Conference Paper

A Comparison of Audio Models for Virtual Reality Video

Davies, S., Cunningham, S. and Picking, R.

This is a paper presented at the IEEE International Conference on Cyberworlds (CW), 20-22 Sept 2017

A Comparison of Audio Models for Virtual Reality Video

Steven Davies
Glyndwr University
Mold Road, Wrexham, LL11 2AW
e-mail: s11005998@mail.glyndwr.ac.uk

Stuart Cunningham & Rich Picking
Glyndwr University
Mold Road, Wrexham, LL11 2AW
e-mail: {s.cunningham | r.picking}@glyndwr.ac.uk

***Abstract*— This paper investigates the relationship between audio models for Virtual Reality (VR) video with respect to the senses of immersion and realism that each model delivers. Mono, Stereo, 5.1 Surround Sound, and a Virtual Spatialised Position configuration was developed for testing in a VR music video and evaluated with a user study. Participants experienced the VR video with these differing audio models as accompaniment a total of four times. Qualitative and quantitative data were recorded to evaluate user experience. The results indicate that no statistical significance was present between the four models in relation to immersion or realism, suggesting that complex audio renderings are not always necessary for effective user experience.**

Keywords- Human senses; interaction; sensory perception; virtual reality; User Interfaces—Interaction styles

I. INTRODUCTION

Recent years have seen a surge in popularity of Virtual Reality (VR) devices and applications being brought into the mainstream consumer marketplace, from dedicated systems, such as the Oculus Rift, HTC Vive etc., through to mobile based alternatives, such as Google Cardboard and Samsung Gear VR headsets. 360° video capture or computer-generated imaging have been popular methods for creating the visual environments therein and much of the focus to date has been upon creating realistic visuals and ensuring that the user's ability to view and navigate the VR environment has been convincing and effective. However, whilst visuals have received the majority of attention thus far, we now focus on the best methods for representing audio in VR environments.

In this paper, we investigate the implications and responses of users to several modes of audio configuration. We present the results of an experiment to determine the perceived effects of various sonic presentation modes in a VR space; specifically, the representation of sound in a 360° music video. It is intended that the outcomes from this experiment will provide insight into how audio can be convincingly presented to the user, whilst factoring in the related overheads associated with each mode of audio implementation and post-production, such as the number of audio assets and processing load/time required to support their inclusion.

II. BACKGROUND

VR Experiences are rarely *natively* tactile; relying on controllers or other devices to serve as middleware to immersive experiences. Indeed, the landscape of VR interaction currently relies on a stepped or layered approach to familiar conventions such as those found in mass market, consumer AV products. Much focus is often paid to the visual aspects of such technologies and while this might seem a natural response to the primary human sense, the audio aspects of an audio-visual relationship provide significant tertiary experience. The aim can be explained simply as "*...3-D sound system(s) [use] processes that either complement or replace spatial attributes that existed originally in association with a given sound source."* [1]

However, the scale of such a consideration is captured in '3-D Sound for Virtual Reality and Multimedia':

"*The idealistic but laudable goal of a 3-D sound system designer ought to involve the notion of complete control and manipulation of someone else's spatial auditory perception*"

Much research exists around the selection of appropriate models of audio for implementation in VR environments. Jin and Best [2] suggest *"...it is not really multi-channel audio that should be the issue per se, but spatial audio because the arguable objective of multi-channel audio systems is the generation of a spatial-audio percept."*

It is important to emphasize that, in this work we are only concerned with the best representation of audio to VR users who are listening via headphones, since we see this as the best-fit companion to the VR headset. Whilst it is entirely possible for the audio presentation of a VR experience to be achieved through the use of speaker systems, this requires considerable additional infrastructure.

Representation of VR sound without headphones is an interesting and valid area of research, as in the work of Grani, Nordhal, and Serafin [3] who have examined user perceptions and experiences of sound in VR presented via a wavefield synthesis system.

