EUROGRAPHICS Workshop on Graphics and Cultural Heritage (2021)
A. Chalmers and V. Hulusic (Editors)

Short Paper

3D sound for digital cultural heritage

A. Mušanović[1], B. Mijatović[2] and S. Rizvić [3]

[1]Radio and Television of Bosnia and Herzegovina
[2]Sarajevo Film Academy, Sarajevo, Bosnia and Herzegovina
[3]Faculty of Electrical Engineering, University of Sarajevo, Bosnia and Herzegovina

Abstract

Virtual Reality enables the users to experience cultural heritage. Time travel through past times is transferring us in virtual environments with 3D reconstructions of cultural monuments inhabited by historical characters. The full immersion in this different reality can be achieved only with proper spatialization of sound. In this paper we discuss the related work in 3D sound implementation for digital cultural heritage applications and compare it with our experiences.

CCS Concepts

• Applied computing → Sound and music computing; Digital libraries and archives; • General and reference → Design; • Human-centered computing → Virtual reality;

1. Introduction

The world is moving into digital. There is no single aspect of our lives not influenced by digital technologies. Smart phones, mobile devices, Virtual and Augmented Reality are surrounding us and extending our reality. One of the advantages of digital age is the possibility to experience cultural heritage. Visits to museums and cultural monuments are enhanced with immersive experiences of events and environments from the past.

Even storytelling, a way of communication as old as the mankind, became digital. Interactive digital storytelling (IDS) communicates the information in an attractive and educational manner. Our group is researching IDS methodology since ten years [RDA*17]. Our digital heritage applications use Virtual Reality (VR) to create immersive experiences of past events, impossible to en- counter in reality. Many of them contain VR video stories as combinations of 360 video, computer animation and actors recorded against the green screen.

In order to be immersive, these VR videos need to have the adjoined 3D sound, so the user can hear the voices from right positions in space. Recording and production of such videos are topics of this paper. The standards for embedding spatial audio format into the linear designs are not developed. We present our solutions for sound post-production of VR videos to make them spatialized, while leaving the music and voice over headlocked. The lack of post-production standards and general solutions makes this work very challenging, particularly if the videos are to be published on YouTube.

We will introduce the readers with sound spatialization concepts and present an overview of methods in sound spatialization for Virtual Reality with their advantages and drawbacks. Then we will describe our methodology in sound recording and post-production. The use of this methodology will be presented through case studies of projects we implemented. At the end, we will offer our conclusions and directions for future work.

2. Related work

2.1. Sound spatialization concepts

Hearing is a very sophisticated device brought by nature. What one perceive as descriptive sound properties is actually a result of intense calculations done subconsciously by the brain. A raw sensation captured by the ears basically contains only simple amplitude modulation of spectral content in temporal domain and without some background processing enforced by the experience one would never realize what is heard and from where.

Distance and direction are complex sound properties which are perceived psycho-acoustically by detecting variations of frequency, timbre and loudness of the sound source along with the reverberation from the surrounding objects. Sense of direction depends on the amplitude, time and phase differences between the sounds perceived by each of the ears. Although in monophonic technique sense of distance can be achieved to some extent by variations in spectral content and reverberation, sense of direction is not possible when using just one speaker and/or just one ear.

When a sound reaches both ears simultaneously with the same loudness it is perceived as centered. As it is heard louder at one of the ears, it is perceived as coming from that side. Different loudness ratio between the sounds reaching each ear allows precise stereo

DOI: 10.2312/gch.20211415

panning of the source. That is the basic principle of the stereophonic technique [url21a], invented by Alan Blumlein in 1931.

Ambisonics is a special surround format which covers the full sphere of the sound field. It was developed during 1970s under the auspices of the British National Research Development Corporation [url21b]. Despite its great potential and wide range of possible applications, it only recently made some commercial success thanks to the current availability of powerful digital signal processing. Ambisonics audio format is based on three dimensional extension of the traditional mid-side (MS) recording technique. While MS technique employs one cardioid or omnidirectional "mid" microphone and one figure-8 diferential "side" microphone, ambisonics contains one omnidirectional "mid" channel and three differential "side" channels designated to left-right, up-down and front-back axes. Encoded sound field is stored as four channel audio file called B-format and the channels are labeled W, Y, Z and X, where W is omnidirectional channel and other three are differential channels for three axes. Currently there are two channel ordering standards: Furse-Malham (FuMa) - WXYZ and Ambisonics Channel Number (ACN) - WYZX [url21c]. Output channels of the most ambisonics microphones are arranged in the FuMa (WXYZ) order while the ACN (WYZX) order is more used by the VR processing and reproduction standards.

