Immersive Audio Recording for Virtual and Augmented Reality

0 INTRODUCTION

This work follows on from [1]...

1 Literature Review

1.1 Recording Techniques

1.2 Microphones

This sections reviews the microphone arrays used in the project

1.2.1 OCT

- OCT Surround - Mics were panned according to the Auro speaker layout - https://docs.google.com/document/d/1-h7T9wjdY2L8d3MzBohdzhVsfJPdRlLM2abT22jpuyE/edit

2 Recording at Abbey Road Studios

The following section describes the planning process and recording session of this project. All microphones arrays used in the recording session are listed and described including their position within the room.

2.1 Planning

To produce a VR experience of a live musical performance an appropriate musical ensemble was required. There were a few prerequisites before deciding upon an ensemble, such as deciding upon a musical genre. The style of music needed to be appropriate for a wider demographic and could not be explicit in its nature. The ensemble must also be well-rehearsed and employ a sense of professionalism to ensure that the time available in the studio was used efficiently and productively. With these conditions in place, a London-based Indie-Pop ensemble called 'Nova Neon' were contacted. Nova Neon are a five-piece band who play a sophisticated style of Indie-Pop using a mix of both anechoic sounds and acoustic instruments. The song chosen for the performance was called 'Close Your Eyes' from their album 'Chroma'. This song was chosen for the hard panned left and right guitars which could be exaggerated within an Higher-Order Ambisonic reproduction.

The recording location of choice was Studio Three at Abbey Road Studios. The room is large enough to comfortably house the 5 piece band and boasts heigh ceilings that allow for a more diffuse sound-field to exist higher in the room which could be taken advantage of in exploring capturing the ambience of the space using dedicated multichannel microphone arrays.

2.2 Microphone and Camera Set Up

As this research is currently of interest amongst audio engineers and professionals in the audio industry, it was decided to invite other collaborators to participate in the research. Dr Hyunkook Lee from Huddersfield University was invited to attend the recording session to set-up the Equal Segment Microphone Array, which he had shown to perform well at capturing audio for VR. The technical director of Schoeps Mikrofone Helmut Wittek was also contacted with an invitation to collaborate. Schoeps agreed to contribute by supplying their ORTF-3D Surround and OCT-9 Surround multichannel microphone arrays (MMAs). By collaborating with Schoeps Mikrofone and Dr Hyunkook Lee, the research was expanded to include a wider selection of multichannel microphone arrays for subjective analysis.

The floor plan for studio three can be seen in Figure 1 annotated with three letters, A (green), B (red) and C (blue) showing the different positions that the Ambisonic microphones, multichannel microphone arrays and 360° video cameras were placed.

The following sections list the microphones that were placed in each of these positions.

2.2.1 Position A

Position A is located in the centre of recording space and in the middle of the musicians. This is where the visuals for the VR experience were captured in 360° using the GoPro Omnirig. The Soundfield ST450 MKII, EM32 Eigenmike, Equal-Segment-Microphone-Array (ESMA) and Schoeps ORTF3D Surround were also set-up in this position to capture direct sound radiating from the instruments.

2.2.1.1 GoPro Omnirig The GoPro Omnirig is a synchronised six-camera array used to capture 360° videos [3]. It was placed at position A at a height of 160cm to the bottom of the array. GoPro number one in the Omnirig was positioned to face a reference point at the back wall where the drum kit would be set-up.

2.2.1.2 Neumann KU100 The KU100 was positioned to be facing towards the reference point (drum kit). Though this would not be used in the listening tests it was included so it could be used to compare against the other recording techniques.

2.2.1.3 EM32 Eigenmike The Eigenmike was placed on its side just above the GoPro Omnirig at a height of 1.81m shown in figure 2. The aim was for the Eigenmike to capture the direct sound from the instruments and provide a soundfield recording 'canvas' on which the spot microphone recordings could be placed within a Higher-Order Ambisonic framework.

2.2.1.4 Soundfield ST450 MKII The soundfield microphone was placed at a height of 2m and placed on its side above the GoPro Omnirig and Eigenmike. It is worth noting that the 'end fire' switch must be activated on the ST450 pre-amplifier if the microphone is placed on its side for recording. The 'end fire' setting adjusts the X, Y and Z axis data to ensure that the correct soundfield orientation is captured.

2.2.1.5 Equal Segment Microphone Array (ESMA) The ESMA is based on the 'four segment array' proposed by Michael Williams in 1991 [4], [5]. It was designed to capture sound from 360° in the azimuth plane using four cardioid microphones positioned in a square arrangement, with a distance of 25cm between each capsule. The angles between each microphone should be 90°, creating a stereophonic recording angle (SRA) of ±45°[4], [5].

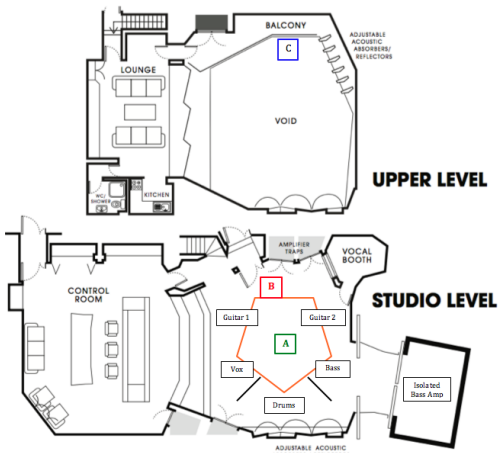

Fig. 1: Preliminary floor plan of the recording session in Studio Three [2]

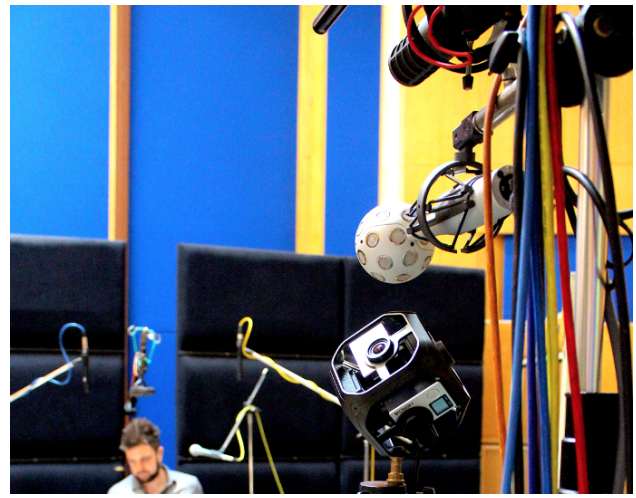

Fig. 2: Placement of the first Soundfield ST450 MKII microphone, Eigenmike and GoPro Omnirig at position A

Through listening tests, Dr Hyunkook Lee of Huddersfield University found that by changing the distance between each capsule to 50cm, the localisation accuracy of sound sources within a VR environment was improved [6]. Each of the four microphones on the azimuth plane were angled down by 45° to face the instruments and optimise direct sound capture from the instruments. The array was positioned at a height of 2.15m, with the upward facing microphone capsules reaching a height of 2.28m. The four additional upward-facing cardioid microphones were added to capture some of the diffuse sound within the recording environment. The four upward-facing microphones were also set-up with distance of 50cm between each capsule. The use of cardioid microphones and the wide spacing between each microphone capsule helps to minimise inter-channel cross-talk, whilst still providing satisfactory localisation of sound sources within the azimuth plane.

2.2.1.6 ORTF-3D Surround The supercardioid microphones used in this array allow for sufficient signal separation between each microphone to avoid unwanted inter-channel cross-talk and comb-filtering effects. The spacing between each microphone follows the principles of the ORTF technique to allow for the required inter-channel time differences and improved spatial resolution to be included into the recording. The ESMA and ORTF can both be seen at position A in figure 3.

