

***Abstract*—With fast technological developments, traditional perceptual environments disappear and new ones emerge. These changes make the human senses adapt to new ways of perceptual understanding, for example, regarding the perceptual integration of sound and vision. Proceeding from the fact that sound cooperates with visual attention processes, the aim of this study is to investigate the effect of different sound design conditions on the perception of cinematic content in immersive audiovisual reproductions. Here we introduce the results of a visual selective attention task (counting objects) performed by participants watching a 270-degree immersive audiovisual display, on which a movie ("Ego Cure") was shown. Four sound conditions were used, which employed an increasing number of loudspeakers, i.e., mono, stereo, 5.1 and 7.1.4. Eye tracking was used to record the participant's eye gaze during the task. The eye tracking data showed that an increased number of speakers and a wider spatial audio distribution diffused the participants' attention from the task-related part of the display to non-task-related directions. The number of participants looking at the task-irrelevant display in the 7.1.4 condition was significantly more than in the mono audio condition. This implies that additional spatial cues in the auditory modality automatically influence human visual attention (involuntary eye movements) and human analysis of visual information. Sound engineers should consider this when mixing educational or any other information-oriented productions.**

Keywords—spatial sound; eye tracking; eye gaze; immersive environments; audiovisual perception; sound systems

I. INTRODUCTION

Science and technology continuously seek for advanced forms of sensory reproduction systems. For example, just for the past two decades interest in virtual reality (VR) technology has been growing, decreasing and growing again, and now VR is a huge sensation in entertainment, arts and research industries. It is predicted already that different types of VR systems can differently influence human mental states and perception, and affect social behavior in the real world [1].

Experiencing audiovisual productions in a deeply engaging space can be extremely surprising for us, since people receive audio events from unusual directions with visual content distributed on much broader displays. We learn to adapt to this, but to what extent?

Regarding spatial information that can be received by our ears, the human auditory system picks up directional cues by means of the pinna, which shape the spectral content of sound, as well as interaural time- and level differences between the two ears [2]. However, we can question our capability to perceive all the spatial cues arriving from immersive audiovisual systems, for example from those with 270- or 360-degree sound [3].

Previous research by Mendonça et al [4] has shown that too much auditory spatial information can have a negative impact on the perception of visual information in immersive audiovisual situations. The perception of visual information is often investigated by means of divided and selective attention tasks. In a divided attention task, participants perform a task, e.g., counting persons, while focusing on various locations in their visual field. In a typical selective visual attention task, the task is performed with focus on a single location [5]. In Mendonça et al [4], participants performed divided and selective visual attention tasks in an audiovisual environment. Visual scenes were recorded on a spherical video camera system, while three sound conditions were used: mono, stereo and Directional Audio Coding (DirAC; see [6]). During the visual attention tasks, the participants wore an eye-tracking device. The results showed that a spatially defined sound outside of the field of view had a negative influence on visual stability, specifically in the selective attention tasks [4]. In this type of attention task, participants usually follow only one event, while ignoring others. In accordance with psychophysical studies [7], Mendonça and colleagues [4] thus showed that people continuously switch their attention between the auditory and visual modality and that sound can automatically and involuntarily guide visual attention.

Following these findings, the goal of the present study was to test how human perception of visual events in a professional short film reproduction can be affected by sound systems with various numbers of speakers, i.e., mono (1 speaker), stereo (2 speakers), 5.1 (5 speakers and a subwoofer) and 7.1.4 (7 speakers and a subwoofer).

Nowadays, surround sound condition 5.1 and spatial audio condition 7.1.4 are becoming increasingly popular and are more likely to be found in regular households [8]. Based on Mendonça et al [4], it is conceivable that an increase in spatial auditory information, i.e., an increase in the number of sound sources (speakers), negatively influences the participants' performance on a visual attention task.

In this study, four groups of thirteen people performed individually different tasks while watching scenes from the VR movie "Ego Cure" [9], with one out of the four audio conditions mentioned above.

We here present the results of a selective visual attention task, from the first scene, in which we asked subjects to count how many falling boxes they could see in the scene. This task was explained before the playback of the movie. The display was divided into three parts: A, B and C, in which A was a non-task related display, B was a task-related display, and C was ambiguous. During the task, the participant's eye gaze was tracked by an eye-tracking system (Tobii Pro Glasses).

