Audio Reproduction in Virtual Reality Cinemas – Position Paper

Luke Reed*[1,2] Philip Phelps‡[1]

[1] The University of the West of England, Dept. Computer Science and Creative Technologies, Creative Technologies Lab°

[2] South West Creative Technologies Network – Immersion Fellowship#

ABSTRACT

Virtual Reality (VR) and 360 film have caught the attention of audiences and content creators and emerged as a new media, however, the market penetration of VR and head mounted hardware has remained slow despite the availability of more affordable mobile options. This has resulted in some audiences turning to VR cinemas, festivals and out-of-home exhibitions. Creating affordable, scalable VR cinemas presents a number of challenges and many of the decisions taken in both developing and facilitating these curated exhibitions directly impact audience's reception of spatial audio soundtracks. This workshop position paper looks to discuss the potential issues and future solutions in the use of current synchronous exhibition applications, the competing formats, standardised Head Related Transfer Functions, headphone build/colourisation, and the on-boarding process.

Keywords: Cinematic VR. 360 film. Spatial audio. 3D Audio. Binaural. Ambisonics. Object Based Audio. VR cinemas. Exhibition. Playback software. HRTF. Headphones. On-boarding.

1 INTRODUCTION

Virtual Reality (VR) devices and experiences have become affordable and available to the public with much anticipation that the current wave of VR will become a mainstream and ubiquitous format. However, adoption in the home has been slower than forecast with the vast majority of consumers in the United Kingdom having little to no intent to purchase the necessary hardware [9] despite demonstrating a desire for the content [1]. This raises significant questions as to the lifespan of VR in the mid-to-long term and prolongs ambivalence as whether or not it will ever cross the "chasm" from the early market into the mainstream [22].

As a result, it could be argued that for the medium to develop beyond fulfilling the appetite of a niche audience consisting of largely young males, who incidentally are already invested for the gaming applications. Cinematic VR (CineVR - also referred to as 360 and 180 film) experiences and technologies might need to occupy a public space outside of the home akin to the theatre, arcade or cinema space in order for the content to reach audiences that want it but are excluded by the technological barrier. With this in mind the considerations of audio in VR cinemas needs to be addressed.

This position paper provides an introduction to the concept of VR cinemas, followed by a meta-analysis of the current state of the art, finally presenting a challenge for the future of the medium and its exhibition platforms.

2 VR CINEMAS

VR cinemas are increasingly popular at film festivals that offer "new technologies", "VR" or "immersive" categories (La Biennale di Venezia [4], Sundance New Frontier [36], Montreal FNC [21]) also emerging as a term for ticketed events hosted by cultural institutions (such as museums, theatres and art galleries) as highly scalable VR exhibitions where participants are equipped with individual head mounted displays (HMDs). VR Cinemas use Mobile VR architecture (Samsung Gear [31], Oculus Go [24], Lenovo Mirage Solo [20]), offering a cost-effective method for event organisers to facilitate playback of VR content on scale. The hardware constraints of Mobile VR platforms generally limits the content to three-degrees-of-freedom (3DOF) such as CineVR requiring participants to remain seated rather than room-scale and six-degrees-of-freedom (6DOF) experiences that provide more interactivity and full body movement but require more space, more sophisticated and expensive hardware, thus inhibiting the scalability of exhibition.

In the case of 3DOF and seated exhibition, the playback of content can be controlled centrally by a server that either streams wirelessly or cues side-loaded content across multiple HMDs for synchronous playback. This offers significant benefits for participant experience not least a simplified on-boarding (introduction) strategy which removes much of the complexity of guiding novice participants through complex user interfaces, gestures and bespoke controllers.

An onboarding process often involves human hosts who introduce participants to the hardware and facilitate correct and comfortable set-up. While this ensures an opportunity for quality control for participants in small groups, it can become a slow and expensive limitation of large scale exhibition, causing participants to wait long periods of time in virtual lobbies and potentially exhausting the 20-40 minute window of opportunity before symptoms of cyber-sickness cause participant to drop-out [28] and require a break from the head-mounted screen experience.

