THE TRANSPARENCY OF BINAURAL AURALISATION USING VERY HIGH ORDER CIRCULAR HARMONICS

M Dring College of Engineering and Technology, University of Derby, UK
Dr B Wiggins College of Engineering and Technology, University of Derby, UK

1 INTRODUCTION

Ambisonics, which is based on the spherical harmonic (SH) decomposition of the sound field to a specific order and subsequent reproduction over loudspeakers or headphones[1,2,3], allows for computationally efficient and mathematically lossless rotation of the sound field which is particularly useful for head-tracked headphone listening. This has resulted in Ambisonics becoming one of the standard delivery formats of spatial audio for 360° video and virtual reality platforms. The decoding and use of Ambisonics in headphone based binaural reproduction was first discussed by McKeag and McGrath who used 1st order Ambisonic recordings to feed head-tracked binaural audio over headphones in 1996[2]. The Ambisonics to binaural decoding algorithm[3], works by using anechoic head related transfer functions (HRTFs) to decode the Ambisonic material to a virtual speaker array with each position specified by a pair of HRTFs. If the HRTFs are anechoic, then rotating the Ambisonic sound field decoding to these fixed positions has the same aural effect as rotating the head of the listener and will, in effect, interpolate new HRTFs that will exactly match the desired HRTF up to the spatial aliasing frequency, which is dependent on the Ambisonic order[4]. If anechoic HRTFs are utilised, then the reverberant field of the room, or sound scene, to be reproduced must be encoded/captured in the Ambisonic B-Format signals fed to the binaural decoder/filters. However, when utilising computer aided design packages or recording/measuring acoustic spaces using microphones, output is limited to low Ambisonic orders and, hence, a low spatial aliasing frequency.

To overcome this issue and obtain a room auralisation to a much higher Ambisonic order than is currently available, the room's Binaural Room Impulse Response/Transfer Function (BRIR/TF) can be rendered/captured. These signals are then decomposed into a set of spherical harmonics to the desired order (if enough spatial samples are taken to achieve that order). However, if the current technique of virtual decoding is used, rotating a sound field will result in the room remaining static, but the reproduced sources rotating within it. For head-tracking to achieve correct cues, it is the room that must rotate.

In this work, instead of capturing the response of the room at multiple locations, one location is used, and the virtual head rotated through a full 360° in the horizontal plane. This allows head rotations to correctly map to a static source, with a source panned into the sound field now rotating the room and associated acoustic response. As long as enough rotations are captured, then very high order circular harmonic encoding/decoding can be implemented, with head rotation, of a reverberant space. Currently, software created for this project works up to 31st order, requiring 63 channels of audio.

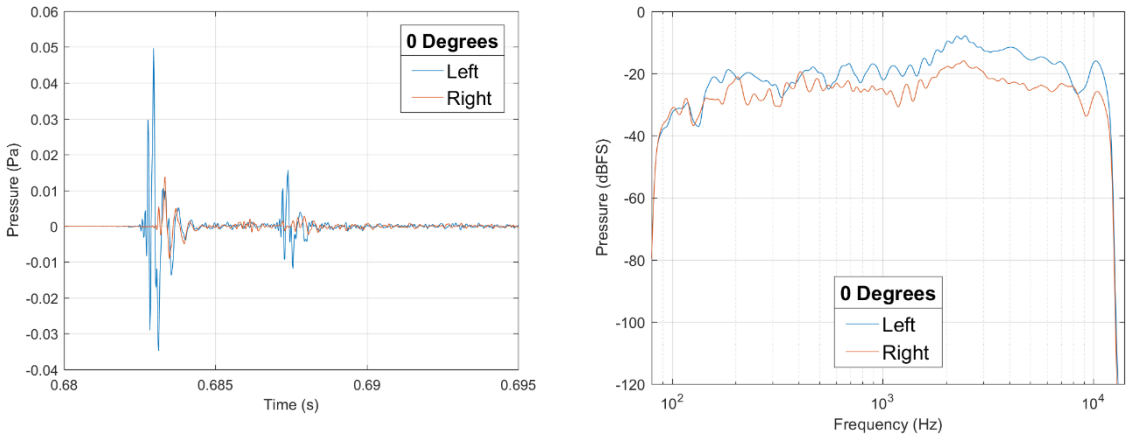
The experimentation process employed in this study was designed to investigate the perceived similarities in spatial audio attributes when representing a static sound source within a sound field using circular harmonics with varying orders over headphones. The experiment uses the ABX method to determine when the order used to present the audio stimuli has become 'spatially equivalent' (i.e. they are perceived the same when compared with a 31st order version). The orders selected for comparison were 1st, 5th, 10th, 15th and 20th; essentially testing for 5 individual null hypothesis (H_0) that a difference between 31st order and 'x' order cannot be determined, setting the significance level at 0.05.

