Author Accepted Manuscript (originally presented at the 144th Audio Engineering Convention in Milan, May 2018)

An Investigation into Spatial Attributes of 360° Microphone Techniques for Virtual Reality

Connor Millns and Hyunkook Lee

Applied Psychoacoustics Lab, University of Huddersfield, West Yorkshire, United Kingdom
connor.millns@hud.ac.uk & h.lee@hud.ac.uk

ABSTRACT

Listening tests were conducted to evaluate perceived spatial attributes of two types of 360° microphone techniques for virtual reality (First Order Ambisonics (FOA) and the Equal Segment Microphone Array (ESMA)). Also a binaural dummy head was included as a baseline for VR audio. The four attributes tested were: source shift/ensemble spread, source/ensemble distance, environmental width and environmental depth. The stimuli used in these tests included single and multisource sounds consisting of both human voice and instruments. The results indicate that listeners can distinguish differences in three of the four spatial attributes. The binaural head was rated the highest for each attribute and FOA was rated the least except for in environmental depth.

1 Introduction

With the ever-growing rise in popularity of Virtual Reality (VR), it is important that content is of the highest quality. For capturing audio for VR, currently the most popular technique amongst practitioners seems to be First Order Ambisonics (FOA), as it can record three-dimensional (3D) audio with a compact, single-point microphone array. Once binauralised the soundfield can easily be rotated in synchronisation with 360° video. However, there are different techniques available for capturing 360° audio. For example, recently a near-coincident array called the Equal Segment Microphone Array (ESMA) with 50cm microphone spacing was proposed as a suitable technique for capturing 360° audio for VR [1]. Near-coincident microphone techniques are often preferred to coincident techniques (e.g. FOA) by professional recording engineers since they tend to provide a more spacious sound in recording.

At present, however, there is a lack of research on the performances of various VR microphone techniques in terms of perceived spatial attributes. Previous studies mainly focused on timbral quality and localisation performance of different Ambisonics microphones [2,3] or naturalness of the reproduced spatial environment [4,5].

From this background, this paper presents an experiment that aims to reveal perceptual difference in spatial attributes for two different types of 360° microphone techniques for VR: FOA and ESMA. Additionally, a binaural dummy head microphone was used as a baseline for binaural spatial quality. The main hypothesis of this experiment is that listeners can perceive spatial difference between the microphone techniques. More specifically, the binaural microphone would produce the most optimal quality as it produces the most accurate interaural and pinnae cues for localisation, whereas the FOA microphone would have reduced spatial fidelity compared to the ESMA due to its coincident microphone arrangement. If any significant perceptual difference is found among the microphone techniques, there may also be some correlation between attributes. Ultimately this research aims to provide recording engineers with more options over the choice of microphone technique depending on the spatial attributes they wish to highlight.

2 Microphone Techniques

An overview of the three microphone techniques used in this experiment will be given in this section.

2.1 Equal Segment Microphone Array (ESMA)

The ESMA is based upon a microphone array design concept called ‘critical linking’, which was proposed by Williams and Le Dû [6]. The method calls for the stereophonic recording angle (SRA) of each pair of microphones in a given array to add up to 360° without any overlap. SRA is the region of the soundfield captured in front of a stereo microphone array that will be reproduced fully wide between two loudspeakers; sound sources recorded at boundaries of the SRA will be localised fully left or right.

Williams [7] introduced the ‘Equal Segment Microphone Array’ for multichannel recording based on the ‘critical linking’ concept with the addition that each pair of adjacent microphones of the array has the same SRA. A near-coincident, four cardioid microphone version of the ESMA, with various microphone spacings, was tested by Lee

[1] for localisation accuracy. The conventional ESMA of the same type by Williams uses 25cm spacing between microphones. However, Lee proposed that 50cm was the appropriate spacing for achieving the SRA of 90° for each stereophonic segment within a quadraphonic loudspeaker reproduction, based on his interchannel time and level difference trade-off model. Listening tests confirmed that the 50cm spacing produced more accurate localisation than 25cm and 0cm.

