A SUBJECTIVE COMPARISON OF VIRTUAL STEREO MICROPHONE TECHNIQUES FOR RENDERING AMBISONICS

A Girijavallabhan University of Derby, UK
B Wiggins University of Derby, UK

1 INTRODUCTION

Ambisonics separates recording (encoding) and reproduction (decoding) by capturing the complete sound field around a point in space for reconstruction over headphones and loudspeaker arrays[1]. The increased spatial recording resolution of higher order microphones, however, has not been complemented, at present, with more flexible options for stereo reproduction which would be a useful addition given that 'standard' stereo is still the prevalent form of listening over both loudspeakers and headphones Currently, binaural decoding is most common for rendering ambisonics over headphones which comes with challenges - localisation blur, individualisation. There are literature examples[2, 3] that exclusively mention binauralisation as 'the' decoding strategy for headphones. This has unfortunately contributed to non-binaural stereo being a forgotten approach despite ambisonic microphones historically being dubbed as 'set and forget' stereo microphones, given their ability to be rendered using virtual microphone arrays extracted after the recording has taken place. Prevalent stereo recording arrays have historically been compared based on their spacing, polar patterns and orientation producing divided subjective opinions for each. This paper documents the generation of near-coincident and spaced arrays for stereo decoding over headphones and loudspeakers and compares them to the standard coincident decodes to ascertain whether they are a useful or preferred addition to the available tools for Ambisonic stereo reproduction. The filters designed in this paper, along with example Reaper projects, can be downloaded from *www.BruceWiggins.co.uk*

2 STEREOPHONIC MICROPHONE TECHNIQUES

A stereo recording is a straightforward way of creating a 'snapshot' of sound as heard in the audience, including the effects of the space[4]. This is often a compromise between imaging accuracy and spaciousness through microphone positioning, where neither is capable of solely creating the desirable illusion of natural spatial acoustics[5]. Two-channel techniques are the most basic stereo recording techniques, where the two independent microphone outputs are mapped directly to the left and the right channels. Two-channel recording techniques can be divided into three categories based on spacing.

2.1 Spaced techniques

The oldest documented recording techniques are based on time-of-arrival stereophony, where microphones are positioned symmetrically along a line perpendicular to and divided by the midline of the sound source[6]. These are reduced derivations of Bell Labs' 'wall of sound,' with the most popular being a spaced pair, AB. Despite the originally intended perfect waveform reconstruction not being possible with two microphones, the human hearing compounds the two sounds of a source originating from the two channels at different times into a single phantom source with spacing between the two microphones affecting the maximum Inter-Channel Time Difference (ICTD). The two microphones not being equidistant from sound sources (unless the source is on the mid-line between them) also results in Inter-Channel Level Differences (ICLD).

2.2 Coincident techniques

Coincident microphone techniques are those where the two microphones are along the median plane such that their capsules are as close as physically possible. Blumlein is the most notable of the coincident technique family, maintaining constant signal power across the recording angle given its 90° angular relationship between the pickup pattern lobes[Error! Reference source not found.]. This results in sound sources located in front of the array being picked up with uniform acoustical power. XY is another prominent technique featuring crossed cardioids at 90°and is attributed to reproducing a spatially compressed version of the original sound sources into half the angle subtended by the loudspeakers[8].

2.3 Near Coincident techniques

Dooley[9] used the term near-coincident to refer to a pair of microphones that are placed close enough to act virtually coincident at low frequencies yet spaced enough to capture significant time differences between them for sources positioned at the extreme. The most notable near-coincident stereo recording technique is ORTF, approximating the spacing and directionality of the human ears which results in the maximum possible Inter-Channel Time Delays (ICTDs) being comparable to that experienced naturally by a listener.

3 COINCIDENT DECODING OF TOA

An ambisonics microphone can be used as a versatile stereo microphone where the FOA channels (except Z) can be combined to provide a basic steerable free-field normalised virtual microphone, controlling the direction and the (first order) polar pattern and angling of the virtual microphone[1]. A crossed pair of figure-of-eight microphones (Blumlein) is the easiest virtual array that can be derived using the sum and differences of the X and Y channels.

$$Left = \frac{\sqrt{2}}{2}(X + Y) \ \text{Right} = \frac{\sqrt{2}}{2}(X - Y) \tag{1}$$

Where X is a Front/Back figure of 8 and Y is a Left/Right figure of 8

A crossed pair of cardioids can similarly be derived using the omnidirectional W channel in combination with X and Y[5].

$$Left = \text{W} + \frac{(X + Y)}{2} \ \ Right = \text{W} + \frac{(X - Y)}{2} \tag{2}$$

Where W is an omni (3dB down from X and Y), X is a Front/Back figure of 8 and Y is a Left/Right figure of 8 (if deriving from AmbiX B-Format, W would need reducing by 3dB).

O3A Virtual Microphone by Blue Ripple Sound[10] is currently the most comprehensive plugin for coincident microphone array decoding, featuring two virtual microphones whose patterns and angling can be altered arbitrarily, to emulate X-Y, Blumlein and mid-side arrays. UHJ (or C-format) is a consumer format for ambisonics delivery, devised by Gerzon for compatibility with media transmission formats primarily based on one and two-channel delivery. The resulting decode is also considered to resemble a ‘coincident’ virtual array.

4 NEAR COINCIDENT AND SPACED DECODING OF TOA

Near coincident and spaced microphone patterns can be obtained from an Ambisonic recording if, instead of simple sum and difference extraction of virtual polar patterns (as in coincident decoding above), a process similar to that originally proposed by McKeag and McGrath[11] is employed (i.e. Ambisonics to Binaural). However, instead of measuring the response of the spherical harmonics sampled using HRTFs, impulse responses for 1st-order microphone patterns at various locations on the XY plane are synthesised for the required number of directions necessary for correct Ambisonic encoding and reconstruction using 3rd-order harmonics (as the 1st-order patterns now also

incorporate directionally dependent time delays, higher order harmonics improve the accuracy of the decoded audio with respect to the spatial aliasing frequency and area of correct reproduction).

In this paper, only horizontal reconstruction is attempted. The number of positions needed to be sampled around the mic array is dictated by the order of Ambisonics required (N). The number of points, equally spaced, around the array, needs to be one more than the number of channels used to encode that Ambisonic order (assuming circular harmonics/horizontal only encoding/decoding as shown in equation (3)).

$$numsamples = 2(N + 1) \tag{3}$$

Where N is the Ambisonic order

Once the required number of points has been determined along with the position of the two microphones, their polar patterns, and which direction the microphones are facing, the impulse response from each sampled point to the two microphones can be calculated. This is shown for an example ORTF mic array below in Figure 1.