Masuch, when discussing the role of sound in the related domain of audio-based gaming, identifies the challenge as "*... designing audio based games [is] to keep the right balance between functionality and aesthetics*". The research notes also that in the experiment on audio only gaming, "*All participants were impressed by the true 360 of action possible with some of the audiogames and by the head-tracking technique*." [4]

Any model, despite its intention to realistically replicate our natural sensory response to sound is, by its very nature, a facsimile. "*Natural versus Virtual Spatial Hearing*" [1] requires multiple component parts as in the Source-Medium-Receiver model and, while exploring Binaural and Stereo enabling technologies, the success of the effect is ultimately dependent on the physiology, in particular the head, positioning and shape of the ears, (HRTF etc.) and psychological perception of the end user.

While ambisonic acquisition can provide increased numbers of sources for improved localisation, the increasing complexity and variation in terms of approaches; requiring user specific Head-Related Transfer Function (HRTF) models, significantly increased processing power for higher order models and further post production modelling [5] has been argued to have mixed successes in testing [6], especially without high specification controlled listening rooms as presentation environments.

Conversely, Williams [7] described Equal Segmentation Microphone Arrays (ESMA) for spatial audio acquisition and more recently, Lee & Rumsey [8] described the binauralisation of ESMA models to create spatialized audio for Virtual Reality and have reported favourable responses from participants using four equidistant microphones in a configuration similar to Williams' Star model, but with four microphones at 90° intervals, in a configuration with capsules 50 cm apart (ESMA 50).

Additionally, the mainstream consumption platforms for VR Video [9][10] have tools and specifications for the audio component of the experience, with each providing portals stipulating technical requirements for spatial audio encoders. However, little documentation is provided to elaborate on preference or best practice for adopting these methods over simpler audio models, such as Mono or Stereo. In fact, negative effects of poor spatialisation in VR experiences are frequently reported by users. [11]

III. Evaluating VR Audio Models

To compare audio models, an experiment was devised in order to gather quantitative and qualitative data from participants in response to their experience of consuming multiple models with the accompaniment of identical visual material in each pass.As noted, various opinions exist in relation to preferences for models of audio recording and playback for VR video. It is postulated that despite the many options available, users will have a preference for a particular system or another. This preference might be in relation to a natural fondness for a particular model or an earned or inferred familiarity might dictate that experiential parameters influence the draw to a particular model.

In order to explore these opinions, and as description of audio phenomena can often display tendencies towards vagueness, a sample group with a high level of auditory acuity was chosen. A convenience sample of 24 undergraduate students studying courses in Sound and Music Technologies were selected alongside students from courses in Computer Game Development. In this sample participants were predominantly male and the majority of participants were under the age of 30.

A VR Video was created using a twin fisheye lens acquisition model. In order to limit any possible influence from custom experimental camera arrays, the commercially available Samsung Gear360 (2016) was used to capture the visual element of the test. Stitching was completed using The Foundry's CaraVR tooklit for Nuke as this stitching method produced media with fewer errors when compared to the ActionDirector software bundled with the Gear360.

A video performance by The Roseville Band was recorded, as accompaniment to a pre-recorded piece of music of their own composition. A synthesised performance was chosen in order to limit time signature drift between each recording, which could impact on experimental results. The performers were arranged at equidistant intervals around the camera, facing the centre of the space.

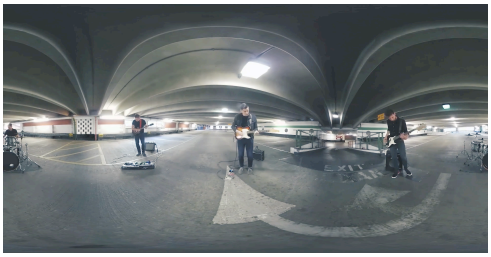

Figure 1. Equirectangular Stitched Test Media

The completed recording was then used as a consistent visual component for each audio model.