2.2 First Order Ambisonics (FOA)

FOA is the simplest form of Ambisonics. An FOA microphone system consists of four subcardioid microphone capsules that are arranged in a tetrahedron. The raw signals from the microphone system are referred to as A-format. The raw signals should be converted into spherical harmonics of W, X, Y and Z, which are called B-format signals and should be decoded for playback over a loudspeaker array. The first commercially available FOA microphone system was developed by SoundField. Presently, there are a number of FOA systems available, such as: the Sennheiser Ambeo, the Core Sound TetraMic, the Twirling720, etc.

2.3 Binaural Dummy Head

Binaural dummy head microphones are modelled on actual human heads in order to replicate the outer ear. The main cues for human auditory localisation are: Interaural Time Difference (ITD) for low frequencies, Interaural Level Difference (ILD) for high frequencies and pinnae filtering. The dummy head was included for comparison in this study as it was considered to provide a baseline quality for 360° audio for VR.

3 Experiments

3.1 Recording

Room impulse responses (RIRs) were first captured in the University of Huddersfield's St. Pauls' concert hall (RT = 2.1s) (Figure 1.). RIRs were taken instead of recording live source material and they were convolved with various sound sources afterwards. This allowed different mic arrays to be placed at the same position and the identical performance could be captured for all of the arrays, thus improving the repeatability. The RIRs for the microphone arrays used are freely available from the MAIR (Microphone Array Impulse Responses) library[1] [8].

The exponential sine sweep method [9] was used to capture the RIRs. The software used for this task was the Applied Psychoacoustic Lab's (APL) HAART (Huddersfield Acoustical Analysis and Research Toolbox) [10]. Four cardioid Neumann KM184 D microphones were used for the ESMA. FOA was captured using the Sennheiser Ambeo microphone and the dummy head used was the Neumann KU100.

Figure 1. St. Paul's concert hall, the recording location.

[1] https://github.com/APL-Huddersfield/MAIR-Library-and-Renderer

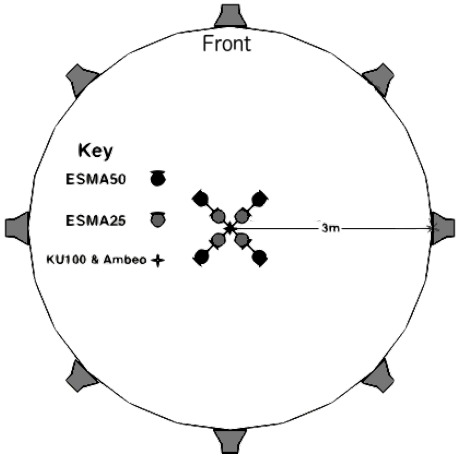

Figure 2. Loudspeaker and microphone array setup for capturing 360° room impulse responses.

3.2 Stimuli Creation

In order to test the influence of sound source type on perceived spatial attributes, a single speech and two different ensemble musical sources were chosen, as presented in Table 1. Two sets of dry multitrack recordings were taken from Cambridge Music Technology[2]; one set was from a barbershop quartet called the Rounders and the other set was from a folk jazz band called Flèche d'or performing a song called 'Swing Bazar'. The Rounders consisted of four male voices singing in close harmony. Flèche d'or consisted of an acoustic guitarist, electric guitarist, an accordion player, double bassist and a violinist. No processing was applied to the multitrack signals, apart from slight level balancing to make each source equally loud. A female Romanian speech sample was recorded in a small recording studio with dry acoustics at the University of Huddersfield using a Neumann U87 microphone. The recording was made to be as dry as possible by placing the microphone close to the performer with acoustic baffles around her.

Each of the sources had different positional arrangements to assess whether the spatial performance holds up with different source azimuth angles. Table 1 presents the list of source conditions used for the tests. 0° is towards the stage in the concert hall, and positive azimuth angle increases are clockwise.

