The Influence of Visual Stimuli on Localisation Accuracy for Audio in 360° Virtual Reality Content.

Peter Rowland-Jones
Derby University
p.rowlandjones1@unimail.derby.ac.uk

ABSTRACT

360° content is now available to use with mobile devices online, using a head mounted display (HMD). This combined with binaural surround sound techniques allows for a more immersive experience. This study investigates the influence of a visual stimulus on the ability to accurately localise sound sources on the azimuth plane. 360° content was created and viewed through Youtube' online spatial audio platform, a HMD & headphones. Multiple conditions were then presented to participants. Results indicate that the presence of a visual stimulus has little effect on participant's ability to localise a sound source & although full head tracking is available, front & back confusion can still occur. Results also indicate that a threshold for audio visual perception bias lies between a 0-15° azimuth offset.

1. INTRODUCTION

360° content has grown in popularity in recent years due to its availability online & the capabilities of personal mobile devices. This improved accessibility has allowed content creators to develop more immersive experiences for users. Spatial audio techniques used with this content allows users to experience a soundfield as they would in real life. This achieved using binaural surround sound techniques.

Binaural techniques rely on the principal that if the ears are presented with the same pressure information that would be found in real life, the auditory system will perceive an immersive soundfield. This lending itself well to virtual reality (VR) technology in terms of realism [1].

Binaural sound makes use of virtual loudspeaker techniques; This being the convolution of speaker feeds with a set of head related impulse responses (HRIR). Each HRIR being used as a filter for each virtual loudspeaker location & interpolation being used between them. This providing the necessary interaural time & level differences (ITD/ILD) found at each ear & the perception of sound from an externalised location [2].

The use of ambisonic audio for virtual reality environments is becoming increasingly popular due to fixed bandwidths and flexibility of playback. With online providers currently accepting 1st order B-Format audio (W,X,Y,Z) [3]. Youtube' spatial audio rendering can then dynamically binauralise audio feeds at the user end, allowing head tracking information to be applied to the audio stream.

In real world environments, a combination of cues are used to localise a sound source, these being auditory & visual among others. This study aims to assess to what extent the addition of a visual aid can influence the accuracy of localising sound within VR experiences.

A recent study into localisation accuracy within 360° media, found improvements using binaural audio compared to stereo; but used a comparison mean opinion score for analysis [4]. Meaning results were influenced by participant's opinions and possible demand characteristics. This study aims to avoid this form of qualitative data. A similar study has recently been conducted analysing the localisation accuracy of moving sound sources, over virtual (Binaural) & real loudspeaker arrays. With conclusions stating that the influence of head-tracking for the test needs to be assessed in terms of front/back confusion [5]. These findings/methodologies were used as a guide for this study.

Section 2 discusses the methods used, detailing content creation, measurements & simulations. Section 3 presents the results from the study and discusses key findings in section 4, while section 5 discusses conclusions and possible future work.

2. METHODOLOGY

Localisation studies commonly employ techniques to avoid any form of biasing, such as inverse filtering with headphone impulse responses [6] or even blindfolding participants. When assessing localisation within VR content, the combination of audio and visual cues is rarely considered. The synchronicity between the visual & auditory stimulus could influence a user's ability to locate sources within the content. This study aimed to assess whether the addition of a visual aid can influence participants ability to lateralise sound & to what extent offset audio can affect this.

2.1 Content Creation

Audio content for the study was created using the digital audio workstation, Reaper. The chosen material being an Anechoic Archimedes vocal extract [7], chosen due to containing spectral content that is familiar to the auditory system [1]. This was then positioned at each required azimuth angle using the WigWare ambisonic panner [8], depending on the condition being created as shown in Fig 1.1.

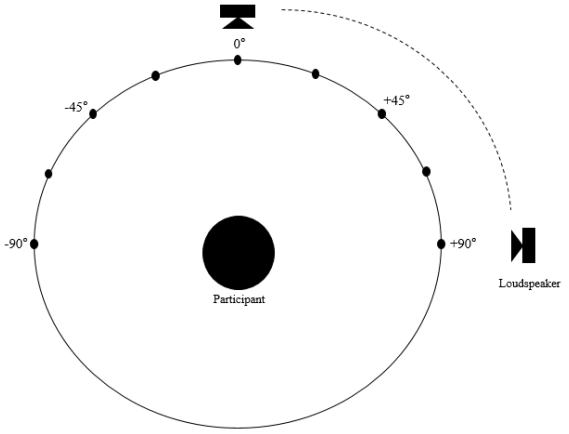

Figure 1.1 Basic layout of azimuth panning for content creation.