2.2.2 Position B

Position B is located to the rear of Studio Three near the entrance. The objective for this position was to capture a 180° view of the musicians and provide a different audio and visual perspective for the VR experience. The Samsung Gear 360 camera, Stereo X-Y pair, OCT-9 Surround, Perceptual Control Microphone Array (PCMA) and Sennheiser AMBEO Ambisonic microphone were placed

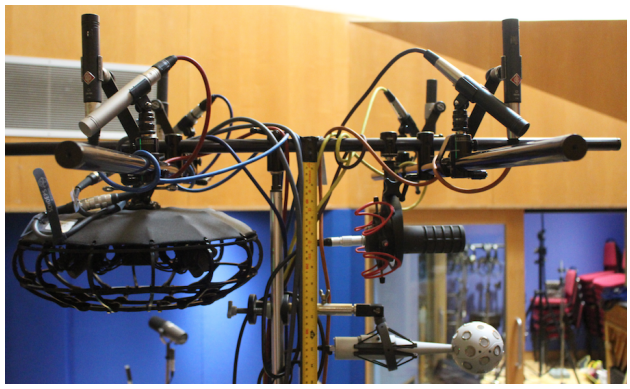

Fig. 3: Picture of the ORTF-3D Surround Array without windshield (top) and the ESMA (left) just underneath. The front of the arrays was positioned facing the reference point at the drum kit.

Fig. 4: Placement of the Sennheiser AMBEO and Samsung Gear 360 camera at position B

at position B to capture both audio and video from this viewing (and listening) position.

2.2.2.1 Samsung Gear 360 The Samsung 360 was used to provide a 180° visual perspective of the live performance with the musicians in front of the viewing position. The camera, multichannel microphone arrays and Ambisonic microphones were directed at the same reference point facing towards the drum kit.

2.2.2.2 Stereo X-Y Pair A coincident stereo X-Y microphone arrangement was set-up at position B as a reference. Two Neumann KM184 cardioid microphones were arranged to produce a stereo recording angle of 115° and positioned to face the drum kit at a height of 1.94m to the coincident point.

2.2.2.3 Sennheiser AMBEO The Sennheiser AMBEO is a B-format Ambisonic microphone similar to the Soundfield ST450 MKII. It is also capable of recording First-Order Ambisonics using a tetrahedral coincident arrangement of four cardioid microphone capsules [7]. The AMBEO is available to purchase at a considerably lower price compared to the Soundfield ST450 MKII and therefore it would be of interest to compare their performance. The AMBEO was placed at position B, just above the Samsung Gear 360 camera at a height of 1.59m to the centre of the microphone grill. The positioning of the AMBEO did cause some occlusion where the Samsung camera was obstructing sound from entering the bottom of the microphone, shown in figure 4. This was not ideal but due to limited space available the positioning was deemed appropriate.

2.2.2.4 Perspective Control Microphone Array (PCMA) Due to space and equipment limitations encountered during the recording session, the PCMA set-up used the front-centre, rear-facing and upward-facing microphones from the OCT-9 Surround. Coincident stereo pairs were set-up for only the front left and front right positions. The PCMA was set-up at a height of 1.82m to the microphone capsules.

2.2.2.5 OCT-9 Surround The OCT-9 Surround microphone array was designed by Günther Theile and Helmut Wittek for the Auro-3D (9.1) loudspeaker arrangement. The front facing section is based on Theile's optimised cardioid triangle surround (OCT-Surround) array, which is used to capture sound for 5.1 multichannel systems [8], [9]. Four upward-facing supercardioid microphones are added to the OCT-surround to create the complete OCT-9 Surround array. Using directional microphones in this way again helps to increase channel separation and avoid unwanted inter-channel cross-talk. The front three microphones are used to capture direct sound whilst the rear and upward-facing microphones capture the diffuse sound and ambience of the recording environment.

2.2.3 Position C

Position C was located behind position A at the rear of Studio Three. At position C, the IRT Cross, Hamasaki Cube and second Soundfield ST450 MKII microphone were placed much higher in the room in contrast to the multichannel microphone arrays located at positions B and C. The aim was for the height arrays and Ambisonic microphone in position C to record more of the diffuse sound existing higher up in the room and capture the ambience of the recording space.

2.2.3.1 IRT Cross The IRT cross or 'atmo-cross' is a multichannel microphone array generally used to capture diffuse sound and direct environmental sounds such as applause and crowd noise. It can be created using four cardioid microphones positioned in a square with 20cm - 25cm spacing between each capsule [10]. The IRT Cross can also be used in combination with other arrays to allow for the capture of both direct and diffuse sound to provide a full spatial representation of a musical performance within

an environment. The IRT Cross was positioned at a height of 3.5m at position C, which helped to reduce direct sound capture and increase the diffuse sound captured.

2.2.3.2 Hamasaki Cube The Hamasaki Cube is a multichannel microphone array specifically designed to capture the reverberation and diffuse sound in performance spaces such as concert halls. Due to space limitations, the dimensions of the Hamasaki Cube were reduced from 1m to 0.7m. Eight Neumann U87 condenser microphones were used to create the Hamasaki Cube, which was positioned at a height of 3m to the microphone capsules on the lower layer of the cube. The microphone capsules on the upper layer of the Hamasaki Cube reached a total height of 3.7m. Given the widely spaced microphone arrangement and the height of the Hamsaki Cube creating sufficient inter-channel time and level differences, it was expected to perform well at capturing the ambience Studio Three.

2.2.3.3 Soundfield ST450 MKII The second Soundfield ST450 MKII Ambisonic microphone was positioned between the front two microphones of the IRT Cross at a height of 3.45m to centre of the microphone grill. This allowed for a First-Order Ambisonic soundfield recording higher up in the recording space in the hope of capturing more of the rooms ambience.

All three microphones at position C is illustrated in figure 5 and the separation between position A and C is shown in figure 6.

Fig. 5: Placement of the second Soundfield ST450 MKII, IRT Cross and Hamasaki Cube at position C

2.2.4 Spot Mics

For an ensemble such as this it is natural when performing for each musician to be spot miked due to the natural variation in dynamics. If only the microphone arrays described above were used then the lead and backing

Fig. 6: Microphones at position A with microphones in position C in the background

vocalists would be drowned out by the drums. As it was also decided that the bass guitar would be recorded in isolation to avoid the lower bass frequencies overpowering other sound sources within the recording space, the bass guitar would not be picked up by the arrays at all. Therefore each instrument required their own spot mic to ensure that the vocals and bass guitar could be mixed in consistently with the rest of the ensemble. The spot microphones were positioned by Abbey Road engineers using their usual recording techniques and work flow. The spot microphones and placement for each instrument are detailed below.

2.2.4.1 Drums Each section of the drum kit was recorded using spot microphones, the details for which are displayed in table 1.

Section	Microphone Model	Polar Pattern	Position
Kick Drum	Shure Beta 52a	Cardioid	Inside the kick drum
	Neumann U47-fet	Cardioid	In front of the kick drum
Snare Drum	Shure Unidyne III 57	Cardioid	Top of snare drum
	AKG 414	Cardioid	Underneath snare drum
Hi-Hat	Shure SM58	Cardioid	Above Hi-Hat
Knee	Sony C38	Cardioid	Above drummer's right knee
Mid Tom	Beyer M201	Cardioid	Above mid tom
High Tom	Beyer M201	Cardioid	Above mid tom
Mono Overhead	Coles 4038	Bi-directional	Above the drummers head
Stereo Overheads	2 x DPA 4011	Cardioid	Standing position for Vocalist
Stereo Drum Room Mics	2 x DPA 4006	Omnidirectional	3 feet in front of the drum kit

Table 1: A table of the spot microphones used for recording the ensemble

2.2.4.2 Bass Guitar To prevent excessive low frequency spill from the bass guitar, an Ampeg B15N Portaflex bass amplifier was placed in an isolation booth and recorded with a Neumann FET 57 microphone. As it would look unnatural to hear the bass but see no bass

amplifier in the VR experience, a 'dummy' bass amp was placed in the room to the left of the bassist with the rest of the musicians. Figure 7 shows the FET mic placement on the dummy amplifier.

