Audio Engineering Society

Convention Paper

Presented at the 147th Convention
2019 October 16 – 19, New York

Sound design and reproduction techniques for co-located narrative VR experiences

Marta Gospodarek, Andrea Genovese, Dennis Dembeck, Corinne Brenner, Agnieszka Roginska, and Ken Perlin

New York University

Correspondence should be addressed to Marta Gospodarek (mo1417@nyu.edu)

ABSTRACT

Immersive co-located theatre aims to bring the social aspects of traditional cinematic and theatrical experience into Virtual Reality (VR). Within these VR environments, participants can see and hear each other, while their virtual seating location corresponds to their actual position in the physical space. These elements create a realistic sense of presence and communication, which enables an audience to create a cognitive impression of a shared virtual space. This article presents a theoretical framework behind the design principles, challenges and factors involved in the sound production of co-located VR cinematic productions, followed by a case-study discussion examining the implementation of an example system for a 6-minute cinematic experience for 30 simultaneous users. A hybrid reproduction system is proposed for the delivery of an effective sound design for shared cinematic VR.

1 Introduction

Virtual Reality (VR) is expanding very fast in the fields of gaming and entertainment and numerous cinematic productions experiment with headsets to deliver new sort of experiences. However, most of these works are designed to engage a single user at a time and do not usually entail one very important feature that exists in cinema and theatre, a sense of social gathering.
Technological developments observed in recent years now enable the creation of different kinds of VR productions that allow co-located large audiences to experience a shared virtual environment presented in specially-designed entertainment spaces [1, 2]. These productions have the goal of bringing back the social aspects of cinema and theatre, which is achieved by designing an experience where the participants are able to see and hear each other as virtual avatars, spatially coherent with their actual physical location. A cognitive impression of being in a shared experience can thus take place and enable a sense of presence, and certain forms of communication and awareness with fellow audience members.

The audio layer is especially important in VR experiences as it affects the subjective senses of *immersion*, *plausibility* and *presence* [3], which are key to the success of co-located immersive VR. Spatial audio techniques, which allow users to match the position of sounds with their respective visual cues, can, in fact, improve these metrics of quality [4], while a poor audio production may affect them negatively [5].
As of today, there is not much literature on the sound

design theory behind this particular style of creative production. This article reviews the factors and principles behind the implementations of audio systems for co-located narrative VR, whether cinematic or theatrical. We propose the use of hybrid reproduction systems made of both loudspeakers and a transparent hearing device (such as nearfield speakers or transparent earphones) in order to address the audio challenges involved.
The second part of this paper illustrates a case-study discussion around the implementation of the audio reproduction system for a short narrative VR art piece, *"Cave"*. The experience gained by the authors through working on this production helped to inform and validate the design principles discussed, as well as identify the technical variables which may affect particular choices. A short survey was conducted to gain formative insights on the effectiveness of the system and to illustrate the challenging aspects that need to be addressed by future work.

2 Background

Immersive co-located VR is a new type of production category which merges some elements of gaming interaction with the linear narrative elements of theatre and cinema. The defining element is the assumption that multiple participants are located in the same room and experience the same virtual content (through virtual or mixed reality devices) under their own individual perspectives and points-of-view, while also being able to see each other in the virtual space. Each participant is rendered in the shared virtual scene as a virtual avatar (usually humanoid), spatially matching their physical location and orientation in real-time, by means of motion-tracking technology.
Since simulating the social setting of a crowd inside a theatre is a goal for these systems, it is important that participants are treated as audience members and feel present as such in the space. [6]. To this goal, the audience is usually placed in "seats" from which unique first-person views are dynamically rendered and the narrative content is placed onto a virtual "stage".
To technically achieve a multi-user visual reproduction, the headsets are connected using a network-synced infrastructure that allows for the simultaneous delivery of the content for all participants [7]. While the narrative content may or may not be linearly progressing (usually it is), the rendering of the audience members' avatars (e.g. their head rotation and off-axis shift) needs to be actively updating close to real-time. Each client device reports its 3D location and orientation to a server, and receives the location of every other device with their respective timestamps [8]. The rendering is finally facilitated at each device through time synchronization signals that make sure that there are no differences in perceived time between participants.

