3D SOUND CAN HAVE A NEGATIVE IMPACT ON THE PERCEPTION OF VISUAL CONTENT IN AUDIOVISUAL REPRODUCTIONS

Catarina Mendonça, Olli Rummukainen, Ville Pulkki

Dept. Processing and Acoustics
Aalto University
P O Box 13000 FI-00076 AALTO, Finland
Catarina.Mendonca@aalto.fi

ABSTRACT

There is reason to believe that sound interacts with visual attention mechanisms. Practical implications of that interaction have never been analyzed in the context of spatial sound design for audiovisual reproduction. The study reported here aimed to test if sound spatialization could affect eye movements and the processing of visual events in audiovisual scenes. We presented participants with audiovisual scenes of a metro station. The sound was either mono, stereo, or 3D. Participants wore eye tracking glasses during the experiment and their task was to count how many people entered the metro. In the divided attention task, participants had to count people entering 3 doors of the metro. In the selective attention task, participants had to count how many people entered the middle door alone.

It was found that sound spatialization did not affect the divided attention task. But in the selective attention task participants counted less visual events with 3D sound. In that condition, the number of eye fixations and time spent in the visual area of interest were smaller. It is hypothesized that, in the case of divided attention, the attention is already disengaged and fluctuating, which could explain why sound did not have any additional effect. In the selective attention task, participants must remain concentrated in only one visual area and competing well-spatialized sounds in peripheral areas might have a negative impact. These results should be taken into consideration when designing sound spatialization algorithms and soundtracks.

1. INTRODUCTION

Over the past decades, spatial sound coding and reproduction have undergone a substantial development. Since the introduction of stereo sound much has changed, and modern sound systems completely surround the user in a naturalistic and highly spatialized environment. However, there is reason to believe that audition guides visual attention. If that is the case, then well spatialized sound, with surrounding auditory events that are not limited to the field of view, might interfere by driving attention to unwanted areas.

There are well documented crossmodal interactions in spatial attention [1][2][3]. Information from different sensory modalities may be integrated preattentively [4] to produce supramodal internal representations of space - independent from sensory modality - that can guide attention [5] [2] [6]. When performing simple visual tasks, such as classifying numbers as odd or even, there is an involuntary engagement of attention caused by novelty and change in the acoustic environment [7].

From the neurophysiological point of view, the presentation of an auditory stimulus in temporal and spatial proximity with the visual target affects the saccade-related activity in the midbrain [8]. In terms of saccades, it is well-established that eye movement is affected by any distractor, either in the visual, haptic or auditory modality [9][10]. In the case of selective attention, the type of attention in which one must attend only one element and disregard any competing stimuli, it has been found that humans consecutively shift their attention between the visual and auditory modality [11]. Furthermore, when such shifts occur the attention to the other sensory modality is inhibited. It is therefore reasonable to question whether in the case of video reproduction the type of sound spatialization can affect how visual events are processed.

The goal of the work reported here was to test if different types of sound spatialization could lead to a different perception of visual events, and different visual exploration patterns, in video reproduction. An experiment was devised in which participants watched the audiovisual reproduction of metro station scenes and were instructed to count how many people entered the metro. There were two tasks - either selective or divided attention; and three sound conditions - mono, stereo and 3D. The participants' eye position was tracked during the experiments and data were analyzed in terms of distribution of eye fixations, number and duration of eye fixations within defined areas of interest.

2. METHODS

2.1. Participants

There were 8 participants taking part in this experiment, ages 20-41, one was female. All participants were naive with respect to the purposes of the study. Participants were either students or researchers. All participants were screened to confirm normal visual function. Two participants wore glasses. No subject reported any hearing impairment and all participants have taken a pure-tone audiometric test in the past 3 years. All participants provided written informed consent. The research project was approved by the Aalto University Ethics Committee.