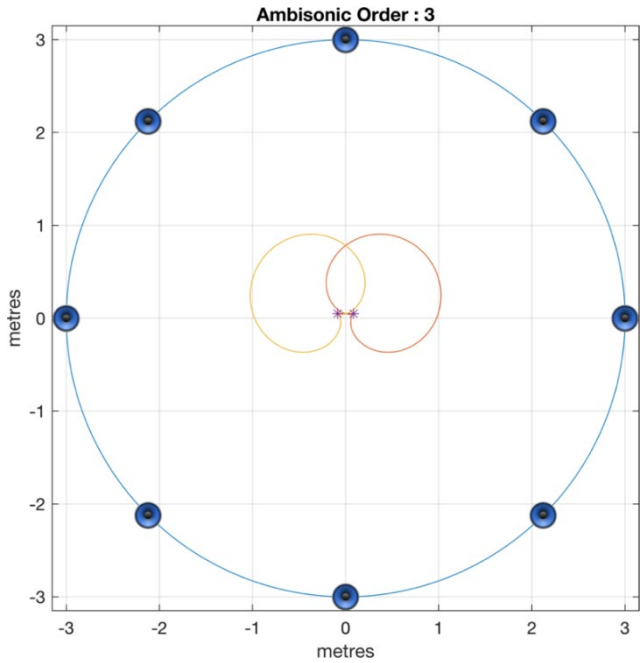

Figure 1 - Simulation geometry and mic patterns/positions for an ORTF microphone array sampled using 3rd-order spherical harmonics

The impulse response from each loudspeaker to each microphone first calculates the delay times from source to receiver with loudspeakers placed at a radius (3m) from the origin. Knowing the loudspeaker distance, the angle from the origin and the coordinates of the microphones, along with the speed of sound, the time delay and the angle from each microphone and loudspeaker can be calculated. The angle is used to determine the relative level, and the time delay for each impulse. This is then converted to a sample delay time by multiplying by the sampling rate (48kHz). As this will be a fractional value, a sinc pulse is used to represent the impulse response at the correct delay and amplitude.

$$dist_m^q = \sqrt{\left(Sp_x^q - Mic_x^m\right)^2 + \left(Sp_y^q - Mic_y^m\right)^2}$$

$$ang_m^q = \tan^{-1}\left(\frac{\left(Sp_y^q - Mic_y^m\right)^2}{\left(Sp_x^q - Mic_x^m\right)^2}\right)$$

$$amp_m^q = \frac{SpDist_m^q}{dist_m^q} \times \left((1 - MP_m) + MP_m \cos\left(\mathrm{ang_m^q} - MA_m\right)\right)$$

$$delay_m^q = \frac{dist_m^q}{c} \times fs \tag{4}$$

Where: q is the speaker index (1 to 8), m is the microphone index (1 or 2), dist is the distance from the speaker (q) to the microphone (m), ang is the angle from the microphone (m) to the speaker (q), amp is the relative level of the speaker (q) to the microphone (m) (where 1, or 0dB is the reference for a distance of the speaker distance to the origin), MP is the microphone pattern (0 (omni) -> 1 (figure of 8)), MA is the orientation of the microphone (m), and delay is the delay, in samples, from the speaker to the microphone.

These values are then used to generate impulse responses from each loudspeaker to each microphone as shown in equation (5).

$$MicIR_m^q(n) = amp_m^q \times sinc\left(n - delay_m^q\right) \tag{5}$$

Where n is the sample index (from 0 to the length of the filter in samples)

4.1 Ambisonics to Virtual Microphone Decoding

Once the Virtual Mic IRs (MicIR) are calculated, they need to be transformed into the spherical harmonic domain. Spherical harmonics of order n and degree m can be described[12] as:

$$Y_n^m(\theta, \phi) = N_n^{|m|} P_n^{|m|}\left(sin(\phi)\right) \begin{cases} cos(|m|\theta) & if\ m > 0 \\ sin(|m|\theta) & if\ m < 0 \\ 1 & if\ m = 0 \end{cases}$$

$$N_n^{|m|} = \sqrt{\left(2 - \delta_m \times \frac{(n - |m|)!}{(n + |m|)!}\right)} \tag{6}$$

Where: P_n^m is the Associated Legendre Polynomial (or Function), n is the (Ambisonic) order, *m* is the degree of the spherical harmonic (-n ≤ m ≤ n) and δ_m is the Kronecker delta function,

$$\delta_m = \begin{cases} 1\ if\ m = 0 \\ 0\ if\ m \neq 0 \end{cases} \tag{7}$$

Due to the relationship between order (n) and degree (m), the number of channels needed to represent the 3D sound scene will be $(n+1)^2$

For this initial investigation and proof of concept, only horizontal synthesis/reconstruction is carried out as the microphones are placed at the same heights (0 degrees elevation). This occurs when the degree (m) is +/- the order (n) as shown in Figure 2.

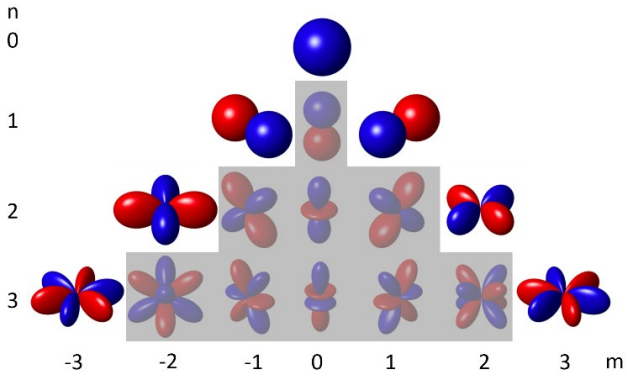

Figure 2 - Spherical Harmonics over the sphere up to 3rd order. Circular (horizontal) harmonics are emphasised.

Ambisonics states that audio, transformed into the spherical harmonic domain, can be reproduced by superimposing the audio from a discrete set of speakers at fixed positions, also transformed into the spherical harmonic domain:

$$s \cdot Y_n^m(\theta, 0) = \sum_{i=1}^{L} g_i \cdot Y_n^m(\theta_i, 0) \tag{8}$$

Where s is the pressure of the source signal from direction (θ, ϕ=0), g_i is the gain of the i^{th} speaker from direction (θ_i,ϕ_i=0).

Re-writing this using vector notation where B is a column vector of Ambisonic channels (encoded sound source, collectively known as n^{th} order B-Format (the LHS of equation (8), although in this case, only the circular harmonics are utilised).

$$B = [Y_0^0(\theta, 0), Y_1^{-1}(\theta, 0), Y_1^1(\theta, 0), \dots Y_n^{m=n}(\theta, 0)]^T s \tag{9}$$

p is a column vector of signals that come from the BRIR sampled positions to reconstruct the sound field (g_i in (8)) and C is the re-encoding matrix that represents the spherical harmonic coefficients for the sampled positions. So, for an 8 sample, and 3rd order (7 harmonics) system it would be a 8 x 7 matrix:

$$C = \begin{bmatrix} Y_0^0(\theta_1, 0) & \cdots & Y_3^3(\theta_1, 0) \\ \vdots & \ddots & \vdots \\ Y_0^0(\theta_8, 0) & \cdots & Y_3^3(\theta_8, 0) \end{bmatrix} \tag{10}$$

The vector form of this equation can now be rewritten more simply:

$$B = C \times p \tag{11}$$

This is stating that the B-format (LHS) can be reconstructed by the summed speakers' output weighted by the spherical harmonic coefficients (RHS).

