Immersive Audio Post-production for 360º Video: Workflow Case Studies

Justin Paterson[1], and Oliver Kadel[2]

[1] *University of West London, London, UK*

[2] *1.618 Digital, London, UK*

Correspondence should be addressed to Justin Paterson (justin.paterson@uwl.ac.uk)

ABSTRACT

Perhaps the most pervasive immersive format at present is 360º video, which can be panned whilst being viewed. Typically, such footage is captured with a specialist camera. Approaches and workflow for the creation of 3-D audio for this medium are seldom documented, and methods are often heuristic. This paper offers insight into such approaches, and whilst centered on post-production, also discusses some aspects of audio capture. This is done via a number of case studies that draw from the commercial work of immersive-audio company, 1.618 Digital. Although these case studies are unified by certain common approaches, they also include unusual aspects such as binaural recording of insects, sonic capture of moving vehicles and the use of drones.

1 Introduction

The term Virtual Reality (VR) is commonly used, but as is increasingly understood, there are various 'immersive' implementations that offer different user experiences and require different types of equipment. Such categories include Augmented Reality, Mixed Reality and Extended Reality although for convenience in this text, they will collectively be referred to as VR for convenience. At the time of writing, such terms are not formally defined, although the interested reader will find some useful explanations from Liu [1]. These technologies are linked by featuring computer-generated content.

However, perhaps the most prolific immersive experience at present is that of 360º video. This is created by synchronizing and joining streams of videos from multiple cameras or a single multi-head camera into a single 'surround' video – a process termed stitching [2]. Although 360º video can feature embedded computer-generated content, it tends not to. This artefact is more akin to conventional video capture except that it allows the viewer to see a 360° field of view from a fixed perspective (although this could be captured from a moving camera), and it also includes elevation panning on the vertical axis, giving two axes of movement. It does not allow the six Degrees of Freedom (6DoF) commonly associated with 'true' VR-type technologies – 6DoF also changes the viewer's perspective in response to head tilting and might also allow navigation within a virtual space to offer a change of viewpoint.

One of the principal reasons for 360º video to proliferate is its adoption by YouTube, although confusingly, it is identified there as 'Virtual Reality'. Such videos have been viewed hundreds of millions of times. Facebook also employs 360º video, and there it is termed 'Facebook 360'. Typically, all such videos might be viewed either on a computer using the cursor to allow dynamic panning around the scene within the video-viewing window, or using a Head-Mounted Display (HMD) that pans the scene in response to the viewer's head movement to produce an immersive effect. Such displays might be bespoke VR devices (e.g. HTC Vive, PSVR, Oculus Rift or Lenovo Mirage Solo), or passive units (e.g. Samsung Gear VR or Daydream View) into which the user can insert a smart phone that streams a stereoscopic version of the video. Most listeners are presumed to be listening on headphones, and so binaural decoding of ambisonic audio is typical.

Audio for 360° video presents unique challenges and at present, neither the workflow nor the tools are standardized, and there are many different approaches to implement differing sonic capture, evocation or responses that serve such a dynamic visual panorama. Tools for creating audio for immersive media are evolving at a rapid pace, and whilst there are numerous emergent examples that offer similar functionality to extant products – often marking market-entrance for companies – occasionally, tools are introduced that facilitate increments in the state-of-the-art of creativity. Deployment of such tools influences workflow and

the end artefact via Actor Network Theory, as presented in music production by Zagorski-Thomas [3], leading to a step forward in creative responses according von Hippel's 'user innovation' model [4]. Broadly speaking, as with conventional video, there are two approaches for creating suitable audio: location recording and post-production and these might both be used in combination.