Contrary to the traditional sound recording techniques which always designate some transducers in the systems as primary/focused and other ones as auxiliary/ambience, ambisonics format is fully isotropic. It means that every sound from any direction is treated equally and it can be further manipulated to sound both "on axis" and "off axis", in traditional sense of miking technique. Ambisonic encoded audio can be freely rotated during the reproduction and every source in the sound field can be focused and unfocused using head related transfer function (HRTF) while changing the virtual distance from the listener and moving all across the sonic sphere, which makes for an excellent platform for interactive sound field manipulation in VR systems.

3. Our methodology of 3D sound implementation

3.1. Sound recording

Filming of 360 videos is not entirely new, however expansion of 360 videos started with commercialization of virtual reality headsets. These devices made possible one of our longest dreams regarding videos – to be a part of one. In his book [Woh19] the author states that VR videos are one of our oldest and most persistent dreams. Filming of space in 360 degrees is definitely one of our newest challenges which we tackled throughout our projects of digitising cultural heritage and bringing it closer to broad audience [RBO*21]. Every director we worked with found it challenging to actually “direct” attention of viewers in full 360 environment. Basic tool for storytelling of every director is framing which is completely absent from 360 videos. Directors must switch their attention to scene setup, blocking and placement of actors and objects in space. Distance from the camera plays a huge part in this as it enhances emotions and immersion into the video [PDSS17].

Camera placement becomes crucial for narration in 360 video because everything is actually in frame. Additional lights can only be added as part of scenery since no crew or equipment can be placed anywhere near the action. The equipment itself must now become part of the scene which leads us to a very demanding part of 360 video – recording of the sound. There are many ways to record the sound on the set, but the most important microphone on the set is the boom mic, operated by the boom operator. It is always the main microphone for the dialog backed up by additional lavalier microphones and ambient microphones. Short explanation of actual recording is very well summarized by Henri Rapp [Rap20]. However, as already pointed, there is no place for a boom operator on 360 film set.

Sound recording must be approached differently in 360 video. Some manufacturers are already making microphones designed to record full ambisonics sphere which we will be analyzing later, but those systems are still new and are not doing all the work. As we said before, placement of actors and objects plays a crucial part of 360 video filming, and for immersive feeling in reproduction of final VR video, placement of microphones becomes of high importance too. In our sets usage of lavalier microphones proved crucial for sound recording. Not only can they be placed on the actors and hidden in their costumes, but those microphones actually move through space which is very important aspect for 360 video in contrast to more common 2D video. Additional microphones placed around the set and concealed inside the set itself can greatly help with recording of sound movement. Extra care should be pointed towards the level of sound recording. Lavalier microphones are usually omni-directional mics with small capsule and have poor noise reduction. It is therefore imperative to record 48000 Hz / 24-bit audio which carries a huge amount of information for post-production of sound. This format allows us to record sound at much lesser gain resulting in much quieter sound that has very little background noise. 24-bit audio will have enough information for dialog reconstruction in post-production even when recorded at -40 db which is way lower than standard -12 db for voice.

3.2. Sound post-production

We have been modelling the ambisonics sound field in Digital Audio Workstation (DAW) software. The location recordings are made using traditional mono and stereo microphone configurations at near-field and far-field positioning, the same way as on a standard film set. The recorded audio then goes through the usual post-production including selecting best takes, dialogue checkerboarding, noise reduction and other restoration, dynamic processing, equalization and other processes needed. Once all the audio sources, music and ambiences are ready and sounding good over ordinary stereo monitoring, the actual building of the ambisonics sound field begins.