The implementation of such cinematic, or theatrical, experiences can exist under different variations. One particular experimental production, "Holojam in Wonderland", shown at the 2017 New York's Future of Storytelling Festival, was portrayed as an "Immersive Mixed-Reality Theatre"[2]. Two live-rendered actors and four audience members shared a virtual reality stage where a theatrical narration took place in a shared environment. While the actors represented the story characters, the audience was represented by avatars of butterflies, and all were allowed to move in 6 degrees of freedom (DOF), explore the virtual world, and interact with a semi-linear progression of events.
The sound was implemented through a quadraphonic loudspeaker system with an additional overhead speaker. The actors' dialogue was presented in dual form as live free-field speech alternated to pre-recorded dialogue lines played from the overhead speaker. This choice served the narrative purpose of simulating one actor's change in size both visually and aurally. No headphones were used as it was necessary for the free-field dialogue to be heard without the effects of occlusion and attenuation of the sound path to the ears, although it is possible to achieve transparent headphone reproduction using hear-through microphones [9].

3 Design factors and principles

3.1 Sound principles for co-located VR theater

Each kind of VR experience requires a different approach to sound design and sound production due to the potentially different modalities of the medium used for the storytelling. Most design implementations in VR are based on the game-audio framework (in case of interactive experiences [10]) or on the cinematic framework (in 360º videos [11]). Co-located VR theater entails a set of design requirements for the audio layer which differs from the other types of VR productions:

Transparent hearing To enable communication within the audience, it is important that users can hear each other during the experience. The use of headphone playback is not appropriate in this context as it impairs free-field listening abilities of the participant. Even open-back headphones are shown to produce occlusion and attenuation effects at the ear canal [12], making the blend between real and virtual sources more difficult to achieve. Although it is possible to equalize this effects with an attentive individualized calibration, a more flexible solution is to employ different kinds of non-obstructive sound reproduction devices such as hear-through earphones or headphones supplemented with microphones that enable transparent hearing [9]. A loudspeaker-only reproduction method would also provide transparent hearing, but likely interfere with other requirements.

Spatial sound The auditory localization of sound objects has to match the visual localization of the sound sources in order to achieve immersion and presence inside the experience [4]. Spatial audio techniques need to be used to ensure proper perception of the sound localization and adequate proximity effects between the far and the near auditory fields.

Cinematic sound design The sound layer has to support the storytelling and reflect a cinematic style of sound design, supporting the full spectrum of sounds which make a compelling experience. The implementation of sound for co-located cinematic VR is merging the approaches from traditional cinema and game audio. The VR narrative is linear, meaning it is played identically for every performance. This format creates an opportunity to design sounds which perfectly match the visual action, without the need for sound randomization which is necessary in games [10]. On the other hand, the experience is also interactive. The user has the ability to modify their orientation and position inside the scene, which means that their point of listening can change.

Individual audio mix When each member of the audience's "virtual seat" corresponds to their position in physical space, the sound mix delivered also must be matched to that position and orientation, and thus differs for each member of the audience. As a result, each member of the audience receives an individual sound mix which represents their point of listening.

3.2 Proposed reproduction system

The biggest challenge is to allow audience members to hear each other while delivering a high quality spatial audio layer. This paper suggests the employment of a hybrid reproduction system for immersive co-located VR experiences. The proposed system consists of a transparent hearing device and a loudspeaker array.

3.2.1 Spatial audio over headphones

Spatial audio content is more easily and flexibly deliverable through binaural audio techniques. Binaural audio technology allows to reproduce spatial sound by encoding auditory cues into a stereo audio signal, thus changing the perceived localization of object sound sources [13].
The cues which depend on the anthropomorphic measurements of the person's head, pinna, and torso are unique for every individual. Head Related Transfer Function (HRTF) characterizes the auditory spatial cues of a person for a defined sound source position. It includes interaural time difference (ITD), interaural intensity difference (IID), and spectral modulations. The limitation of binaural sound in most VR productions is the use of non-individualized HRTFs, which can cause distortions in perceived sound image such as front-back confusions, distortions in localization on the vertical plane, and weak externalization [14]. Adequate reverberation (coherent to the visual environment) and head-tracking can, to some extent, mitigate the drawbacks of using non-personalized HRTFs. Headphone playback is the most common way of delivering spatial audio, mostly because it can ensure a perfect separation between the two binaural channels. The drawback of headphone reproduction is that even open-back headphones introduce significant attenuation of real-world sound sources [12]. The resulting coloration takes away from the plausibility of the experience, and in the case of the immersive co-located experience where the interaction between the audience members during the experience is crucial, a system which enables delivery of audio signals without impairing user's normal free-field hearing is necessary. One of the solutions to that problem are earphone drivers coupled with acoustically transparent earpieces [15]. Another way of delivering the audio are nearfield open ear devices mounted in front of the ears, oriented towards the entrance of the ear canal. However, the small size of transducers