II. Experiment

A. Method

1. Participants

Fifty-four participants (12 females, 42 males) participated in this experiment on a voluntary basis. The data of two participants were discarded due to technical failure with Max/MSP interface in the beginning of the experiment. The participants were in between 20-50 years of age (31.68 ± 6.84 years). All participants had normal vision or corrected-to-normal vision. Those with corrected-to-normal vision were in a range from around +3 farsighted and around -7.5 nearsighted. All participants reported having normal hearing and were naive regarding the purpose of the study. Besides the 52 participants who participated in the experiment, three additional participants, who were familiar with the experiment and the stimuli, took a part in an informal pilot study to secure the functioning of the experimental setup. These informal data were not included in the statistical analysis of the experiment. All 52 participants gave informed consent as to their participation after the purpose and procedure of the experiment was explained to them.

2. Settings

The tests were performed in an immersive audiovisual environment at the Department of Signal Processing and Acoustics at the Aalto University School of Electrical Engineering. The setting was situated inside an acoustically treated space and comprised three HD video projectors and 40 loudspeakers (Genelec 1029). The screens were 2.5 × 1.88 m each and pentagonal. The resolution of the video material was transformed into 4320 × 1080 pixels, more specific details can be found in [10].

The audio system arrangement for mono, stereo, 5.1 and 7.1.4 was built based on the information of azimuths, elevation and distance from the audio control booth used for mixing and mastering. It was equipped with Avid Pro Tools 12 HDX2 system and Mac OS, 16 channels Avid S6 control surface, a 2 × 2.7 m projection screen and a 9.1.4 Dolby Atmos speaker setup with Genelec speakers.

During mixing, plugin Spanner was one of the main panning instruments for complete independent control of every channel. We used 12 loudspeakers out of 40 available for this test. Participants interacted with the system via a tablet computer with a touch screen interface. This interface was developed specifically for an experiment with a Max/MSP visual programming language.

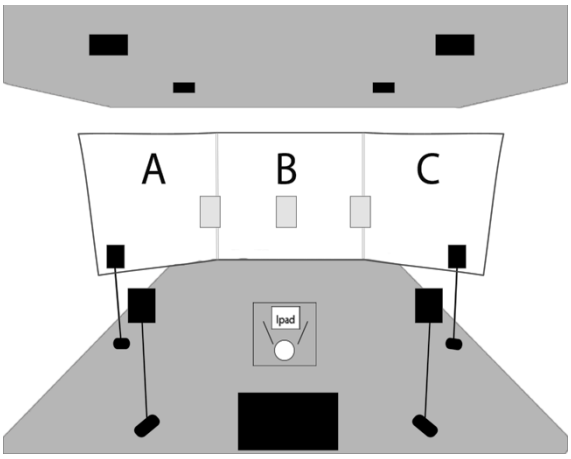


Fig. 1. Graphical representation of the audiovisual environment in which the experiment took place. There were three separated display areas (A, B, C) and sound system configurations consisting of mono, stereo, 5.1 and 7.1.4.

3. Stimuli

The visual stimulus content was a 9-second long scene from a movie provided by author Synes Elischka, a Doctoral Candidate at the Department of Film, Television and Scenography in Aalto University. The video movie is called "Ego Cure" and it is the first VR film made in Finland [9]. It was shot on 360 RED Epic cameras with a 6K high-quality footage. The sound was recorded with boom microphones and standard audio recording techniques for movie production.

The movie was a pilot project produced in 2017 by Aalto University ELO Film School Helsinki in collaboration with Aalto University Media Lab. Originally it was created for a VR headset and binaural headphone listening mixed by sound designer Jorma Kaulanen. In total, over sixty people including professional freelancers and students were involved in the production of Ego Cure.

The visual stimulus was accompanied by four different audio conditions: mono, stereo, 5.1 and 7.1.4. All conditions were professionally mixed and mastered in a sound studio. The idea was to create less point-like and more diffused audio outside the task-related area, i.e., the display (B) that participants needed to view in order to perform the task. Parameters of the studio (azimuths, elevation and distance) were provided and closely replicated in the experimental audiovisual environment.

4. Test procedures

Each participant was seated in the center of the audiovisual system and had a tablet computer with a touch screen (Apple iPad Mini) in their hands. During the whole experiment, participants wore Tobii Pro 2 eye-tracking glasses which allows participants to move their head freely while watching the displays. This eye-tracking device recorded sound, images, and gaze coordinates for data analysis.