2 METHOD

2.1 The Modelled Room

A room known to the authors was chosen as a space which avoids symmetry in dimensions (10.4m x 5.84m x 2.75m) and construction materials (average absorption coefficient = 0.36); favouring a diffuse field at all positions within the room. A single listener seat is positioned at the coordinates x, y, z = 2.9, 5.25, 1.2 and the sound source positioned at the coordinates x, y, z = 4.93, 3.23, 1.2. The result of this is a sound source at 45 degrees off axis at a distance of 2.87m relative to the forward-facing listener seat position.

Using the tools within the acoustic modelling software, 72 response files were generated which represents a rotation at the receiver location every 5 degrees. The response files are converted into a binaural impulse response (BIR) through convolution with a KEMAR dummy head related transfer function (HRTF) defined within the software, therefore resulting in 72 separate binaural room impulse response (BRIR) files for a source at 45 degrees. A truncated example showing the direct and first reflection at 0 degrees rotation of the head is shown in Figure 1.



(a) **(b)**

Figure 1: (a) BRIR and (b) Magnitude of BRTF at 0-degree head rotation

2.2 Binaural Conversion

Following the work first described by McKeag and McGrath[2] and a more complete higher order description of Ambisonics to Binaural conversion by Politis and Poirier-Quinot[3], a MATLAB script was developed to convert the 72 BRIR files into the 63 circular harmonic equivalent responses (31st order). The script determines the horizontal only spherical harmonic coefficients required to represent the sound field every 5 degrees and the resultant decoder values needed to reproduce a sound field based on the pseudo inverse of this matrix[3]. The BRIRs are weighted by the decoder coefficients with left and right ear responses then summed to construct the circular harmonic impulse responses that represent the simulated sound field for a static sound source at 45 degrees. If a sound source is 'panned' into this sound field, at 0 degrees, the sound field will reproduce the room and the modelled static source at 45 degrees. If a sound source is panned away from that position, then the result will be the rotation of the head within that field. The three pairs of BRIRs generated for a 1st order example are shown in Figure 2, with an example of the reconstruction accuracy of the direct and first few reflections in the time domain shown in Figure 3.