The four models tested in this experiment were created by working with multitrack audio from the studio recording by the band, mixed or processed to create distinct models for testing. The models used were Mono, Stereo, 5.1 Surround (binauralised in Unity3D), so chosen for both for its wide adoption and its similarity to the PCMA model [13] and finally a spatial model created in the Unity 3D game engine using individual audio emitters with position and distance mapping to mimic the configuration of the performers from the video. The spatial model shares similarities with a First Order ambisonic implementation, with sources positioned in a 'square' configuration, but this model is distinct and is perhaps best described as Virtual Spatial Positioning (VSP). This model positions sound sources at regular intervals (90°) in keeping with ESMA modelling, but because this model was virtually created rather than recorded with ESMA 50 or ESMA 25 arrays, an added benefit is that the arrays are not captured in vision. This design also leans strongly towards the 4.0 Binaural model created by the BBC R&D experiments with the *Last Night of the Proms 2016* [13]

It might be effective to describe the experiment as an ABX test, with four variables. However, in describing comparison in audio tracks in relation to moving image, Chion establishes the method of *forced marriage.* [14]

"Taking care to cut out the original ... show them the sequence several times, accompanied by these various musical pieces played over the images in an aleatory manner... we begin to see the image in all its potential signification and expression."

In this experiment, musical melody and structure remained constant, as the recording features an 'as live' performance, with constant points of synchronisation throughout. The intention was not to seek 'comical juxtapositions', or 'surprise'; rather, the aim was to explore audio models in order to record and explore responses to discover if a preference or rejection of a particular model would also provide valuable information for establishing a framework for best practice in VR video.

The piece used for playback was the finished mastered track the band had produced for release as a commercial single. This mastered track was used in original form as the

Stereo model for testing. The stereo track was converted to mono using automated conversion in the Adobe Audition software, with secondary adjustments to match the overall level of the track.

The 5.1 Surround model and Virtual Spatial Position (VSP) versions were created through remixing of the mastered component elements from the original recording session, which consisted of 15 individual tracks, containing Vocal, Lead Guitar, Rhythm Guitar, Bass, Drums and some synthesised electronic drones, described in the session as 'ambiences'. A mastering engineer was enlisted to produce a mix with spatial positioning that was in keeping with the visual media. The method devised to marry the two components used the established mapping of the Dolby 5.1 standard. The equirectangular video was overlaid against the position of the speakers with the central point of the video aligned against the Centre speaker position. The equidistant position of each performer then mapped to the remaining speaker positions (L, R, Ls, Rs). The LFE channel was not utilised. This mix used 90° spread for each performer-relative stem track and used directional bias towards their relative position against the visual media. Whilst this was in some ways an instinctual response to the mixing process in keeping with the positional placement of performers, it is again, in essence, a simulated ESMA [15] or PCMA [12] model, as indicated in the surround panning in Figure 2.

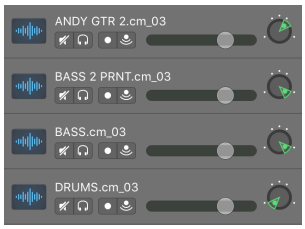

Figure 2. Stem Track positioning, relative to Equirectangular Projection

The resultant media was combined in Unity3D, texturing the inside of a sphere with the equirectangular video and using the Oculus SDK to create a first-person controller with camera (or player prefab) placed at the centre of the virtual sphere. For each model, audio emitters were positioned at appropriate positions:

For the Mono and Stereo models, the emitters were positioned at the same coordinates as the player prefab and were nested so that the audio emitters would remain relative to the viewport.

For the 5.1 configuration audio emitters were placed in fixed positions relative to the L, C, R, Ls, Rs mix positions and for the VSP model, audio emitters were placed in fixed positions relative to the visual position of each musician.

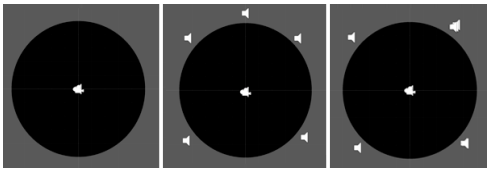

Figure 3. Mono (*left*) 5.1 (*centre*) & VSP (*right*) Unity3D Scene Design

As with previous experiments [16], questions were constructed to explore concepts of immersion and realism. Two central questions were asked of participants; 'The AV experience in model 'x' is immersive', and 'The AV experience in model 'x' is realistic' with participants asked to score each model against a Likert scale: Strongly Agree; Agree; Neither Agree Nor Disagree; Agree; Strongly Agree. In addition, sections for qualitative analysis were provided for each model. Finally, participants were asked to score each model in order of preference from Highest to Lowest, with an overall section for qualitative feedback relating to their experience of the experiment as a whole.

