Audio Engineering Society

Convention Paper

Presented at the 149th Convention
2020 October 21–24, New York, USA

The Sense of Auditory Presence in a Choir for Virtual Reality

Louise Bryce[1], Mark Sandler[1], Lars Koreska Andersen[2], Ali Adjorlu[2], Stefania Serafin[2]

[1] *Queen Mary University of London. Mile End Rd, Bethnal Green, London, United Kingdom*

[2] *Aalborg University, Copenhagen. A. C. Meyers Vænge 15, 2450 København, Denmark*

Correspondence should be addressed to Louise Bryce (l.bryce@qmul.ac.uk)

ABSTRACT

This paper investigates which auditory parameters influence the sense of presence in an immersive environment of children with low functioning autism (LFA) and social anxiety. The auditory parameters investigated were Spatialized, Authoritarian (e.g. the ability to hear a conductor) and Ambisonic. A 360-degree video of the Danish National Children's choir shown in the VE enables the participants to sing and dance together with them, while simultaneously being exposed to the different parameters. We discuss how the different audio parameters affect the plausibility of the environment to children. While asking questions in relation to presence, the four participating children were having difficulties grasping the meaning, even when adjusted to their preference and level of understanding. However, the observational data yielded a change in their behaviour and participation with the virtual choir, while differentiating with the changing parameters. This could indicate that presence through auditory stimuli can be observed through the behavioural patterns of children with LFA and social anxiety.

1 Introduction

Exposure therapy is a method of behavioural therapy designed to give those with certain disorders, e.g. social anxiety, the ability to overcome their fears by placing them in the situation inducing them. Virtual Reality (VR) has become a popular way to deliver such therapies [1], and even more so with the recent democratization of both hardware and software technologies. Head-mounted Displays themselves have become an affordable piece of technology, being used to deliver serious experiences like exposure therapy and learning tools as well as recreational experiences like games. This paper examines a form of virtual exposure therapy for children with low-functioning autism (LFA) and the role of auditory stimuli on the experience. In a previous paper [10], a virtual environment was created in Unity. The application allows children and psychologists to cohabit virtual space containing a choir. 360-degree video was used to capture the experience, with eight different video options for the psychologist to choose from. In this paper we extend the application to examine how different audio environments affect participants' presence in the virtual environment.

Specifically, we extend the previous application in order to provide the psychologist with volume controls. The level of the volume is interactively controlled by the psychologist. Our ultimate goal is to examine if there is a common preferred audio mix between participants.

2 Related Work

Social Anxiety Disorder (SAD) can affect the lives of those who suffer from it immensely [2]. It refers to those who fear negative results from social interaction, causing many to stop themselves from being in social situations altogether. Exposure therapy was used to begin treating SAD in the mid-1980s. Compared to other phobias this form of exposure therapy came much later due to the difficulties that surround putting people into appropriate social situations [3], a factor overcome by VR.

Exposure therapy allows a psychologist to slowly introduce feelings of anxiety within the patient, while also repeating these feelings over many sessions. This allows the patient to become accustomed to situations and to become better at dealing with feelings of anxiety. VR has given high success rates in exposure therapy in previous studies [4] as it provides a setting which can be heavily controlled and adapted to individual needs as well as allowing for simulations that are impossible to exist in real life, or are highly improbable in a real-world environment. For example, treating a fear of heights (acrophobia) is easier in VR. Due to recent advances it is now easier to record 360-degree video than ever. This allows for video and surround sound to immerse the user, although this usually has limited interaction capabilities over graphically generated 3D worlds.

Audio is an integral aspect of user experiences in VR [5]. Spatialized audio outside of the users field of vision can often be used to induce emotion in users, an important factor in exposure therapy particularly within the social anxiety setting. Making the audio as realistic as possible in such environments is something that requires further study and a different set of methods to those used in other forms of media [6]. Virtual environments can be a safer way to practice certain tasks without putting physical objects or people at risk [7]. While designing audio for virtual reality experiences it is important to keep within the aesthetics of an experience while also keeping the experience functional [8]. While many VR experiences intend to achieve realism, others do not directly strive for this. These, however, do not always have to be directly linked and many different aspects such as the recording and delivery of an audio experience can directly affect how an experience is perceived.