The participants were asked to perform a visual selective attention task: counting how many falling boxes they could see in the movie scene (Fig. 2 for a snapshot). The total number of falling boxes was 17. Fifty-two participants were

randomly divided into four groups and had one sound condition during the task.



Fig. 2. Snapshot of a scene from the movie "Ego Cure" used in the experiment. The task-relevant information (falling boxes) appeared on the right edge of display B (red line).

Before the start of the experiment each participant practiced a trial with no audio conditions and completely different visual stimuli. Still pictures with white circles on a black background were used as examples to familiarize participants with the Max/MSP interface so as to prevent questions and mistakes.

None of the participants saw the original material used for the main experiment. The playback of used audiovisual stimulus was under control of the participant. When the participant was ready, he/she could touch a "Play video" button, read the task, watch the video and input the total number of counted boxes. The average duration of the experiment was around 15 minutes per participant.

B. Results

Fig. 3 shows an example of a gaze plot with the distribution of the participants' fixations around the displays (blue dots) To analyze the eye-gaze data, the screen display was divided in three parts: A, B and C. A was a non-task related display, B was a task-related display and C an ambiguous display (see Figure 1). Since the task-relevant information (falling boxes) appeared on the right edge of task-relevant display B, we assumed that the display bordering that area, display C, would also be checked for any falling boxes. Hence, we called display C "ambiguous".

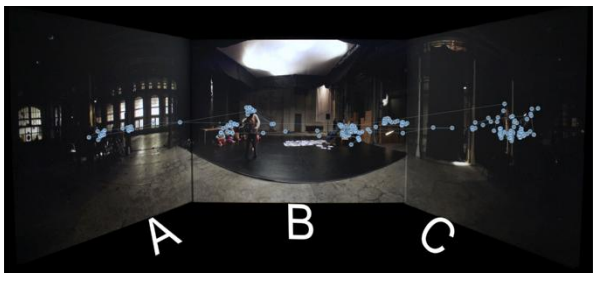


Fig. 3. Gaze plots (blue dots) from one participant in audio condition 7.1.4.

The task of the participant was to count the number of falling boxes in task-relevant display B. There were 17, but most participants reported a far lower number on average (Fig. 4). A non-parametric Friedman test showed that the audio condition (mono, stereo, 5.1, or 7.1.4) had no significant influence on the perceived number of falling boxes (X^2=3.7, df=3, p=0.31).

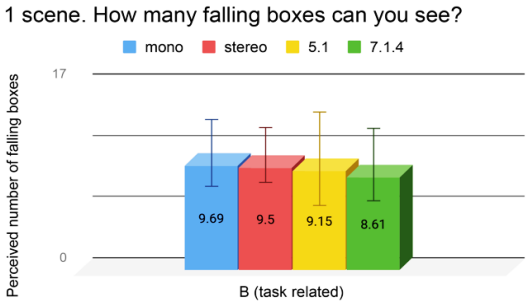


Fig. 4. Perceived number of falling boxes in mono, stereo, 5.1 and 7.1.4 audio conditions (1scene).

For each of the three displays (A, B, C), we counted whether a participant had gazed at the display (Fig. 5). The dichotomous data were subjected to the non-parametric Cochran Q test to compare the data for the four sound conditions. The number of conditions (k=4) times the number of participants (n=13) for each sound condition yielded a sufficient sample size.

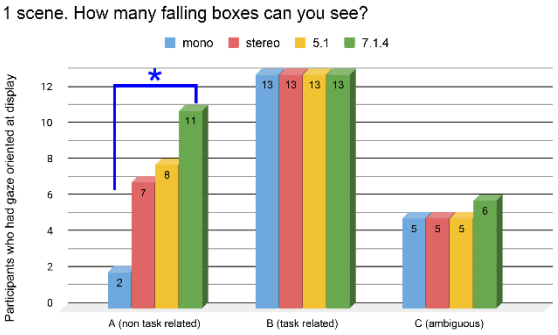


Fig. 5. Gaze oriented at the display in one scene in areas A, B or C for mono, stereo, 5.1 and 7.1.4 audio conditions.