Location recording often uses ambisonic capture to represent a 3-D sound field that is readily panned in response to head tracking (this only requires a computationally cheap multiplication by a scalar) – binaural capture does not readily afford this facility. Conveniently, a single microphone can be employed for such recordings and alongside a range of First-Order Ambisonics (FOA) devices, the Zylia Pro offers 3rd-order capture, and the Eigenmike®, 4th order. Bates et al. performed a useful comparison of the performance characteristics of various ambisonic microphones [5], [6]. Additional spot microphones might also be used, but the all-encompassing visual capture of 360° video creates problems if these (or the crew) come into shot.

Post-production also broadly aligns with conventional approaches. However, one difference is that monaural sounds might be panned in three dimensions via binaural synthesis. Such placement might be static – head-locked, or dynamic where the sound is either locked to a moving on-screen object (emitter), or simply be non-diegetic. Stereo sources can also be panned, although localization might be less precise. Dry sources can be the most sympathetic to such placement, and tools that synthesize early reflections (e.g. dearVR Pro) can greatly assist with localization, perception of elevation and externalization during headphone playback. Tools are developing that tie sound localization to moving emitters. These are simplifying what had been somewhat arduous automation; for instance, Audioease 360pan suite gives easy control over ambience, and dearVR Spatial Connect offers a 3-D-audio production environment tied to a DAW from within VR.

This paper attempts to provide insight into relevant workflows by presenting a number of case studies created by UK company '1.618 Digital', a "creative sound-design studio offering audio production and post-production services, immersive & spatial-audio solutions for 360 video content, interactive VR/AR media and intelligent audio branding" [7].

2 Malaria: Life on the front line

1.618 Digital worked with VR City and Comic Relief to produce audio for a 360º video – 'Malaria: Life on the front line' – depicting the effects of malaria in Uganda, to raise awareness of the disease and how it can be treated [8]. The director expressed a vision of having mosquitoes that appeared as realistic as possible in order to evoke empathy among the public and increase the donation rate to charity. The film was shot in Uganda in February 2018. Ironically, since that was the dry season, it was not possible to find mosquitos in sufficient quantity for audio capture; however, upon return to the UK, the London School of Hygiene & Tropical Medicine offered to provide the team with recording access to their insect stock.

Binaural recording was employed in order to make the effect as visceral as possible. A Neumann KU-100 dummy head placed in a mosquito enclosure lined with a soft fabric. Several recordings were made with different numbers of mosquitoes (10 to several hundred) – this affected sound intensity and timbre. Such variation was required for different scenes of the film, from a big swarm flythrough to a silent moment with a single mosquito landing on the listener's head. The latter was harder to capture because mosquitoes tend to quickly settle without being stimulated by edible substances; however, according to the lab, mosquitoes react very strongly to human sweat, and are instantly attracted to the source. Thus, the problem was overcome by using sugar solution and a sock from one of the crew members to choreograph the flight trajectory of the mosquitos across the head, and encourage them to land on the dummy head's ear. One of the insects even flew inside the ear.

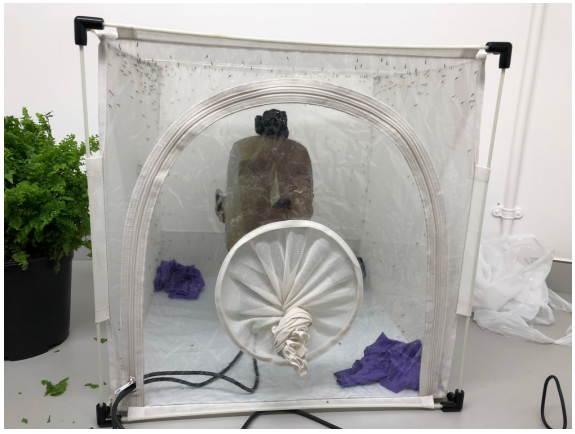

Figure 1. The dummy head in the mosquito enclosure. Note the sock placed on top of the head, with sugary cloths in the corners.

These recordings were head-locked so the end-listener not could not inadvertently move their head and miss the feeling of the mosquito approaching the ear. The final mix also included ambisonic atmospheres, 3-D dynamically panned sounds, and a musical score. Voiced by David Tennant, the narration was recorded later in the studio and was also head-locked. The project was nominated for a VR Social-Impact Award at VR Awards 2018.