To calculate what should be output from the speakers to reconstruct the original B-Format (*B*), we rearrange the equation so that *p* is the subject of the formula. Where C^{-1} is now the decoding matrix needed. Given that the matrix, *C*, is not usually square, the pseudo inverse is needed (C^+).

$$p = C^+ \times B \tag{12}$$

4.2 Virtual Microphone Decoding

Virtual Microphone decoding simulates each speaker using a pair of MicIRs to represent each speaker/sampling position/Microphone combination. This will add an extra step to the equations presented previously (convolving each speaker output with a MicIR pair). Using the vector notation already used above in (12).

$$\begin{bmatrix} Left \\ Right \end{bmatrix} = \begin{bmatrix} C^+ \times MicIR_1 \otimes B \\ C^+ \times MicIR_2 \otimes B \end{bmatrix} \quad (13)$$

Where *Left* and *Right* are the left and right headphone/loudspeaker signals being generated, C^+ is the decoding matrix, *MicIR* are the Microphone Impulse Responses and *B* is the Ambisonically encoded audio to be reproduced. The important point here, and one that makes the decoding of Ambisonics computationally efficient, is that the multiplication of the decoding matrix and the MicIRs can be pre-calculated resulting in a pair of IRs per circular harmonic, rather than one pair per sampled position, which will reduce the number of convolutions needed[13].

$MicIR_{1,2}$ will be a one-row vector (IR) per speaker position and Mic. For example, an 8-position decode, and MicIR data for each position that contains NN samples (1024, for example), the $MicIR_1$ matrix would be an 8 x 1024 matrix:

$$\begin{bmatrix} MicIR_1^1(0) & \cdots & MicIR_1^1(\mathrm{NN}-1) \\ \vdots & \ddots & \vdots \\ MicIR_1^8(0) & \cdots & MicIR_1^8(\mathrm{NN}-1) \end{bmatrix} \quad (14)$$

Assuming the B format audio will be provided later, the virtual microphone decoding filters can be pre-calculated as the product of C^+ and $MicIR_{1,2}$. This will result in a pair of filters per Ambisonic channel, rather than a pair of filters per speaker/sampling position (and, due to Ambisonic decoder theory) there will always be fewer Ambisonic channels than speakers.

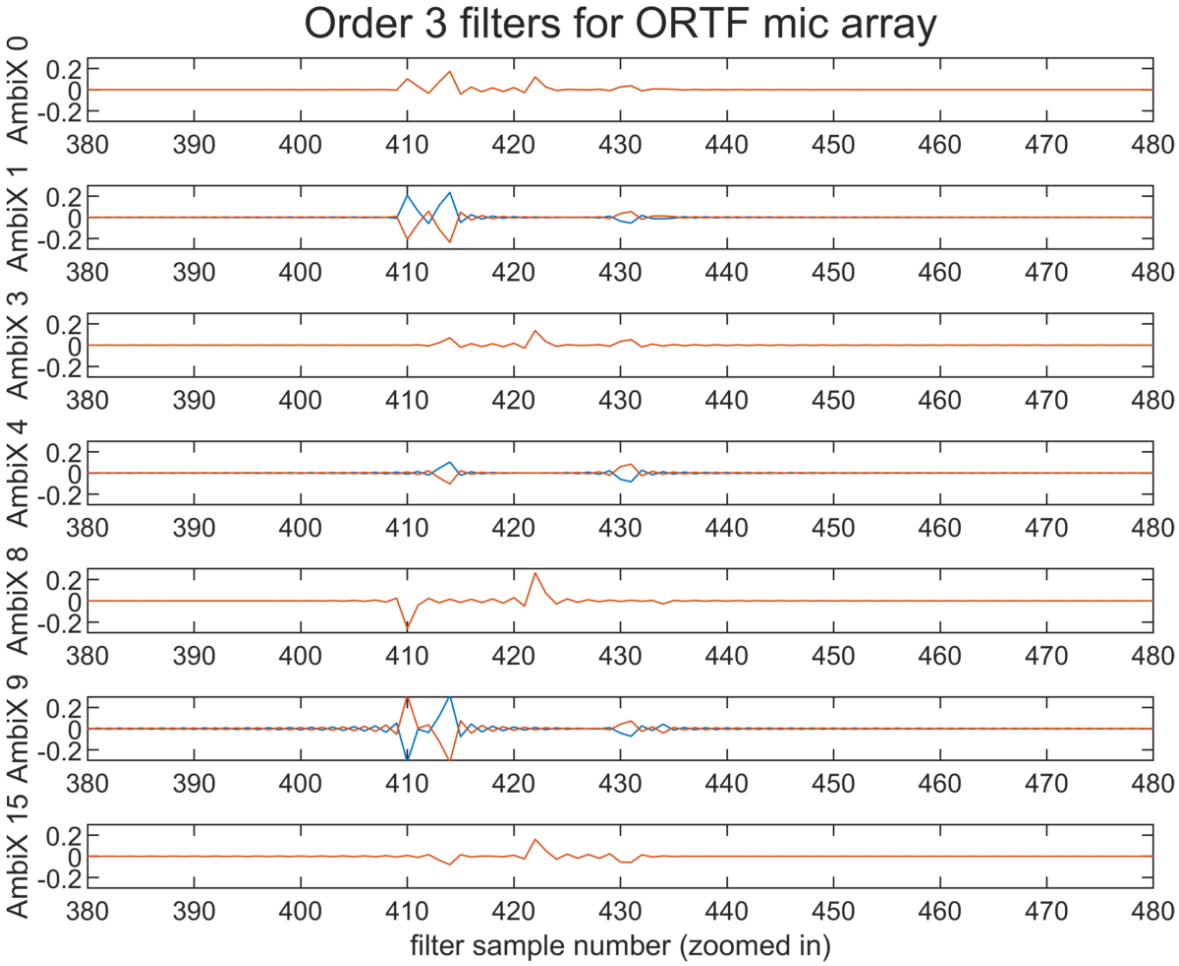

Figure 3 - The three stereo decoding filter pairs needed for a 3rd order virtual microphone decode using freefield simulated IR responses. Blue represents microphone 1 and orange represents microphone 2 responses. Filters are 1024 points but only show samples 380-480.