The difference between the number of participants who had gazed at display A (non-task related), display B (task-related), and display C (ambiguous) in the mono audio condition was statistically significant (Q=16.17, df=2, p=0.001**). Post-hoc tests (p<0.05) showed that more participants looked at the task-related display than at the other two displays. In the stereo condition (Q=9.45, df=2, p=0.009**), significantly more participants looked at the task-relevant display than at the ambiguous display. Interestingly, the difference in the number of participants who looked at the task-relevant display and the non-task relevant display was not significant. The same was found for the 5.1 condition (Q=9.8, df=2, p=0.007**) and the 7.1.4 audio condition (Q=11.14, df=2, p=0.004**). In both cases, significantly more participants looked at the task-relevant display than at the ambiguous display. The difference between the task-relevant display and the non-task relevant display, however, was not significant.

Related to the research question, we analyzed whether audio condition influenced the number of views at each of the display areas. Fig. 5 shows that, regardless of audio condition, all 13 participants had looked at the task-related display B. Fewer participants on average had looked at the ambiguous display C, but a non-parametric Cochran's Q test did not show any significant difference between audio conditions (Q=0.22, df=3, p=0.974). For non-task related display A, audio condition affected the number of participants who viewed at

the display. Cochran's Q test showed a significant difference ($Q=12.60$, $df=3$, $p=0.006$). Only 2 out of 13 participants had looked at display A when the task was done in a mono audio condition, whereas 11 out of 13 participants had looked at display A when the audio condition was 7.1.4. A post hoc non-parametric sign test showed that this difference was statistically significant with Bonferroni correction ($0.05/6=0.0083$) on the alpha- level ($p<0.004$). No other statistically significant differences were found, although the stereo condition (7 out of 13 participants) and the 5.1 condition (8 out of 13 participants) attracted a considerable number of participants to non-task related display A as well.

More detailed eye-tracking data are shown in figures 6-13. They show that as the audio condition involved more speakers, from mono audio to 7.1.4, the averaged viewing locations (n=13) over displays A, B, and C became more and more diffuse. The colors of the heat maps illustrate the number of total fixations for 13 participants per condition. Transparent green: 100-200 fixations; Solid green: 200-300 fixations; yellow: 300-400 fixations; orange: 400-500 fixations; red: above 500 fixations.

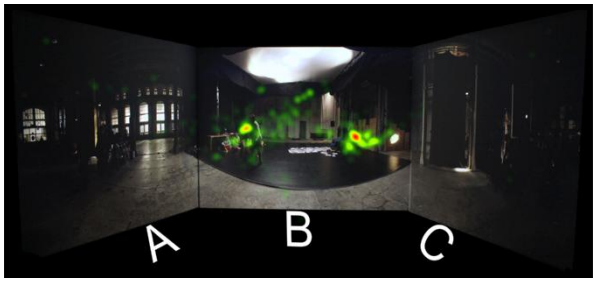


Fig. 6. Heat map for the mono audio condition before boxes start falling (1 scene).

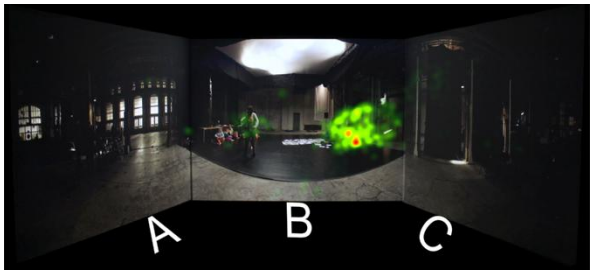


Fig. 7. Heat map for the mono audio condition while boxes are falling (1 scene).

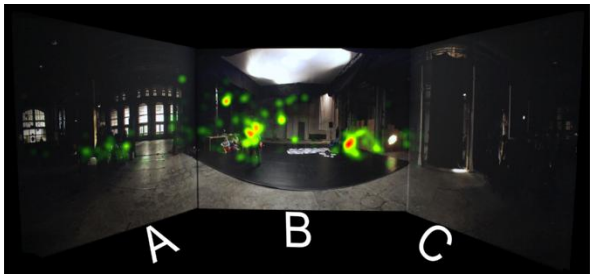


Fig. 8. Heat map for the stereo audio condition before boxes start falling (1 scene).