There remains an opportunity for exhibition developers to streamline much of this process through an onboarding routine within the headset and diverting time and resources away from real world while facilitating users to construct their understanding within the VR world. This would provide an opportunity to teach the participant about simple control and gestures if available, and make alterations and personalised optimisations to the listening experience as will be discussed later.

Synchronous playback across multiple HMDs requires the use of a third-party application for control (e.g. Showtime VR [33], VR Sync [37]) or a bespoke solution being developed at which point support for spatial audio soundtracks must be understood and facilitated by the developer. A number of free-to-use software frameworks are available to developers creating bespoke playback software and options are dominated by Google Resonance [29] and Facebook 360 Audio [10]. Development of bespoke playback software remains a complex problem when using tools such as Unity3D and the Android SDK. Such tools and platforms provide a series of challenges (such as limited channel numbers and constrictive security measures [5]), all of which limit the

* Luke.Reed@uwe.ac.uk
‡ Philip3.Phelps@uwe.ac.uk
° http://uwecreativetechnologies.com/
# https://swctn.org.uk/immersion/

functionality of the more complex behaviours required to improve upon the current offering of exhibition software.

3 Competing formats

There are a number of commonly used and competing formats for delivering spatial audio in use. These formats can be loosely arranged into three categories: channel-based, scene-based and object-based [30].

Channel-based audio (CBA) refers to discrete audio channels and often correspond to specific loudspeakers within the listening environment (e.g, 5.1, 7.1). Within cinematic VR the audience receive the audio via headphones and thus limiting channel based approaches in VR generally to stereo. While conventional channel based formats could be played through virtual speakers few players support this function natively and in combination with the low spatial resolution, lack of height and redundant channels (centre & LFE) mean that other methods of spatial representation are favoured.

Two channel soundtracks afford a fixed binaural recording that provides some impression of spatial depth, however, in the case of 360° media as there is no head-tracking taking place a particular listener orientation is assumed which has the potential to create confusion and disorientation in audience members where activity and movement in the sound field and visual component is incongruent. For example, a car passes from right to left at 0°. If the audience is viewing at this approximate angle the binaural effect works as intended and the audio and image are perceived as cohesive. If the audience were to look outside of the field of view of the car pass, for example 135°, the audio cue would likely cause audio-visual disconnect and momentary confusion as they search for the causality of the audio event (Fig. 1).

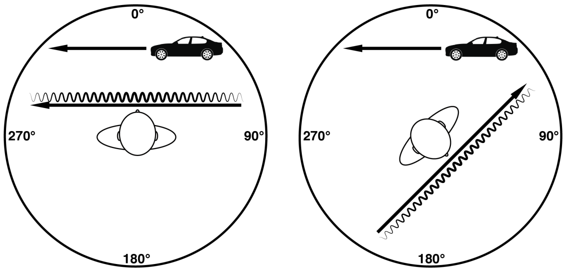


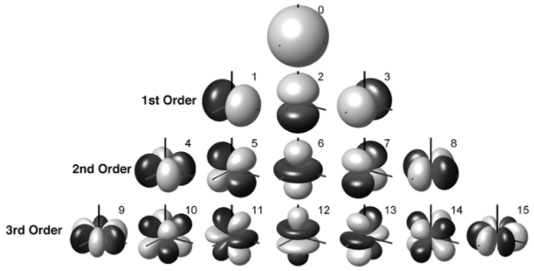
Figure 1. Fixed binaural audio-visual disorientation in 360° media.

Despite this limitation of fixed binaural soundtracks, the simple and commonplace two channel format is both easier to create and computationally cheap to playback as all binaural effects are baked-in to the mix. Furthermore there remains numerous examples in arts media in which simple stereo fulfils the needs of the experience, such as in dance, abstract and music videos in which externalisation is not necessarily critical (e.g. The Guardian's Celestial Motion [8]). As 180° media becomes more prevalent (such as Google's VR180 [38]) we will likely see binaural audio become increasingly used as the limited field of view would restrict front-back inversion found in 360° media.

Scene-based audio (SBA) formats are a method of encoding and representing three-dimensional (3D) sound fields for a fixed point in space. They provide an ideal method for 3DOF media as the listener cannot move or translate along axis, but rotation around axis such as pitch and yaw can be derived from accelerometer and gyro sensors within the HMD, enabling a 3D audio rendering to be orientationally transformed to align with the listener's orientation within a spherical virtual video projection.