So, taking the 8 position, 3rd order example, where C^+ is a 7 x 8 matrix (assuming C, before inversion, is a 8 x 7 matrix):

$$VMicIR_L = \begin{bmatrix} D_0^0(\theta_1,0) & \cdots & D_0^0(\theta_8,0) \\ \vdots & \ddots & \vdots \\ D_3^3(\theta_1,0) & \cdots & D_3^3(\theta_8,0) \end{bmatrix} \times \begin{bmatrix} MicIR_1^1(0) & \cdots & MicIR_1^1(\mathrm{NN}-1) \\ \vdots & \ddots & \vdots \\ MicIR_1^8(0) & \cdots & MicIR_1^8(\mathrm{NN}-1) \end{bmatrix}$$

$$VMicIR_R = \begin{bmatrix} D_0^0(\theta_1,0) & \cdots & D_0^0(\theta_8,0) \\ \vdots & \ddots & \vdots \\ D_3^3(\theta_1,0) & \cdots & D_3^3(\theta_8,0) \end{bmatrix} \times \begin{bmatrix} MicIR_2^1(0) & \cdots & MicIR_2^1(\mathrm{NN}-1) \\ \vdots & \ddots & \vdots \\ MicIR_2^8(0) & \cdots & MicIR_2^8(\mathrm{NN}-1) \end{bmatrix} \quad (15)$$

Where $D=C^+$, $VMicIR_{L,R}$ are the Virtual Mic Ambisonic decoding filters (one filter per circular harmonic) comprising 7 pairs of impulse responses of length NN samples as shown in Figure 3

Once the decoding filters have been calculated, the audio output can be derived by convolving the Ambisonically encoded audio by these filters.

$$\begin{aligned} Left &= [VMicIR_L \otimes B] \\ Right &= [VMicIR_R \otimes B] \end{aligned} \quad (16)$$

5 OBJECTIVE ANALYSIS OF DECODES

The listening test (described shortly) used three musical pieces (3rd order Jazz, Choral and Classical recordings). Decorrelation analysis was conducted as suggested by Rumsey and Lewis[14] for comparison with subjective measures. MATLAB's xcorr function was used to calculate the peak cross-correlation coefficients of the 6 decodes for each music material and headphone and loudspeaker monitoring. The peak IACC values (Figure 4) were computed directly from the left and right channels to objectively analyse the stimuli for headphone playback.

IACC	Jazz	Choral	Classical
Blumlein	0.289	0.308	0.589
ORTF	0.783	0.562	0.864
AB	0.809	0.310	0.872
UHJ	0.908	0.739	0.883
XY	0.937	0.847	0.958
MONO	1.000	1.000	1.000
IACC (LS)	**Jazz**	**Choral**	**Classical**
Blumlein	0.334	0.362	0.480
ORTF	0.642	0.503	0.644
AB	0.709	0.191	0.710
UHJ	0.843	0.745	0.802
XY	0.867	0.854	0.876
MONO	0.942	0.967	0.948

Figure 4 - Peak IACC coefficients for both headphone (HP) and loudspeaker (LS) computed for the 6 decodes for Jazz, Choral and Classical.

Decorrelation analysis for loudspeakers was obtained using BRIRs for each loudspeaker, captured via Farina's swept sine wave technique[15] using binaural microphones on a KEMAR 45BB-6 dummy head[16] as shown in Figure 5. The peak IACC coefficients were computed using the binaural output obtained by convolving the two channels from each of the renders separately with the BRIRs.

The relative peak IACC coefficients for Blumlein, ORTF and XY matched the IACC analysis of corresponding stereo microphone arrays by Conceição and Furlong[17] with their correlation increasing in that order. The coefficient for AB matched their findings, except in the case of the choral material where Blumlein interestingly had a higher peak IACC, likely due to increased rear reverberant energy (in the church setting) along with spectral variance in the source. It was apparent that the choral material would be ideal for evaluating spaciousness given its low IACC/ICCC values.

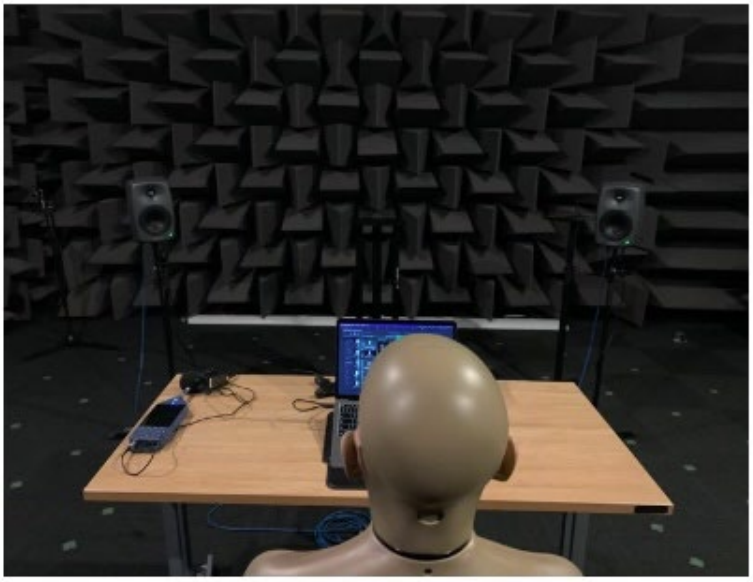

Figure 5 - A KEMAR 45BB HATS dummy head positioned to capture the BRIRs for separately capturing HRTFs representing the two loud-speaker positions.

6 LISTENING TESTS METHODOLOGY

6.1 Test design

This study was best suited for BS.2132 based on ITU recommendations as no known reference existed and spatial and timbral differences were expected between the stimuli owing to the combined effect of differences in polar patterns, spacing and orientation of the virtual microphones. BS.2132-0[18] adopts the general subjective assessment of multiple sound systems, requiring less control of test parameters. BS.2132-0 defines 'overall subjective quality' as a global attribute that needs to be evaluated to quantify the judgement of the subjects of stimuli to an internal reference. The recommendation while offering some examples of main and sub-attributes with possible descriptions, suggests referring to BS.2399-0[20] for spatial attributes, which draws on the elicitation work from the likes of Rumsey[21] and Lorho[22] for spatial attributes.

A within-subjects design with each subject rating all the stimuli involved in the experimental design was preferred for its increased efficiency and controllability. A compromised approach was devised, where the headphone-based listening test was administered remotely online to maximise participation. The loudspeaker-based listening test given its logistics and variables, was administered under controlled conditions in the hemi-anechoic chamber (MS037) at the University of Derby's Markeaton Street campus.

6.2 Test delivery

Go Listen, an end-to-end online listening test platform developed by Barry et al.[23] was chosen for conducting both listening tests. The online listening test had a total of 48 participants, consisting of

12 students of audio, music, sound, and acoustics while 15 of them were professionals in these 4 sectors. The in-person listening test had 10 participants, consisting of students and staff from the University of Derby. Out of the 5 students, 4 were postgraduate students of audio engineering, and the 5 members of staff were all directly involved in audio.