The B-Format output was then remapped using the WigAmbiRemap JS plugin. The Furse-Malham channel ordering was converted to ACN (Ambisonic Channel Number ordering - 0,1,2,3) with semi-normalisation applied (Fig1.2). This process allowed the audio to be presented in ambiX format, compatible with the FFMPEG stitching process & Youtube.

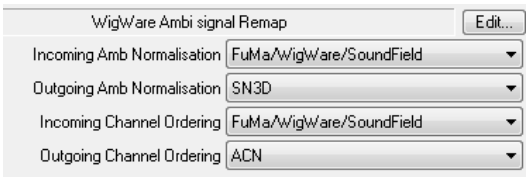

Figure 2.2 WigWare Ambi Signal Remapping for ACN.

To create the images for the study, a 'Ricoh Theta' camera [10] was used, providing equirectangular images. Each image contained visuals of the chosen test environment and markers on the floor with 15° spacing, spanning over -90° to +90°. The camera was placed at 1.3m height with all markers at a 2m distance. While loudspeaker locations were at a 2m distance and 1.15m height. These heights chosen to account for an approximate head height of a participant sat on a chair during the test. This was then accounted for using distance parameters on the ambisonic panner and near field compensation at 2m to account for low frequency addition.

The audio/visual content was then stitched together using FFMPEG with the code found below; providing a multi-channel .mov file. An example of which can be found in fig 1.3.

ffmpeg -loop 1 -i PS.jpg -i ambiX.wav -map 1:a -map 0:v -c:a copy -channel_layout 4.0 -c:v libx264 -b:v 40000k -bufsize 40000k -shortest PSandAmbiX.mov **[9]**

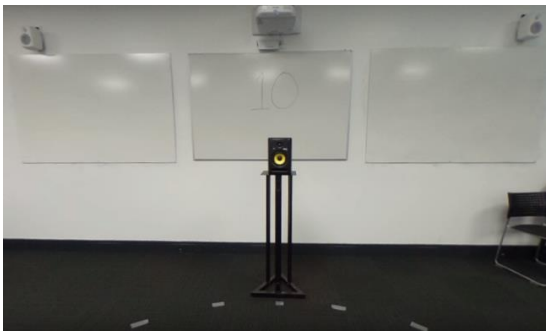

Figure 3.3 Resulting .mov file for HMD display

Then, using the spatial-media-master plug-in provided by Youtube for use with Python27, injection was completed to mark the files as spherical and containing spatial audio, ready for upload to Youtube and use with a HMD. (Fig 1.4)

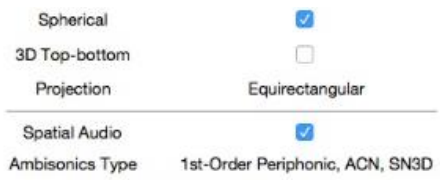

Figure 4.4 Spatial Media-Master online injection plugin.

2.2 Test Procedure

The task for participants was to localise each sound source by stating which angle they perceived it to be emitting from. The study covered two areas; Participants ability to localise a sound source with & without a visual stimulus, then assessment of the extent to which this occurs. This completed using a Nexus 5 mobile phone, Skullcandy crusher over-ear headphones & a head mounted display, these shown in figure 2.1.

Figure 2.1 Head mounted display, mobile device and headphones used during testing.

16 videos were created, each with a 5 second silence to begin, then the audio played through twice. During which the participants notified the researcher of the perceived location. 5 different azimuth angles were produced, 0°, 45°, 90°, 270° & 315°. Each with and without the visual stimulus. A further 3 conditions were created with the visual stimulus remaining in its main location (0°,45°,315°) while the audio was panned to a 15° and 30° offset. This to assess to what extent the visual stimulus would influence sound source localisation accuracy. All conditions were presented to participants in a randomised order and completed using a Youtube playlist to avoid the need to stop between conditions. The room used for the study was the same as in the visual content to allow for a higher sense of realism. Participants also received a small training period prior to the test to allow for configuration of equipment and familiarisation of procedure. An example of the final set up can be found in Figure 2.2.