Table 1. A description of the stimuli used in the four listening tests

Source No.	Source Type	Source Position
1	Single Speech	0°
2	Single Speech	45°
3	Single Speech	90°
4	Barbershop Quartet	-45°, 45°, 135°, -135°
5	Barbershop Quartet	0°, 90°, 180°, -90°
6	Jazz Instrumental	-45°, 0°, 45°, 135°, -135°

The FOA RIRs were recorded in A-format, converted to B-format then decoded into a loudspeaker format. In this experiment a quadraphonic loudspeaker array (speakers at -45°, 45°, 135° and -135° azimuths) was used. The Ambix VST plugin decoded the RIRs into the loudspeaker format. The quad Max rE [11] and Mode Matching (MM) [12] decode configurations from the SADIE database[3] were used for the decoding of the B-format signals.

[2] http://www.cambridge-mt.com/ms-mtk.htm

Convolving the RIRs from the dummy head with the sound sources produced a 2-channel binaural output. The FOA and ESMA RIRs convolved with sound sources produced 4-channel stimuli for the quadraphonic setup, these were then binauralised. The Head Related Impulse Responses (HRIRs) used for the binauralisation were the diffuse field compensated KU100 HRIRs from the SADIE database [3]. Table 2 shows the list of microphone array conditions. In total, 30 stimuli were created (5 microphone array conditions x 6 source conditions).

Table 2. Microphone techniques used for the recording.

Microphone Technique	**Conditions**
FOA	Max rE
	Mode Matching (MM)
ESMA	25cm spacing
	50cm spacing
Dummy head	Facing 0°

3.3 Listening Test Design

Spatial attributes tested in the current experiment were chosen based on the ones described by Rumsey [13]. Two attributes were related to the sound source while the other two attributes were related to the environment. The attributes and their definition are presented in table 3. These attributes were selected because they can highlight fundamental spatial differences between the techniques. Furthermore, low-level attributes form the basis for higher level spatial attributes, e.g. environmental width and depth can be considered components of listener envelopment [13]. The source shift attribute was only tested for the single source conditions (source no.1-3 from table 1) and ensemble spread was tested only for multisource conditions (source no.4-6 from table 1).

Table 3. Attributes under test and their definitions.

Attribute	**Definition**
Source Shift/Ensemble Spread	Amount of perceived shift/spread of the sound image.
Source/Ensemble Distance	Amount of perceived distance between the listener and the sound source
Environmental Width	Perceived horizontal width of the space in which the sounds are located.
Environmental Depth	Perceived front/back depth of the space in which the sounds are located.

Each of the four attributes were tested in a separate session. One session would include six trials (source conditions) with five conditions (microphone technique). Subjects rated the conditions on a continuous scale with only two labels at the extremes of the scale, e.g. nearer (bottom of the scale) and further (top of the scale) for source/ensemble distance. An adapted version of the universal listening test interface generator called 'HULTI-GEN'[4] [14] was used as the graphical user interface for the test (Figure 3.).

[4] https://www.york.ac.uk/sadie-project

[4] www.hud.ac.uk/apl/resources

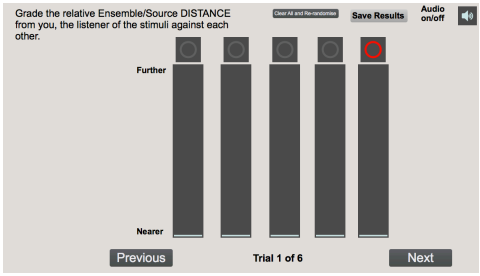

Figure 3. An example of HULTI-GEN graphical user interface used for the listening tests.

Test, trial and condition order were randomised for each subject to reduce the chance of order bias occurring. Subjects were briefed on the definition of the spatial attribute before each test. Prior to the main test, subjects were asked to listen to all stimuli for a familiarisation purpose.

The listening tests took place at an ITU-R BS.1116 [15]-compliant critical listening room at the APL of the University of Huddersfield. Sennheiser HD650 headphones were used for the test and they were amplified through an Apogee Groove digital-to-analogue converter.