All the coincident decodes of the B-format recordings were rendered using the Virtual Microphone plugin. The UHJ decode was rendered using the aXMonitor plugin[24] in its UHJ Super Stereo (IIR) setting, due to its description of providing a balanced left-right distribution of sources without clustering at the extremes of the stereo spread.

7 ATTRIBUTES

A two-stage approach was followed as recommended in BS.2132-0 where 3 excerpts (jazz, choral and classical) were rated for overall quality followed by select excerpts for four prevalent spatial attributes – spaciousness, envelopment, naturalness, and balance. A multitude of spatial attributes in addition to the large number of timbral attributes could have been employed. The three impression attributes spaciousness, envelopment and naturalness were chosen owing to their recurrence in spatial audio studies as the 3 most popular attributes. While localisation was not a focus of the study, the balance attribute was chosen as a subjective impression to describe sound stage distribution.

7.1 Overall quality

The overall quality attribute was used to represent the overall preference or impression of the stimuli from participants, as defined in BS.2132-0. This is analogous to the Basic Audio Quality (BAQ) from BS.1116, but without a reference for comparison. Three trials were conducted for overall quality, featuring an excerpt each from the 3 music materials.

7.2 Spaciousness

The spaciousness attribute follows the reverberance attribute from BS.2399 but is defined as a perceived sense of space as previous evaluation studies found ‘reverberance’ to be a more ambiguous term. An excerpt from the choral material was used due to its low correlation values.

7.3 Envelopment

The envelopment attribute definition provided was derived from BS.2399, with an emphasis on horizontal spread to make a clear distinction from the engulfment attribute which is related to vertical spread. While BS.2399 specifies envelopment as being achieved by the dry sources alone as well as through reverberance, just the ensemble (source) was considered in this case without reverberance given the separate spaciousness attribute which was considered a better measure of the environment. The effect of reverberance was understood to be minimised through increased correlation, which favoured the use of an excerpt of the jazz material given its high correlation.

7.4 Naturalness

The naturalness attribute was not found in BS.2399 but was derived from previous spatial audio evaluation literature from the likes of Rumsey[25]. Despite its ambiguity given the individualised frame of reference, this attribute was chosen as a general frame of reference for being present in the audience of the musical performance. Whilst defining a clear visual dimensional reference position in the audience would have been ideal, this was not possible given the lack of details of the specific position of the TOA microphone. An excerpt from the jazz material was chosen considering its minimal number of individual sources in the trio.

7.5 Balance

The balance attribute definition provided was derived from BS.2399 based on the skewness of the left-to-right sound stage, with consideration for holes or gaps. This was in line with the popular description of a 'hole in the middle' associated with spaced techniques. This attribute while not popularly employed in spatial audio evaluation, was chosen in an attempt to generalise the spatial distribution of each decoder. A jazz excerpt was used in this case considering it consisted of a manageable number of elements in a trio, making for a clearer distinction of positions.

8 HEADPHONE TEST RESULTS

The response ratings for the headphones test were statistically analysed in SPSS 28[26] using an RMANOVA considering its higher sample size. In this case, the effect of the virtual microphone techniques and genre (music material) on the preference ratings were tested using a two-way RMANOVA.

8.1 Overall Quality

A moderate effect size was observed for decoder x material. Bonferroni post hoc tests for the overall quality attribute across 3 stimuli returned these significant differences –

- ORTF had a higher mean rating than AB and mono.
- UHJ had a higher mean rating than AB and mono.
- XY had a higher mean rating than AB and mono.
- Blumlein had a higher mean rating than mono.

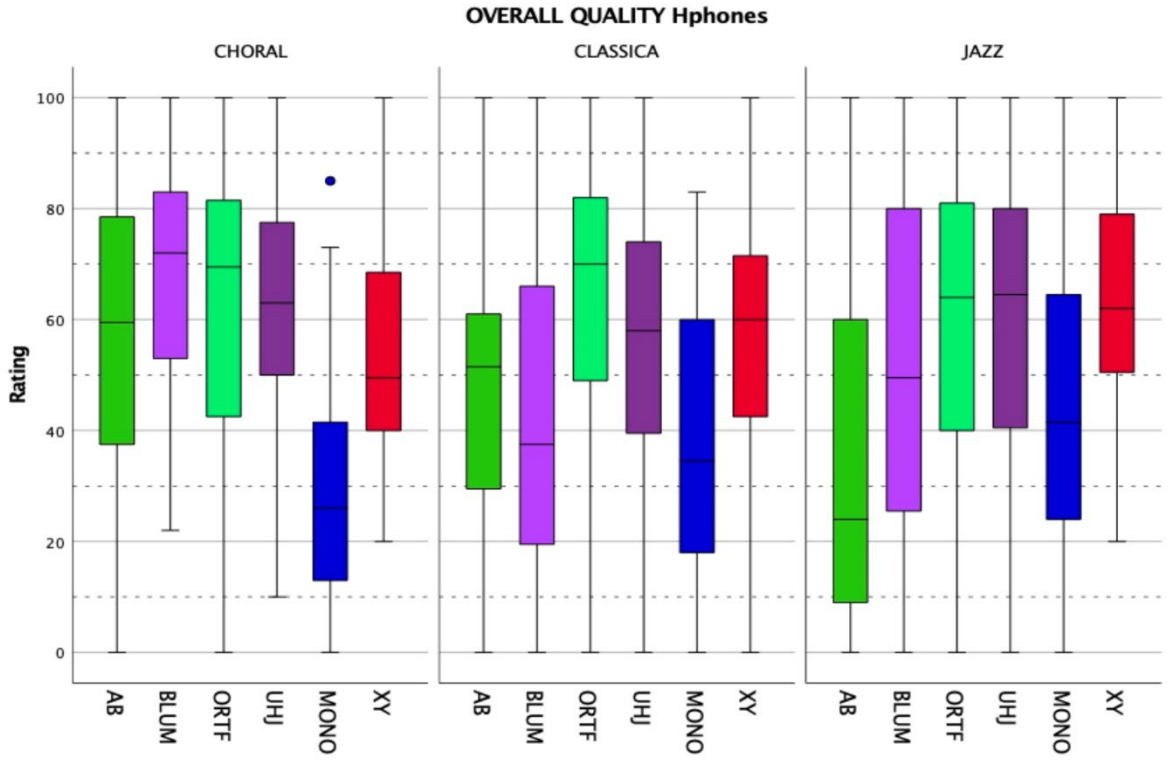

Figure 6 - A box plot showing the medians of the decoders with standard deviations across 3 stimuli for Overall Quality – headphones.