Ambisonic audio represents the 3D sound field through spherical harmonics at increasing orders of resolution (Fig. 2). This spatial resolution can be scaled down through the simple subtraction of channels, meaning a soundtrack may be produced in third-order ambisonics (3OA) (16 channels) but need to be exhibited in second order ambisonics (2OA) (nine channels) or first-order ambisonics (1OA) (four channels). This kind of flexibility is ideal considering the fluidity of formats and playback technologies during VR cinema's nascent years.

Figure 2. Spherical harmonics up to the third order with Ambisonic Channel Number [39].



A number of variations on the ambisonic format exist which differ in terms of their channel ordering and normalisation method. Currently the AmbiX (Ambisonics eXchangable format) has seen widespread adoption as it utilises the infinitely scalable Ambisonic Channel Numbering convention (0123), in place of the traditional Furse Malham (FuMa) lettering sequence (WXYZ) and the SN3D normalisation method ensuring that no signal exceeds the zero-order, omnidirectional centre signal [23].

Currently a number of audio production tools exist with most functioning up to 3OA or other higher order ambisonic (HOA) layouts (Facebook Spatial Workstation [10], Blue Ripple O3A [27], SPARTA [35], IEM [16]) while most exhibition applications remain limited to 1OA and 2OA (Samsung VR, YouTube, Oculus Gallery). Two exceptions are found which support 3OA; Vive Cinema, which currently is only available on HTC Vive's tethered HMDs powered by high specification PCs and therefore not as cost-effective and scalable as the mobile platforms currently used; and Jump Inspector, an application from Google's early Jump toolkit which is no longer supported [17].

The main criticism of ambisonics lies in its inability to reproduce discrete sources. Mach1Spatial is a competing proprietary format consisting of 8 directional virtual speakers arranged on vertices of a cube [11], it employs a Virtualised Vector Based Panning (VVBP) strategy, similar to Vector Based Amplitude Panning (VBAP) or Spatial PCM Sampling (SPS) to place sources within the sound field [19]. The lack of the omnidirectional centre results in each virtual speaker able to be treated as part of a VVBP cluster or a discrete channel providing the potential for greater diversity and differentiation of sound sources and clearer localisation when compared to 2OA [12], however, this comes at the expense of more complex rotational transformations necessary for 360 media. Currently SamsungVR player and few others are capable of reproducing Mach1Spatial despite a number high profile hollywood productions making use of the format (e.g. Alien Covenant: In Utero, Mr Robot VR).

Facebook's acquisition of Two Big Ears now branded as FB360 Spatial Workstation has subsequently has seen their Hybrid Higher Order Ambisonic (HHOA) (also known as .TBE format) become widely adopted in the cinematic VR community. This is in part due to the free-to-use tools and strong online support community. The .TBE format consists of 8 channels of 2OA with the second order vertical harmonic (channel 6 or R omitted). In addition the format affords head-locked stereo channels enabling two sound stages to be utilised by content creators to separate sources like music and voice-over from the ambisonic decoders and binauralisation, as such this secondary audio sound stage has become an integral part of the CineVR form enabling a non-diegetic or in-the-head narrative space and must be adopted in any future conventions.

Object-based audio (OBA) involves the use of channels of discrete audio tagged with metadata such that the positioning of sources occurs in run-time. OBA for spatial audio is far more focused towards addressing the variations in loudspeaker arrangement in conventional cinema and home environments (e.g. Dolby Atmos), as in VR exhibition loudspeakers are virtual and can in theory be arbitrarily re-arranged at will. Depending on the number of objects the channel count and therefore data bandwidth and computational load required is the major limiting factor of OBA in VR Cinema as each object would need its own emitter or virtual speaker to be instantiated, in comparison to SBA where this fixed and consistent throughout run-time.

Of particular note is the application of OBA to allow audience members to define certain parameters of the sound track, such as language and mix balance. This function could help determine a workflow for music and effects (M&E) stems in conventional film and TV that is crucial for dubbing international VR media. For example, dialogue material could be swappable spatialised objects whereas all other elements of the soundtrack could be bounced into a fixed stem or bed (head-locked stereo or HOA) sidetracking the need to package multiple higher order mixes and inflate file sizes.