Fig. 7: Neumann FET 57 placement on the 'dummy' amplifier

2.2.4.3 Electric Guitars An Orange Crush 60R amplifier and a Fender Blues Junior amplifier were placed on the right and left of the central position with both guitarists respectively. Both amplifiers were placed on chairs to raise the height and angled up to allow a clear path for sound to travel towards the multichannel microphone arrays and Ambisonic microphones. Neumann U87 condenser microphones were also placed centrally in front of the amplifiers. Direct signals were also taken from both guitarists pedal boards. Effort was made to match the volumes of both guitar amplifiers to create a balanced sound within the room which would be captured by the multichannel microphone arrays. Figure 8 shows both of the electric guitar amplifiers as described.

2.2.4.4 Vocals The lead vocalist was recorded using a Shure SM7B microphone and positioned to the right of the drum kit, shown in figure 9. Acoustic panels were placed between the drums and vocalist to minimise any spill picked up by the vocal microphone. Backing vocals provided by guitarist 2 were also recorded using a Shure SM7B, whilst the drummers backing vocals were recorded using a Shure SM58. Microphone directivity was crucial to reject as much off-axis sound as possible.

(a) Neumann U87 placement on Guitar 1 amplifier

(b) Neumann U87 placement on Guitar 2 amplifier

Fig. 8: Spot microphones on the electric guitar amplifiers

Fig. 9: Picture showing the placement of the Shure SM7 and acoustic panels

2.3 Recording Process

Once the equipment was set-up for the recording session, each of the microphone channels from each multichannel microphone array and Ambisonic microphone were connected to tie lines in the live room and routed to a channel on the SSLJ 9000J 96-channel mixing console in the control room. The spot microphones used the pre-amplifiers on the mixing console as did each channel of the Hamasaki Cube. For the Ambisonic microphones (Soundfield ST450 and Sennheiser AMBEO), a stepped pre-amplifier was required to ensure that the gain levels set for each of the W, X, Y and Z channels were identical. There were thirty six AMS Neve Montserrat pre-amplifier channels and twelve AMS Neve 1081 pre-amplifier channels available for the session. Care was taken to ensure that each microphone array and Ambisonic microphone used the same pre-amp model for each of their individual channels.

The session was recorded on a ProTools HD rig [11] at a sampling rate of 48kHz and a bit-depth of 24. This was the practical option as there were in excess of ninety channels being used and the file sizes had to be taken into consideration. Recording at 48kHz/24bit allowed for high-quality recordings whilst ensuring that the file sizes were still practical to work with in the post-production process. The Eigenmike was recorded onto a separate laptop that was synchronised to the ProTools HD session and time-stamped, allowing for the Eigenmike to be recorded simultaneously with the ProTools HD session.

A 5.1 surround system comprising of five Bowers and Wilkins 800D speakers was used for monitoring in the control room. Where possible, the arrays were 'folded down' to the 5.1 system for monitoring. The spot microphones were monitored in stereo as would be the case in a conventional recording session. Although the monitoring system did not allow for 3-D audio reproduction, it was useful

to listen and switch between each of the different channels and different microphone arrays in real-time. Timbral qualities such as the brightness of the microphones could be identified whilst monitoring and it provided an early indication to how each of the different multichannel microphone arrays and Ambisonics microphones might sound after the implementation stage.

3 Post-Processing and Workflow Set Up

Once the recording at Abbey Road Studios was complete, the recording session was assessed using 5.1 surround sound system with a PC running ProTools 12 in the media suit at the University of York. In a normal audio editing situation the best bits of every take could be split and merged together. However as this would cause discontinuity in the video and the audio, coupled with the impracticality imposed by the abundance of audio tracks, a single take must be used. It was decided that take eight of 'Close Your Eyes' would be used as the track for testing. This take was exported and condensed into a smaller project where each track was exported as a WAV file ready to be imported into Reaper for Ambisonic processing.

3.1 Video Post-Production

Before the audio could be mixed within an Ambisonic framework, the 360° videos needed to be produced. This was to ensure that the sound sources were accurately aligned with the visual in the VR environment, i.e the audio of the lead vocals needed to sound as though they were coming from the lead vocals. As two different 360°cameras were used during recording, two different methods of spherical video production were used.

3.1.1 Video Stitching and Editing

3.1.1.1 Position A - 360° perspective GoPro Omni The GoPro Omni rig comes paired with a software suite including Omni Importer that can be used for easily importing 6 individual video files that can then be stitched together and edited to create a spherical video ready for use in a VR workflow using Kolor's AutoPano Video software. The videos files from the six GoPro cameras for take eight were imported using the Omni Importer software and stitched together in AutoPano Video. With time and experience it is possible to optimise the video stitching process to prevent the 'ghosting' effect in which subjects/objects placed too close to the GoPro Omni rig are distorted as they are split between two of the GoPro cameras and part of the subject/object is lost within a blind spot. It can be seen in the video that the guitarist have fallen victim to such an effect.

Once stitched together the video was exported as an mp4 file and imported into Adobe Premiere Pro [12] for further editing. The colour balance was adjusted to add some vibrancy to the visuals, before a title was added at the beginning of the video. Fades were also applied at the beginning and end of the video to stop the video starting and ending abruptly. The editing process in Adobe Premiere was simple but necessary to produce a video suitable for use in the VR experience and listening tests.

3.1.1.2 Position B - 180° perspective Samsung 360 Videos shot using the Samsung 360 camera can be stitched and processed using Samsungs dedicated software 'Gear 360 Action Developer'[13]. As the camera only uses two fish eye lenses it is not possible to adjust where the software stitches the two videos together meaning that any ghosting effect that occurs can not be removed. As the Samsung 360 was placed further away from the musicians and the front facing lens was directed towards the entire ensemble, ghosting did not affect any subjects of interest. Once the video was stitched the same post-processing in Adobe Premiere was applied to the video.

3.1.2 Video Playback and Head-Tracking

Kolor's GoPro VR Player [14] can be used to open and preview 360 videos with built in head-mounted display (HMD) support allowing the video to be navigated using either the HMD's motion sensors or by clicking on the video and dragging. In the case of this project an Oculus Development Kit 2 (DK2) [15] headset was utilised. Using an external piece of software it is possible to send the head-tracking data from GoPro VR Player to Reaper to synchronise the soundfield rotation required for head movement. This is explained in section **??**.

3.2 Audio Post-Production

3.2.1 Creating a Higher-Order Ambisonics Template in Reaper

Reaper [16] is (at the time of writing) currently the only DAW that allows for up to 64 channels per track making it perfect for Higher Order Ambisonic (HOA) production. A Reaper template utilising a 36 channel bus for each of the microphone arrays was produced, allowing the microphone arrays to be encoded up to Fifth-Order Ambisonics. Each of the 36 channel buses for each of the microphones arrays were then routed to a 36 channel track containing a sound field rotator and Fifth-Order binaural decoder.

The Eigenmike was imported and processed on its own 36 channel track and routed to a sound field rotator and 3rd decoder. As the Eigenmike was placed on its side during the recording session, the orientation was corrected by applying an additional sound field rotation.