Figure 2.2 Final setup for testing.

During a pilot test, it was found that synchronicity between the participants perceived angle and the real-world markers would be lost. Possibly due to the use of streaming over the web & processing restraints of the mobile device. This meaning that results acquired this way were open to inconsistencies. To overcome this the markers within the visuals were used to allow participants to notify the researcher of perceived location. These markers placed at spacings of 15° with participants able to state between two, giving twenty-five increments to choose from (7.5° per division).

2.3 Participant Selection

Participants used within the study were a mixture of trained listeners & general population, this to gain a representation of the consumer base using VR. All were asked to state if they had any form of hearing impairment to which 1 participant claimed slight tinnitus. The total number of participants was 10, with 7 male & 3 female. Each test lasted approximately 20 minutes, resulting in a total of 4hours to complete testing; accounting for consent/debriefs.

3. RESULTS

A chronbach alpha analysis was completed to assess the scale reliability, resulting in ($\alpha = 0.44$). This meaning that the scale used within the design may not hold high internal consistency of results. The following sections present the results found for each test. This being the accuracy of localisation for with and without visuals and accuracy over offset audio-visual stimulus.

3.1 With & Without Visuals

An ANOVA was used to assess the accuracy of participants localisation over each defined angle. Fig 3.1 displays the results for the test conditions with a visual stimulus (F-Value – 66.28, P – 2.77) suggesting support of the null hypothesis that all means are the same. Fig 3.2 displays results for the same conditions without visuals (F- 288.33, P - 7.27), this also suggesting the support of the null hypothesis. As can be seen, participants were able to localise each position with high levels of accuracy regardless of visual stimulus. The only location providing some level of variance being 0°.

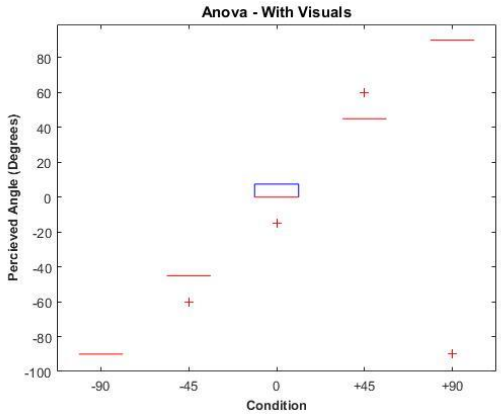

Figure 3.1. Anova With visuals for 0 °,45 °,90 °,270 °,315°. (F- 66.28, P- 2.77)

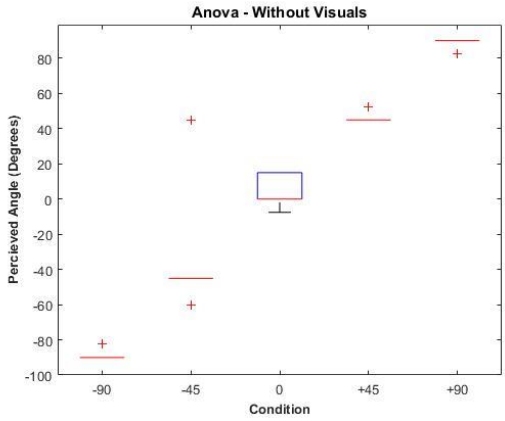

Figure 3.2. Anova Without visuals for 0 °,45 °,90 °,270 °,315°. (F- 288.33, P- 7.27)

Fig 3.3 displays the results found when comparing the levels of variance between the 0° azimuth with and without the visual stimulus. (F-Value – 1.17, P – 0.208). The overall means of each show that in both conditions the averages were the same, although more variance occurs without the visuals. The resulting p-value being >0.05 cannot certify that this is due to the visual aid & could be due to demand characteristics.