The new MPEG-H standard [15] enables the combination of HOA, CBA and OBA providing a potential pipeline for a interchangeable spatial audio formats and user selected modification. While VR Cinema continues to favour mobile VR systems, the high channel numbers provided in MPEG-H are arguably excessive and go beyond the hardware limitations of the systems in use. An optimised profile and best practice guidelines should be developed quickly once the format becomes more widely available for content creation.

4 Head Related Transfer Functions

Perceptual cues that enable the localisation of sources in the sound field can be synthesised through signal processing. The Interaural Time Difference (ITD), Interaural Level Difference (ILD) and spectral cues together are known as a Head Related Transfer Function (HRTF). The use of headphone based reproduction of spatial audio, combined with spatial audio formats, and orientationally head-tracked HRTF signal processing, can provide the listener with a greater sense of dynamic envelopment and presence within 360 media. Currently most VR playback software use a single standard HRTF based on average physiology, however there are many variables derived from the individual's anatomy (size and shape of the torso, head and pinnae) that affect that individual's unique externalisation and localisation cues. This would require bespoke measurements to improve spatial audio reproduction over headphones but this is complex, time consuming and expensive to deploy in the consumer environment.

A vast array of HRTF datasets exist across the LISTEN, CIPIC, FIU and MIT/KEMAR databases. The MARL-NYU project [2] has amalgamated these into a standardised repository that addresses the differences in capture formats (e.g. sample rate, sample length, phase, amplitude, angle increment), standardising a wealth of available data into a format that is comparable side by side - albeit at the expense of the fidelity advantages of some datasets - for example those that are captured at high sample rates (FIU, 96kHz) or narrower angle increments (CIPIC, ~5°).

A standard to facilitate interchange of HRTFs exists in the Spatially Oriented Format for Acoustics (SOFA) [3, 34] which allows the specifics of each capture to be described in the metadata, potentially rendering the MARL-NYU repository efforts redundant. Unfortunately, the ability to deploy SOFA files remains broadly unimplemented in playback software. This may be due to the lack of consumer level availability to HRTF measurement or selection methods - or indeed public awareness of HRTFs as an influential factor in improving an overall listening experience.

While access to acoustic measurement of HRTFs requires significantly prohibitive time and resources, other methods of individualisation have been evaluated to produce satisfactory results, most notably the selection and use of non-individualised HRTFs through perceptual feedback and adaptation [13, 18]. A guided process for user-selection of the most suitable HRTF from a range of non-individualised HRTFs might significantly improve perception of audio externalisation in a headphone based VR experience at exhibition. If such a guided HRTF selection process was integrated into an automated onboarding process it could simultaneously teach the user of its impact whilst adapting the HRTF model from a standardised database like the MARL-NYU repository and/or a collection of SOFA files. This process would need to be designed with the user experience and entertainment in mind so as to not disengage audience members through onboarding.

5 Headphones

Headphones play a critical role in the reproduction of VR spatial audio as they act as the final filter through which a listener receives a simulated audio environment. While open backed headphones are often described as providing a more natural sense of externalisation, within VR cinemas it is common to use closed backed models as these isolate the user from the external environment and prevent bleed from multiple simultaneous users. It has also been found that noise cancelling headphones which provide "flat" acoustic responses (minimal driver / enclosure resonance across the audio frequency range) excel at natural sounding externalisation in addition to providing good elevation and front / back discrimination cues [7]. While noise cancelling headphones would inflate the cost per seat in VR cinemas, the application of correction filters for 'normal' headphones through additional equalisation processing is as yet broadly unimplimented in playback software.

Headphone Transfer Functions (HpTFs) can be applied to equalise frequency response in a similar way to HRTFs through the convolution of a Headphone Impulse Response (HpIR). Furthermore, a large database of HpTFs exist as the Princeton Headphone Open Archive (PHOnA) [26] and makes use of the SOFA file format, laying down the foundation of a working process [6]. A VR cinema operator may find that headphones commonly used in VR exhibition already appear in the PHOnA database and may not require bespoke measurements at all.