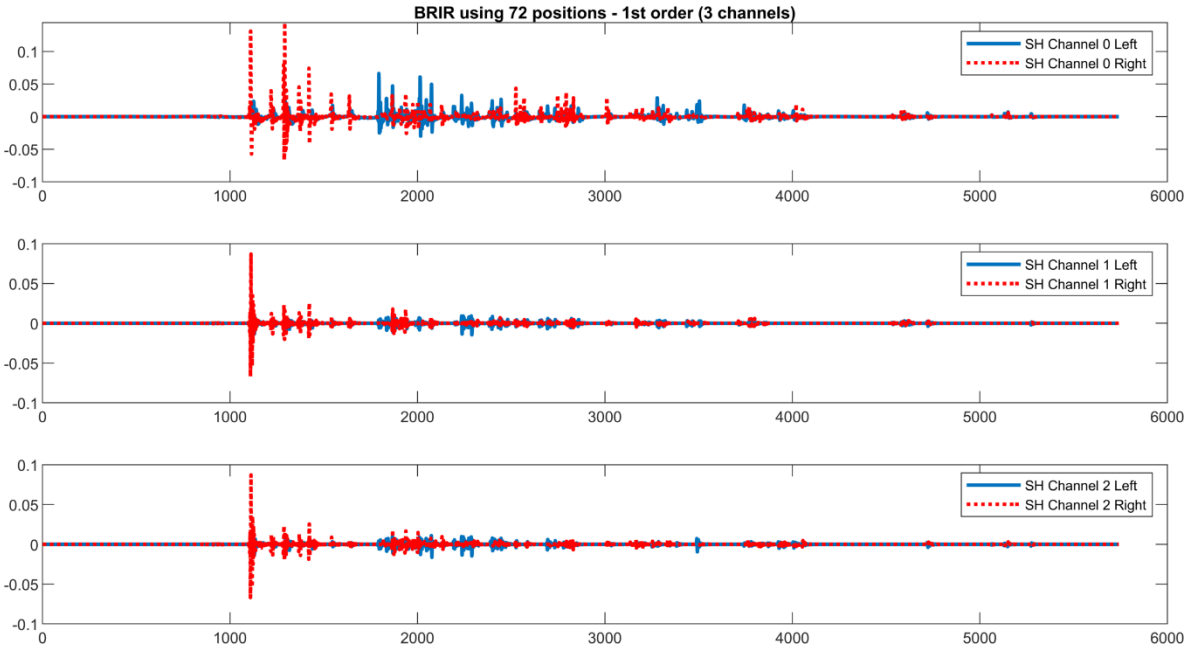


Figure 2: Horizontal Spherical Harmonic (SH) BRIRs for a 3 channel, 1st order auralisation for a fixed source at 45 degrees

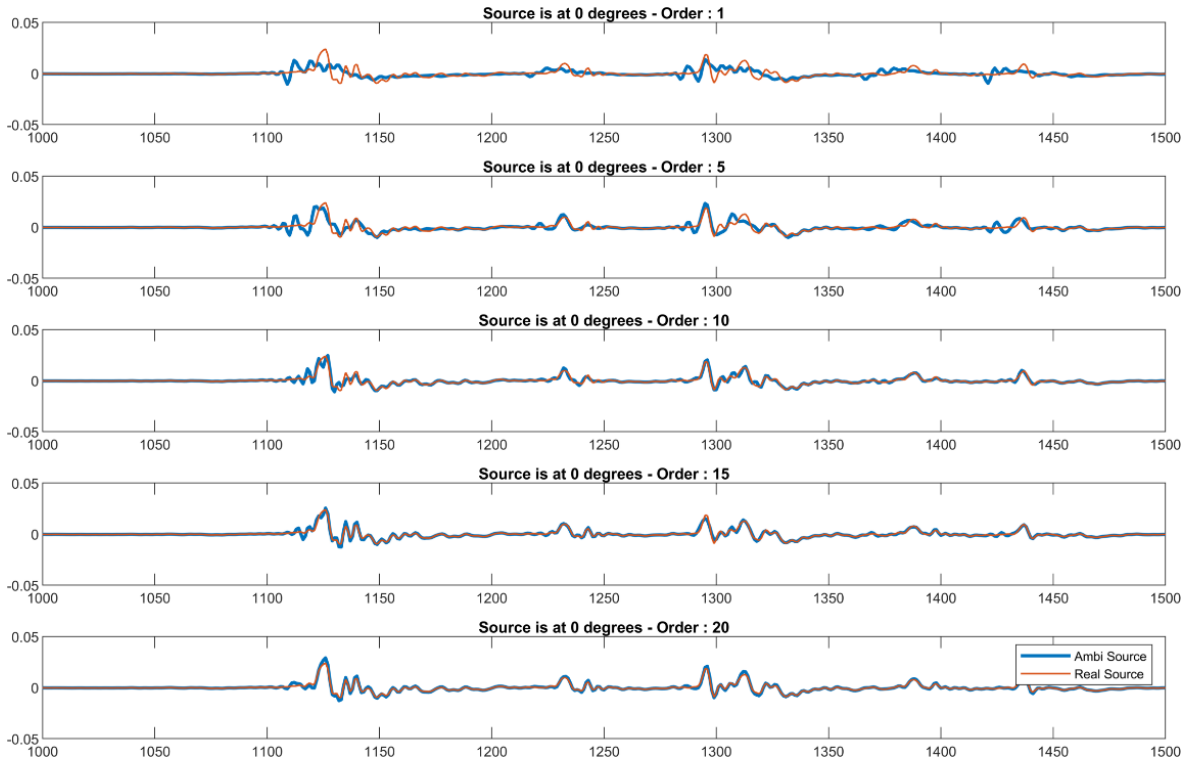


Figure 3: Time domain reconstruction of a source (left BRIR channel shown) at 1st, 5th, 10th, 15th and 20th order (red line) and the original data (blue line)