3.4 Subjects

11 male and female subjects took part in the listening tests. These subjects were staff and postgraduate researchers and undergraduate students from the University of Huddersfield's music technology courses. The staff and postgraduate researchers and two undergraduate students had extensive amount of listening test experience in spatial audio evaluation. Five undergraduate students did not have previous listening test experience, although they all had several years of training in sound recording and critical listening. All listeners reported to have normal hearing.

4 Results and Discussion

Firstly the Shapiro-Wilks' test of normality was performed on the data obtained from the listening tests. The results showed that the data was not normally distributed ($p < 0.05$). Therefore, the non-parametric Friedman test was used to examine the main effect of microphone technique for each test condition. If significant effect was found, then the Wilcoxon signed-rank test was used for pairwise comparisons of the microphone techniques. The results of the Wilcoxon test were Bonferroni corrected.

Results are plotted using the median and notch edges (Figure 4-7). If the notch edges are not overlapping, then it is high likely that there is significant difference between the two conditions [16].

4.1 Source Shift/Ensemble Spread

The results for Source Shift/Ensemble Spread for each source condition are plotted in Figure 4. The Friedman test suggests that there is a significant difference between the microphone techniques for source shift/ensemble spread ($p < 0.05$), for all source conditions except no. 1 and 5. It can be seen that the dummy head was consistently rated to have the most shift/spread. For both single speech trials (source no. 2 and no.3) ESMA50 was rated to have greater source shift than the two FOA decodings. Additionally, there was a significant difference between ESMA50 and the Max rE for the source no. 2.

Significant difference in source shift could be interpreted as potential localisation inaccuracy for the microphone techniques. Since the dummy head is the baseline in this experiment, it can be said that any significant deviation could be considered inaccurate. ESMA50 exhibits no significant difference with the dummy head. However, the perceived shift for the Max rE FOA was significantly narrower from the dummy head and ESMA50, suggesting that the technique could suffer from localisation inaccuracy.

Significant source conditions for ensemble spread (ensemble source no.4 and no.6) seem to correlate with the significant source conditions for source shift (speech source no.2 and no.3). Microphone techniques with the most amount of source shift also displayed the most amount of ensemble spread.

4.2 Source/Ensemble Distance

The results for source/ensemble distance for each source condition are shown in figure 5. For the single source conditions (no.2 and no.3) the dummy head was rated to be the most distant, followed by both ESMAs. The FOA decodings were rated to be the closest sounding. The dummy head was significantly more distant than FOA and so too was ESMA25 significantly more distant than the Max rE from source condition no.2. It appears that ESMA50 was also significantly more distant than FOA, although the Wilcoxon test does not suggest so (possibly due to the Bonferroni correction being too conservative). Furthermore the spread of data for ESMA50 was quite large and there were also outliers present. The same problem affects MM for source condition no.3, as there was only significant difference between the dummy head and FOA, not between ESMA and FOA (Figure 5). Interestingly, for ensemble distance the dummy head was rated the closest for the source condition no.6. This contrasts with the results for the single speech conditions. Both ESMA and FOA were consistent across stimuli type.

The results show a similar trend that occurs in the source shift/ensemble spread test for source condition no.2 and no.3, where the dummy head had the highest rating, followed by ESMA and FOA the lowest. The coincident design of the FOA microphone may have affected the rating. Coincident microphone arrays exhibit higher amounts of channel coherence than near-coincident or even spaced microphone arrays. This high channel coherence produces a high Interaural Cross Correlation Coefficient (IACC) value of the resulting ear signals, which typically produce nearer sound images [17].

Results suggest that distance perception for the dummy head is affected by source azimuth angle. As the sound source moves away from the centre, an increase in distance perception can be seen for the single source conditions (no.1-3). Figure 5 shows source no.1 (azimuth 0°) was rated the nearest and source no.3 (azimuth 90°) the furthest. This trend continues with ensemble conditions, where no.4, with two central sound sources (0° and 180°), was rated nearer than no.5, with a quadraphonic arrangement. Source condition no.6 was the nearest rating for the dummy head, this could be due to the condition lacking 90° sound sources and instead including a 0° sound source.