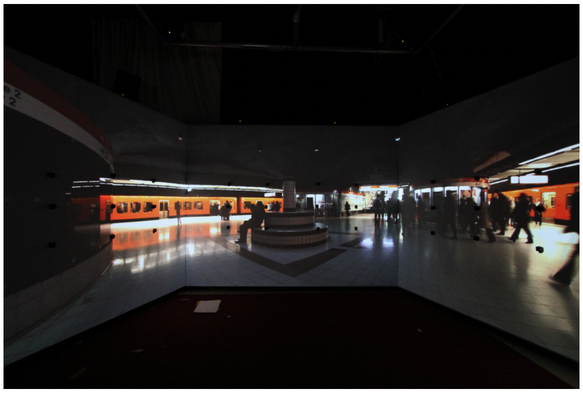


Figure 1: A snapshot of a scene.

2.2. Settings

The experiment took place in a large audiovisual environment, which consisted of three HD video projectors and 29 loudspeakers (Genelec 1029). The loudspeakers were set in a spherical configuration at five elevation levels around the observation position. The loudspeaker grid was more dense behind the projection screens. The image was projected onto three nearly acoustically transparent screens, 2.5 x1.88 m each, following the shape of the base of a pentagon. The participant was seated in the center of the system, at 1.72 m from the center of each screen and 2.1 m from the loudspeaker grid. The total visual resolution of the setup was 4320 x 1080 pixels, resulting in inter-line distance of 3.5 arcmin at the used viewing distance. The projection area extended to the ground and the setup was built in an acoustically treated room. During the experiment there was a noise floor of 35.6 LAeq. Further details of the implementation are found in [12]. The participants provided their responses through a tablet computer with a touch screen (Apple iPad 2).

The stimulus contents were recorded from a metro station in Helsinki using an A-format microphone (Soundfield SPS200) and a spherical video camera system (Ladybug 3). The visual scene was recorded and reproduced at 16 frames-per-second due to computational demands.

During the whole experiment, participants wore the eye tracking device Tobii Glasses, which recorded image and sound and the coordinates of the subject's gaze, while allowing for free head movement. To be able to analyze all trials together, across participants, there were 30 infrared markers placed over the screen and gaze coordinates were analyzed with respect to those markers.

2.3. Stimuli and Procedure

The stimuli consisted of 30-second long audiovisual scenes from a metro station (see Figure 1 for a snapshot). There were three different scenes, but in all of them the same metro was stopped and people entered through the doors. There was a different number of people entering the metro doors in each scene. There were three different sound conditions: mono, stereo and 3D . In the mono condition all the scene sound was summed up and presented from only one loudspeaker, placed above the listener's head. This loudspeaker position was chosen to present both non-spatialized and spatially uncorrelated sound with regard to the image. In the Stereo condition, sound was presented through two loudspeakers, placed to the left and right of the central screen, each summing up the spatial information of the corresponding hemisphere. The 3D sound condition, used Directional Audio Coding (DirAC) [13] and sound was reproduced through the entire spherical grid of loudspeakers. In that sound condition DirAC was used to derive the A-format microphone signals (24 bits/48 kHz) to obtain the loudspeaker signals, which resulted in an ecologically valid reproduction of the original sound scene.

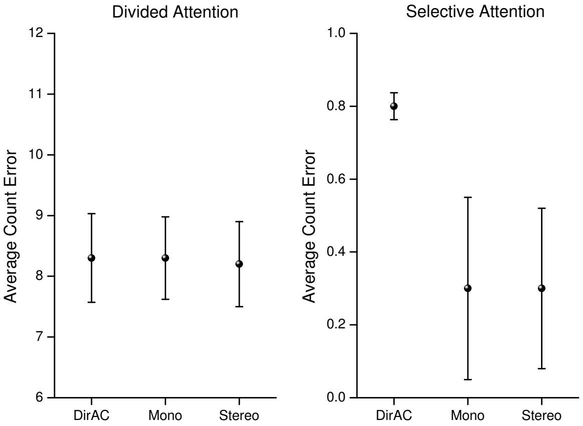


Figure 2: Average count error per experimental condition. Bars display Standard Error of Mean.

There were two tasks. In half of the trials, participants were asked to count how many people entered all three visible metro doors. This was a divided attention task, since it could only be achieved by paying attention to three interest areas at once. The other half of the trials, participants had to count how many people entered the middle door alone. This was a selective attention task, as participants had to disregard all other events and focus only on one area. In the divided attention task, the total number of people entering the doors was 22, 24 and 28 in each of the three scenes. In the selective attention task, the number of people entering the middle door was 6, 8 and 11 in each scene.