DAW’s submix busses, aux busses and master bus are set up for four channels labeled W, Y, Z and X (Figure 1). In order to properly monitor the ambisonics audio over the headphones during work, one must use a plugin on the master bus which converts the ambisonics to binaural. It is mandatory to have it bypassed before the master rendering in order to prevent doubled binaural encoding. The resulting 4-channel ambisonics encoded audio file is ready for binaural decoding by a playback engine. Each audio source which is meant to be spatialized must pass through the dedicated ambisonics panner plugin in order to properly encode its position in the sound field sphere. Most of the needed plugins are available in open source and freeware form. As the technology is relatively new, still experimental and in the constant development, the efficiency of the spatial panning, amount of phase-related colouring and the compatibility with the binaural projection varies among different implementations. So, we are at constant researching for new and improved solutions which would bring even better sounding end results. Unfortunately, there are still many bugs and incompatibility problems between the DAW hosts and various plugin implementations, as well as some unpredictable cross-platform results. There is a vast room for further research and improvements of this technology for the time to come. Fortunately, some newer DAW hosts recently started to include the ambisonics panners as standard, working in similar manner as the standard surround panners.

Tracking the audio sources across the sound field is the most important step for obtaining an accurate sonic representation of the sound sources displayed in the visual. This part of the post-production process is also the most time consuming, especially when there are lot of moving sound sources in the scene. All the movements of the sound sources in the scene have to be tracked as accurately as possible using DAW's automation curves. The position of every spatialized audio source is defined with three-dimensional vector coordinates while the origin is placed at the virtual center of the listener's head. Depending on the panner implementation, one can use Cartesian or polar coordinate system. In our experience, a polar coordinate system is more efficient to use in linear productions, as the anomalies in tracked distance are more tolerable than the anomalies in sound direction. Once the direction of the sound source is tracked as accurately as possible, actual modelled distance can be approximated within a reasonably wide margin of convincing ranges. Fortunately, ears are a bit less sensitive to a distance than eyes. Reference VR video is imported in the DAW project as standard video file and previewed as a planar-projected "unfolded" version of the original spherical video. Hence, all straight lines are shown as arches on the preview and the sense of size proportions and perspective is skewed. Six parts of the projected picture represent six side of the cubical space around the scene. While DAWs implement a direct playback of VR videos, all coordinates of the sound sources have to be determined by referencing to that weird-looking and surreal visual representation. Pure direction is relatively easier to track when using polar coordinates, but accurately judging distance and elevation requires some skills and trial-and-error cycles of previewing draft renders on the VR headset. Fortunately, some developers recently started to implement DAW plugins which enable a rotatable VR video preview, as the one included with the fb360 Spatial Workstation. Some of them even offer the solutions for tracking the sound source directly within the headset using VR controllers, but there are still some compatibility and stability issues with different DAWs and platforms.

Tracking the sound sources only applies to linear designs for VR video production. When working with Unity, accurate directional tracking is a non-issue, as all the sound sources are already linked to their parent assets. Once the visual object moves into the scene, the direction of the linked sound sources automatically moves along, so drawing separate automation curves for the sounds is not needed.

As in the standard film production, layering additional foley sounds, effects and ambiences is a good way to enhance the realism of the scene events and the emotional impact. Adding judiciously amount of appropriate reverberation and early reflections can greatly increase the realism of localization and sense of the projected space. The Doppler effect makes the simulated moving of the sound sources more convincing. Even a very small amount of pitch shifting during the move can make it more realistic.

Headlocked sound elements such as music and narration have to be routed over the separate stereo output bus in the DAW, so they bypass the ambisonics panners and the binaural encoder inserted for monitoring the spatialized sources. When rendering the master, the end result consists of six channels of audio: the four-channel ambisonics audio file for the spatialized content and the two-channel stereo file for the headlocked content.