8.2 Spaciousness

A medium to large effect size was observed. Bonferroni post hoc tests for spaciousness returned the following significant differences ($p < 0.05$) –

- The mean rating of AB was higher than that of ORTF, XY and mono.
- The mean ratings of Blumlein, ORTF, UHJ and XY were all higher than that of mono.

8.3 Envelopment

A large effect size was observed. Bonferroni post hoc tests envelopment returned the following significant differences ($p < 0.05$) –

- The mean rating of Blumlein was higher than that of AB, UHJ, XY and mono.
- The mean rating of ORTF was higher than that of UHJ and mono.
- The mean ratings of AB, UHJ and XY were all higher than that of mono.

8.4 Naturalness

A small effect size was observed. Bonferroni post hoc tests for naturalness only revealed that UHJ had a significantly higher mean rating ($p < 0.05$) than mono.

8.5 Balance

A large effect size was observed. Bonferroni post hoc tests for balance returned the following significant differences ($p < 0.05$) –

- The mean rating of mono was significantly higher than that of AB, Blumlein, ORTF and XY.
- The mean rating of UHJ was significantly higher than that of Blumlein, ORTF and XY.
- The mean rating of AB was significantly higher than that of Blumlein, ORTF and XY.
- The mean rating of XY was significantly higher than that of Blumlein.

9 LOUDSPEAKER TEST RESULTS

The loudspeaker test with a sample size lower than recommended for a RMANOVA, was deemed as best suited for a non-parametric Friedman's ANOVA. The only exception to this was the overall rating across three stimuli material which required a two-way RMANOVA as Friedman's ANOVA is limited to one-way analysis.

9.1 Overall Quality

Box plots for the overall quality results can be seen in Figure 7. A moderate effect size was observed for decoder X material. Bonferroni post hoc tests for the overall quality attribute across the 3 stimuli material however did not return significant differences after Type 1 error adjustment. The follow-up Bonferroni post hoc test results showing a significant difference from RMANOVA (Friedman's ANOVA was found to produce the same results) separately for the overall quality of jazz and choral stimuli material are provided below for additional clarity. Classical material did not return significant differences.

- ORTF had a higher mean rating than Blumlein (Jazz).
- Blumlein had a higher mean rating than mono. (Choral).

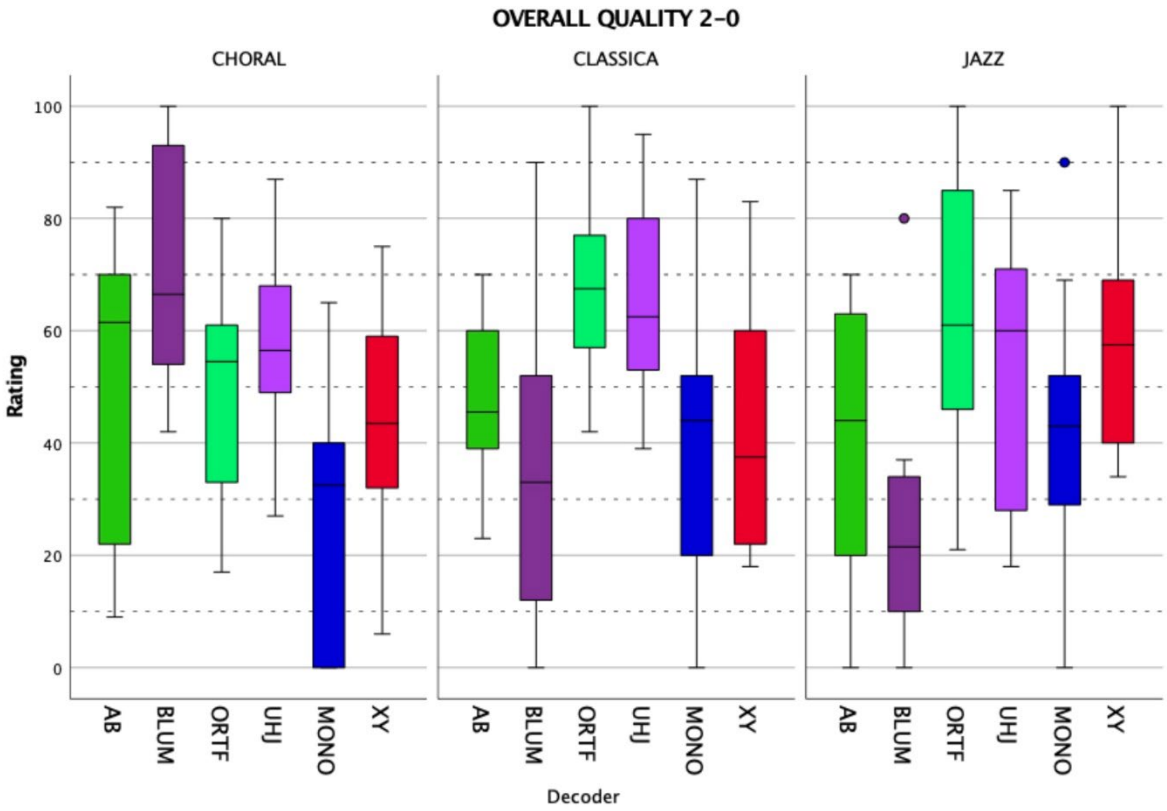

Figure 7 - A box plot showing the medians of the decoders with standard deviations across 3 stimuli types for Overall Quality – loudspeakers.

9.2 Spaciousness

A small effect size was observed. Dunn-Bonferroni post hoc tests however did not reveal significant differences between any decoder pairs after adjustment for Type 1 error.

9.3 Envelopment

A small effect size was observed. Dunn-Bonferroni post hoc tests showed that Blumlein was ranked significantly higher than mono ($p = 0.019$) after adjustment for Type 1 error.

9.4 Naturalness

A small effect size was observed. Dunn-Bonferroni post hoc tests returned no significant differences between any decoders.

9.5 Balance

A small effect size was observed. Dunn-Bonferroni post hoc tests showed that UHJ was ranked significantly higher than Blumlein ($p = 0.019$) after adjustment for Type 1 error.

10 DISCUSSION

10.1 Headphone Listening Test

The hypothesised preference for near-coincident decoders was not conclusively proven, but ORTF's mean rating was the highest and significantly higher than that of AB. UHJ and XY had the next highest

mean, higher than Blumlein. The preference for near-coincident over spaced was conclusively proven, with XY and UHJ also being conclusively higher rated than AB. Blumlein interestingly showed the greatest variance in overall rating across music material, with ORTF showing the least variance.