3D Microphones are becoming more affordable and widely used; the Sennheiser Ambeo VR is a popular example [9]. Mono microphone recordings, when placed individually, do not contain spatial information. However, in virtual environments they can be placed and mixed to give the user a sense of spatial presence. Using ambisonic microphones or arrays can give better spatial reproduction as the room itself is also recorded, allowing for better immersion. Using these in conjunction with well mixed mono sources can give the most convincing, immersive experiences [8].

3 Capturing the Choir

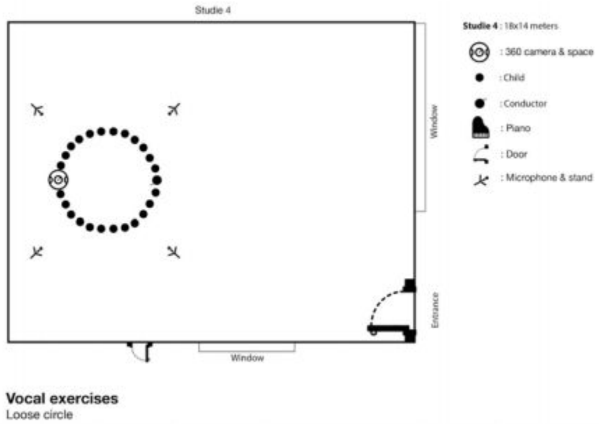

Figure 1. The recording plan for one of the video options

For this experiment the Danish National Children's Choir was recorded in their rehearsal hall in Copenhagen, Denmark. The video was recorded using an Insta 360 Pro Camera, while the audio was recorded using the Ambeo microphone, four spot microphones and an extra microphone on the conductor. It should be noted that, due to issues during the time of the recording, the conductor microphone is unavailable in the first video session. The children perform two traditional Danish children's choir songs - *Hvis jeg var en cirkushest* and *Tarzan Mamma Mia* - accompanied by a pianist. Each of these were recorded three times, with the camera placement varying. A vocal warmup exercise was also recorded with two different placements, leading to a total of eight video and audio choices. The camera placements allow for the participating child (see section 4) to immerse themselves in different areas of the choir, with each area providing different situations with differing anxiety levels. For example, the participant can be placed in the middle of the choir, close to other children. Alternatively, they can be placed in the back row with more distance between themselves and other children which may reduce anxiety.

4 Method

In a previous paper [10] the video capturing, virtual reality implementation and the possibility of exposure therapy were introduced. While audio was present in the shared VE, the different audio sources were pre-mixed by the researchers and were not studied in detail.

This methodology will discuss the use of audio as the active and focused parameter in the shared VE.

4.1 Surround Sound Mixing

FMod Studio is an industry standard piece of middleware (a piece of software that provides extra services and capabilities), which is used in conjunction with game engines to deliver audio to the user. It allows sound designers to work in a more comfortable and traditional audio setting [11].

FMod was used to mix and deliver this auditory experience due to its 3D audio capabilities and ability to connect with the visual delivery system, Unity. The frameworks used in the previous study [10] were changed so that FMod could be used in conjunction with the video delivery system.

UI controls were given to the psychologist to control the volume of individual tracks (see figure 2), in order to maintain total control over the shared VE. Volume buttons were implemented for the conductors microphone, the ambisonic microphone and a single button controlling the four spot microphones to prevent the spatialisation from being put off balance.

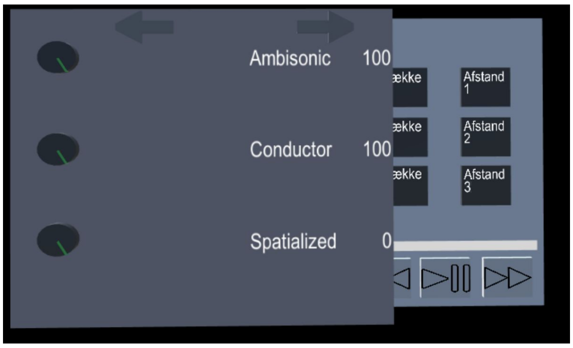

Figure 2. The user controls for the psychologist, showing the volume knobs

Events are the main controllable aspect of FMod, allowing for multiple tracks of audio to be played Therefore, for every video option in this VE, an event has been created. Each of these events contains all six tracks for each video, these all play together when the event is called. Similarly, this can pause and stop the audio.