in this type of reproduction device often leads to a non-linear frequency response and attenuation in the low frequency region [16]. To mitigate the frequency response problem, a hybrid reproduction system with loudspeakers is proposed.

3.2.2 Loudspeaker playback

Loudspeaker playback is broadly used in cinema production. Surround speaker systems enhance the perception of envelopment and spaciousness of a sound scene, and enable designers to create an impression of movement of the sound sources around the listening space. A sub-woofer speaker provides energy at low frequencies, which are especially important in cinematic sound design where emotional impact is greatly enhanced by the use of low frequency sound effects [17]. The limitation of a speaker-only system is that the subjective localization of sound sources is not very accurate. Furthermore, it is hard to create a convincing virtual source positioned closer to the listener than the physical position of the loudspeaker. Besides that, in surround speaker setups the sweet spot is usually limited to the central seating position [18]. Adding more speakers to the setup can enhance the immersion and allow for a more accurate trajectory of movement of the sound sources, but the rendering of near-field sound sources is still limited. The use of wavefield synthesis techniques would allow designers to create a very realistic soundfield around the listening area, but its implementation is very expensive and requires acoustic treatment of the performance space [19].
The proposed use of a hybrid reproduction system can take advantage of both types of reproduction methods and deliver high-quality convincing and cohesive sound. Hear-through earphones enable transparent hearing and deliver an individual mix of binaural audio to each user, providing an accurate match with the visuals. The speaker system improves the experience by providing a full frequency spectrum of sound and enhances the 3D auditory scene with far-field sounds, which can improve the externalization [20].

3.3 Technical challenges

3.3.1 Delay

An audio signal played simultaneously through earphones and loudspeakers will not reach a listener at the same moment in time. Signals from loudspeakers arrive to a listener delayed, and the delay will depend on the distance of the listener from the speaker. This issue might be especially important if the sounds played through the device have a short temporal structure (significant amount of transients) where the delay can be perceived by ear [21]. Delay adjustments at the binaural device, for some of the seating positions, might be necessary in larger spaces. Contrarily, this issue is not salient for sounds with longer temporal structures.

3.3.2 Distance attenuation

For each doubling of distance from the source, the intensity of a signal in free field decreases by 6 dB [22]. Depending on the distance of each listener to each speaker in a chosen configuration, the signal may be attenuated to a different degree. This seat-dependency must be taken into account during the mixing stage to ensure a proper sound level for each of the more sensitive positions.

3.3.3 HRTFs rendering

When listening to loudspeakers, listeners perceive sound through their own natural HRTFs. The situation is different with spatial sound on nearfield devices: a listener would be usually delivered sound processed through generalized HRTF filters, which are likely non-ideally tuned to their personal spatial cues response. This might cause a problem if too similar sounds were to be played through both the earphones and the loudspeakers, as the HRTFs may color the signal spectrum and create a timbre mismatch between the two delivery methods [23]. In an ideal situation, individualized HRTF filters, measured in the listening room for each seating positions, would deliver the highest possible spatial audio quality. But this is simply unfeasible, a more efficient workaround is to keep the content distinct for the two reproduction systems.