AB was rated significantly higher than the majority of the other decoders (ORTF, XY and mono) for spaciousness, with the highest mean rating. It can be inferred that ORTF and XY were not sufficiently spaced and decorrelated compared to AB and suffered from excess off-axis rejection. The mean ratings mostly matched the relative peak ICCC coefficients except for AB which, despite exhibiting the same decorrelation as Blumlein, had a higher mean spaciousness rating. The difference however was not significant, indicating that Blumlein's antiphase rear pickup made up for the lack of relative spacing when compared to AB. This matched Greisinger's[28] calculations and Lim's[27] findings of these stereo arrays showing high similarity in spaciousness, which Lim attributed to the effect of the polar patterns. This also retained Lim's theory of the spacing between microphones not being directly correlated with perceived spaciousness. However, Lim's finding of there being a lack of difference in spaciousness between arrays did not hold true. The means of Blumlein, ORTF and UHJ were similar, suggesting that the spacing used by ORTF could be compensated by mapping of rear spatial information to the front, for a similar sense of space to be reproduced. All 5 decoders having a significantly higher rating than mono meant that mono reproduced a clearly compromised sense of space captured in the B-format. A mild outlier rating of 0 was observed with spaciousness for AB from a subject who had also rated mono at 0, which was interesting considering they had identified as an audio professional.

Blumlein had the highest mean rating for envelopment, significantly higher than all other decoders except ORTF The results did not prove the argument from Stevens et al.[29] about UHJ offering high envelopment. Blumlein and ORTF exhibited high similarity, indicating that their differences in polar pattern and angling can be compensated using the 17 cm spacing. AB, UHJ and XY showed high similarity in means. All the decoders were significantly more enveloping than mono, which showed the least envelopment as expected.

Naturalness was found to be ambiguous as explained by Rumsey[21], considering its sole result of UHJ being significantly higher rated than mono. This is particularly interesting and was analogous to the finding of Stevens et al.[29] that UHJ, despite mapping rear spatial information to the front in an unnatural manner, was found 'ecologically valid' in comparison to surround-sound setups. AB had high similarity when compared to UHJ with the second highest mean. Blumlein, ORTF and XY showed high similarity in means against mono, which was also interesting, which potentially might have been due to jazz recordings in mono being prominent historically.

Mono exhibited a significantly better balanced distribution of sources than the others (except UHJ) due to central concentration without any skewing. UHJ's MS-like pattern resulted in a similarly even distribution alongside AB, both of which were rated more highly than Blumlein, ORTF and XY. While AB's result was unexpected, the distance from the microphone could have a role to play in this where the diffused nature of distant recordings could have contributed to this. XY offered a significantly better balance distribution than Blumlein. Two mild outlier ratings between 60 and 80 were observed with balance for the Blumlein decode.

10.2 Loudspeaker Listening Test

The lower sample size appeared to produce minimal significant results with small effect sizes for the loudspeaker test. The overall quality rating across all material witnessed no significant differences. Individually, the jazz trial showed a preference for ORTF over Blumlein while the choral trial favoured Blumlein over mono potentially due to a combination of the widespread and diffuse rear reverberant energy. The mean overall ratings however had ORTF and UHJ ranking the highest. While this could be a result of the low sample size, this could also be attributed to the crosstalk involved not favouring ORTF as much as the direct playback over headphones (ORTF can be seen as a simple head simulation). UHJ's similarly high ranking to ORTF is an interesting observation, potentially showing

UHJ being more effective for speaker reproduction than headphones (as designed). Blumlein had the lowest mean rating amongst the stereo decodes, unlike the headphone test (AB).

Spaciousness did not yield significant differences in means but Blumlein exhibited the highest mean rating slightly higher than AB, which was unexpected considering that Blumlein had a higher IACC. This would suggest significant rear reverberant energy being present. Blumlein offered the highest mean rating for envelopment, significantly higher than that of mono. While naturalness did not yield significant differences, UHJ exhibited the highest mean rating as with the headphone test. Blumlein unexpectedly exhibited the lowest mean rating, which conflicted with the conventional association of the technique with realism for loudspeaker reproduction. The balance attribute most favoured UHJ, significantly higher rated than Blumlein which had the lowest rating. Mono's unaltered sound stage distribution resulted in a high rating, but UHJ ranking the highest in means was an interesting observation, unlike on headphones.

11 CONCLUSIONS

This study subjectively compared 4 prevalent stereo microphone techniques – AB, Blumlein, XY and ORTF as virtual arrays to decode TOA recordings, along with UHJ and mono. Commercial plugins were used for the coincident decodes (XY, UHJ and Blumlein) while a custom rendering approach was employed for spaced (AB) and near-coincident (ORTF) decodes. Listening tests were conducted separately over headphones and loudspeakers evaluating overall preference and four prevalent spatial attributes – spaciousness, envelopment, naturalness, and balance. While the hypothesised absolute preference for near-coincident decoding was not conclusively proven in the results, ORTF exhibited the highest mean rating for preference over both headphones and loudspeakers (shared with UHJ). The preference for near-coincident over spaced was conclusively proven for headphones, along with other differences between specific decoder pairs. The loudspeaker test however did not return significant differences for preference, presumably due to a low sample size but also from the inter-channel crosstalk not favouring ORTF as much. It was also observed that correlation across channels did not directly influence preference ratings, but largely matched the spaciousness ratings. Spaciousness for headphones revealed the strong sense of environment reproduced by the spaced AB as with previous literature, with Blumlein and UHJ being comparable given their front-mapped rear spatial information. Blumlein exhibited the highest mean rating for envelopment generally, ranking higher than all except ORTF for headphones. Naturalness despite its ambiguity across participants, interestingly revealed a preference for UHJ over mono on headphones and had the highest mean ratings in both tests. Balance showed Blumlein's tendency to skew the sound stage distribution, with UHJ and mono offering the least skewness. It is suspected that a larger sample size could potentially bring out more significant differences, but the results obtained here could also be said to apply to the particular recording contexts used in this test meaning that more recording experiments would be required to conclusively draw preferences. The work does point to spaced and near-coincident arrays being a useful addition to the arsenal of available decoding techniques for Ambisonics.

12 FURTHER WORK

Increasing the sample size for the listening tests would be a vital improvement to improve conclusively, especially in the case of the loudspeaker test. A controlled headphone test in person would also be more suited to maintain consistency and effective participation. In line with participant feedback, popular music material more suited to listener interests would be more incentivised to use as stimuli. It would also be highly effective to employ a controlled recording setup with well-documented positions of sound sources across material, to increase transparency. Variables such as listener experience and increased participant training need to be explored, which were excluded from this study to reduce complexity.

The elicitation of additional spatial and timbral attributes needs to be employed to find which might potentially show a greater correlation with preference ratings. In addition to the 4 arrays compared in

this study, there exists an endless variety of two-channel and multi-channel recording techniques that would also be beneficial to compare as virtual decode techniques, for stereo and surround sound reproduction. A dual-band decoding requires further investigation for usability, and shuffling could be applied for the Blumlein array to test for improvements along with comparisons to current binaural decoding strategies over headphones.

In terms of the decoding methods, including energy vector or other optimisation methods, looking at the spatial aliasing artefacts for each decode and investigating windowing of the sinc pulse based fractional delay line are to be investigated.