There were a total of 6 experimental conditions, corresponding to the three sound conditions and two tasks. Each scene*sound*task combination was repeated randomly three times per subject. Since all scenes were analyzed together, there were a total of 9 repetitions per condition per subject.

Each trial started by stating, in the screen, which task participants had to perform. The screen could state either ”ALL DOORS” or ”MIDDLE DOOR”. When participants were ready, they clicked a ”Continue” button in the tablet. When the trial ended, participants were instructed to input in the tablet the total number of events counted. Participants were expressly instructed to provide the most honest count of how many people were seen entering the metro. They were additionally told that the researchers were not interested in the correct answer but in the true counting process throughout the experiment. There were two breaks during the experiment. The total duration of the experiment was approximately 50 minutes per subject.

3. RESULTS

Results are presented in this section, first with respect to the data obtained from the counting task, and second with respect to eye tracking data.

3.1. Counting task

The average error in the counting task per experimental condition is shown in Figure 2. In the divided attention task it was found that participants performed very similarly in all sound conditions. The average count error was 8.2 (SD = 2.1), 8.1 (SD = 1.9) and 8.2 (SD = 2.0) events in the DirAC, Mono and Stereo conditions, respectively. In a One-Way Repeated Measures ANOVA it was found that there were no significant differences in the results of the different sound conditions (F(2) = 0.02, p = 0.98).

In the selective attention task results followed a different pattern. The mean count errors were smaller across participants. The errors were 0.8 (SD = 0.1), 0.3 (SD = 0.7) and 0.3 (SD = 0.6) in the DirAC, Mono and Stereo sound condition, respectively. A One-Way Repeated Measures ANOVA revealed that results differed across sound condition (F(2) = 7.25, p = 0.004). In a post-hoc Scheff test it was found that both Mono and Stereo results differed from DirAC results, but they did not differ from each other.

In sum, the type of sound spatialization had an effect on the counting of visual events per scene, but only in the selective attention task. In the divided attention task participants performed similarly across all conditions.