3.3.4 Room acoustics

When playing back sound on speakers, the room acoustics influences the end signal as it reaches listeners' ears. The acoustic character of the room might significantly differ from the one given to the virtual sound layer played at the earphones (for example, when using artificial reverberation). The acoustic mismatch might negatively impact the perceptual auditory integration

between the two reproduction elements, affecting the smoothness and capability of immersion into the experience. For this reason, the exhibition room should be acoustically treated to reduce reflections. If this cannot be accomplished, it is desirable to adjust the reverb of the virtual content to be similar or even slightly longer than the actual room reverb, in order to minimize the perceived reverberation mismatch between the two reproduction systems.

3.4 Sound design

The sound-design style for co-located cinematic VR is based on traditional cinematic approaches with the addition of spatial audio processing. The main stems necessary for film audio soundtracks include dialog, music, and sound effects (foley, sfx, backgrounds and ambiances). In contrast to the stereo or surround cinematic mix, there are more audio formats available to the sound designer in shared narrative VR. The audio layers can be reproduced using different spatial audio techniques, even concurrently: as audio objects using binaural rendering, as Ambisonics files that capture the whole sphere of sound around the listener [24], as traditional stereo on headphones, as surround formats through VSS processing on headphones [25], or as a channel-based mix played back on speakers.
Each sound layer has different requirements in terms of spatial processing and diffusion (see Table 1). The dialogue and foley require very precise scene placement, achievable with binaural rendering. Distance cues such as level attenuation, direct sound to reverb ratio, as well as radiation pattern, need to be added to ensure a realistic sound change during the character's movement [22]. Background and ambiance sounds are usually more diffused. They can be reproduced in Ambisonics format on headphones to allow the rotation of the soundfield according to the listener's orientation, as static stereo tracks, or in surround format on speakers. The music can be reproduced as either diegetic or non-diegetic using different spatial audio formats [26]. When using binaural techniques, the music will be perceived as coming from within the virtual space, thus diegetic. When instead using stereo or surround speaker playback it will more likely to be perceived as coming from the background.

4 *"Cave"*: A case study

An early version of the proposed sound system was implemented for a six-minute virtual reality co-located narrative piece called *"Cave"* [1, 8], that took place in a single, multi-user, virtual environment. The story involved one main character, one supporting character, and a virtual mammoth. The experience was prepared for a 30-member virtual audience, separated into two groups in the thrust stage format (Fig. 1). The audience could see each other as avatars inside the virtual experience while the position and orientation of their heads were tracked using the headsets' IMUs, so that the avatars' heads moved inside the virtual space accordingly. The VR experience was built using the Unity game engine [27], and was executed on standalone headsets for each audience member, using a smartphone as the control unit. The game networking service Photon [28] was used for sending all data and signals between all devices [8].

Layers	Binaural	Ambisonics	Surround	Stereo
Dialogue	✓			
Foley	✓			
Ambiances		✓	✓	✓
Music	✓	✓	✓	✓

Table 1: Suggested audio techniques for audio layers.

Fig. 1: Audience avatars in the VR experience *"Cave"*. (© and Art: Kris Layng, 2018)

4.1 Design choices

The audio layers used in the experience consisted of dialogue, music, sound effects, and ambiances. The sound effects and the dialogue materials were treated as point source objects, connected to a visual component in the three-dimensional space. While the original content of these materials was in mono format, a dynamic binaural rendering engine tool (Steam Audio) transformed the sound objects into responsive stereo binaural streams, responding to the spatial relationship

between characters and listeners. Additional distance cues were tuned separately for the foley, while the dialogue track had to maintain constant intensity in order to keep it intelligible. As the rendering was individual per-device, each audience member was able to get a unique sound perspective into the scene.
The ambiance sounds were created both from a mix of stereo recordings and sound objects. Important and constant background sound, such as the wind noise in the entrance to the cave or water stream, were positioned at diffuse point sources within the scene, while more general and sporadic sounds, such as drops of water, were rendered in stereo and not given specific spatial positions. The music track was exported as non-spatialized stereo in order to give it a sense of separation from the dialogue and sound effect layers.