3 Share the road: The WheelSwap VR Experience

Ford Motor Company and Happy Finish created two complementary films [9], [10] as part of Ford's 'Share the road: The WheelSwap VR Experience' campaign to create an empathy-driven interactive experience that promoted positive relationships between cyclists and drivers, allowing the viewer to see through the eyes of both a cyclist and a driver. Filming took place in Barcelona, Spain. 1.618 Digital was tasked to design and mix both realistic 3-D audio and hyper-real effects during the post-production stage in order to enhance the impact of the films.

All scenes were captured by the filming team without any production audio, the exception being driver reactions, which were multi-take wild tracks. Those recordings included gasps and sighs to demonstrate typical responses from both cyclists and drivers.

In a subsequent session, lavalier microphones with radio transmitters were attached near key sonic areas of the vehicles, for instance the bicycle transmission and the car exhaust. Attempting to (head-lock) pan these beneath the listener helped to enhance the sound field's authenticity and contriving them to appear closer than reality added intensity to the sense of travelling. High-frequency content was filtered and direct-to-synthetic-reverberant ratio was adjusted according to distance to increase authenticity [12].

In post-production, a commonly employed workaround is the repurposing of assets from different projects in order to make the file-sourcing process more efficient, so A-format FOA exterior atmospheres from locations with similar sonic characteristics were selected. FOA is generally deemed sufficient for this purpose. The FOA was converted to B-format using a Sennheiser Ambeo encoder (designed to work with the Ambeo VR microphone), and its low-cut filter, virtual microphone positioning, soundfield rotation and ambisonic filter correction were used at this stage to fine tune the creative effect.

As is common in 360º video, these were subsequently up-mixed to 2nd-order ambisonics (the limit of FB 360 Spatial Workstation at that time) to be mixed with a subsequent layer of spot effects at superior horizontal resolution, as discussed in [11]. These spots were emitter-tied Foley and SFX, dynamically panned in 3D according to the visual representation. Effects ranged from footsteps to passing cars and the flapping of pigeon wings. All were object-tracked with automation of azimuth and distance in the FB 360 3D panner within Pro Tools Ultimate. Nowadays (later in 2018) most productions would utilize 3rd-order ambisonics over 16 channels [$(N+1)^2$] to further enhance the spatial resolution of a program.

The driver/cyclist utterances were requested to be unnaturally prominent to exaggerate the visual events in the fashion of "sound shots" [13, p. 43]. Obvious music production was generally avoided, however tonal textures helped the narrative by underlining emotions such as anxiety around near-miss accidents.

This project won the (VR Category) Drum DADI Award 2018.

4 Elbphilharmonie 360° | A cultural landmark where all music meets

Google Arts and Culture partnered with the Elbphilharmonie concert hall in Hamburg, Germany to produce a promotional (360º) video [14] in 2016. The brass band 'Mute' performed around the venue alongside a string ensemble. The main creative aim of this project was to create a FOA mix that combined live diegetic recordings (captured alongside the video) with multiple studio-recorded stems, emitter-sync'ed to the players on film to facilitate a more pronounced spatial effect during end-listener head rotation.

The studio recording session was approached in a fairly conventional fashion with single monoaural

spot-microphones, and room/distant pairs – although the latter were only used for the (standalone) stereo mix. For the video, the ambience was recreated in 3-D during post-production to aid localization. During filming, each instrument was recorded with individual radio microphones installed in acoustically non-ideal locations, since they had to be hidden from view of the 360º camera. The musicians played to a metronome via an earpiece in order to stay in time with the original studio track.

In post-production, the stems of the individual instruments had to be mixed much 'closer' to the viewer to gain the desired effect. They were treated with a convolution reverberation captured in the concert hall to attempt to impart a sense of locality, although the constantly changing filming location rendered this rather arbitrary.