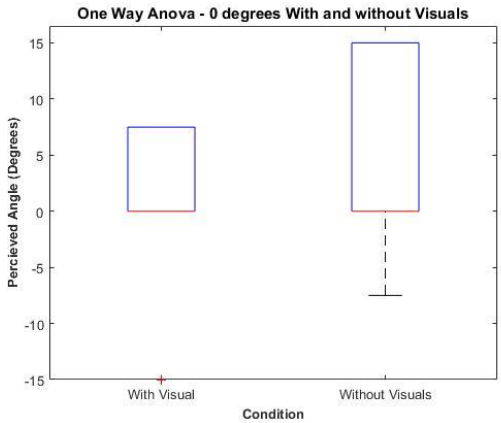

Figure 3.3. Anova for 0° with & without visuals. (F- 1.17, P – 0.208)

3.2 Error Rate

Fig 3.4 displays results found for variance of perceived sound source location at 0° for a visual aid with 15/30° audio offsets. (F-Value – 6.08, P – 0.0036) With a p value <0.05 suggesting rejection of the null hypothesis that the visual aid has no influence. Looking at the mean average of the first two conditions shows that both were localised at 0°, although higher levels of variance were found in the 1st condition. This possibly due to the previously mentioned demand characteristics. Although when looking at fig 3.5 below for the 45° offset comparisons; The mean averages show the same behavior of localising the sound source from the visual location rather than the 15° offset. Both plots then show a range of variance of approximately 15°-30° for the final location. This displaying a possible threshold for negative impact of visual aid being within a 15°offset.

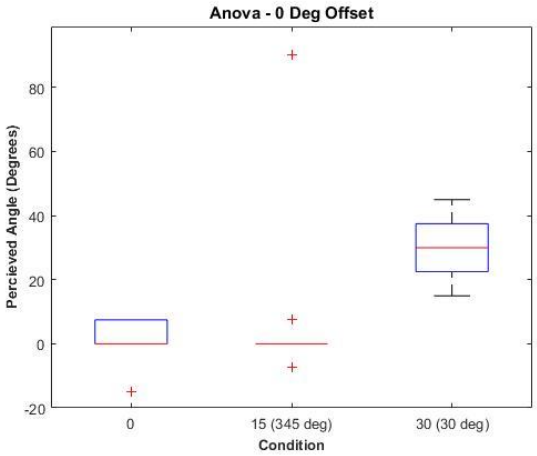

Figure 3.4. Anova 0° offset (345°, 30°) with visuals. (F- 6.08, P-0.0036)

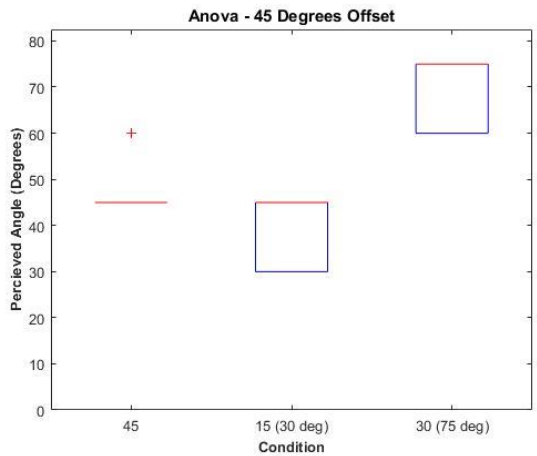

Figure 3.5. Anova 45° offset (30° & 75°) with visuals. (F – 12.79, P – 0.0001)

4. DISCUSSION

As can be seen from Fig3.1 & 3.2 the influence of a visual stimulus at differing azimuth angles show to have little significance on participant's abilities to lateralise sound. Possibly due to the markers themselves within the 'no visuals' testing being a visual factor that participants used to anchor their perceptions. Another influential factor also being the lack of reflections used for the auditory stimulus. Binaural room impulse responses were not used, to concentrate on the influence of the visual aid. But the lack reflections to distance the sound from the user was noted during testing, as participants could be seen using the ILD between the ears to lateralise the audio by turning their heads.

Issues due to front back confusion can be seen at 90° (Fig 3.1) with visuals and 315° without (Fig 3.2) due to outliers in data. This possibly due to the HRTF data used by Youtube not suiting participants, but could also be caused by the randomised nature of the conditions meaning participants were already facing this direction before the next test was completed.