4.2 Audio workflow

The audio implementation for the project was done in the Unity Engine using the Steam Audio plugin [29] for sound spatialization. The IMU tracking data was utilized to affect the individualized mix for each of the participants. The audio stems were designed and edited in Pro Tools [30] following a traditional linear workflow, as in film post production, using a 2-D video rendering of *"Cave"* provided for reference for all editing. The mixing stage was split between Pro Tools and Unity. All equalization, compression, and limiting was applied in Pro Tools prior to Unity to be sure that all stems blended well together before being spatialized. Those were mixed "dry" so that the reverb could be rendered in Unity based on the listeners' locations. Once the stems were imported into Unity, they were attached to the corresponding visual object or rendered in stereo format. Spatialization, reverberation and additional equalization for stems was programmed per-user using the Steam Audio plug-in, while the Unity audio mixer was used to introduce general changes in the intensity of the audio layers. Finally, the synchronization between visual and audio layers was implemented by using *timeline* playback tool, for both visuals and audio tracks.
One general constraint related to audio production is the availability of appropriate tools for efficient audio editing, mixing, implementation and monitoring in VR. The workflow suffered from a non-unitary approach due to the lack of proper audio editing and monitoring tools within the game engine. This fact entailed long back-and-forth process between the digital audio workstation and the game engine in order to achieve a satisfactory result on the final reproduction system.

Fig. 2: Headset with prototypes of nearfield open-ear device from Bose. (Photo: Eric Chang, 2018)

4.3 Reproduction system

The sound reproduction system consisted of one single speaker with subwoofer placed in the middle of the stage, and a prototype nearfield open-ear device produced by Bose Corporation specifically for *"Cave"*, which allowed transparent hearing. The devices were mounted in front of the ears and were oriented to project towards the entrance of the ear canal (Fig. 2).
The speaker unit was used mostly for the sound effects of the virtual mammoth and it was made sure that the physical position of the speaker matched the virtual position of the mammoth's avatar as seen by all audience members. In this implementation, one speaker was sufficient because the visual object for which the sound effects were rendered was static. With more moving elements, more speakers would be necessary to ensure proper localization of the sounds chosen to come from the loudspeakers. A system-wide calibration was performed to achieve a proper blend between the nearfield devices and the loudspeakers, and to ensure that the levels would be comfortable for each member of the audience.

5 Survey

To explore the efficacy the sound implementation, we developed a questionnaire offered to all 1,927 users immediately after watching the experience. We received 374 responses (a 19% response rate), of which 317 were complete and used to provide richer insights into user experiences.

Impact of Familiarity with VR on Audio Experience

Item	Low Familiarity	Medium Familiarity	High Familiarity
I enjoyed the *"Cave"* experience	5.88 (1.12)	5.72 (1.26)	5.92 (1.19)
I enjoyed the audio elements of the experience	5.76 (1.02)	5.77 (1.18) 5	6.06 (1.0)
I understood some audio elements were spatialized (placed in the room)	5.32 (1.54)	5.66 (1.50)	5.97 (1.38)
I felt the audio spatialization helped me feel immersed in the experience	5.79 (1.27)	5.90 (1.25)	6.14 (1.10)

Table 2: Judgments based on the user's familiarity with VR. The questions were given on a Likert scale (1, Strongly Disagree to 7, Strongly Agree).

The questions took one of four formats: i) 7-point Likert-type scale, ranging from 1 (Strongly Disagree) to 7 (Strongly Agree), ii) single choice response, potentially including an "Other" option with short text entry, iii) multiple selection response, potentially including an "Other" option with short text entry, and iv) Open-ended text response.

5.1 Respondent profile

Over half of the participants reported their age as under 35, with 57 (18%) reporting 18-24, and 117 (37%) reporting 25-34; among those over 35, 65 (21%) reported 35-44, 42 (13%) reported 45-54, 18 (6%) reported 55-64, 10 (3%) reported 64+, and 8 (<3%) preferred not to give an age. Participants were asked where they had been seated in the audience from a list of 5 areas; they came from a relatively even distribution of the areas with the fewest responses from the right front row (57, or 18% of the sample) and the most responses from the left front row (76, 24% of the sample).

Participants reported how long they had used virtual reality technology, and a single self-reported value for level of expertise with virtual reality. Based on these responses, we created three categories of familiarity with VR: High familiarity participants (86, 27%) had used VR for a year or longer, and rated themselves a 6 or 7 (Extremely proficient); Low familiarity participants (88, 28%) had used VR for less than a year, and rated themselves a 3, 2, or 1 (Not at all proficient); and Medium familiarity participants (143, 45%) provided any other combination of time and rating of expertise.