Parameters are parts of the sound that can be changed while an event is running, this is traditionally referred to automation. For this study, the knobs as seen in fig. 2 control the loudness level of tracks in FMod.

When recording, the ambisonic microphone was centred at the child to give the best impression of the room and the four spot microphones are placed as they were in the room. The psychologist has the ability to change these audio levels due to the nature of the VE. For its spatialisation of audio, FMod uses a great amount of distance attenuation - the change of audio loudness depending on the distance of the listener from the source. In this study, the users cannot move around in the VE and are static in the space - only able to look around. Initial levels of these alongside the other microphones are presented as the same loudness level to allow the child to feel comfortable.

4.2 Virtual Environment and Tasks

In order to study the presence effects of the audio parameters, the VE should not induce any unintentional effect on children with LFA. The displayed 360 video therefore puts the participants in the back row of the choir, avoiding exposure to any social anxiety. To keep consistency and to avoid introducing miscommunication, the program Discord, a Voice over Internet Protocol (VoIP), was used to give clear communication between the test participant and the researcher.

The task for the participant is to experience the three different parameters, while answering questions from the researcher whom they share the VE with.
As in fig 2., the child was also given a single UI knob to control their own volume within the VE, to ensure their own comfort.

4.2.2 LFA and its effect on participants

Since the test participants are children diagnosed with a form of autism, it was important to consider what effect it would have on the approach of the methods and measurements. Due to ethical reasons, information about the participants was limited to observations from the researchers, and the restricted information from the counsellor who facilitated the participants at site. Using this information, the participants were defined according to DSM-V [12] severity level. The included participants were scored between level 2 and level 3 and in combination with the participants being of intellectual impairment, categorized them with low-functioning autism (LFA). Additionally, in our given age group children are much more impulsive than reflective, even their social background can have an effect on their ability to reflect [13]. With the two aforementioned complications, it is assumed that verbal reflection of the questions asked is exceptionally difficult for the participants, the methods were produced accordingly.

In correlation with the assumptions given for the participants, considerations on measuring presence is a necessity, as normal measurements in presence used a self-reporting assessment system. However LFA in children, is often accompanied with an intellectual deficiency and atypical limited verbal language [14], which can lead to misunderstanding and communication errors. It can therefore be argued that *The Independent Television Commission - Sense of Presence Inventory* Presence Questionnaire [15] is too extensive for the target group, and should therefore be severely reduced, in order to measure the magnitude of presence within the participating children. The simplification of the questionnaire comes down to two out of five [16] different areas which is interesting in correlation with the change in auditorial parameters.

•An individual's level of spatial presence is related to the participants feeling of being there, in accordance with the changing audio parameters.

•The ecological validity or the naturalness of the experience, which relates to whether the experience felt like a real-life scenario or if the experience gives an unnatural feeling, like watching a movie.

4.3 Participant Questionnaire

Since children have a very limited relationship with numerical values and abstract thought, it is imperative for the interviewer to quickly create an impression of the child's response, and take that consideration for the following questions [13]. Therefore, three pre-planned questions were created for the researcher to start from. The presented questions were translated from Danish to English.

Individuals level of spatial presence:

1. Do you feel like you are part of this virtual world?
2. Do you feel like part of the experience?
3. Do you feel we are here in the experience?

The ecological validity or the naturalness of the experience -

1. How true do you feel this whole thing is?
2. How realistic do you feel this is?
3. Do you think it's like reality, or is it just a movie?