Other influential factors found were participants who wore glasses had to remove them to complete the study. Although the HMD provided a focus lens, this was still found to be unsatisfactory as they found the device uncomfortable to wear, possibly skewing results.

5. CONCLUSION

This study aimed to assess the influence of a visual aid on the accuracy of localisation on the azimuth plane for VR content. To assess this, two types of test were conducted. A standard comparison between participants ability to localise with & without a visual aid, & assessment of the degree to which a sound source can be offset from a visual aid and be perceived as emitting from the visual origin. To complete the tests, a mobile device was used with a HMD & headphones. 360° content was created with equirectangular images and binaurally synthesised audio. These then presented over the Youtube 360 Platform online. Results showed that the influence of a visual aid on ability to accurately localise a sound source can be insignificant. Sound directly in front of a user can also cause slight variation in perceived angle, but testing methodologies would have to be improved to assess the real world significance of this. When assessing the degree to which audio can be offset from a visual stimulus, findings showed that between 0°-15° offset from the visual aid can cause perceptions to be localised visually rather than using auditory cues. Issues such as front & back confusion can still be noted within the results, while previous hypothesis were that due to the addition of head tracking within the VR content, this would not occur.

6. FUTURE WORK

Future work would include improving the accuracy of this testing methodology, by increasing the range of values a participant can choose from for perceived angle; and the capture of results. Possibly by using a handheld controller to point within the VR content & a dual screen display for the researcher. Then the use of more participants to allow a greater real world comparison of results. Also the use of binaural room impulse responses for the addition of reflections & distance perception, to assess its influence on the localisation accuracy for VR. If successful, investigations into the influence of visual stimulus on lateralization of sound over elevation could be completed.

7. REFERENCES

[1] G, Ballou, 2015. Handbook For Sound Engineer. 5th ed. New York & London: Taylor & Francis Group. 71,72,73.

[2] Holman, T (2008). *Surround Sound - Up and Running*. 2nd ed. Oxford: Elsevier. Pg 85,86,88,91

[3] Kares, J, Larcher, V (2016). Streaming Immersive Audio Content . In Audio For Virtual & Augmented Reality . Los Angeles , 30/09/2016 - 01/10/2016. USA : Audio Engineering Society . 1 -8

[4] Niwa , K, Ochi, D, Kameda, A, Kamamoto, Y, Moriya, T (2016). Smartphone-Based 360 Video Streaming/Viewing system including acoustic immersion. . In 141st Convention. Los Angeles , 29/09/2016 - 02/10/2016. USA : Audio Engineering Society . 1,2,4,5,6.

[5] Hughes , S, Kearney, G (2016). Moving Virtual Source Perception in 2D Space . In Audio for Virtual & Augmented Reality . Los Angeles , 30/09/2016 - 01/10/2016. USA : Audio Engineering Society . 1 - 9.

[6] Olive, S, Welti, T, Mcmullin, E, (2013). A Virtual Headphone Listening Test Methodology . In 51st International Conference. Finland , 22-24/10/2013. USA: Audio Engineering Society . 1 -10.

[7] Bruce Wiggins. 2017. BBC Anechoic Material. [ONLINE] Available at: https://unimailderbyac-my.sharepoint.com/personal/seng009_derby_ac_uk/_layouts/15/onedrive.aspx?id=%2Fpersonal%2Fseng009_derby_ac_uk%2FDocuments%2FWork%2FAudio%2FTest%20CDs%2FAnechoic%20Recordings&FolderCTID=0x012000E975FD09CB71B84CAFDD65FAEF602689. [Accessed 10 April 2017].

[8] Bruce Wiggins. 2017. The Blog Of Bruce - Wigware. [ONLINE] Available at: https://www.brucewiggins.co.uk/?page_id=78. [Accessed 12 April 2017].

[9] Wiggins, B. 2017. Youtube, Ambisonics & VR. [ONLINE] Available at: https://www.brucewiggins.co.uk/?p=666. [Accessed 10 April 2017].

[10] Ricoh Theta 360. 2017. Theta. [ONLINE] Available at: https://theta360.com/uk/. [Accessed 17 April 2017].