13 ACKNOWLEDGEMENTS

Many thanks to Bo-Erik Sandholm (https://www.ohti.xyz) for inspiring this work with the request “I’m looking for a method to extract virtual spaced omnis or cardioids from Ambisonic recordings” and for providing the audio samples for the listening tests.

14 REFERENCES

1. Skeet, M. (1983) ‘Ambisonics (EMM Apr 1983)’, Electronics & Music Maker, (Apr 1983), pp. 22–23.
2. Arteaga, D. (2023) Introduction to Ambisonics. Available at: https://doi.org/10.5281/zenodo.7963105.
3. Zotter, F. and Frank, M. (2019) Ambisonics. Cham: Springer Nature (Springer Topics in Signal Processing). Available at: https://doi.org/10.1007/978-3-030-17207-7.
4. Bartlett, B. (2014) Recording music on location: capturing the live performance. 2nd ed. Burlington, Mass: Focal Press. Available at: https://doi.org/10.4324/9781315777078.
5. Rumsey, Francis. (2001) Spatial audio. Oxford: Focal (Music technology series.).
6. Streicher, R. and Dooley, W. (1984) ‘Basic Stereo Microphone Perspectives: A Review’, in Audio Engineering Society Conference: 2nd International Conference: The Art and Technology of Recording. Available at: http://www.aes.org/e-lib/browse.cfm?elib=11662.
7. Rayburn, R.A. and Eargle, J. (2012) Eargle’s microphone book: From Mono to stereo to surround: A guide to microphone design and Application. Waltham, MA: Focal Press/Elsevier.
8. Bernfeld, B. and Smith, B. (1978) ‘Computer-Aided Model of Stereophonic Systems’, in. Audio Engineering Society Convention 59, Audio Engineering Society. Available at: *https://aes2.org/publications/elibrary-page/?id=3033* (Accessed: 26 July 2023).
9. Dooley, W.L. and Streicher, R.D. (1982) ‘M-S Stereo: A Powerful Technique for Working in Stereo’, Journal of the Audio Engineering Society, 30(10), pp. 707–718.
10. Blue Ripple Sound (2023) Plugin: O3A Virtual Microphone. Available at: http://www.blueripplesound.com/plugin/o3a_virtual_microphone (Accessed: 11 June 2023).
11. McKeag, A., McGrath, D. “Sound Field Format to Binaural Decoder with Head-Tracking.” 6th Austrailian Regional Convention of the AES, Melbourne, Australia. 10 – 12 September. Preprint 4302. 1996
12. Kronlachner, M., & Zotter, F., “Spatial transformations for the enhancement of Ambisonic recordings”. In Proceedings of the 2nd International Conference on Spatial Audio, Erlangen. 2014
13. Wiggins, B. Paterson-Stephens, I., Schillebeeckx, P., “The analysis of multi-channel sound reproduction algorithms using HRTF data.” 19th International AES Surround Sound Convention, Germany, p. 111-123. 2001
14. Rumsey, F.J. and Lewis, W. (2002) ‘Effect of Rear Microphone Spacing on Spatial Impression for Omnidirectional Surround Sound Microphone Arrays’, in. Available at: *http://www.aes.org/e-lib/inst/browse.cfm?elib=11421* (Accessed: 1 September 2023).
15. Farina, A. (2000) ‘Simultaneous Measurement of Impulse Response and Distortion with a Swept-Sine Technique’, in AES Convention.

16. Gras, 45bb-6 Head & torso for ear- and headphone test, 2-CH CCP (no date) GRAS. Available at: https://www.grasacoustics.com/products/head-torso-simulators-kemar/product/504-45bb-6 (Accessed: 27 March 2024).
17. Conceição, M. and Furlong, D. (2013) 'Influence of Different Microphone Arrays on IACC as an Objective Measure Of Spaciousness', in. Audio Engineering Society Convention 134, Audio Engineering Society. Available at: *https://aes2.org/publications/elibrary-page/?id=16786* (Accessed: 24 August 2023).
18. ITU-R (2019b) 'ITU-R BS.2132-0 – Method for the subjective quality assessment of audible differences of sound systems using multiple stimuli without a given reference'. Available at: https://www.itu.int/rec/R-REC-BS.2132-0-201910-I/en.
19. ITU-R (2015) 'BS.1534-3: Method for the subjective assessment of intermediate quality level of audio systems'. Available at: https://www.itu.int/rec/R-REC-BS.775/en.
20. ITU-R (2017) 'ITU-R BS.2399-0– Methods for selecting and describing attributes and terms, in the preparation of subjective tests'. Available at: https://www.itu.int/pub/R-REP-BS.2399.
21. Rumsey, F. (2006) 'Spatial audio and sensory evaluation techniques – context, history and aims'.
22. Lorho, G. (2005) 'Individual Vocabulary Profiling of Spatial Enhancement Systems for Stereo Headphone Reproduction', in. Audio Engineering Society Convention 119, Audio Engineering Society. Available at: *https://aes2.org/publications/elibrary-page/?id=13360* (Accessed: 10 July 2023).
23. Barry, Dan & Zhang, Qijian & Sun, Pheobe & Hines, Andrew. (2021). Go Listen: An End-to-End Online Listening Test Platform. Journal of Open Research Software. 9. 20. 10.5334/jors.361.
24. SSA Plugins (2023) 'aXMonitor - Ambisonic Monitor - Ambisonic to Stereo & Binaural Decoder • SSA Plugins', SSA Plugins. Available at: https://www.ssa-plugins.com/product/axmonitor/ (Accessed: 3 June 2023).
25. Rumsey, F. (2002) 'Spatial Quality Evaluation for Reproduced Sound: Terminology, Meaning, and a Scene-Based Paradigm', J. Audio Eng. Soc., 50(9).
26. IBM (2023) IBM SPSS Software 28.0 | IBM. Available at: https://www.ibm.com/spss (Accessed: 22 August 2023).
27. Lim, W. (2013) 'An Objective Comparison of Stereo Recording Techniques through the Use of Subjective Listener Preference Ratings', in. Audio Engineering Society Convention 135, Audio Engineering Society. Available at: *https://aes2.org/publications/elibrary-page/?id=17045* (Accessed: 28 June 2023).
28. Griesinger, D. (1985) 'Spaciousness and Localization in Listening Rooms: How to Make Coincident Recording Sound as Spacious as Spaced Microphone Arrays', in. Audio Engineering Society Convention 79, Audio Engineering Society. Available at: *https://aes2.org/publications/elibrary-page/?id=11461* (Accessed: 7 July 2023).
29. Stevens, F., Murphy, D. and Smith, S. (2017) 'Ecological Validity of Stereo UHJ Soundscape Reproduction', in. Available at: http://www.aes.org/e-lib/inst/browse.cfm?elib=18641 (Accessed: 2 August 2023).