In sum, this sample of conference attendees comprised relatively young, technically-savvy professionals. Although participants were not randomly selected from the audience, they viewed the experience from all areas, and had varied expertise with virtual reality.

5.2 User experience

All participants were asked for their judgments of the experience as a whole, and specific questions on audio quality and spatialization. Participants enjoyed *"Cave"* and the audio elements of the experience, regardless of experience with VR (F's < 2.3, p's > .01). Participants at all levels of VR experience also reported understanding that audio was spatialized, and the spatialization contributed to feeling immersed in the experience (Table 2).

However, responses did differ based on seating for *"I enjoyed the audio elements of the experience"*. Participants in the right back row had lower reported ratings (M = 5.45, SD = 1.22) than other 4 locations (Means > 5.7), a small but significant difference, F = 3.51, p = .008, η^2 = .044.

Most participants indicated that they enjoyed the score, effects, and foley effects; a smaller but substantial number of participants reported enjoying the dialogue (Table 3).

6 Discussion

In the presented case study of sound design and reproduction for immersive co-located virtual reality theatre, the cumulative effects of real-world elements (including seating, networking, and delivery of visual elements of virtual reality) and the implementation of a hybrid reproduction system delivered an effective

Which elements (if any) did you enjoy?	
Element	Percentage of participants
The score	81.1%
Effects	61.5%
Foley	61.2%
Dialogue	35.5%

Table 3: Percentage scores for different audio layers

sound experience for this shared virtual art piece.
The results of the survey suggest that the presented approach can be sufficient for delivering an immersive audio layer given the defining elements of this particular experience, although the results are descriptive for this convenience sample, and not intended to generalize to a wider population. Participants, in general, enjoyed the audio elements of the experience. However, the audience members seating in the right back row gave significantly lower ratings of enjoyment, although they did not indicate an impact on their understanding of spatialization, and feeling that the spatialization helped them feel immersed in the experience. It is likely that these lower ratings were affected by audio glitches found to occur during several of the showings due to jittery network connections in some devices.
Participants gave lower scores to the dialogue when evaluating the different elements of the audio layers (Table 3). We noticed that the D/A converters on the headsets introduced a significant amount of sound distortion which affected mostly the quality of the dialogue rendering and might be reflected in the results of the survey. This indicates that the quality of headset's audio hardware should be taken into account when choosing a device for VR production.
Most of the participants noticed that the sound was spatialized and felt it helped them to feel immersed in the experience, which suggests that the sound implementation and reproduction was successful to enhance the immersion. However, having a control group evaluating a reference audio track should allow for more robust empirical results which was not possible within the context of this production. Also, more specific evaluation questions could bring more conclusions about the perception of sound during the experience.

The implementation described in the case study was limited to a single speaker and subwoofer. Adding more speakers surrounding the audience may further improve the immersion and allow the reproduction of more layers of audio other than sound effects, e.g. music or ambiances. Playing music tracks through the loudspeakers could indeed help with better separating the background music and the dialogues. We also noticed that some of the instrument tracks were perceived as coming from within the scene even though they were rendered in stereo.
Our implementation did not take into account the acoustics of the performance space due to the production limits. This resulted in sounds played through the speakers to have a different acoustic characteristics than the binaural layer. Furthermore, informal investigation revealed that even though the audience could see and hear each other inside the experience, their voices were not fully perceived as though they were coming from the same space as the action. Placing microphones above the audience may solve that problem. The sound from microphones would be processed in real-time through the same reverb processor used within the experience, to ensure consistency between the sounds of the real-world natural environment and virtual scene.
Another challenge we encountered during production was the asynchronous playback between speakers and nearfield devices. Even a small delay would cause perceptible distortions of those sounds played through both systems. We solved this problem by removing all of the transient sounds from the speaker playback, leaving only the sound effects of a longer temporal structure.