Reflecting is difficult for the participants, which points towards a reduction in the Likert scale measurements normally used in the ITC sense of presence Inventory questionnaire. By simplifying the answer possibilities to yes/no data, we can correlate it with the participants' observed behaviour and hopefully get a more honest picture of the effect. Circular boxes, coloured green and red, respectively yes and no, were implemented for the participant to answer through (see figure 3). The converted Likert scale to yes/no buttons. These are instantiated in front of the child at the command of the psychologist, and can be removed as well

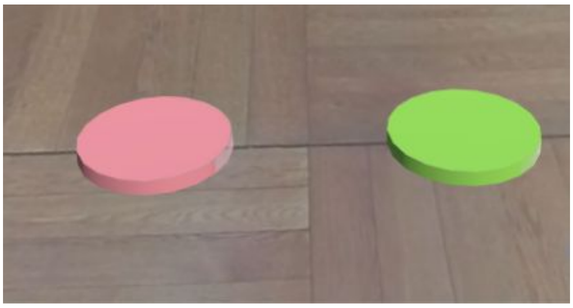

Figure 3: The children could choose yes or no by pressing these buttons in the VE

A total of 4 participants were recruited (3 male, 1 female) through the schools counsellor, with ages ranging from 8 to 13. All of them had tried VR before, and were comfortable with Oculus Rift. It should be noted that the 3rd participant is wheelchair bound, and was therefore restricted in movement.

4.4 Experimental Location

The test was conducted at Skovmoseskolen, Rødovre municipality in Copenhagen, Denmark. Using two laptops with the appropriate hardware for them to be Virtual Reality Ready, using Oculus Rift as the Head Mounted Display(HMD) with the rifts' built-in headphones. It was run in a small 5 by 7 foot room, familiar to the participants as the designated 'VR classroom'. Two researchers were present during the test, one as the designated researcher while the other took observational notes. A Counsellor representing the school was present as well, acting as a calming presence so the participant was not alone with two unknown researchers.

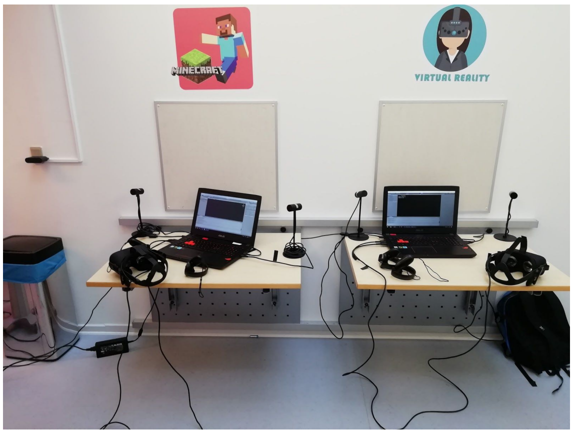

Figure 4: The test setup

The counsellor would bring in the children one at the time, where the leading researcher would greet and introduce them to what they were going to experience. In the corner of the room, another researcher would greet upon entering, but silently take observational notes during the test. The leading researcher would then randomize the order of which the parameters would be tested among all of the test participants. Before the questions, each participant had time to settle in and be comfortable with the HMD and the VE. For each parameter, the two presence questions were asked, and the participant could answer through the two circular shapes, either green for yes, or red for no. When the last question was asked, the participant could in their own time stop the test, and the counsellor would escort them out.

5 Results

The following section will present the results gathered through the test, both the interview and observation data. Abbreviations describes the outcome of the test conductors' questions, and the participants correlating answers. Participants 1 to 4 will be abbreviated to P1, P2, P3 & P4. For example: Participant 1 in Condition A (Ambisonic test) answered G (green) for the first question, and R (red) for the second question.

5.1 Observational Data

The data presented is both the observational data from the researcher taking notes, and from in-game recordings.

P1: Before entering the shared VE, the participant was excited to test the application. Although when entering the VE, their gazes were mostly downwards, by the end of the test these were more towards eye level.