4.3Environmental Width

Significant difference was found for four source conditions (no.2, no.3, no.4 and no.6 from Figure 6.). Pairwise analysis of the microphone conditions only revealed significant difference between the dummy head and the Max rE FOA. However, a general trend can be seen in Figure 6 where the dummy head was rated the widest, ESMA in middle and FOA the narrowest, with the exception of source condition no.3.

The results in Figure 6 uphold the trend seen in the previous two tests, that dummy was rated highest and FOA the lowest. This again could be due to the potentially high channel coherence of the FOA microphone producing a high IACC value. Hidaka et al. [18] found that a high IACC value suggests a narrow sound image.

4.4 Environmental Depth

No source conditions in the environmental depth test exhibited any significant difference between the microphone techniques. Subjects ratings were inconsistent across the source conditions and no trend can be discerned (Figure 7). The reason for such inconsistency could be because environmental depth is not a familiar attribute to the subjects and perhaps also hard to hear. Rumsey [13] proposed this attribute in his scene-based paradigm, but states that the attribute is not based on any elicitation test and that environmental width tends to dominate it perceptually.

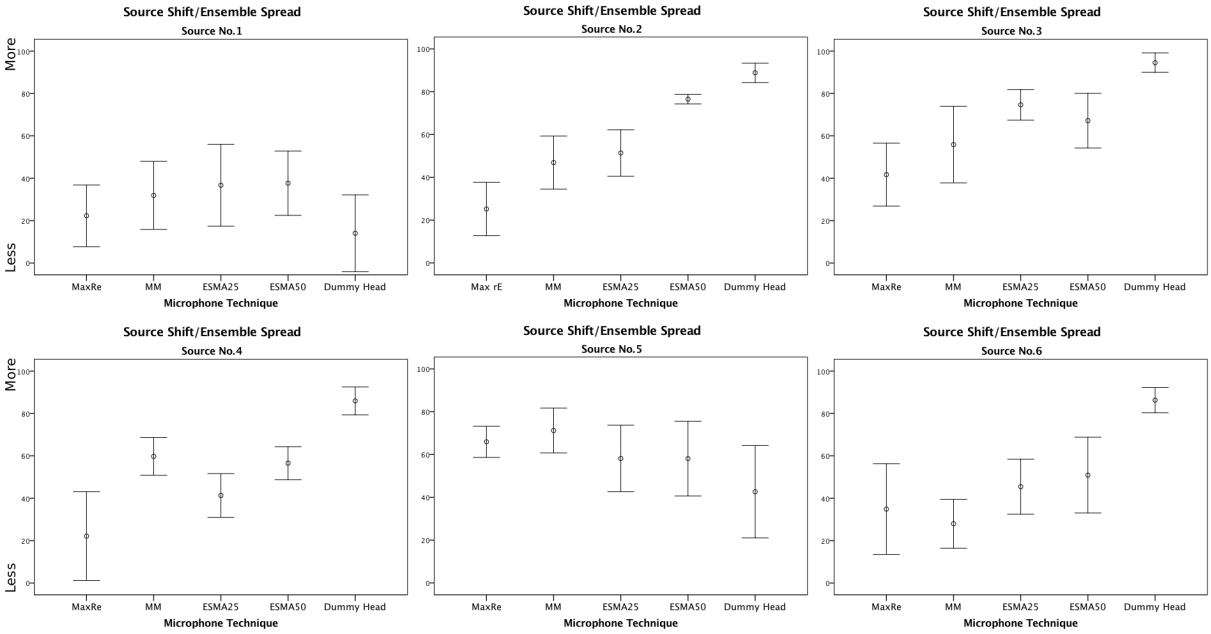

Figure 4. Median and notch edge plots for the source shift/ensemble spread test.