The B-Format recordings produced by the Soundfield ST450 MKII microphones created four tracks for each of the W, X, Y and Z information channels. The separate W, X, Y and Z channels were assigned to a four channel 'parent' track in Reaper. The Soundfield ST450's also required their own 4 channel First-Order Ambisonic tracks and decoders.

Once the template was produced, the raw audio files were imported and grouped into the appropriate microphone groups. The following sections describe the processing of the raw audio files to produce a Fifth-Order

head-tracking binaural mix.

3.2.2 Treatment and Ambisonic Encoding

Before anything was encoded into an Ambisonic format, Reaper 'ReaEQ' equalisation plugins were applied to each of the microphone tracks. It is important to EQ the tracks before Ambisonic encoding to prevent spherical harmonic distortion. In some instances where the microphones were placed close to the instruments, subtle filtering was applied such as a 6db reduction below 110Hz on the drum kit to reduce the overpowering bass drum in an attempt to increase the clarity of the mix. Individual microphone channels within an array were treated with the same equalisation to provide consistency for the listening tests. The spot microphones were treated by Mirek Stiles from Abbey Road Studio before being imported into the Reaper project. This was to ensure that the spot microphone recordings were of good tonal quality as a standard mix should be.

Inserted in the FX chain next was a Fifth-Order AmbiX encoder. The AmbiX encoders use the ambiX Ambisonic format, with the Ambisonic Channel Number (ACN) ordering and SN3D normalisation conventions [17]. The AmbiX encoder plug-ins provide a graphical interface from which one can position mono or stereo sound sources to specific azimuth and elevation angles around the VR environment, shown in figure 10.

The spot microphones were duplicated and split into two groups. The first group was used to encode the spot microphones relative to the video from viewing position A and the second for viewing position B. This was done by opening the video in GoPro VR Player and setting the view to the reference 0° position, in this case the centre of the drum kit. The individual spot microphones were then placed in the 3-dimensional space using the ambiX encoder plug-ins and using the Oculus DK2 as a visual reference. A B-Format impulse response of St Albans Cathedral [18] was obtained from the OpenAir online impulse response library to create a convolution reverb effect on the lead and backing vocals. Using the AmbiX MCFX Convolver4 plug-in [19], [20] inserted onto 4-channel auxiliary track, it was possible to achieve a 3-D reverb effect that suited the VR experience better than a standard mono or stereo reverb effect. The vocal tracks could then be sent individually to the B-Format reverb bus to achieve the required wet/dry ratio.

Each microphone arrays was encoded in the same way that they should naturally be listened to. For example, the OCT surround channels were encoded in the same position that the speakers are designed to be placed as described in section 1.2.1. For other microphones arrays that are not designed to be fed straight to a specific speaker layout, each microphone channel was encoded into a position relative to the listener as the microphones were placed in the room.

For the raw SoundField ST450 recordings it was necessary to convert the Soundfield's B-Format recordings from the Furse-Malham format to the ambiX format using the AmbiX Converter plug-in [19], [20], which corrected the

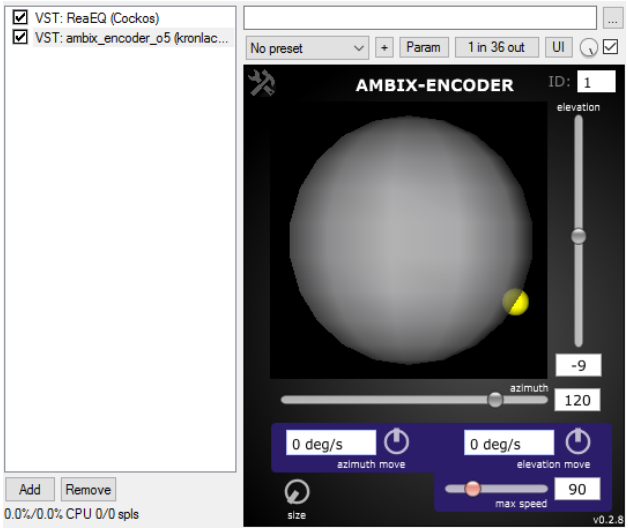

Fig. 10: Screenshot of the Fifth-Order AmbiX encoder for a mono sound source (Kick Drum)

channel sequence and normalisation before the decoding stage. Apart from the format conversion, no other Ambisonic encoding was required as the Soundfield ST450 MKII microphone captured the full spherical soundfield with sound sources already in the correct locations.

Before importing into the reaper project, the Eigenmike recording was encoded into Third-Order Ambisonics using the EigenUnits plug-in by mh acoustics [21]. This then meant the 36 channel 3rd order Eigenmike tack could be imported into the Reaper project without he need to encode the track in real time. As the Eigenmike was mounted on its side during the recording session, the soundfield had to be rotated by -90° in the elevation plane to compensate for this positioning using an AmbiX Third-Order Rotator [19].

3.2.3 Binaural Decoding

Each of the microphone array channels and spot mics were routed to a final mixing bus used for 5th order decoding. The Egienmike and Soundfield microphones were sent to a 3rd order and 1st order mix bus respectively. An ambiX Binaural Decoder plugin was inserted onto each of the mix buses. The plug-in requires a configuration files that contains information regarding where the HRTF data set to use is stored, the virtual loudspeaker layout that should be used and provides a decode matrix to multiple the signal vector by [22]. Configuration files for the following Ambisonic orders and speaker layouts were taken from the SADIE database [23]:

Order	Speaker Layout	Configuration File
First	50 point Lebedev Grid	”h5-v5-50_point_Lebedev_Grid-MaxRe-pinv”
Third	32 point Lebedev Grid	h3-v3-Lebedev-Grid-26-Speakers-MaxRe-pinv
Fifth	8 point Cube	h1-v1-Cube-MaxRe-pinv

As Google have recently adopted the KU100 measurements from the SADIE database to use in their YouTube360 [24] online platform, the same HRTF data set was used in the binaural decoding process. The HRTF set first needed to be converted from 44.1kHz/16-bit as downloaded from the SADIE database to 48kHz/24-bit to match the settings of the Reaper project using the r8brain V1.9 [25] application.

3.3 Audio Visual Synchronisation

When turning your head in the real world, sound sources remain static in 3D space and your perspective of the world is rotated both visually and sonically. When listening to a binaural decode over headphones however, when you turn you head the headphones and the audio being played through them stay in place on your head. If something was encoded you sound like it was coming from your right, it would sound like it was coming from your right no matter where you turned your head. To create a truly immersive audio visual environment the sound field must be rotated as the visuals are.

This can be achieved by using a sound field rotation plug-in and feeding it real time rotation data.

SpookSyncVR [26] is built in Max MSP as a stand-alone application that allows data exchange between GoPro VR Player and Reaper using Open Sound Control (OSC). Using SpookSyncVR it is possible to gather the X (yaw) and Y (pitch) positional data from the headset and transfer the information to Reaper so that the sound field can be rotated accordingly. For this an Ambix Rotator plug-in was inserted into each of the three decoder tracks just before the binaural decoding plug-in for which the yaw and pitch data from GoPro VR Player were assigned to control the values of the yaw and pitch data of the rotator plug-ins. By designating the GoPro VR Player as the ’master’ and Reaper as the ’slave’ it was possible to synchronise, play and stop the audio and video together.