It should be noted that the overall effectiveness of an HpTF is affected by differences in measurement of listener's pinnae and placement and that it has been found that incorrect applications of HpTF affect the externalisation negatively [32]. It is suggested that similar perceptual testing methods during onboarding might help to equalise the audible anomalies in for users with specific attention focused between 100-1600Hz (critical for front-back confusion) and 4-7kHz (critical for horizontal localisation) [14].

Finally, the overall build of the headphone must be thoroughly considered in terms of practicality of use in a VR cinema context. Headphone hardware used in public exhibitions of any kind go through considerable abuse and as such must be robust and repairable, ideally modular. Some VR Cinemas make use of headphones with controllable in-line amplifiers to shift volume control away from interactions with the HMD and avoiding opportunities to interrupt playback. The most fragile weak point with any HMD using external headphones is the physical jack connection which can wear under the strain of heavy cables and movement during viewing. Common failures include audio dropouts, mono fold-downs and irregular left right balance. Fortunately a range of creative commons, inexpensive 3D-printed cable clips and managers can be used to relieve such issues [25].

6 Conclusions

This position paper concludes with a criteria for future VR exhibitors to challenge and guide developments.

1. Support for native HOA formats up to at least 3OA including stereo head-locked.
2. Support for an optimised MPEG-H format for mobile VR platforms.
3. Develop and integrate a thorough and entertaining onboarding method that:
    a. enables non-individualised HRTF selection and adaptation using perception methods.
    b. enable equalisation or HpTF to flatten the colouration of headphones used.
4. Support for SOFA files.
5. Consider, adapt and maintain the build of the headphone hardware.

In this paper the current state of the art has been discussed with respect to VR exhibition of 3DOF media with spatial audio in public spaces. While many advancements have been made, there is still much ground to cover so that the consumer level experience benefits from the wealth of understanding available. It is hoped that this position paper will help to create healthy debate and direct the future of audio in VR cinema exhibitions.

References

[1] C. Allen and D. Tucker, Immersive Content Formats for Future Audiences, Digital Catapult, 2019.

[2] A. Andreopoulou, and A. Roginska, Towards the Creation of a Standardized HRTF Repository, *Audio Engineering Society Convention 131*, October 2011.

[3] Audio Engineering Society, AES standard for file exchange - Spatial acoustic data file format, *AES 69-2015*, 2015.

[4] La Biennale di Venezia, *Venice Virtual Reality*, 2018. [Online]. https://www.labiennale.org/en/cinema/2018/lineup/venice-virtual-reality

[5] M. Binelli, D. Pinardi, T. Nili, and A. Farina, Individualized HRTF for playing VR videos with Ambisonics spatial audio on HMDs, *International Conference on Audio for Virtual and Augmented Reality,* August, 2018.

[6] B. Boren, M. Geronazzo, P. Majdak, and E. Choueiri, PHOnA: A Public Dataset of Measured Headphone Transfer Functions, *Audio Engineering Society Convention 137*, 2014.

[7] B. Boren and A. Roginska, The Effects of Headphones on Listener HRTF Preference, *Audio Engineering Society Convention 131*, 2011.

[8] Celestial Motion – a virtual dance experience, *The Guardian VR* [Online]. https://youtu.be/R2Tx0Zuw8DA

[9] Z. Emmanuel, Virtual Reality – UK – December 2018, Mintel Group Ltd, 2018.

[10] Facebook Audio 360, *github.io* [Online]. https://facebookincubator.github.io/facebook-360-spatial-workstation/

[11] A. Farina, Introducing SPS format and its first practical implementation, called Mach1 [Online]. http://www.angelofarina.it/SPS-conversion.htm

[12] A. Farina. A. Amendola, L. Chiesi, A. Capra and S. Campanini, Spatial PCM Sampling: A New Method for Sound Recording and Playback, *AES 52nd International Conference*, 2013.

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

[14] P. Gutierrez-Parera and J. Lopez, Influence of the Quality of Consumer Headphones in the Perception of Spatial Audio, *Applied Sciences*, vol. 6, no. 4, p.117, 2016.

[15] J. Herre, J. Hilpert, A. Kuntz, and J. Plogsties, MPEG-H 3D Audio—The New Standard for Coding of Immersive Spatial Audio, *IEEE Journal Of Selected Topics In Signal Processing*, vol. 9, no. 5, pp. 770-779, August 2015.