P2: The participant waved to the researcher in the beginning, and would dance with the children in the video. Although less movement was recognized in condition B than the other two conditions. The participant related the application to TV, mostly because he watches a lot of TV in general. Before entering the shared VE, the participant was reluctant towards the researchers, but showed excitement towards the counsellor. The participant did not engage in dancing or singing during the test, but did have a noticeable look around in the shared VE. Although there were minor differences in movement between condition B and the rest. P2 did not dance with the choir, but was more observant.

P3: Before entering the shared VE, the participant was excited to test the application. Being wheelchair bound, the participant was constricted in movement, but that did not affect the possibility to use her hands to be part of the choir. When the red and green circles instantiated, the participant started gazing towards those and interacting with them instead of the whole experience.

P4: The participant seemed ridden with anxiety, as his gaze was fixated on the ground most of the time, and his shoulders hunched up. When spoken to outside the shared VE, the participant only answered in thumbs up and down. However, when entering the shared VE and starting the test, the participant completely changed behaviour and demeanour. The music made him mumble to the rhythm and through the changes of the audio, more towards singing. The change was mostly noticeable during condition C (Spatialized) where the participant danced and sang with the choir. When exiting the shared VE, he went back to the behaviour and demeanour he displayed when entering the test room.

6 Discussion

The interview data has shown that there is no support for the hypothesis on whether children with LFA can differentiate between audio parameters in a shared VE setting, or even recognize the difference in presence. Even though we reduced what was already considered an extensive questionnaire to two questions converting two out of five areas of presence, the participants did not comprehend the questions asked.

This could be connected to the participants' different perception and interpretation of the world surrounding them, in correlation to diagnose LFA and how we underestimated the participants' understanding. There are multiple possible causes for this, firstly it may be their deficit in multisensory perception [17] and their willingness to attend and report global processing information [18]. This way of processing information of general stimuli can, besides changing their way of interpreting questions, contribute to the fact that there were no real significant effects between the audio parameters.

Secondly, the 360 video was largely overstimulating making it the focus for the local information process in the participants, and therefore hard to differentiate what actually made them believe they were affected by presence. Or, the audio parameters were sufficient and fitting for the visuals for the test participants to have a hyporesponsive relationship [19] with the audio parameters.

Observing the participants yielded a difference in their physical body demeanour. With the current sample-size and the methods used, it is not possible to point towards the change being correlated to audio in accordance to presence. There is the possibility that this could be time spent in the shared VE which progressively increases their engagement with the application, or that the researcher was engaging the participants to join in with the choir. Confidential persons, in the eyes of the test participants, can be a key factor into facilitating more concrete answers about abstract concepts such as presence. As it was observed that the test participants responded more extensively towards the counsellor, using the approach of having a more normal conversation with smaller breaks in between could be the more preferable method [13].

7 Conclusions and Further Work

This paper has explored the use of spatial audio for exposure therapy, the results thus far have been inconclusive and so further studies should attempt to use a more tailor-made approach towards measuring the effects of audio. In considering the exposure of children with LFA to spatial audio, the field of research is in its infancy. And, while an attempt has been made, it was discovered that different perspectives and correlating approaches have a heavy influence on the measurements. The lack of experience the children had with 360-degree video may have caused confounding issues. Another area of interest to explore is how the psychologists themselves use the interchangeable audio parameters to help the child in the environment.

8 Acknowledgements

The authors would like to thank the children and conductors of the Danish National Children's Choir for allowing us to record their rehearsal. Thanks to Skovmoseskolen for providing us with test participants, and for being so considerate towards us. This work was partially supported by the ESPRC and AHRC centre for doctoral training in Media and Arts Technology, grant number EP/L016 32X/1.8.

9 References

[1] T. D. Parsons and A. A. Rizzo, “Affective outcomes of virtual reality exposure therapy for anxiety and specific phobias: A meta-analysis, ”Journal of behavior therapy and experimental psychiatry, vol. 39, no. 3, pp.250–261, 2008.

[2] I. M. Aderka, S. G. Hofmann, A. Nickerson, H. Hermesh, E. Gilboa-Schechtman, and S. Marom, “Functional impairment in social anxiety disorder", Journal of Anxiety Disorders, vol. 26,no. 3, pp. 393–400, April 2012.