Until recently, we did not have a technical possibility to implement headlocked elements in our linear designs for VR videos, due to inability to combine both the spatialized and non spatialized audio in the same playout on the platforms we used. The biggest practical issue of this was unwanted HRTF rotation of the music, narration and other undiegetic elements along with the rest of the sound field. Inside the margin of about 45 degrees from the preset azimuth angle the narrowing and skewing of the music's stereo image sounded somewhat tolerable, but once the listener's head reaches angle of 90 degrees to the music's azimuth, stereo image of the music collapses to mono, soundstage details from the music are lost and some phase cancellation on certain frequencies occurs. Having these limitations, we had to be creative in order to minimize the unwanted artifacts. In some scenes the music was still spatialized but panned with respect to probability what direction would listener look toward in particular moment, which is usually the actor who is speaking or some other event which takes attention. In our projects Baiae and Roman Heritage in the Balkans [url21q], dedicated to the ancient Greek and Roman historical periods, the background music for the scenes is arranged for a small ensemble consisting of few ancient musical instruments. The music is not placed in the sound field in form of ordinary stereo mix (which actually appears in the ambisonic sound field as planar object). Instead, every single instrument is spatialized and placed into the scene. That effectively made the music diegetic. When the listener is looking at the actor who is speaking, the instruments are arranged in slight arched formation behind the listener, making a virtual invisible ancient "band" playing the background music in the same room and under the same acoustic environment. As the formation of the instrument group is not planar, stereo image of the music never fully collapses to mono during a head rotation.

Additional problem is a non-diegetic narration which should be unspatialized and headlocked by definition. Fortunately, there is a workaround which enables headlocking of any mono sound source, even when using only the ordinary 4-channel ambisonics without the separate headlocked channels. A source is routed directly to the W omnidirectional output channel, bypassing the panners. As nothing of the same source is routed to any of three differential channels Y, Z and X, the sound remains effectively unspatialized and headlocked.

Now with Oculus Quest we can use a bit buggy but working solution for embedding ambisonics audio together with separate headlocked stereo for video playing on the headset, using the fb360 Encoder with a working version of GPAC. We are still researching or a viable headlocking solution for our YouTube VR videos as the currently proposed method appears to work inconsistently on some platforms. Although the embedded audio is properly spatialized, some video players show wrong angles of the encoded sound sources and some others play just static ambisonics image fixed into initial position without the HRTF rotation.

4. Case studies

Our team has worked on a number of digital heritage projects, ranging from storytelling VR videos, VR gameplay applications with videos to fiction films in VR. All of them created different challenges for directing, camera work [MR21] and sound. While experimenting with ambisonic sound format through our projects, we substantially increased complexity when it comes to sound recording.

Among our first projects were VR Simulation of Mostar cliff diving [SRH*20] and Sarajevo War Tunnel VR [RBB*19]. Through series of videos the user learns about cultural monuments, and at the end is presented with quiz about what he/she has learned. As a reward for completing the quiz the user can either virtually jump from the Mostar Bridge or walk through the tunnel from wartime Sarajevo. To make videos more immersive we decided to record them in 360 degrees with different actors and participants to guide users through the videos. Videos are simple and narrators in videos are wearing lavalier microphones for sound recording. No additional effects or foley sounds were added to the videos. To increase immersion into 360 videos for Roman Heritage in Balkans, Old Crafts Virtual Museum Baiae Dry Visit projects we decided to record locations without actors and then add actors in postproduction. This way actors could change costumes and roles in videos. This also meant filming on green screen with shotgun microphone and lavalier mic on actors in the studio. Sound conditions were perfect for recording. These applications were made in Unity which gave us opportunity to use ambisonic sound format supported by Unity and place sound sources all around 360 video. Additional sound foley and effects were also implemented around the scene giving more overall depth to the sound. The effect was enthralling and proved ambisonic format has great use in 360 videos and VR applications [SRC*20].

Motivated by these new discoveries we started preparing our first VR Experimental film. The film is happening in an improvised courtroom where the solicitor of revision is trying to convince the judge and jury to rehabilitate the dissident writers, while the solicitor of history is trying to prevent him together with the jury. The story is told in a satirical way. In the courtroom we had two main actors (solicitors of revision and history), a judge, a clerk and several audience members. We used one Tascam DR-70d 4 channel recorder and four lavalier microphones and placed two mics on the main actors (one each), one between the judge and the clerk, and the fourth in the middle of audience members, since they had fewer lines of text. Our sound technician now had to pay close attention on who is speaking to be able to record more distant audience members. This produced some noise which had to be repaired in post-production. However, this insured that we have the sound recorded on exact places, and with help from camera spatial sound we were able to produce our first 4 channel ambisonic video and export it directly from Adobe Premiere Pro to Youtube [url19].