In order to reduce bias, each audio model was assigned a code letter for the purposes of testing: A (Mono); B (Stereo); C (5.1); and D (VSP). The study featured $4! = 24$ participants. Each participant experienced each model once, in a unique sequential order in which every order permutation of the four stimui was tested from A,B,C,D to D,C,B,A. When a participant began the study they were allocated one of the unique sequences.

The experiments were conducted with an Oculus Rift (DK1) HMD as previous studies [16] had revealed that HMD's represented the most 'natural' and 'immersive' method for experiencing VR Video. Sennheiser HD 206 Headphones were used for the tests.

IV. Results

In response to the question of Immersion, there was a mixed response to each model. However the quantitative data collected was analysed for statistical significance using a Friedman test. The results regarding Immersion were not statistically significant $\chi^2(3) = 6.383$, $p > 0.05$, although overall, the Mono Model scored the highest, and the Stereo and VSP models finishing joint lowest.

Results obtained from the question regarding Realism were also analysed for statistical significance using a Friedman test, which indicates that there is no significant difference in participants' perception of Realism across the four different audio models $\chi^2(3) = 4.631$, $p > 0.05$.

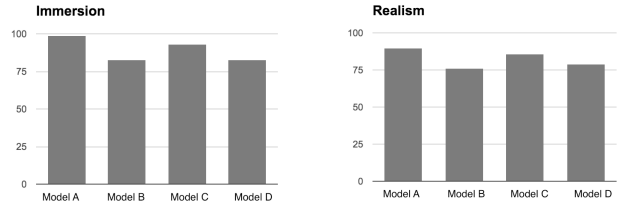

Figure 4. Questionnaire Results: Immersion (*left*) and Realism (*right*)

The qualitative data provided interesting results in response to the question of 'Immersion'. Participants occasionally described the Mono model as providing spatial feedback, for example: "*When I look at the drums I can hear them more*" and that the model offered "*Good separation*" between instruments. The assumption here must be that the psychology of perception has altered or influenced participant's experiences of listening. Chion [15] postulates that "*What do I see of what I hear and what do I hear of what I see?*" are valuable questions. In this example, the act of seeing, or 'active participation' [17] by interacting and controlling the direction of view serves to alter the perceived recognition of elements of the soundtrack.

The response to the VSP Model was the most polarised. Some found the directionality of sound to provide a high level of immersion. For others, increased directionality had hyper-real results, with feedback such as "*clipped*", "*phased*" and "*processed*" being recorded for this model.

V. Conclusions

Over the period of research there has been a significant increase in attention in VR video as well as great advancement in technologies both dedicated and conducive to VR. Techniques for producing VR Video have become increasingly innovative but these methods remain largely unstandardised with commercialism and mass-market adoption as primary focus. While ambisonic models increase in complexity and resolution, the VSP model tested with a horizontal resolution similar to a First Order model was not favourable to participants. As the majority of users do not have their own unique HRTF models, it may be advisable to postpone or limit spatial audio encoding for VR Video experiences where the presentation conditions cannot be controlled or verified.

Despite the significant increase in human and computational effort to produce spatially placed audio emitters for VR Video, the results suggest that the effect on users nominal as either PCMA/5.1 placement or VSP. In fact, physiological and perceptual characteristics can mean that a more positive effect is observed by providing the more computationally efficient Mono audio model [18] with no spatial cues. This allows the participant to imbue media stimuli with personal preference, much like the method observed in previous experiments for exploring virtual spaces undirected, rather than receive soundscapes that have been environmentally designed.

Also, as qualitative user responses to the non-spatial media were positive, an argument for defaulting to the least complicated, and least processor intensive model could be argued, especially as user perception in these cases often exhibited perceived spatialisation.

VI. Future Work

The experiment aimed to identify a best practice model for the creation and implementation of audio for VR video. In essence, the results here suggest that individuals bring with them a level of preconception as to the role of audio in a virtual experience. Some feel that this should be spatially representative; while others record that the soundscape should be reminiscent of commercial music, as in the case of streaming or broadcast. Higher Order ambisonic models may improve the effect of spatialisation and as the platforms available to users begin to support Third Order models. It would be interesting to compare the response to higher order models, compared to First Order, ESMA or VSP models.