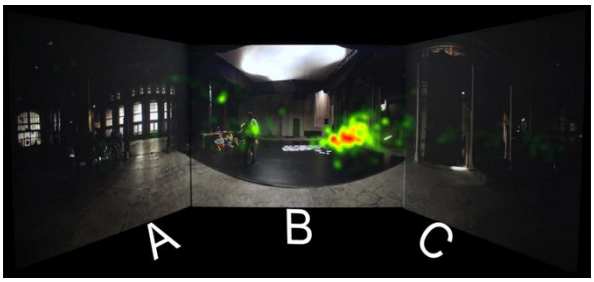


Fig. 9. Heat map for the stereo audio condition while boxes are falling (1 scene).

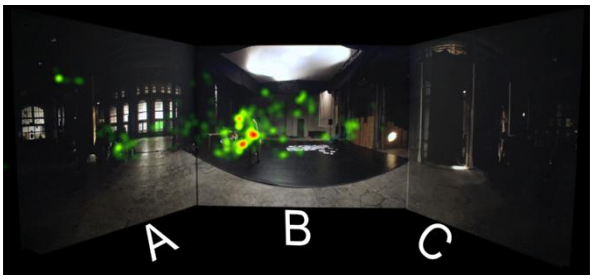


Fig. 10. Heat map for the 5.1 audio condition before boxes start falling (1 scene).

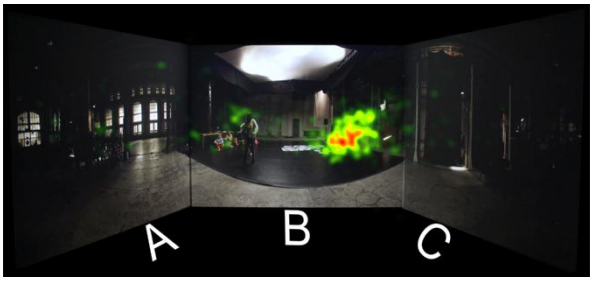


Fig. 11. Heat map for the 5.1 audio condition while boxes are falling (1 scene).

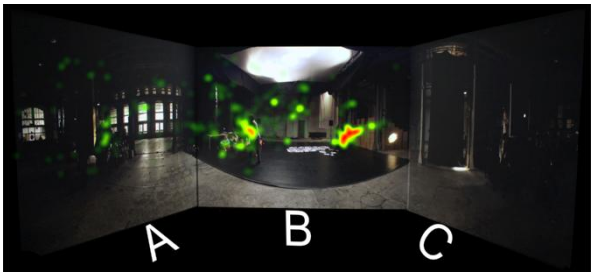


Fig. 12. Heat map for the 7.1.4 audio condition before boxes start falling (1 scene).

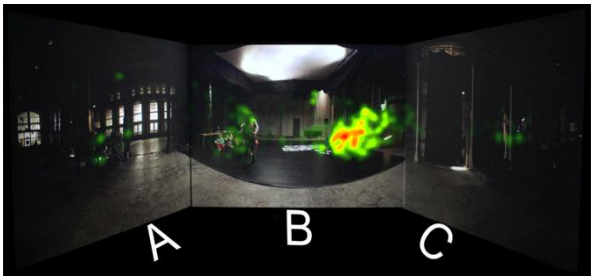


Fig. 13. Heat map for the 7.1.4 audio condition while boxes are falling (1 scene).

The analyzed data and heat maps in the scene clearly show a larger fixation distribution for stereo, 5.1 and 7.1.4 audio conditions compared to mono audio, especially on unwanted, non-task related display (A).

III. Discussion & Conclusion

In this study participants watched a 270- degree immersive audiovisual display, on which a movie was played. The participants were asked to perform a selective visual attention task while wearing an eye-tracking device and listening to one of four audio conditions. These audio conditions employed an increasing number of loudspeakers, i.e., mono, stereo, 5.1 and 7.1.4. The eye-tracking device (Tobii Pro glasses) recorded the participant's eye gaze during the task. Although it is mainly agreed that audio spatialization has positive effects on the perception of presence [11] and quality [12], the objective of this experiment was to further investigate whether an increase in spatial auditory information, i.e., an increased number of sound sources (speakers) used in regular households, would influence the participants' performance on a visual attention task.

Previous research has shown that 3D sound might have a negative influence on the perception of visual events outside the field of view, specifically in a selective attention task [4]. In this research not only task errors but also the number and duration of eye fixations within specific areas of interest were analyzed. In accordance with their findings, in the present study, which employed different type of scenes, tasks, sound systems and data analysis methods, we found that sound systems with more than one speaker attracted the participants' attention to a non-task related display. Compared to the mono audio condition, significantly more participants watched the display unrelated to the visual selective attention task in the 7.1.4 audio condition. The 5.1 audio and the stereo condition showed the same tendency but not significantly here.