4 Listening Tests

Two rounds of listening tests were conducted for viewing position A (test 1) and B (test 2). The procedure was identical however the data used, such as video and microphone configurations used were different for each. Participants were recruited from the University of York and Abbey Road Studios for both tests, all of who were required to have some previous experience with mixing/producing and/or spatial audio. The number of participants recruited for each test were as follows:

	UoY	Abbey Road	Total
Test 1	15	4	19
Test 2	29	9	38

4.1 Attributes Focus Group

The aim of the listening test was to assess the performance of each microphone array configuration for a VR environment in terms of its spatial and timbral quality. Due to the subjectivity of such a test, a focus group was assembled with the purpose of producing a list of mutually agreeable adjectives to use to describe certain spatial and timbral attributes. The attributes chosen to use within the listening tests are shown in table2.

Attribute	Description
Spatial	
Locatedness	How easily you can locate a sound source within the VR environment
Sense of Space	How well the space where the recording was made is perceived
Externalisation	Perception of sound coming from all around your head
Envelopment	Whether the sounds are perceived to originate inside of outside of the head
Timbral	
Full	Abundance of low frequencies present
Bright	Abundance of high frequencies present
Flat	Lack of high and low frequencies present
Rich	The mix sounds good with both high and low frequencies
Realistic	The sounds heard in the VR experience are realistic (sound like real instruments) and timbral characteristics have been preserved.
Loud	The perceived level sounds high

Table 2: Table of Spatial and Timbral Attributes

4.2 Procedure

Using an Oculus DK2 headset and a pair of Audio Technica MH50x headphones, participants were presented with an 80 second VR sample of the recording session which included the songs intro, verse and chorus using one of the microphone arrays appropriate for the samples viewing position. Once the clip was finished participants would answer a questionnaire that was spit into two main sections. The first asked them to rate on a scale of 1 - 10 each of the spatial audio attributes listed in table 2 where 1 indicated they did not experience that particular attribute well and 10 being that they experience that attribute very well. The second section asked them to select as many of the timbral

attributes they felt best described the overall timbre of the clip. Participants were also asked to rate on a scale of 1 - 10 how much they enjoyed the VR experience. This procedure was repeated a number of times depending on the number of microphone arrays to be presented (8 times for test 1 and 7 times for test 2). The order in which samples were presented was randomised per participant.

To ensure a uniform understanding of the list of attributes that were used in the questionnaire a short training exercise was conducted before each test. This involved taking the participants through each of the attributes with audio examples.

5 Analysis Overview

All test data was recorded using a Google Form and collated into a large document upon finishing the testing period. All data was then imported into MATLAB for analysis. Table 3 shows the average spatial attribute results for each microphone across each spatial attribute. The highest scores are highlighted in green. The directional OCT microphone array from microphone position and viewing position B scored highest for 'Locatedness' and scored the highest along with the Hamasaki Cube, a diffuse-field microphone array located at position C for 'Sense of Space'. The Hamasaki Cube when viewing from position A scored the highest for 'Externalisation' and 'Envelopment' which results in the highest over all spatial attribute average score. A break down of timbral attribute scores is covered in section 5.4

The following six sections will be used to further break down and analyse the data:

- Analysis 1: Does viewing position affect Spatial Attribute rating?
- Analysis 2: Does the choice of microphone array affect Spatial Attribute score?
- Analysis 3: What is the effect of using Directional or Diffuse-Field Arrays?
- Analysis 4: Does perception of timbre change with difference viewing positions?
- Analysis 5: Is there a correlation between spatial attribute score and selected timral attributes?
- Analysis 6: Is there a correlation between the enjoyment rating and spatial audio attributes?

This will be followed by an summary section with conclusions draw from each analysis section.

5.1 Analysis 1: Does viewing position affect spatial attribute scores?

To measure the effect that viewing from a different position within the room might have on the participants perception of the given spatial attributes, the test data was split into four groups containing the average results for each spatial attribute, each of which was split into two sub groups corresponding to the data obtained when viewing from position A or B. The data is illustrated in figure 11.

It can be seen that the average spatial attribute scores for viewing position A and B are close with the difference in overall mean score being 0.02. Running a Two-Sample T-Test between each of the four spatial attribute groups (e.g A vs B for locatedness etc) indicates no statistical significance. Running the same test for the averaged combined spatial attribute score (average of all four spatial attribute scores) also indicates no statistical significant between viewing position. This is made clear in figure 12 illustrating the overall similarity in the distribution of scores for viewing position A and B.

Conclusion

The bar chart indicates that the average spatial attribute scores are extremely close with an overall average for both being 7.1. The Two-Sample T-Test returned $p > 0.05$ for all data groups indicating that the probability of these results recurring is not unlikely and therefore the results shown are not statistically significant.

Viewing Position	Mic Position	Microphones	Locatedness	Sense of Space	Externalisation	Envelopment	Average
A	A	Spots	7.26	6.16	6.47	7.11	**6.75**
		Eigen	7.21	7.42	6.84	7.53	**7.25**
		ESMA	7.63	6.74	6.95	7.37	**7.17**
		ORTF	7.58	6.74	6.68	7.16	**7.04**
		ST450	7.53	6.74	7.00	7.32	**7.14**
	C	IRT	7.42	7.21	6.79	7.21	**7.16**
		Hamasaki	7.26	7.26	7.58	7.84	**7.49**
		ST456	6.84	7.26	6.89	7.42	**7.11**
B	B	Spots	7.08	6.16	6.66	6.87	**6.69**
		OCT	7.76	7.53	7.18	7.42	**7.47**
		PCMA	7.53	7.18	6.84	7.47	**7.26**
		Ambeo	7.53	6.95	6.47	7.45	**7.10**
	C	IRT Cross	7.45	6.32	6.55	7.16	**6.87**
		Hamasaki	7.34	7.53	7.03	7.34	**7.31**
		ST450	7.47	6.89	7.18	7.03	**7.14**

Table 3: Table containing the average spatial attribute scores for all microphone with on over all average spatial attribute score. Highest scoring microphones are highlighted in green.

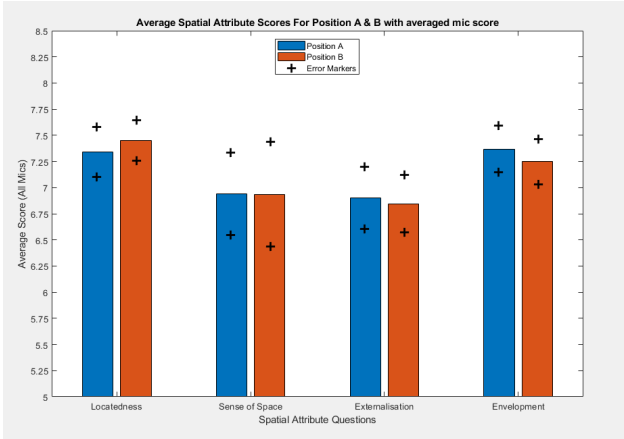

Fig. 11: Bar chart showing average spatial attribute score for each spatial attribute at viewing position A and B

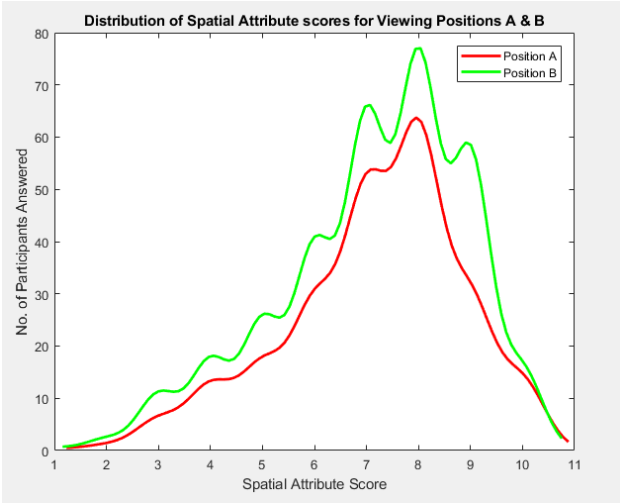

Fig. 12: Histogram showing the distribution of spatial attribute scores for viewing position A and B. This indicates that viewing position has little effect on spatial qualities in a VR environment.