6.1 Future work

Although the discussed production work helped to elucidate and expand the sound-design theory for this type of narrative VR experiences, more empirical work is required to investigate and validate the best approaches for an effective delivery of sound. While the discussion of the principles is mostly derived from professional experience and qualitative critical perspectives, the assessment of the proposed technical systems, the factors and the challenges involved can benefit from both commercial production analysis and laboratory experiments. Further insights can be gained by literature advances in similar applications such as multi-player VR interactions and spatial audio technology.
In practice, future productions would benefit from a revised questionnaire linking and addressing the accuracy of sound localization, device type (VR vs AR),

seating position, sound source externalization, acoustic treatment and matching, and quality of interaction between audience peers. Having controlled conditions in a laboratory experiment would create robust conclusions about the importance of each one of the single elements which compose the proposed hybrid system.

7 Conclusion

This paper presented a discussion around the principles, factors and limitations of the sound-design theory related to the novel field of co-located narrative VR experiences. A first draft of this theory, reviewing technical challenges and proposing a solution based on hybrid reproduction systems, has been derived from practical experiences within prototype productions. The experience with the production of "Cave" is discussed as a platform were some of these principles were investigated and addressed to achieve insights which inform the authors' proposed framework.
Having this base to work upon, future empirical data will help to validate, sharpen and define the guidelines that may drive the creative choices of VR sound designers. It is reasonable to expect, that in the near future, technological advances are likely to affect and update the current conversation.

8 Acknowledgments

"Cave" premiered at ACM/SIGGRAPH 2018 on August 12–16 in Vancouver, Canada. The production was created and presented by a large team of people and supporters besides the authors: Kris Layng, Sebastian Herscher, Thomas Meduri, Tatiana Turin, Jess Bass, Hanako Greensmith, MaYaa Boateng, Kat Jeffery, Ryan Shore, Eric Chang, Paloma Dawkins, Mia Mamede, Susan Darvishi, Marcus Z. Guimaraes, Pasan Dharmasena, Cynthia Allen, Lily Cushman, Connor DeFanti, Rowan Smith, Angela Yu, Dan Lazar, Brandon Buikema, Bose, Google Daydream, Lenovo, Future Tech and NYU Future Reality Lab.

References

[1] Layng, K., Perlin, K., Herscher, S., Brenner, C., and Meduri, T., "CAVE: Making Collective Virtual Narrative," *Leonardo*, 52(4), pp. 349–356, 2019.

[2] Gochfeld, D., Brenner, C., Layng, K., Herscher, S., DeFanti, C., Olko, M., Shinn, D., Riggs, S., Fernández-Vara, C., and Perlin, K., "Holojam in Wonderland: Immersive Mixed Reality Theater," *Leonardo*, 51(4), pp. 362–367, 2018.

[3] Kobayashi, M. and Ueno, K., "The Effects of Spatialized Sounds on the Sense of Presence in Auditory Virtual Environments: A Psychological and Physiological Study Abstract," *Presence*, 24(2), pp. 163–174, 2015.

[4] Brinkman, W.-P., Hoekstra, A. R. D., and van Egmond, R., "The effect of 3D audio and other audio techniques on virtual reality experience," *Annual Review of Cybertherapy and Telemedicine*, pp. 44–48, 2015.

[5] Zhao, J., Zhang, B., Yan, Z., Wang, J., and Fei, Z., "A study on the factors affecting audio-video subjective experience in virtual reality environments," in *International Conference on Virtual Reality and Visualization, ICVRV 2017*, pp. 303–306, IEEE, 2017.

[6] Diemer, J., Alpers, G. W., Peperkorn, H. M., Shiban, Y., and Mühlberger, A., "The impact of perception and presence on emotional reactions: a review of research in virtual reality," *Frontiers in psychology*, 6, p. 26, 2015.

[7] Churchill, E. F. and Snowdon, D., "Collaborative virtual environments: an introductory review of issues and systems," *Virtual Reality*, 3(1), pp. 3–15, 1998.

[8] Herscher, S., DeFanti, C., Vitovitch, N. G., Brenner, C., Xia, H., Layng, K., and Perlin, K., "CAVRN: An Exploration and Evaluation of a Collective Audience Virtual Reality Nexus," *UIST 2019: 32nd ACM User Interface Software and Technology Symposium*, 2019.

[9] Rämö, J. and Välimäki, V., "An allpass hearthrough headset," in *2014 22nd European Signal Processing Conference (EUSIPCO)*, pp. 1123–1127, IEEE, 2014.