[16] IEM Plugin Suite, *Institute of Electronics, Music and Acoustics* [Online]. https://plugins.iem.at/

[17] Jump Inspector APK (final released versions), *angelofarina.it* [Online]. http://www.angelofarina.it/Public/Jump-Videos/Jump-Inspector-APKs/

[18] F. Katz and G. Parseihian, Perceptually based head-related transfer function database optimization, *The Journal of the Acoustical Society of America,* vol. 131 no. 2, pp. EL99–EL105, 2012. doi: 10.1121/1.3672641

[19] Mach1, VVBP Standards for Spatial Audio and Agnostic Format Conversions - White paper [Online]. http://www.qdepartment.com/__Mach1/research/Mach1SpatialSystem-WhitePaper_180523.pdf

[20] Mirage Solo with Daydream, *lenovo.com* [Online]. Availabe: https://www.lenovo.com/gb/en/smart-devices/virtual-reality/lenovo-mirage-solo/Mirage-Solo/p/ZZIRZRHVR01

[21] Montreal Festival du Nouveau Cinéma, *2018 Programme Virtual Reality*, 2018. [Online]. https://nouveaucinema.ca/en/films?type=Virtual+Reality

[22] G. Moore, Crossing the Chasm, Harper Collins, New York, 1991.

[23] C. Nachbar, F. Zotter, E. Deleflie, and A. Sontacchi, "AmbiX - A Suggested Ambisonics Format", *Ambisonics Symposium,* June, 2011.

[24] Oculus Go, *developer.oculus.com* [Online]. https://developer.oculus.com/go/

[25] Oculus Go cable strain relief, *Thingiverse.com*, 2019. [Online]. https://www.thingiverse.com/thing:3105566..

[26] The Princeton Headphone Open Archive (PHOnA), *Princeton.edu*. [Online]. https://www.princeton.edu/3D3A/Phona.html..

[27] Pro Audio Products, *blueripplesound.com* [Online]. http://www.blueripplesound.com/product-listings/pro-audio

[28] L. Rebenitsch and C. Owen, Review on cybersickness in applications and visual displays, *Virtual Reality*, vol. 20, no. 2, pp. 101-125, 2016. doi: 10.1007/s10055-016-0285-9

[29] Resonance Audio, *github.io* [Online] https://resonance-audio.github.io/resonance-audio/

[30] F. Rumsey, Spatial audio: Channels, objects, or ambisonics?, *Journal Audio Engineering Society*, vol. 66, no. 11, pages 987-992, November 2018. http://www.aes.org/e-lib/browse.cfm?elib=19873

[31] Samsung Gear VR Specifications, *samsung.com*, [Online]. https://www.samsung.com/global/galaxy/gear-vr/specs/

[32] D. Schonstein, L. Ferr, and B. Katz. Comparison of headphones and equalization for virtual auditory source localization. *Acoustics '08 Paris*, pp. 4617–4622, 2008.

[33] Showtime VR [Online]. https://showtimevr.eu/

[34] Spatially Oriented Format for Acoustics (SOFA) Conventions Wiki, *Sofaconventions.org*, 2018. [Online]. https://www.sofaconventions.org/mediawiki/index.php/SOFA_(Spatially_Oriented_Format_for_Acoustics)

[35] Spatial Audio Real-time Applications (SPARTA), *Aalto University* [Online]. http://research.spa.aalto.fi/projects/sparta_vsts/plugins.html

[36] Sundance Institute, *2018 Sundance Film Festival's New Frontier: Crossroads of Film, Art and Technology*, 2018. [Online]. http://www.sundance.org/blogs/news/2018-festival-new-frontier

[37] VR Sync [Online] https://vr-sync.com/

[38] VR180, *google.com* [Online]. https://vr.google.com/vr180/

[39] F. Zotter (Adapted L. Reed), Spherical Harmonics up to Ambisonic order 5 as commonly displayed, sorted by increasing Ambisonic Channel Number (ACN), aligned for symmetry, 2013. [Online]. https://en.wikipedia.org/wiki/Ambisonic_data_exchange_formats#/media/File:Spherical_Harmonics_deg5.png