5.2 Analysis 2: Does the choice of microphone array affect Spatial Attribute score?

Breaking down the data showing in figure 11, the average spatial attribute score across all used microphone configurations is shown in figure 13. The Anderson-Darling test was used to determine that not all of the sample data (participants scores per microphone) is normally distributed. Due to non-normally distributed data the Kurskal-Wallis (K-W) ANOVA was used to determine whether the average scores between any of the samples were significantly different. All groups returned $p > 0.05$ other than 'Sense of Space' which returned $p = 0.0227$. As determined in section 5.1 there is no statistically significant difference between the data due to different viewing positions A and B. Therefore the data was separated according to their viewing positions and another K-W test was conducted on each group. This indicated a significant difference within the group of data from viewing position B, returning $p = 0.0035$.

Using MATLABs *multcompare* function, a post-hoc test was conducted to determine that the sample data for two microphone arrays, OCT and Hamasaki Cube are significantly different to the sample data for the spot microphones, circled in red and blue respectively in figure 13.

Conclusion

Looking at the scores for the OCT and Hamasaki cube across all spatial attributes, it is apparent that they typically perform well among other microphone configurations. It is therefore probably more appropriate to look at this result as a significantly poor performance from the spot microphones whilst compared to microphone arrays that consistently perform well as opposed to an significantly exceptional performance from the two microphone arrays. Across all other spatial attributes it has been shown that the difference between performance is not statistically significant.

5.3 Analysis 3: What is the effect of using Directional or Diffuse-Field Arrays?

Section 5.2 revealed no significant difference between using any of the different microphone arrays apart from when it comes to a 'Sense of Space'. However the difference was only found between using the spot mic mix against mixing the spot mics with either the OCT array or the Hamasaki Cube. As both of these microphone arrays belong to different groups (OCT is used as a directional array and the Hamasaki Cube as a diffuse field array) and were not significantly different from each other, it can also be stated that the use of directional or diffuse field arrays is also not statistically significant.

Analysing the bar chart in figure 13 however it is possible to come to some conclusions about particular microphone configurations. For example, looking at the scores for Sense of Space, the three diffuse field microphone in position C whilst viewing from position A can be said to objectively perform worse than the three of the directional microphones at position A (ESMA, ORTF, ST450). However as the Eigenmke scores higher than all of them, drawing a conclusion that one microphone type is superior would be a incorrect. A more collated visualisation of overall microphone configuration performance can be seen in figure 14, highlighting the narrow lead of the OCT microphone configuration.

Conclusion There appears to be no significant effect of using either a directional or diffuse-field microphone array providing that spot microphones alone are not used

5.4 Analysis 4: Is there a difference in perception of timbre with difference viewing positions?

The effect of different viewing positions on participants perception of timbre can be assessed by comparing the data collected for microphones that were shared across both viewing positions which includes the spot microphones and microphone arrays from position C. Figure 15 shows the percentage of participants that selected each timbral attribute for each microphone. Table 4 presents the corresponding data showing the percentage difference between each microphone pair (each microphone at both positions) calculated by subtracting the results for microphones at position B from the same microphones at position A. For each timbral attribute, the average results column indicates whether the attribute was selected by more participants for viewing position A (positive number) or position B (negative number). By looking at the table it can be seen that across all microphones the timbral attributes 'Realistic' and 'Loud' share the trend that they were selected by an equal or greater percentage of participants for viewing position A. The timbral attribute 'Realistic' experienced the greatest variation between viewing positions with an average difference of 19%. Though the increased frequency of participants selecting 'Loud' is only by a minor amount, it is possible that the consistency is caused simply by the

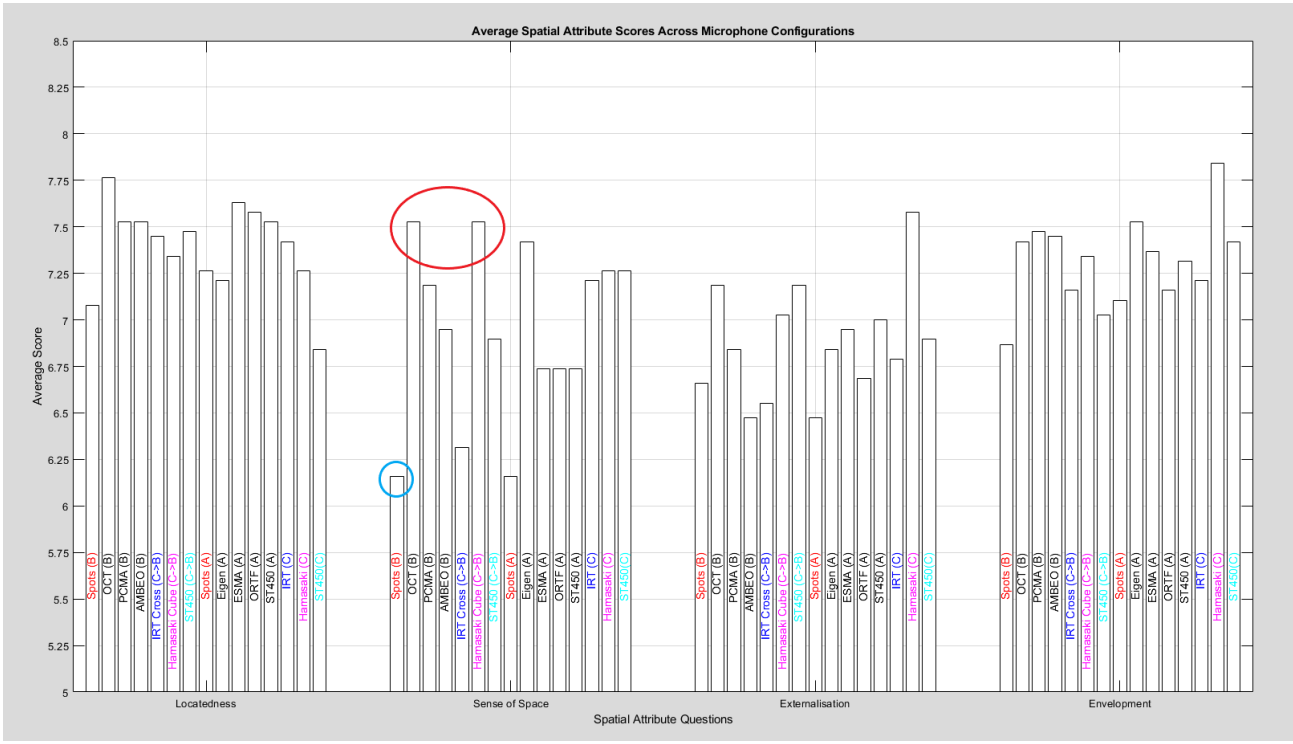

Fig. 13: Bar chart of the average spatial attribute score across all microphone configurations where *microphone(X)* indicates the microphone configuration and location. The microphone names are displayed on their corresponding bar where (C− >B) indicates a microphone from position C whilst viewing from B and (C) indicates a microphone from position C whilst viewing from position A. Microphones that were shared across both viewing positions are highlighted in matching colours.

fact that viewing position A is closer to the musicians thus increasing the possibility for participants to perceive the audio as 'Loud'.