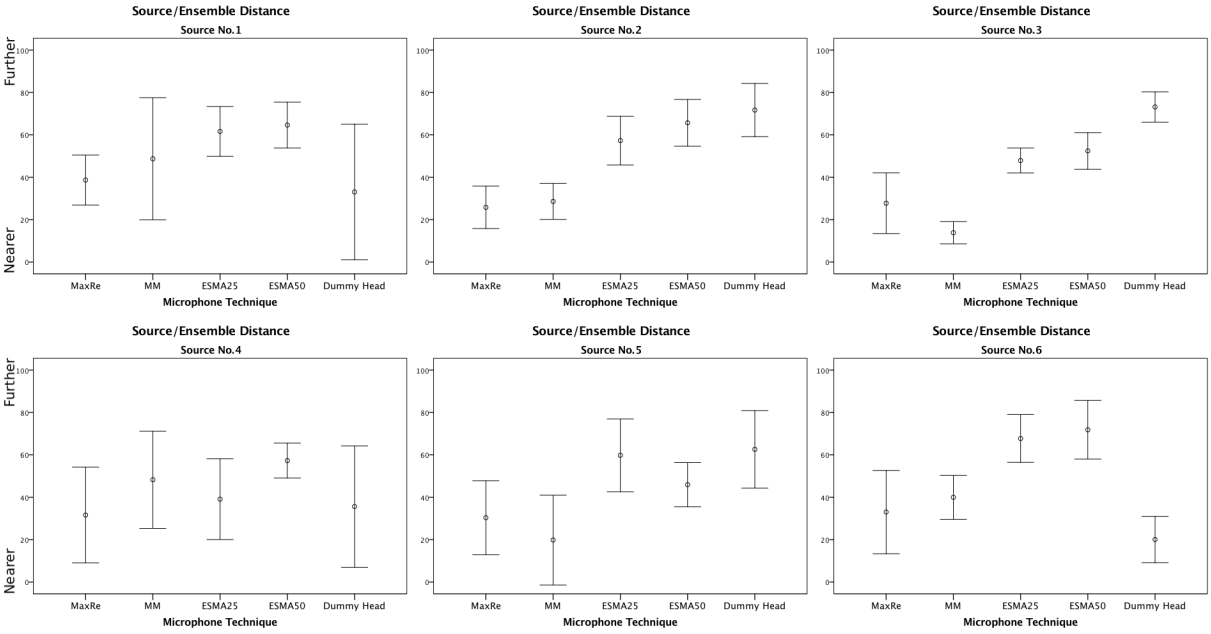

Figure 5. Median and notch edge plots for the source/ensemble distance test.