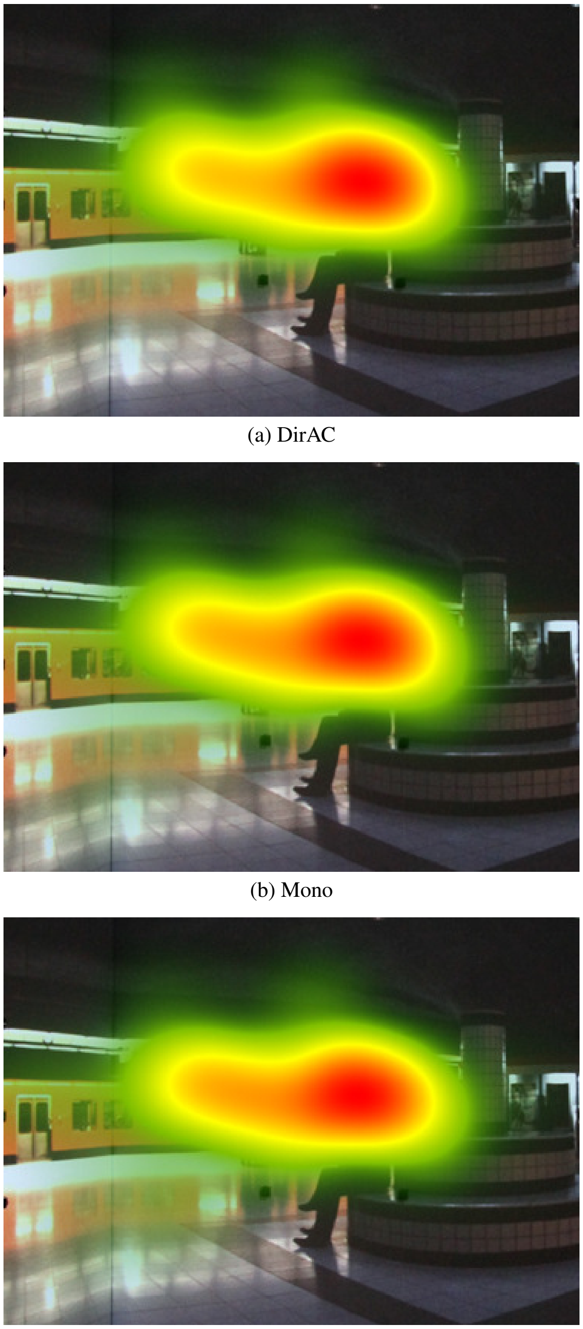
3.2. Gaze

In a first exploration of the eye tracking data, heat maps were computed for each experimental condition. This was done by counting the number of fixations in each screen area and plotting different counts with different colors over a scene snapshot.

All eye tracking data was grouped into two classes: saccades and fixations. The saccades correspond to when the eyes move and the fixations correspond to when the eyes are static. The fixation values from the divided attention task are presented in Figure 3. In this task, gaze patterns were very similar across sound conditions. The visual area with more than 400 fixations was wider in the DirAC condition (34.38 visual deg), narrower in the Stereo condition (32.54 deg) and similar in the Mono condition (32.74 deg).

In the selective attention task the heat maps were more concentrated than in the divided attention task (Figure 4). In the DirAC condition, the visual area with more than 400 fixations was the widest (28.07 deg). In the Stereo and Mono conditions that area was smaller (23.38 deg and 22.55 deg, respectively).

Due to software limitations, no additional quantification of the heat maps is possible.Therefore, to be able to directly quantify and compare conditions, areas of interest were defined in each scene. The areas of interest were used to compute the total fixation count and duration within those areas. The areas of interest were defined by selecting a screen area, as presented in Figure 5. In the divided attention task, three areas of interest were defined, corresponding to each of the three doors. In the selective attention task only one area was defined, corresponding to the middle door. The choice of areas is arbitrary. In this case, several areas were tested, to obtain a sufficient amount of visits for all participants during the experiment. It is noteworthy that participants vary considerably in their gaze pattern.

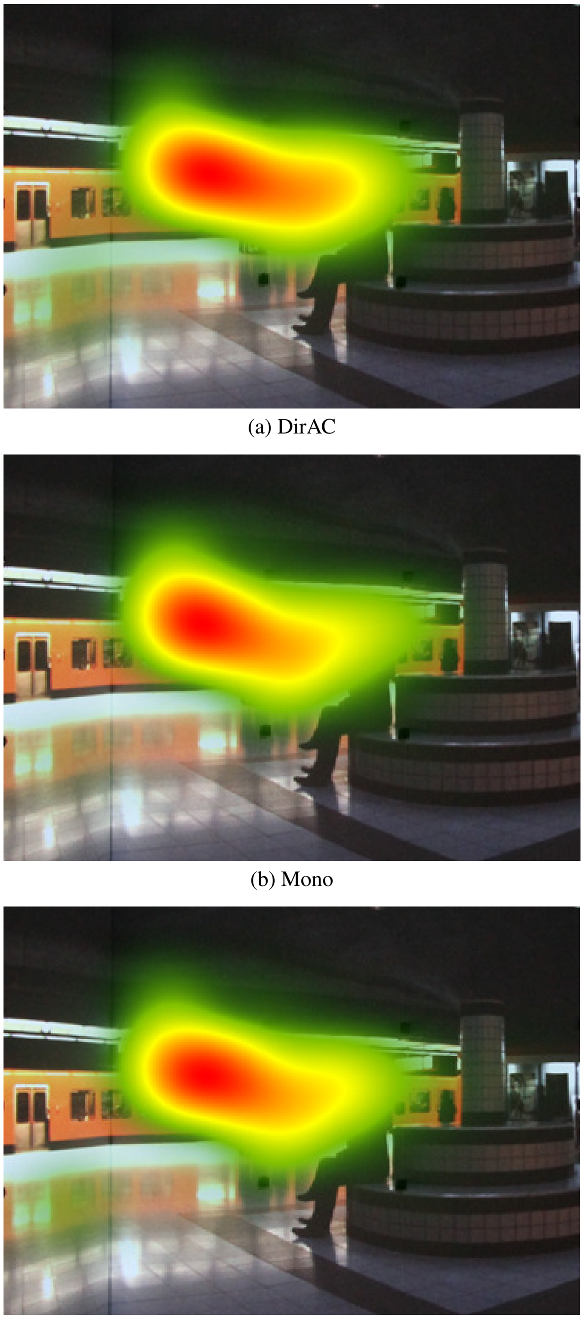


(a) DirAC

(b) Mono

(b) Stereo

Figure 3: Heat maps in the Divided Attention task. Colors represent the number of total fixations across 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.