2.2.1 Diffuse Field Equalisation

Representing sound fields with fewer harmonics will reduce the spatial accuracy with incorrect reconstruction occurring above the spatial aliasing frequency[4]. The frequency response of the system above this frequency will not match the original, and will vary between orders, influencing the listener during perceptual experiments. The diffuse field response can be obtained from averaging the frequency responses of the circular harmonic BRIR's created in section 2.2; the equalisation filter is calculated as the inverse to the obtained response. The averaging and resultant filters are specific to the harmonic order assigned to the process and therefore need to be convolved with the audio output of the same order during the listening stage of the experiment.

2.3 Listening Test

The purpose of this work was attempting to discover what order of circular harmonics would be required before listeners could no longer discern a difference or further improvement in the spatial attributes of the system. The audio stimuli chosen was Vogue – Madonna (1990)[5]; the authors felt it essential to ensure a wide bandwidth of frequency content was used so that judgements given by the listeners were considered across the frequency spectrum and avoid unnecessary repetition of testing for varying narrow band stimuli. The project was structured so the stimuli was routed to six separate filter channels, each representing the orders selected for testing, which were then processed with the diffuse field equalisation dependent on the order selected for playback.

2.3.1 Presenting the Circular Harmonic Impulses

The presentation of audio stimuli requires convolution with the individual circular harmonic impulse responses to create the perceptual experience over headphones for the conditions outlined above. In order to make the convolution of audio stimuli computationally effective and minimise the efforts needed for track management within the software the 3rd party plug-in, X-volver developed by Angelo Farina[6] was used. X-volver is a matrix convolver, enabling a greater number of input channels (maximum of 32) than output channels to be processed in a single instance of the plug-in; therefore, not being constrained to an equal number of input and output channels.

For this project it enabled a maximum of 32 circular harmonic impulses to exist in a single plug-in to receive from an audio stimulus track and output to 2 channels for binaural presentation over headphones. Two instances of the plug-in for convolution purposes were needed to implement the 31st order system which required a total of 63 impulse responses separately for the left and right ears.

2.3.2 Head Tracking and Ambisonic-Panning

A key aspect of presenting spatial audio binaurally is the tracking of the sound scene when rotating the head to match with that experienced in real life. Stable binaural synthesis that is free from unwanted artifacts is achieved using head-tracking devices with low latency (<30ms) and high angular resolution (2°)[7]. Considering these requirements when choosing a suitable head-tracking device, this project employed the MrHeadTracker[8] device as recommended by Rumsey[7].

The head tracker creates and sends MIDI data that is received by a WigWare Very High Order Ambisonic (VHOA) Horizontal Panner (Figure 4). As head movements occur, the panner will mirror the position of the head in the horizontal plane relative to the audio stimulus being passed through the diffuse field equalisation filters. The panner is written to encode the received audio and output up to and including 35th order, although this can be adjusted by use of a slider to a minimum of 1st order with up to 31st order utilised in this project. When the order is changed the number of output channels is altered based on the equation:

$$No.of\ Channels\ = (2 \times Order) + 1 \quad (1)$$

Therefore a 31st order panner will output to 63 channels which are fed to the 63 BRIR pairs stored in the two X-volver instances mentioned in section 2.3.1. As the order is changed within the VHOA Panner the audio is processed with the appropriate set of impulse responses and decoded to present the listener with the stimulus within the simulated sound field in the correct orientation.