Conclusion

In terms of timbre there appears to be little difference made by watching the performance from either position. The attributes effected most by changing viewing positions are a sense of whether things sounded realistic for which position A showed a preference. This increased difference however is mostly influenced by the large difference in spot microphone scores where there is a 39.5% increase from the low 39.5% of participants regarding the spot mics at position B to sound realistic to the 79% of participants regarding the spot mics at position A to sound realistic.

It is hypothesised by the author that this is due to the unnatural lack of room acoustics heard when positioned at a distance where the sound of the room is expected to be heard. As the participant is standing much closer to the

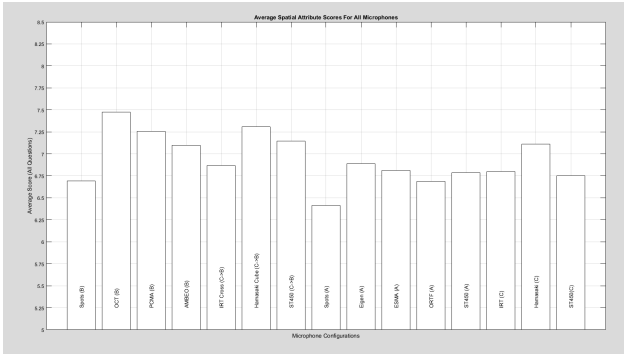

Fig. 14: Bar chart showing the average spatial attribute score across all microphones.

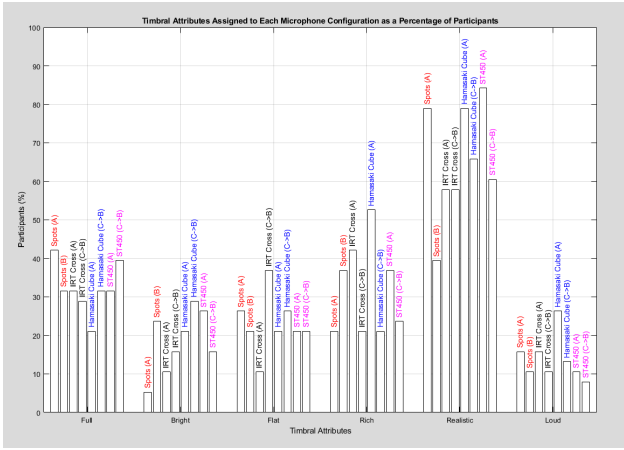

Fig. 15: Bar chart showing the timbral attributes chosen for each microphone array shared between viewing position A and B as a percentage of participants

Attributes	Microphones (%)				Averaged
	Spots	IRT Cross	Hamasaki Cube	ST450	
Full	10.53	2.63	-10.53	-7.89	-1.32
Bright	-18.42	-5.26	-7.89	10.53	-5.26
Flat	5.26	-26.32	-5.26	0	-6.58
Rich	-15.79	21.05	31.59	13.16	12.5
Realistic	39.47	0	13.16	23.68	19.08
Loud	5.2	5.2	13.16	2.63	6.58

Table 4: Table showing the percentage difference between each pair of microphones calculated by $A - B$

musicians at position A where the direct to reverb ratio is naturally much higher, the lack of room acoustics may be less noticeable and still be perceived as realistic. As a consequence of a lack of room acoustics present in the spot microphone mix, sound sources can be perceived more spatially isolated from one another. The unnaturalness of this is possibly less obvious when fully surrounded by the musicians as experienced at position A as this introduces an unnatural separation of the musicians for a listening experience. When viewing from position B however, as the musicians can be seen closer together, the unnatural sound source separation perceived may stand out to participants and further affect their judgement.

5.5 Analysis 5: Is there a correlation between spatial attribute score and selected timbral attributes?

Correlation coefficients were calculated for each combination of spatial attribute score against timbral attribute score. Most calculations returned weak correlations with $p > 0.05$ other than two timbral attributes specifically when comparing against the 'envelopment' spatial attribute. Figure 16 shows the line of best fit for all timbral attribute scores against the spatial attribute scores for 'envelopment'. The dashed lines indicate a statistically significant correlation as found with the timbral attributes 'Full' and 'Realistic'. The graph indicates that there is a significant positive correlation between the increase in sense of envelopment in the virtual environment with the sense of the virtual environment sounding realistic.

The data also indicates a negative correlation between participants sense of envelopment and their perception of the timbre sounding 'Full'. The reason as to why this is is not exactly clear. It could be said that participants perception of 'Full' may mean an unnatural abundance of bass frequencies which may sound unrealistic. If this is the case, as we can see from the positive correlation between a sense of envelopment and a sense of the environment sounding realistic, that an unrealistic sounding environment would lead to a decrease in participants sense of envelopment.

Conclusion

The only spatial attribute to show a significant correlation with timbral attribute data is 'Sense of Space' with a positive correlation for a perception of the mix sounding 'Realistic' and a negative correlation for the perception of the mix sounding 'Full'.

5.6 Analysis 6: Enjoyment Rating

The purpose of both VR and music is primarily to entertain. If users do not enjoy what they are experiencing then the cause should be addressed. For this purpose participants were asked to rate on a scale of 1-10 how much they felt they enjoyed the experience. Figure 17 shows the line of best fit for average enjoyment score against average spatial attribute score for each microphone for both viewing position A and B.

Both graphs show a strong positive correlation between the scores for spatial attributes and enjoyment ratings with a correlation coefficient $r = 0.81$ and $r = 0.96$ for viewing position A and B respectively.

Conclusion

This suggests that when provided with higher quality spatial audio for an experience such as the one presented, viewers are more likely to enjoy the experience. It is important to address that this VR experience differs from the currently most popular form of VR experience available which is gaming. VR games offer the possibility of movement and interaction which may distract users from the quality of spatial audio. However, in an experience such as the one presented in this paper, where the focus of the experience is listening to music and lateral movement is not an option, the users focus is not taken away from the spatial audio quality. It is therefore important for listening to music in VR that the spatial audio quality is high to ensure the listening experience is enjoyed.

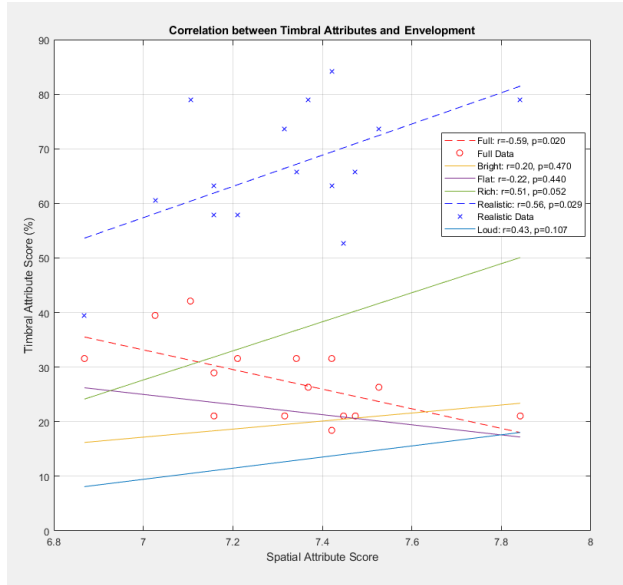

Fig. 16: Graph showing correlations between Timbral Attribute and Envelopment scores where dashed lines indicate a statistically significant correlation for which their individual data points have also been plotted.

6 Analysis Summary

The analysis of the data can be summarised as follows:

Analysis 1: Viewing the performance from position A (360° perspective) or position B (180° perspective) does not significantly influence participants judgement on the proposed spatial attributes.