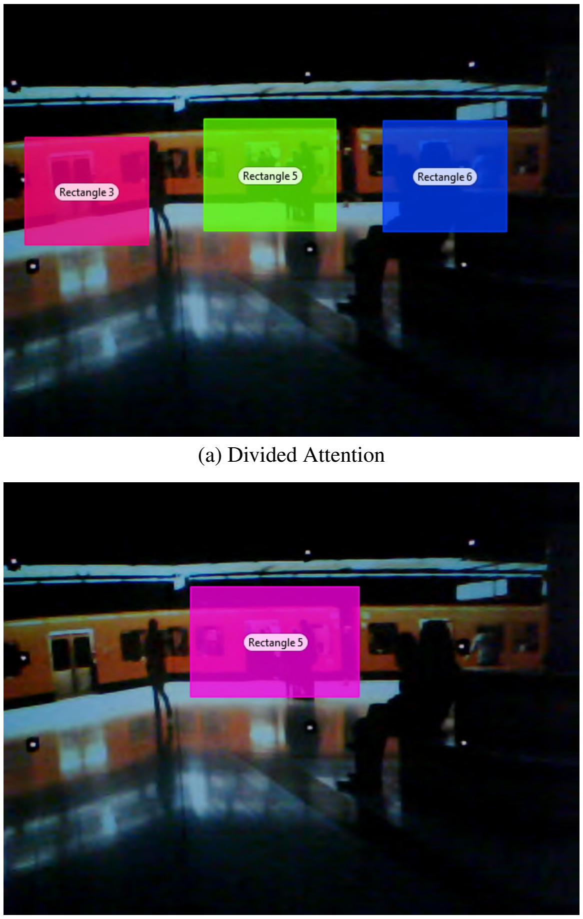


(a) DirAC

(b) Mono

(b) Stereo

Figure 4: Heat maps in the Selective Attention task. Colors represent the number of total fixations across 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.



(a) Divided Attention

(b) Selective Attention

Figure 5: The areas of interest in each task type.

The average total number of fixations varied across conditions. In the divided attention task Stereo had the highest amount of fixations (654, SD = 610), followed by DirAC (536, SD = 510) and then by Mono (424, SD = 608). In the selective attention task, the Stereo sound condition had again the highest number of fixations (2340, SD = 1722), followed by Mono (1903, SD = 1655) and then by DirAC (1675, SD = 1066).

The average of the total amount of time spent in fixations within the area of interest per participant is presented in Figure 6. In the divided attention task, the Stereo condition had the highest total amount of time spent in fixations within the area of interest (27.63 sec, SD = 21.17), followed by DirAC (25.26 sec, SD = 25.37) and by Mono (17.37 sec, SD = 25.37). The high standard deviation levels reveal a high variability between participants. For instance, some participants had consistently brief total fixation durations (less than 5 sec), while others fixated for very prolonged times (above 50 sec) It is important to note that participants were, however, stable from trial-to-trial and condition-to-condition. For example, participants with briefer fixations kept this pattern throughout the test. Such within subject consistency is

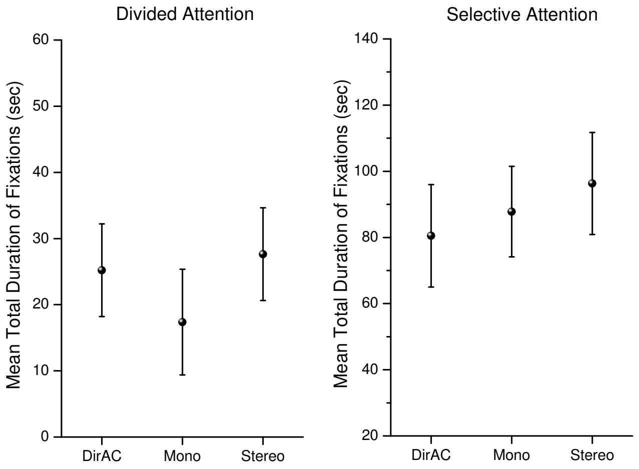


Figure 6: Mean of the total amount of time (sec) spent in fixations per participant within the area of interest in each experimental condition. Bars represent Standard Error of Mean.