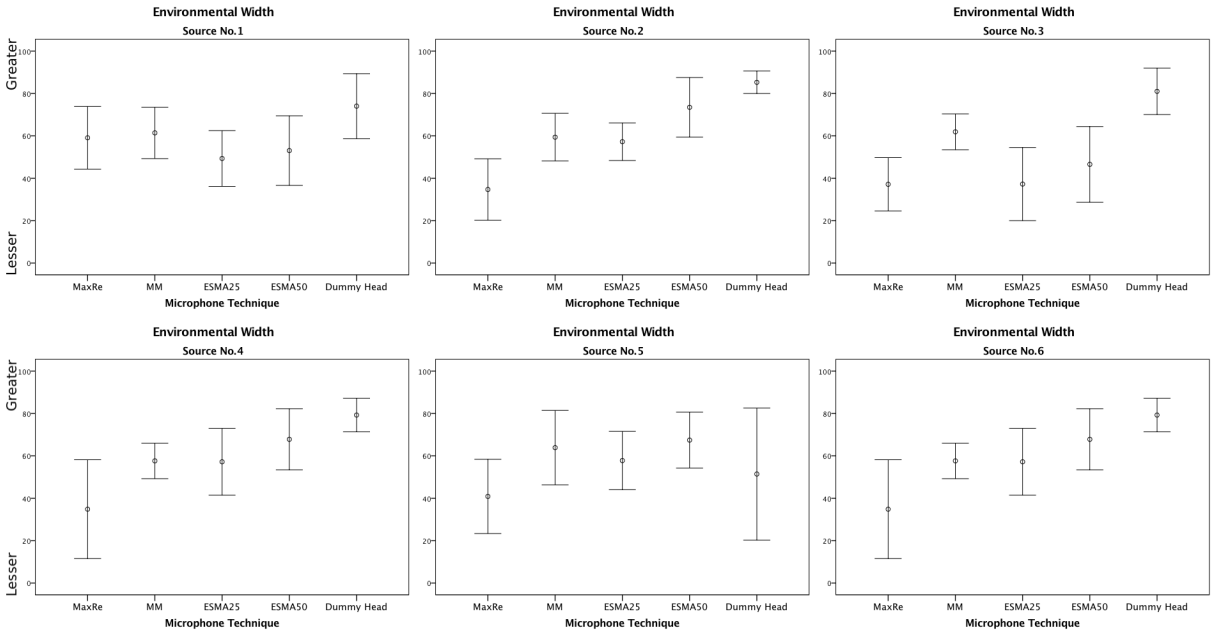

Figure 6. Median and notch edge plots for the environmental width test.

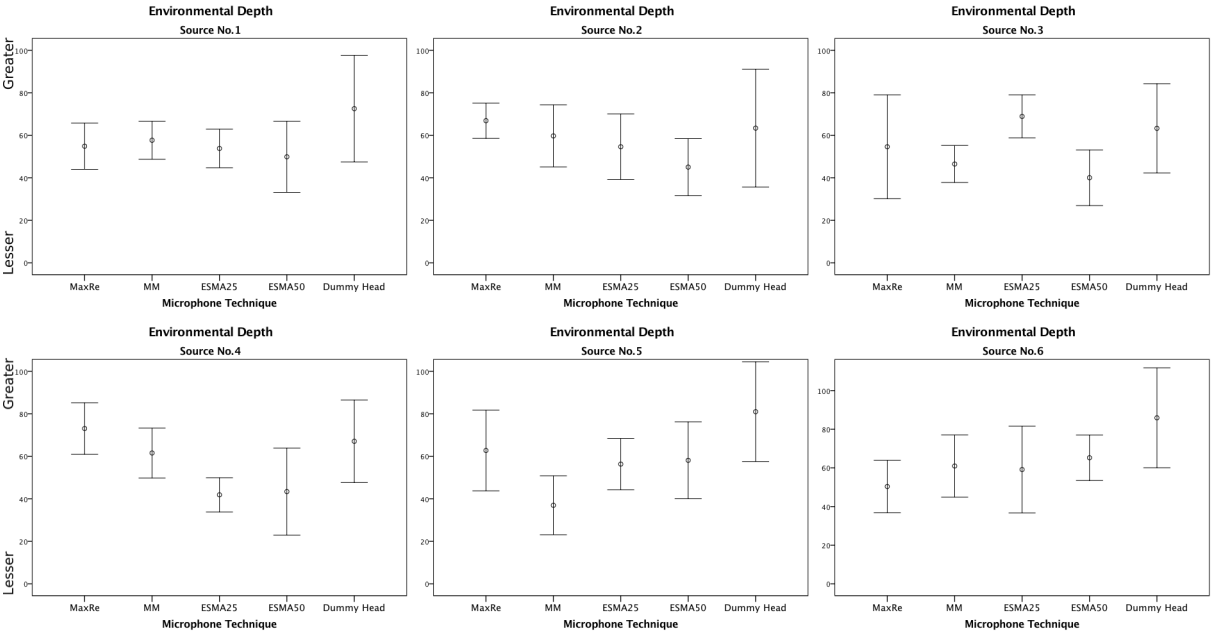

Figure 7. Median and notch edge plots for the environmental depth test.

5 Conclusion

This paper described a series of listening tests that were conducted to examine perceptual differences between two VR optimised microphone techniques for four spatial attributes. Significant differences were found for source shift/ensemble spread, source/ensemble distance and environmental width. There was no significant difference found in the environmental depth test. A general trend can be seen where the dummy was rated the highest, then ESMA in the middle and FOA being rated the lowest for the tests that showed a significant effect.

The present study indicates that spatial attributes for VR audio can be controlled with the choice of microphone technique. A future study will investigate how these spatial differences contribute to the perception of more global attributes, such as immersion and listener preference.