A variation of Thiele's room-related balancing technique [15] was applied to ensure that spatial quality remained accurate between location and studio microphones – controlling the direct sound and early reflections. This was done in the FB 360 system of that time, which modelled room reflections, but as Bates and Boland observe [16], its modelling was somewhat inaccurate (only the first few early-reflections are time-compensated, and are produced solely in the W-Channel of the B-format signal, thus compromising directivity) and so a heuristic approach was taken by applying further time-shift fine-tuning by listening to find the most aesthetically pleasing multi-mic composite. Having said that, with the irregular spot-recording approach and the constraints of 2016 tools (including the FOA limitation) the end result might be considered as less than ideal, although the principle of its approach remains valid.

5 Motion platform VR-experience installation: Atakule Tower

This is a project under development at the time of writing (2018). In Ankara, Turkey, Atakule Tower is a landmark building, for which a visitor experience is being installed. Akin to a flight simulator, this is a 'human transporter pod' on a motion platform. The pod emulates aerial drone-journeys from around the tower to some of Ankara's most famous landmarks, providing a mixed-reality experience that combines both high-resolution video footage with high-quality CG elements overlaid as VFX, paired with both linear and interactive spatial audio, with additional (real-time) voice communication between pairs of users, and two modes of haptic feedback.

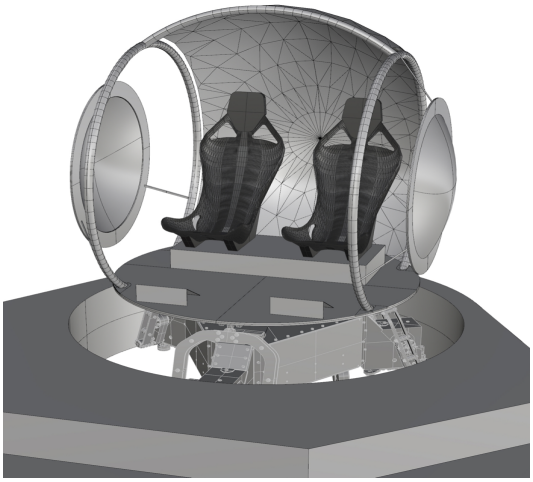

Figure 2. A schematic of the pod.

Telemetry data was captured from the drone flights to recreate the movement of the on-board Red Digital Cinema camera (which had a 260-degree lens). These movements were both refined and augmented in post-production. The data was then implemented in the Unity game engine, and this provided a control signal to make the pod tilt accordingly.

Figure 3. Preparing the drone.

Capturing audio whilst operating the drone was impractical due to its excessive noise levels. Instead, the four key locations that the drone visited were later recorded on the ground in FOA. Many wild tracks of interesting city sounds were also captured as a palette for future spotting. Again, these FOA atmospheres were up-mixed, this time to 3rd-order and then emitter-tied spot effects from the wild tracks were added. Thus, four scenes were created to represent their respective locations, and each carried a recognisable and dynamic sonic signature. Further, audio head-tracking was implemented so that visitors could 'look around' whilst flying.

To make the sound of the flying pod more 'cinematic', an interactive score of synthesizer sounds had key parameters modulated (grouped in parallel sets) in response to each of the flight trajectory movements: turning, tilting, slowing down and acceleration. Each aerial manoeuvre thus carried its own signature sound, and these were also panned in 3-D, aligned with the changes of direction. In considered disregard for the physics of first-person perception, a Doppler effect was added at times to realize what Chion describes as Schaefferian "outside of space" [17, p. 186]. Although a cinematic cliché, the effect is still emotive when triggered from changes of speed or direction.

As well as the motion of the pod, a haptic feedback device – Subpac – added additional vibration at the back seat. It was observed that toggling vibration on/off added more impact than uninterrupted operation at different intensity levels (the haptic experience being continuous due to the movement of the pod), so a separate audio track was created just to trigger the Subpac at certain points.