These findings must be practically implemented and taken into account by audio production professionals. Sound field practitioners should properly study the usage of new technology in order not to scare and overpower the influence of audio experience on visual perception. Instead it seems recommendable to concentrate on simplicity and accuracy of reproduction techniques. Mono and stereo setups will less distract users' visual attention from task-relevant displays.

So how can we learn to completely focus our visual attention when listening to materially layered structure sounds? How much time do we need to adapt to these perceptually complex conditions? These questions are still open and require further investigation. It is important to develop specific rules and regulations while mixing and mastering recorded material for immersive audiovisual environments, so the user is not distracted by the waterfall of events coming simultaneously from several directions.

Acknowledgments

The research leading to these results has received funding from the European Union's Horizon 20202 research and innovation programme under the Marie Sklodowska-Curie grant no. 659114 and Aalto ARTS Supported Internship Programme. The procedures described here have been approved by the Aalto University Ethics Committee, Finland.

We would like to thank to all 54 persons, who voluntarily participated in the experiment and made this research possible. Very special gratitude goes to Aalto University and its Department of Signal Processing and Acoustics, Department of Media and Department of Film, Television and Scenography. We are also grateful to Kyushu University and Department of Human Science, especially to the PhD students of Perceptual Psychology Laboratory, who provided important comments to improve this manuscript. Thanks go to the following people for their technical assistance: Juha Tanhuanpaa, Tarmo Simonen, Aleksi Oyry, Vilppu Pekkanen and Janne Riionheimo. And also, great appreciation to Javier Gómez Bolaños and Archontis Politis for their enormous advice during experiment development, to Jorma Kaulanen for provided audio material, to Antti Ikonen for his mentoring, to Professor Ville Pulkki and Professor Yoshitaka Nakajima for their guidance.

References

[1] T. Mazuryk and M. Gervautz, "Virtual Reality History, Applications, Technology and Future History," Virtual Real., no. January, pp. 58–60, 1997.

[2] D. W. Batteau, "The role of pinna in human localization," Proc. R. Soc. London, Ser. B 168, pp. 158–180, 1967.

[3] H. Schulze, The Sonic Persona: An Anthropology of Sound. Materializing listening: Dispositives of Surround Sound, Bloomsbury Academic; Reprint edition, pp. 181-19, 2018.

[4] C. Mendonça, O. Rummukainen, and V. Pulkki, "3D sound can have a negative impact on the perception of visual content in audiovisual reproductions," in 21st International Conference on Auditory Display (ICAD2015), Graz, Austria, pp. 1-6, 2015.

[5] Y. Liu and C.D. Wickens, "Visual scanning with or without spatial uncertainty and divided and selective attention," Acta Psychologica, vol. 79, no. 2, pp. 131-153.

[6] V. Pulkki, "Spatial sound reproduction with directional audio coding," Journal of the Audio Engineering Society, vol. 55, no. 6, pp. 503–516, 2007.

[7] S. Shomstein and S. Yantis, "Control of attention shifts between vision and audition in human cortex," The Journal of Neuroscience, vol. 24, no. 47, pp. 10 702–10 706, 2004.

[8] Dolby Laboratories, Inc., [white paper], Dolby Atmos: Next Generation Audio for Cinema, 2012.

[9] S. Elischka, "Ego Cure", [VR film]. Aalto ARTS University, Department of Film, Television and Scenography: Aalto University ELO Film School Helsinki, 2015.

[10] J. Gómez Bolaños and V. Pulkki, "Immersive audiovisual environment with 3D audio playback," in Audio Engineering Society 132nd Convention, Budapest, Hungary, pp. 1–9, 2012.

[11] P. Larsson, A. Väljamäe, D. Västfjäl, A. Tajadura-Jimenéz, and M. Kleiner, "Auditory-induced presence in mixed reality environments and related technology," in The Engineering of Mixed Reality Systems. Springer, pp. 143–163, 2010.

[12] A. Väljamäe and A. Tajadura-Jimenéz, "Perceptual optimization of audio-visual media: Moved by sound," Narration and Spectatorship in Moving Images, 2007.