References

[1] H. Lee, "Capturing and Rendering 360° VR Audio using Cardioid Microphones," *Presented at the Conference on Audio for Virtual and Augmented Reality* (2016).

[2] E. Bates, S. Dooney, M. Gorzel, H. O'Dwyer, L. Ferguson, and F. M. Boland, "Comparing Ambisonic Microphones – Part 2," *Presented at the 142nd Convention of the Audio Engineering Society (2017),* convention paper 9730.

[3] C. Guastavino, V. Larcher, G. Gatusseau, and P. Bossard. "Spatial audio quality evaluation: Comparing transaural, ambisonics and stereo," *Proceedings of the 13th Inter- national Conference on Auditory Display,* 2007.

[4] K. Ooi, J. Y. Hong; B. Lam; Z. T. Ong; W. S. Gan, "Validation of binaural recordings with head tracking for use in soundscape evaluation," *presented at the Inter-Noise and Noise-Con congress,* 2017

[5] G. Zalles et al., "A Low-Cost, High-Quality MEMS Ambisonic Microphone', *presented at the 107th Convention of the Audio Engineering Society* (1999), convention paper 9857.

[6] M. Williams and G. Le Du, "Microphone array analysis for multichannel sound recording," *presented at the 107th Convention of the Audio Engineering Society* (1999), convention paper 4997

[7] M. Williams, "Migration of 5.0 multichannel microphone array design to higher order MMAD (6.0, 7.0&8.0) with or without the inter-format compatibility criteria," *presented at the 124th Convention of the Audio Engineering Society* (2008), convention paper 7480

[8] H. Lee and C. Millns, "Microphone Array Impulse Response (MAIR) Library for Spatial Audio Research," *Presented at the 143rd Convention of the Audio Engineering Society* (2017)*,* convention e-Brief 356.

[9] A. Farina, "Simultaneous measurement of impulse response and distortion with a swept-sine technique," *presented at the 108th Convention of the Audio Engineering Society,* (2000), Preprint 5093.

[10] D. Johnson, A. Harker and H. Lee, "HAART: A new impulse response toolbox for spatial audio research," *presented at the 138th Convention of the Audio Engineering Society* (2015), engineering brief 190.M. A. Gerzon, "Practical Periphony: The Re- production of Full-Sphere Sound," *Presented at the 65th Convention of Audio Engineering Society* (1980).

[11] A. Heller, R. Lee and E. Benjamin, "Is My Decoder Ambisonics?," *Presented at the 125th Convention of Audio Engineering Society* (2008).

[12] F. Rumsey, "Spatial Quality Evaluation for Reproduced Sound: Terminology, Meaning, and a Scene-Based Paradigm," *J. Audio Eng. Soc.*, vol. 50, No. 9, pp. 651–666 (2002 September).

[13] C. Gribben and H. Lee, "Toward the Development of a Universal Listening Test Interface Generator in Max," *Presented at the 138nd Convention of the Audio Engineering Society* (2015)*,* convention e-Brief 187.

[14] International Telecommunications Union. (2014). Recommendations ITU-RBS.1116-2: Methods for the Subjective Assessment of Small Impairments in Audio Systems including Multichannel Sound Systems.

[15] R. McGill, J. W. Tukey, and W. A. Larsen. "Variations of box plots," *The American Statistician*, Vol.32, No.1, pp. 12-16 (1978).

[16] S. Sakai, N. Asahi, T. Gotoh, F. Yoshii and H. Akiya, "On the Simplified Measurement of Interaural Crosscorrelation Coefficient and Its Effects on Sound Imaging and Quality," *Presented at the 67th Convention of Audio Engineering Society* (1980).

[17] T. Hidaka, L. L. Beranek, and T. Okano "Interaural cross-correlation, lateral fraction, and low- and high-frequency sound levels as measures of acoustical quality in concert halls," *J. Acoust. Soc. Am.,* vol. 96, no. 2, pp. 988-996 (1995 August).