Figure 4. The Subpac S2 Headphone Subwoofer. (image from [18])

The greater audio mix was produced in Pro Tools in a linear fashion and rendered for implementation in a 3rd-order 16-channel system. However, in order to offer a quasi-6DoF interactive-experience, a number of aircraft control-panel sounds were implemented as discrete objects in Unity. Consequently, visitors could get close to the control panel and perceive sound more realistically, dissociated from 'external' sounds. This was achieved with the Unity native-3-D mode which utilizes a simple low-pass filtering and loudness attenuation on an emitter depending on a distance, again based on a simplified version of Zahorik et al. [12]. The Blue Ripple 'O3A Zoom' plugin [19] allowed each location's external environment to be perceived as smaller and in the distance, and then gradually spread and envelop the listener as the pod approached. It is worth noting that although the above arrangement does not offer 'true' self-navigating 6DoF over acoustically significant distances in a virtual space, in the cockpit of the pod, the illusion is quite effective. Current research in 6DoF audio is still rather emergent, but the interested reader is referred to [20]–[22].

The pod was designed to accommodate two 'passengers', which automatically meant employing two HMDs powered by two PCs with high-specification GPUs. Both computers were synchronized, running the experience through live rendering in Unity. Despite both people technically seeing the same image and experiencing the same movements of the motion platform simultaneously, each had freedom to move their heads independently, and so each needed unique head-tracked control of the audio feeds. Further, the concept was to facilitate real-time communication between the passengers, layered on top of the native installation audio mix. This was first tried via Unity; however, it was quickly evident that the high latency rendered this method inadequate. Instead, an analogue solution was implemented. Audio from two aviation-style headsets with inbuilt microphones were combined in an analogue mixer and then fed back into headphones via two audio interfaces, latency free. This simple addition enabled the passengers to communicate and share their experience in the same metaverse. This social aspect greatly ameliorated the feeling of isolation so prevalent in many VR experiences [23].

6 Conclusions

Contemporary and emergent multimodal formats present exciting opportunities for creative interpretation, yet bring with them new sets of challenges and problems. Whilst it is not possible to

be overly prescriptive in presenting solutions, the collation of some approaches in projects representative of current commercial activity might serve as a useful primer. 360º video differs from CGI-based immersive formats in that the viewer expects something more akin to reality, and the corresponding audio must respond to this. As is so often the case in two-dimensional viewing, a literal sonic representation carries insufficient dramatic impact, and directors often seek hyper reality, and this might require innovative solutions employing hybrid technologies. Despite 360° video's fixed-point perspective, its audio is further complicated if the camera is moving and there is a clear trajectory in current tool development to accommodate this. Perhaps it is actually fortunate that at present, working in such media presents so many challenges, since it precipitates evolving creative responses – free from the trappings of homogeneity that lurk in stereo DAW-based production.

References

[1] J. Liu, "The difference between AR, VR, MR, XR and how to tell them apart," *Hacker Noon*, 02-Apr-2018. [Online]. Available: https://hackernoon.com/the-difference-between-ar-vr-mr-xr-and-how-to-tell-them-apart-45d76e7fd50. [Accessed: 08-Nov-2018].

[2] F. Nielsen, "Surround video: a multihead camera approach," *Vis. Comput.*, vol. 21, no. 1, pp. 92–103, Feb. 2005.

[3] S. Zagorski-Thomas, *The Musicology of Record Production*. Cambridge, UK: Cambridge University Press, 2014.

[4] E. von Hippel, "Lead Users: A Source of Novel Product Concepts," *Manag. Sci.*, vol. 32, no. 7, pp. 791–805, Jul. 1986.

[5] E. Bates, M. Gorzel, L. Ferguson, H. O'Dwyer, and F. M. Boland, "Comparing Ambisonic Microphones – Part 1," in *Audio Engineering Society Conference: 2016 AES International Conference on Sound Field Control*, 2016.