Analysis 2: When compared to a spot microphone only mix, the OCT and Hamasaki Cube were found to be statistically significant with regards to the spatial attribute 'Sense of Space'. This can be seen as a significantly poor performance from the spot microphones at viewing position B. Though little statistical significance was found between the different microphone arrays, by analysing figure 13 in section 5.2 we can determine which microphones are overall the best choice with regards to this listening test. The two microphone arrays that showed statistical significance, the OCT and Hamasaki Cube also happen to consistently be among the top scoring microphone arrays across all spatial attributes.

Analysis 3: There appears to be no significant difference between using either a direct or diffuse-field microphone array so long as spot microphones alone are not used.

Analysis 4: Timbre itself seems to be unaffected by the viewing positions. However participants perception of the mix sounding 'realistic' and 'loud' appears to be the most affected with a higher score for viewing position A. It is hypothesised that this is due to the unnatural room acoustic and sound source separation caused when listening to the spot mic mix that skews the result this way.

Analysis 5: A statistically significant positive correlation was found between 'Sense of Space' and the perception of the mix being 'realistic'. A significant negative correlation was also found between 'Sense of Space' and the perception of the timbre sounding 'Full'.

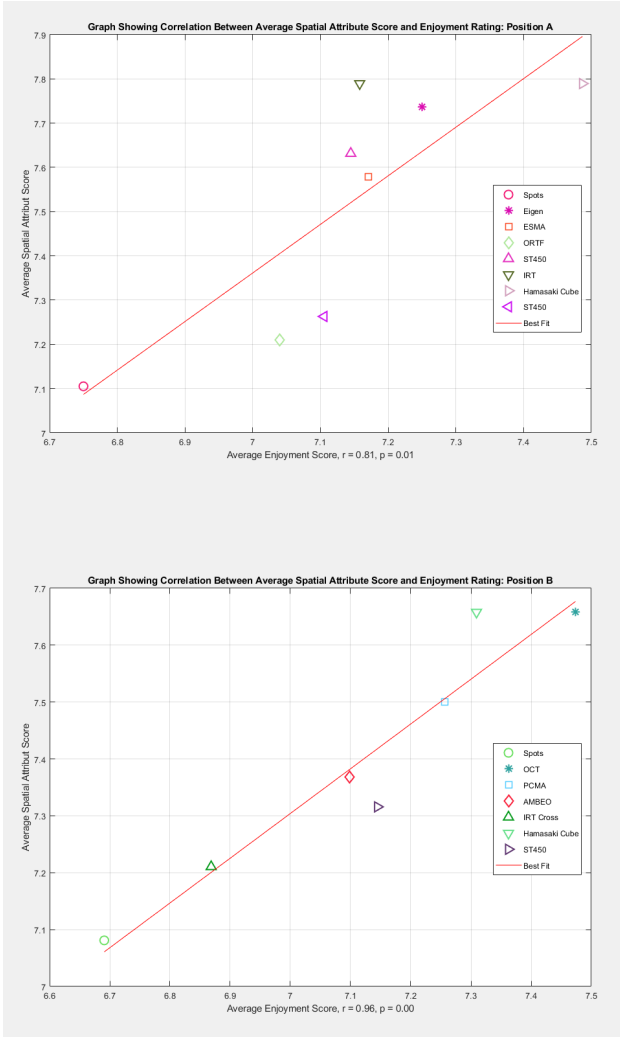

Fig. 17: Two graphs showing a positive correlation between enjoyment rating and spatial attribute scores for viewing positions A (Top) and B (Bottom).

7 REFERENCES

[1] H. Riaz, M. Stiles, C. Armstrong, A. Chadwick, H. Lee, G. Kearney, "Multichannel Microphone Array Recording for Popular Music Production in Virtual Reality," presented at the *Audio Engineering Society Convention 143* (2017 Oct), URL http://www.aes.org/e-lib/browse.cfm?elib=19333.

[2] abbeyroad.com, "Studio Three," Available at: https://www.abbeyroad.com/studio/studio-three, online; Accessed on 13/08/17.

[3] gopro.com, "GoPro Omnirig 360 Camera," Available at: https://shop.gopro.com/virtualreality/omni- online; Accessed on 20/08/17.

[4] M. Williams, "Multichannel Sound Recording Practice Using Microphone Arrays," presented at the *Audio Engineering Society Conference: 24th International Conference: Multichannel Audio, The New Reality* (2003 June).

[5] M. Williams, "Microphone Arrays for Natural Multiphony," presented at the *Audio Engineering Society Convention 91* (1991 Oct).

[6] Hyunkook Lee, "Capturing and Rendering 360 VR Audio using Cardioid Microphones," Available at: http://eprints.hud.ac.uk/29582/1/AES.com, online; Accessed on 9/06/17.

[7] Sennheiser.com, "Sennheiser AMBEO VR Microphone," Available at: http://en-uk.sennheiser.com/microp online; Accessed on 12/06/17.

[8] G. Theile, H. Wittek, "Principles in Surround Recordings with Height," (December 2011).

[9] G. Theile, "Multichannel Natural Music Recording Based on Psychoacoustic Principles," (IRT. München, Germany. October 2001).

[10] Schoeps Mikrofone, "IRT Cross," Available at: http://www.schoeps.de/en/products/irt-cross-set, online; Accessed on 12/06/17.

[11] Avid.com/ProToolsHD, "ProTools HD," Available at: http://www.avid.com/pro-tools-hd, online; Accessed on 20/08/17.

[12] Adobe.com, "Adobe Premiere Pro Video Editing Software," Available at: http://www.adobe.com/uk/products/premiere.html, online; Accessed on 21/08/17.

[13] samsung.com, "Samsung Action Developer," Available at: https://resources.samsungdevelopers.com, online; Accessed on 29/03/18.

[14] kolor.com, "GoPro VR Player," Available at: http://www.kolor.com/gopro-vr-player/, online; Accessed on 20/08/17.

[15] Oculus.com, "Oculus Development Kit 2 (DK2) Headset," Available at: https://www3.oculus.com/en-us/dk2/, online; Accessed on 21/08/17.

[16] reaper.fm, "Reaper Digital Audio Workstation," Available at: https://www.reaper.fm/, online; Accessed on 20/08/17.

[17] E. D. C. Nachbar, F. Zotter, A. Sontacchi, "Ambix - A suggested Format," presented at the *Ambisonics Symposium 2011* (2011 June).

[18] Openair.lib, "St Albans Cathedral B-Format Impulse Response," Available at: http://www.openairlib.net/auralizationdb/content/lady-chapel-st-alba online; Accessed on 22/08/17.

[19] M. Kronlachner, "Plug-in Suite for Mastering the Production and Playback in Surround Sound and Ambisonics," presented at the *Audio Engineering Society Student Design Competition: 136th AES Convention, Berlin* (2014 April),

[20] Matthiaskronlachner.com, "AmbiX v0.2.7 Ambisonic plug-in suite," Available at: http://www.matthiaskron online; Accessed on 22/08/17.

[21] mhacoustics.com, "em32 Eigenmike," Available at: https://mhacoustics.com/, online; Accessed on 12/06/17.

[22] M. Rumori, "Girafe - A Versatile Ambisonics and Binaural System," presented at the *Ambisonics Symposium 2009* (2009 May).

[23] York.ac.uk, "Google adopt SADIE filters for VR pipeline," Available at: https://www.york.ac.uk/sadie-pr

[24] YouTube.com, "YouTube Virtual Reality," Available at: https://www.youtube.com/channel/UCzuqhhs6 online; Accessed on 19/06/17.

[25] Voxengo.com, "r8brain Sample Rate Converter," Available at: https://www.voxengo.com/product/r8brai online; Accessed on 22/08/17.

[26] spook.fm, "SpookSyncVR," Available at: http://www.spook.fm/spooksyncvr/, online; Accessed on 20/08/17.

THE AUTHORS