[3] P. Anderson, B. O. Rothbaum, and L. F. Hodges, “Virtual reality exposure in the treatment of social anxiety,” Cognitive and Behavioral Practice, vol. 10, no. 3, pp. 240–247, June 2003.

[4] P. Emmelkamp, M. Krijn, A. Hulsbosch, S. de Vries, M. Schuemie, and C. van der Mast, “Virtual reality treatment versus exposure in vivo: a comparative evaluation in acrophobia,” Behaviour Research and Therapy, vol. 40, no. 5, pp. 509–516, 2002.

[5] N. D. Hai, N. K. Chaudhary, S. Peksi, R. Ranjan, J. He, and W.. Gan, “Fast HRTF Measurement System with Unconstrained Head Movements for 3D Audio in Virtual and Augmented

Reality Applications," Journal of the Audio Engineering Society, pp. 6576–6577, 2017.

[6] A. Rana, C. Ozcinar, and A. Smolic, "Towards Generating Ambisonics Using Audio-visual Cues for Virtual Reality," pp. 2012–2016, 2019.

[7] D. Johnston, H. Egermann, and G. Kearney, "Soundfields: A mixed reality spatial audio game for children with autism spectrum disorder," in Audio Engineering Society Convention 145, 2018.

[8] S. Davies, S. Cunningham, and R. Picking, "A comparison of audio models for Virtual Reality video", Proceedings - International Conference on Cyberworlds, ACM SIGGRAP, ,pp. 150–153, 2017.

[9] W. Woszczyk and P. Geluso, "Streamlined 3D sound design: the capture and composition of a sound field," in AES 145th International Conference, 2018.

[10] S. Serafin, A. Adjorlu, L. Andersen, and N. Andersen, "Singing in virtual reality with the danish national children's choir," 2019.

[11] S. Horowitz, The Essential Guide to Game Audio: The Theory and Practice of Sound for Games. Routledge, mar 2014. [Online]. Available:https://www.xarg.org/ref/a/041570670X/

[12] A. Speaks, "Dsm-5 diagnostic criteria" New York: NY. Author retrieved, August, vol. 10, p. 2014, 2014.

[13] T. Bjørner, "Research design," in Qualitative Methods for Consumer Research. Hans Reitzels Forlag, 2015, pp. 17–53.

[14] H. Brentani, C. S. d. Paula, D. Bordini, D. Rolim, F. Sato, J. Portolese, M. C. Pacifico, and J. T. McCracken, "Autism spectrum disorders: an overview on diagnosis and treatment,"Brazilian Journal of Psychiatry, vol. 35, pp. S62–S72, 2013.

[15] J. Lessiter, J. Freeman, E. Keogh, and J. Davidoff, "Across media presence questionnaire: The itc sense of presence inventory," Presence: Teleoperators & Virtual Environments, vol. 10, no. 3, pp. 282–297, 2001.

[16] S. Wallace, S. Parsons, and A. Bailey, "Self-reported sense of presence and responses to social stimuli by adolescents with asd in a collaborative virtual reality environment," Journal of Intellectual & Developmental Disability, vol. 42, no. 2, pp. 131–141, 2017.

[17] T. G. Woynaroski, L. D. Kwakye, J. H. Foss-Feig, R. A.Stevenson, W. L. Stone, and M. T. Wallace, "Multisensory speech perception in children with autism spectrum disorders,"Journal of autism and developmental disorders, vol. 43, no. 12, pp. 2891–2902, 2013.

[18] K. Koldewyn, Y. V. Jiang, S. Weigelt, and N. Kanwisher, "Global/local processing in autism: Not a disability, but a disinclination,"Journal of autism and developmental disorders, vol. 43, no. 10, pp. 2329–2340, 2013.

[19] J. L. Kleberg, E. Thorup, and T. Falck-Ytter, "Visual orienting in children with autism:Hyper responsiveness to human eyes presented after a brief alerting audio-signal, but hyporesponsiveness to eyes presented without sound," Autism Research, vol. 10, no. 2, pp. 246–250, 2017.