[10] Horowitz, S. and Looney, S. R., *The essential guide to game audio: the theory and practice of sound for games*, Routledge, 2014.

[11] Paterson, J. and Kadel, O., "Immersive Audio Post-production for 360° Video: Workflow Case Studies," in *2019 AES International Conference on Immersive and Interactive Audio*, York, UK, 2019.

[12] Gupta, R., Ranjan, R., He, J., and Gan, W.-S., "On the use of closed-back headphones for active hear-through equalization in augmented reality applications," in *International Conference on Audio for Virtual and Augmented Reality*, Redmond, WA, USA, 2018.

[13] Begault, D. R. and Trejo, L. J., "3-D sound for virtual reality and multimedia," 2000.

[14] Guezenoc, C. and Séguier, R., "HRTF Individualization: A Survey," *Audio Engineering Society Convention 145*, 2018.

[15] Martin, A., Jin, C., and Andre, V. S., "Psychoacoustic Evaluation of Systems for Delivering Spatialized Augmented-Reality Audio," *Journal of the Audio Engineering Society*, 57(12), pp. 1016–1027, 2009.

[16] Gutierrez-Parera, P., Lopez, J. J., and Aguilera, E., "On the influence of headphone quality in the spatial immersion produced by Binaural Recordings," in *Audio Engineering Society Convention 138*, Audio Engineering Society, 2015.

[17] Whittington, W., *Sound design and science fiction*, University of Texas Press, 2007.

[18] Rumsey, F., "Basic Psychoacoustics for Surround Recording," in *AES 22nd UK Conference*, pp. 1–9, University of Surrey, Guildford, UK, 2007.

[19] Boone, M. M., Verheijen, E. N. G., and van Tol, P. F., "Spatial Sound-Field Reproduction by Wave-Field Synthesis," *J. Audio Eng. Soc*, 43(12), pp. 1003–1012, 1995.

[20] Mueller-Tomfelde, C., "Hybrid Sound Reproduction in Audio-Augmented Reality," *Audio Engineering Society 22nd International Conference on Virtual, Synthetic and Entertainment Audio*, pp. 1–6, 2002.

[21] Scharine, A. A., Cave, K. D., and Letowski, T. R., "Auditory Perception and Cognitive Performance," in C. E. Rash, M. B. Russo, T. R. Letowski, and E. T. Schmeisser, editors, *Helmet Mounted Displays - Sensation, Perception and Cognitive Issues*, chapter 11, pp. 391–490, U.S. Army Aeromedical Research Laboratory, Fort Rucker, Alabama, 1999.

[22] Shinn-Cunningham, B., "Distance cues for virtual auditory space," *Proceedings of the First IEEE Pacific-Rim Conference on Multimedia*, pp. 227–230, 2000.

[23] Takanen, M., Hiipakka, M., and Pulkki, V., "Audibility of coloration artifacts in HRTF filter designs," in *Audio Engineering Society Conference: 45th International Conference: Applications of Time-Frequency Processing in Audio*, Audio Engineering Society, 2012.

[24] Gerzon, M. A., "Periphony: With-height sound reproduction," *Journal of the Audio Engineering Society*, 21(1), pp. 2–10, 1973.

[25] Pike, C. and Melchior, F., "An assessment of virtual surround sound systems for headphone listening of 5.1 multichannel audio," in *Audio Engineering Society Convention 134*, Audio Engineering Society, 2013.

[26] Neumeyer, D., "Diegetic/nondiegetic: A theoretical model," *Music and the Moving Image*, 2(1), pp. 26–39, 2009.

[27] "Unity Real-Time Development Platform | 3D, 2D VR & AR Visualizations," `https://unity.com/`, (Accessed on 08/01/2019).

[28] "Multiplayer Game Development Made Easy | Photon Engine," `https://www.photonengine.com/`, (Accessed on 08/01/2019).

[29] "Steam Audio," `https://valvesoftware.github.io/steam-audio/`, (Accessed on 08/01/2019).

[30] "Pro Tools - Music Software - Avid," `https://www.avid.com/pro-tools`, (Accessed on 08/01/2019).