accounted for in the One-Way Repeated Measures ANOVA. Still, the results for this statistical test revealed no significant differences between sound conditions (F(2) = 0.242, p = 0.791).

In the selective attention task, the Stereo condition had again the highest fixation duration (96.3 sec, SD = 69), followed by Mono (87.8 sec, SD = 61.3) and DirAC had the lowest fixation time (80.5 sec, SD = 73.8). The within-subject variability was smaller in this task. As a consequence, the One-Way Repeated Measures ANOVA revealed significant differences between sound conditions (F(2) = 3.97, p = 0.004). However, in a post-hoc Scheff test it was identified that this difference was only significant between the DirAC and the Stereo condition.

4. DISCUSSION

The aim of this paper was to explore if different sound spatialization levels could be related to different visual exploration and processing of an audiovisual reproduction. To tackle this, an experiment was implemented in which several scenes with similar audiovisual events were presented. There were three types of sound: mono, stereo and 3D (DirAC). There were two tasks: divided attention and selective attention. In both tasks, participants had to count how many people entered the metro on the screen.

The count errors can be interpreted as a measure of the ability to process visual events in a scene. It was found that, in the divided attention task, the sound had no significant impact on the count errors. It can be argued that, in that task, the locus of attention was already fluctuating very frequently. Also, the counting task became very difficult, since whenever participants looked at one door they could have missed events in another door. Therefore, the sound did not have any additional impact. But in the selective attention task, which was considerably easier, sound did have a significant effect. In that case, the DirAC sound condition had significantly higher count errors compared to the other sound conditions. This finding shows for the first time that the type of sound spatialization can affect how visual contents are perceived in a video reproduction.

In a second step, the distribution of eye fixations was analyzed. This can be taken as a measure of spread of attention. Comparing the heat maps obtained for each experimental condition it was again observed that there were no clear differences across sound condition in the divided attention task. However, in the selective attention task the DirAC sound condition had a larger area of frequent fixation. These results are in line with those from the counting task. They show that the patterns of visual attention are affected by sound spatialization when the subject is trying to attend to only one type of event while ignoring others. The fact that the visual attention is more spread out in the 3D sound condition can mean that the more spatialized sound led to more visual exploration or less focused attention.

Finally, the number and duration of fixations within specific areas of interest was analyzed. This analysis can be taken as a measure of stable attention. In the divided attention task it was found that the lowest number of fixations and fixation duration was in the DirAC and Mono conditions. As discussed above, while DirAC was highly spatialized, the Mono condition was an awkward case in which all sound was presented from the ceiling. This highly directive sound from an area outside the field of view could have caused some disturbing effect over attention. This adds to our interpretation that spatially defined sound outside the field of view can have a negative effect on visual stability, by creating a distraction. In the selective attention task, once again, the DirAC condition had significantly worse results, with significantly less time spent in the area of interest. These results further confirm the effect of spatial sound over visual processing of audiovisual scenes.

Taken altogether, these findings suggest that we may have to rethink our approaches to spatial sound for audiovisual reproduction. More realistic spatialization might not always be best. But we should also avoid the temptation to assume that traditional impoverished sound spatialization such as Stereo is better for the perception of visual contents. There are several factors to have in mind at this level. Firstly, the study reported here used only one type of task (counting events). Ideally other tasks should be implemented, like speech intelligibility, event detection and scene interpretation, to find out if the effect is also observed in those tasks. Also, further tests are needed to analyze if this finding also applies to smaller reproduction setups, such as those found in regular households. It is generally accepted that sound spatialization has positive effects on the feeling of presence [14] and quality [15] [16]. So the tradeoffs of different approaches should be carefully compared. Finally, maybe the next best approach will be to change spatialization algorithms to provide for better audiovisual perception. It is plausible that the interactions found here are mostly due to highly salient and point-like auditory events outside of the area of visual interest or in its periphery. With careful additional studies we may be able to identify the spatial areas in which spatial sound interacts with the visual processing of the scene. The simple solution could be to have more diffuse and less point-like sounds outside that critical area.