[6] E. Bates, S. Dooney, M. Gorzel, H. O'Dwyer, L. Ferguson, and F. M. Boland, "Comparing Ambisonic Microphones—Part 2," in *Audio Engineering Society Convention 142*, 2017.

[7] "1618 Digital," Nov-2018. [Online]. Available: https://1618digital.com/. [Accessed: 08-Nov-2018].

[8] "Malaria: Life on the Front Line," *Sport Relief*, 2018. [Online]. Available: https://www.sportrelief.com/malaria. [Accessed: 08-Nov-2018].

[9] Ford UK, "Ford WheelSwap: Life from a Driver's Perspective | Ford UK," 2018. [Online]. Available: https://www.youtube.com/watch?v=hl3Zz2xAFj4&feature=youtu.be. [Accessed: 08-Nov-2018].

[10] Ford UK, "Ford WheelSwap: Life from a Cyclist's Perspective | Ford UK," 2018. [Online]. Available: https://www.youtube.com/watch?v=3k2JgHaoQwI&feature=youtu.be. [Accessed: 08-Nov-2018].

[11] S. Favrot, M. Marschall, J. Käsbach, J. Buchholz, and T. Weller, "Mixed-Order Ambisonics Recording and Playback for Improving Horizontal Directionality," in *Audio Engineering Society Convention 131*, 2011.

[12] P. Zahorik, D. Brungart, and A. Bronkhorst, "Auditory distance perception in humans: A summary of past and present research," *Acta Acust. United Acust.*, vol. 91, pp. 409–420, 2005.

[13] M. Chion, *Audio-Vision : Sound on Screen*. New York: Columbia University Press, 1994.

[14] Google Arts and Culture, "Elbphilharmonie 360° | A cultural landmark where all music meets," 2016. [Online]. Available: https://www.youtube.com/watch?v=evcpxACKldQ. [Accessed: 10-Nov-2018].

[15] G. Theile, "Multichannel Natural Recording Based on Psychoacoustic Principles," in *Audio Engineering Society Convention 108*, Paris, France, 2000.

[16] E. Bates and F. Boland, "Spatial Music, Virtual Reality, and 360 Media," in *Audio Engineering Society Conference: 2016 AES International Conference on Audio for Virtual and Augmented Reality*, 2016.

[17] M. Chion, *Sound: An Acoulogical Treatise*. Duke University Press, 2015.

[18] T. Hertsens, "Shakin' it with the Subpac S2 Headphone Subwoofer," *InnerFidelity*, 27-Oct-2015. [Online]. Available: https://www.innerfidelity.com/content/shakin-it-subpac-s2-headphone-subwoofer. [Accessed: 12-Nov-2018].

[19] "O3A Manipulators," *Blue Ripple Sound*. [Online]. Available: http://www.blueripplesound.com/products/o3a-manipulators. [Accessed: 12-Nov-2018].

[20] D. Rivas Méndez, C. Armstrong, J. Stubbs, M. Stiles, and G. Kearney, "Practical Recording Techniques for Music Production with Six-Degrees of Freedom Virtual Reality," in *Audio Engineering Society Convention 145*, 2018.

[21] E. Stein, M. Walsh, G. Shi, and D. Corsello, "Audio rendering using 6-dof tracking," US20170366914A1, 21-Dec-2017.

[22] A. Plinge, S. J. Schlecht, O. Thiergart, T. Robotham, O. Rummukainen, and E. A. P. Habets, "Six-Degrees-of-Freedom Binaural Audio Reproduction of First-Order Ambisonics with Distance Information," in *Audio Engineering Society Conference: 2018 AES International Conference on Audio for Virtual and Augmented Reality*, 2018.

[23] M. Gutierrez, F. Vexo, and D. Thalmann, *Stepping into Virtual Reality*. Springer Science & Business Media, 2008.