Battle on Neretva VR project was a combination of video and 3D rendered scenes in Unity in an educational game application about one of the most famous battles of the WWII in Bosnia and Herzegovina. We took similar approach in microphone placement for videos that were a part of the application. Each scene contained the commander (main character) who was giving orders, and three partisans of whom two had two replicas. We placed lavalier microphones on each of them and on fourth channel we had additional boom mic hidden inside the scene to help with sound recording. Combination with camera's spatial sound in post-production gave us satisfactory results again, and as we already pointed, Unity has a native sup- port for these formats so we did not have any problems in the final version.

The most demanding work in sound domain came in our last project. It is a VR film about Bosnian poet Ilhamija and his execution for criticizing the harsh Ottoman rule. Field recording of the film was in a way easier that of those before. We only had two characters in the scenes which are otherwise rather quiet and empty. Only two lavalier microphones were used to record dialogues. However we also recorded narrative parts with shotgun microphone and decided to lay music throughout the film. That proved to be quite challenging in post-production because we had to use a mix of spatial sound and headlocked sound. Exact step of post-production was described in the previous Section and the end result was again very satisfactory. The music and narration are heard in full range all the time around the user, while other sounds float around the space. This gives even more cinematic feel while watching the film that is happening around us bringing us closer to a fully immersive video experience.

5. Conclusions and future work

The sound is usually ranked lower than image in production of films. This unwritten rule of undermining sound production in films has a long history, and is usually seen in films made by students and productions with lower budget. But even though sound is often overlooked in production of films, it is surely important as the picture itself. We are seeing the same pattern in VR films. As the quality of sound rises, the quality of VR video rises too. WE encountered it in our projects, since sound in VR just got its 3rd dimension opened up. In our next projects we aim to use even more microphones positioned around the set in combination with ambisonic microphones to see the results. Additional sound effects of changing distance from the user, and camera moving away from sound sources or closer to them will surely increase the immersion into the video. 3D sound was never this interesting to research and develop, as it is a part of the pioneering work in developing a new standard of the film language grammar - the VR film.

References

[MR21] Mijatović B., Rizvić S.: Virtual reality video in digital cultural heritage applications. Virtual Archaeology 2021. 4