In sum, data reported in this study show for the first time that 3D sound might hinder the perception of visual events in video reproduction by dispersing visual attention. Future studies should further explore these findings with different tasks, different scenes and different reproduction setups. This study should be the first step to a thorough analysis of the best parameters in spatial sound design for optimal content perception of video reproductions.

5. ACKNOWLEDGMENT

This work was supported by the project ”Audiovisual Space” by the Academy of Finland and by the European Research Council under the European Community’s Seventh Framework Programme (FP7/2007-2013)/ERC grant agreement No. 240453.

6. REFERENCES

[1] J. Driver and C. Spence, “Crossmodal attention,” *Current opinion in neurobiology*, vol. 8, no. 2, pp. 245–253, 1998.

[2] M. Eimer and J. Velzen, “Crossmodal links in spatial attention are mediated by supramodal control processes: Evidence from event-related potentials,” *Psychophysiology*, vol. 39, no. 4, pp. 437–449, 2002.

[3] C. Spence and J. Driver, *Crossmodal Space and Crossmodal Attention*. Oup Oxford, 2004.

[4] J. Driver and C. Spence, “Cross–modal links in spatial attention,” *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, vol. 353, no. 1373, pp. 1319–1331, 1998.

[5] J. Green, W. Teder-Sälejärvi, and J. McDonald, “Control mechanisms mediating shifts of attention in auditory and visual space: a spatio-temporal erp analysis,” *Experimental Brain Research*, vol. 166, no. 3-4, pp. 358–369, 2005. [Online]. Available: http://dx.doi.org/10.1007/s00221-005-2377-8

[6] J. J. Green and J. J. McDonald, “An event-related potential study of supramodal attentional control and crossmodal attention effects,” *Psychophysiology*, vol. 43, no. 2, pp. 161–171, 2006.

[7] C. Escera, K. Alho, I. Winkler, and R. Näätänen, “Neural mechanisms of involuntary attention to acoustic novelty and change,” *Journal of cognitive neuroscience*, vol. 10, no. 5, pp. 590–604, 1998.

[8] M. A. Frens and A. V. Opstal, “Visual-auditory interactions modulate saccade-related activity in monkey superior colliculus,” *Brain Research Bulletin*, vol. 46, no. 3, pp. 211 – 224, 1998. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0361923098000070

[9] C. Rorden and J. Driver, “Does auditory attention shift in the direction of an upcoming saccade?” *Neuropsychologia*, vol. 37, no. 3, pp. 357 – 377, 1999. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0028393298000724

[10] B. D. Corneil and D. P. Munoz, “The influence of auditory and visual distractors on human orienting gaze shifts,” *The Journal of neuroscience*, vol. 16, no. 24, pp. 8193–8207, 1996.

[11] 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.

[12] J. G. Bolaños and V. Pulkki, “Immersive audiovisual environment with 3d audio playback,” in *Audio Engineering Society Convention 132*. Audio Engineering Society, 2012.

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

[14] P. Larsson, A. Väljamäe, D. Västfjäll, A. Tajadura-Jiménez, and M. Kleiner, “Auditory-induced presence in mixed reality environments and related technology,” in *The Engineering of Mixed Reality Systems*. Springer, 2010, pp. 143–163.

[15] D. Strohmeier and S. Jumisko-Pyykko, “How does my 3d video sound like?-impact of loudspeaker set-ups on audiovisual quality on mid-sized autostereoscopic display,” in *3DTV Conference: The True Vision-Capture, Transmission and Display of 3D Video, 2008*. IEEE, 2008, pp. 73–76.

[16] A. Väljamäe and A. Tajadura-Jiménez, “Perceptual optimization of audio-visual media: Moved by sound,” *Narration and Spectatorship in Moving Images*, 2007.