Finally, this experiment concentrated on creating mixes from existing source material. Whether different results would be achieved by recording natively spatial recordings would also be an interesting comparison.

It will be important to continue to assess new audio models and evaluate their effectiveness in the communication of experiences in Virtual and Mixed Reality as these technologies continue to develop.

VII. Acknowledgements

Many thanks to the participants in the experiment. Thanks also to Jonny Mak, Steffan Owens, The Roseville Band and Sophie Edwards for their continued support.

VIII. References

[1] Begault, D. (2000). *3-D sound for virtual reality and multimedia*. 1st ed. Springfield, Va.: NTIS.

[2] [2]Jin, Craig, et al. "Perceptual spatial-audio coding." *Proceedings of the International Conference on Auditory Display (ICAD 2003)*. 2003.

[3] [3] Grani F, Nordahl R, Serafin S. Multimodal interactions, virtual reality and 360 movies: applications using Wavefield synthesis. In Proceedings of the Audio Mostly 2016 2016 Oct 4 (pp. 20-27). ACM.

[4] Masuch, M., & Röber, N. (2005). Leaving the Screen New Perspectives in Audio-only Gaming.

[5] Lee, H. (2017). In: *AES UK North Workshop on Capturing and Rendering Audio for VR*.

[6] Menhorn, J. (2017). *Audio Design for VR – Ustwo's Land's End*. [online] Designingsound.org. Available at: http://designingsound.org/2016/03/audio-design-for-vr-ustwos-lands-end/ [Accessed 25 Mar. 2017].

[7] Posthorn.com. (2017). *Posthorn | The "Williams Star" Surround Microphone Array*. [online] Available at: http://www.posthorn.com/Micarray_williamsstar.html [Accessed 27 Mar. 2017].

[8] Lee, H. and Rumsey, F. (2013) 'Level and Time Panning of Phantom Images for Musical Sources' *Journal of the Audio Engineering Society* , 61 (12), pp. 978-988. ISSN 15494950

[9] Facebook360.fb.com. (2017). *Spatial Workstation*. [online] Available at: https://facebook360.fb.com/spatial-workstation/ [Accessed 15 Mar. 2017].

[10] Support.google.com. (2017). *Use spatial audio in 360-degree and VR videos - Computer - YouTube Help*. [online] Available at: https://support.google.com/youtube/answer/6395969?co=GENIE.Platform%3DDesktop&hl=en [Accessed 15 Mar. 2017].

[11] Menhorn, J., Gould, R., Menhorn, J., Meyer, B. and Klier, M. (2017). *Designing Sound – Art and technique of sound design*. [online] Designingsound.org. Available at: http://designingsound.org/ [Accessed 25 Mar. 2017].

[12] Lee, H (2011) 'A New Multichannel Microphone Technique for Effective Perspective Control'. *In: 130th Audio Engineering Society AES Convention, 13-16 May 2011*, London, UK

[13] Parnell, T. (2017). In: *AES UK North Workshop on Capturing and Rendering Audio for VR*.

[14] Chion, M. (2013). *Audio-vision*. 1st ed. Armand Colin.

[15] Lee, H (2013) 'Apparent Source Width and Listener Envelopment in Relation to Source-Listener Distance'. *In: AES 52nd International Conference - Sound Field Control, 2nd-4th September 2013*, Guildford, UK , pp. 1-6

[16] Steven Davies, Stuart Cunningham, and Mike Wright. 2015. Immersive RGB+D zoetropic projection for touchscreens and HMDs/wearables. In *Proceedings of the 12th European Conference on Visual Media Production* (CVMP '15). ACM, New York, NY, USA, , Article 14 , 2 pages. DOI=http://dx.doi.org/10.1145/2824840.2824852

[17] Monaco, J. (2009). *How to Read a Film*. 1st ed. New York: Oxford University Press.

[18] Martin Naef, Oliver Staadt, and Markus Gross. 2002. Spatialized audio rendering for immersive virtual environments. In *Proceedings of the ACM symposium on Virtual reality software and technology* (VRST '02). ACM, New York, NY, USA, 65-72. DOI=http://dx.doi.org/10.1145/585740.585752