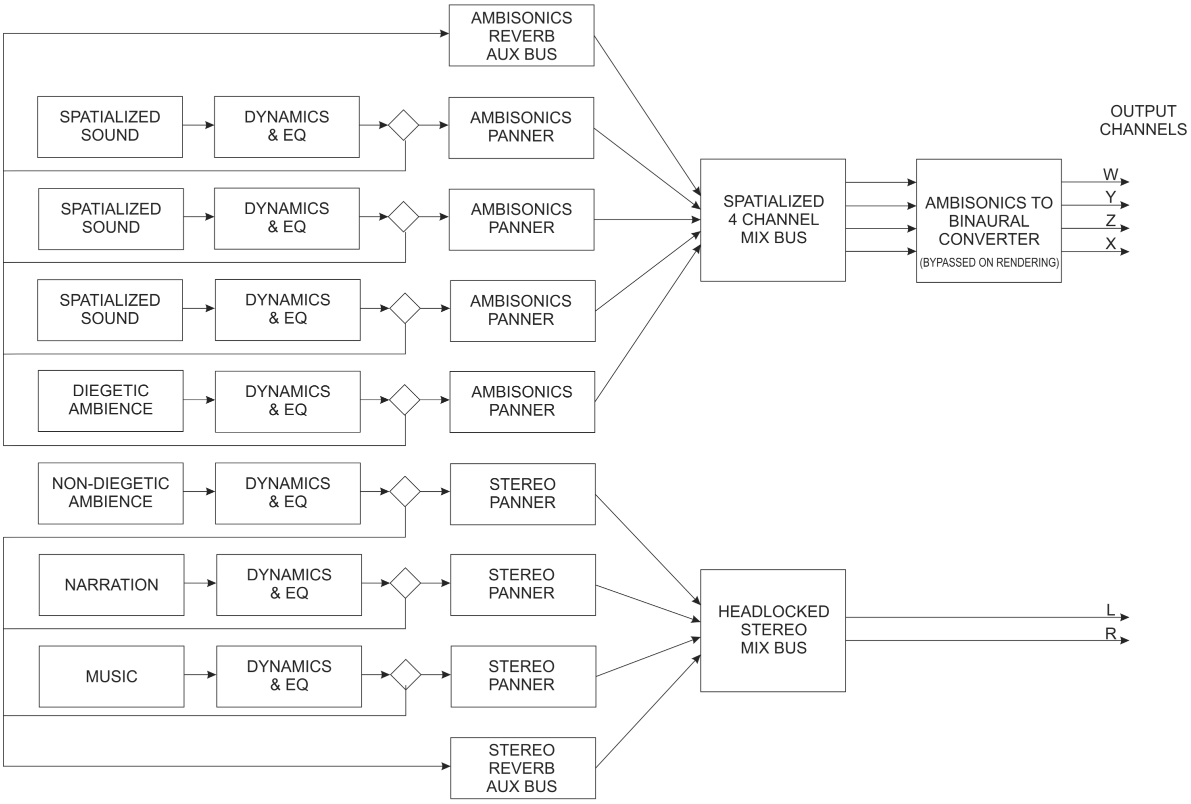

Figure 1: *Signal flow of a typical ambisonics mix with headlocked elements*

[PDSS17] Pope V. C., Dawes R., Schweiger F., Sheikh A.: The geometry of storytelling. In *Proceedings of the 2017 (CHI) Conference on Human Factors in Computing Systems* (may 2017), ACM. URL: https://doi.org/10.1145%2F3025453.3025581, doi:10.1145/3025453.3025581. 2

[Rap20] Rapp H.: Everything you need to know about recording production sound for film, Jul 2020. URL: https://nofilmschool.com/what-you-need-know-about-recording-production-sound. 2

[RBB*19] Rizvic S., Boskovic D., Bruno F., Petriaggi B. D., Sljivo S., Cozza M.: Actors in VR storytelling. In *2019 11th International Conference on Virtual Worlds and Games for Serious Applications (VS-Games)* (sep 2019), IEEE. URL: https://doi.org/10.1109%2Fvs-games.2019.8864520, doi:10.1109/vs-games.2019.8864520. 4

[RBO*21] Rizvić S., Bošković D., Okanović V., Kihić I. I., Prazina I., Mijatović B.: Time travel to the past of bosnia and herzegovina through virtual and augmented reality. *Applied Sciences 11*, 8 (2021). URL: https://www.mdpi.com/2076-3417/11/8/3711, doi:10.3390/app11083711. 2

[RDA*17] Rizvic S., Djapo N., Alispahic F., Hadzihalilovic B., Cengic F. F., Imamovic A., Okanovic V., Boskovic D.: Guidelines for interactive digital storytelling presentations of cultural heritage. In *2017 9th International Conference on Virtual Worlds and Games for Serious Applications (VS-Games)* (sep 2017), IEEE. URL: https://doi.org/10.1109%2Fvs-games.2017.8056610, doi:10.1109/vs-games.2017.8056610. 1

[SRC*20] Skola F., Rizvic S., Cozza M., Barbieri L., Bruno F., Skarlatos D., Liarokapis F.: Virtual reality with 360-video storytelling in cultural heritage: Study of presence, engagement, and immersion. *Sensors 20*, 20 (2020). URL: https://www.mdpi.com/1424-8220/20/20/5851, doi:10.3390/s20205851. 4

[SRH*20] Selmanović E., Rizvic S., Harvey C., Boskovic D., Hulusic V., Chahin M., Sljivo S.: Improving accessibility to intangible cultural heritage preservation using virtual reality. *J. Comput. Cult. Herit. 13*, 2 (May 2020). URL: https://doi.org/10.1145/3377143, doi:10.1145/3377143. 4

[url19] Jul 2019. URL: https://youtu.be/w3vKIQ2NFW0. 4

[url21a] Jul 2021. URL: http://en.wikipedia.org/wiki/Stereophonic_sound. 2

[url21b] Jul 2021. URL: http://en.wikipedia.org/wiki/Ambisonics. 2

[url21c] Jul 2021. URL: http://en.wikipedia.org/wiki/Ambisonic_data_exchange_formats. 2

[Woh19] Wohl M.: *The 360 video handbook: A step-by-step guide to creating video for virtual reality (VR)*. Vrrrynice. com, 2019. 2