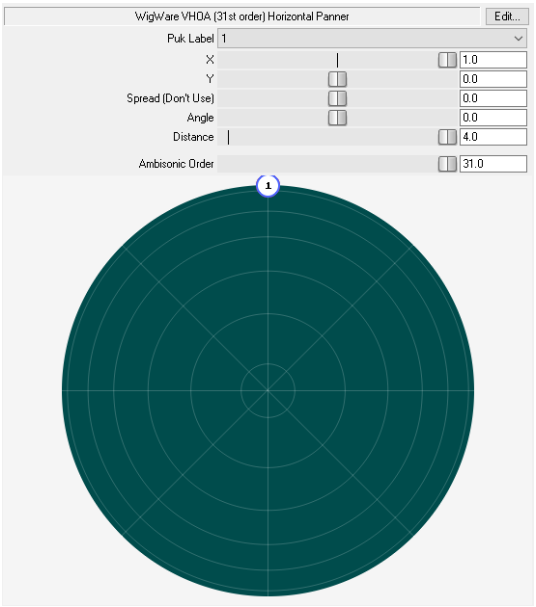


Figure 4: WigWare VHOA Ambisonic Panner Plugin

2.3.3 ABX Testing

An ABX test and graphical user interface (GUI) was designed using MATLAB and utilised Open Sound Control (OSC) to communicate with the Reaper software, shown in Figure 5. OSC communication was linked to software playback controls and also adjustments of the order slider in the VHOA Horizontal Panner.

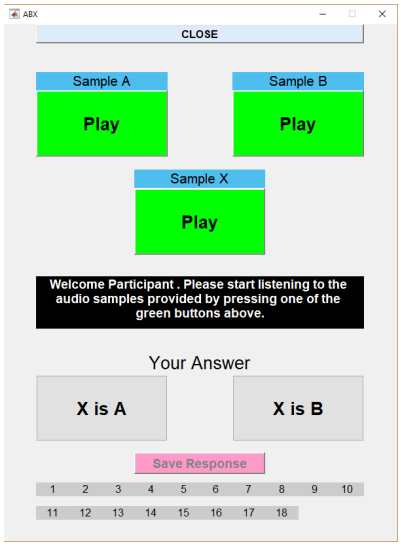


Figure 5: ABX Graphical User Interface

The ABX method is a well-established test method used to determine whether there is a perceived audible difference between two stimuli. To judge if the listener can perceive a difference, critical analysis must be based on both the number of correct answers and the confidence level (in percent) which is derived from the 'p-value'. For psychoacoustic experiments, a 95% confidence level is required to indicate that a perceptual difference between the audio stimuli exists[9].

Subjects were presented a total of 18 trials, with each trial having A or B assigned to the 31st order stimuli. The other choice in each trial consisted of 1st order (2 trials), 5th order (4 trials), 10th order (4 trials), 15th order (4 trials) and 20th order (4 trials) versions of the stimuli. The experiment was not time limited, however the selection of orders and number of trials was chosen as to not induce fatigue, with all participants completing the process within 30 minutes. The GUI was written to randomise the order for each trial to ensure the test conformed to the 'double-blind' requirements. 26 individual participants took part in the experiment; however, the data for Participant 5 did not save successfully and therefore was not used during analysis, resulting in a total of 50 1st order trials and 100 trials for orders greater than 5.

Rumsey[10] states the need to ensure the definitions used for subjective attributes have clarity and proposed a scene-based paradigm to separate attributes associated with the source, environment, and the scene. Using terminology defined by Rumsey[10] in the evaluation of spatial audio, the listeners were asked to make their ABX judgements on the following:

- Source focus – degree to which individual sources can be precisely located in space.
- Source stability – degree to which individual sources remain stable in space with respect to time.
- Scene skew – degree to which a spatial audio scene is skewed from a stated reference position.

These attributes were chosen to elicit responses that will be influenced by the performance of the system when head movement is involved. These attributes were not tested separately. When the head is rotated, the expectation is such that with increased orders and, hence, spatial aliasing frequency, the reproduced sound field ('the scene') and the audio stimulus placed within it ('the source') will elicit an improved response in the stated attributes. To ensure the listeners investigated these attributes and encourage natural interaction with the audio stimuli, it was important to position the sound source at an angle other than on-axis. As previously stated, the source was simulated at a position of 45 degrees to the front-left of the listener; purposely to drive the instinct of moving the head so that the source appeared at 0 degrees, therefore revealing the spatial attributes under test.

3 RESULTS

In this study, the listeners were able to either correctly or incorrectly perceive a difference in the spatial attributes between the 31st order reproduction and the other order presented for each trial. The results that follow will therefore observe the classification of 'correct' (i.e. a difference between the orders was perceived) or 'incorrect' (i.e. a difference between the orders is not perceived).

3.1 Results by Harmonic Order

The responses (as a percentage) given by all participants separated into the five orders investigated are shown in Figure 6. The figure observes the contributions of the 25 individuals together, therefore 50 trials at 1st order and 100 trials for the higher orders.

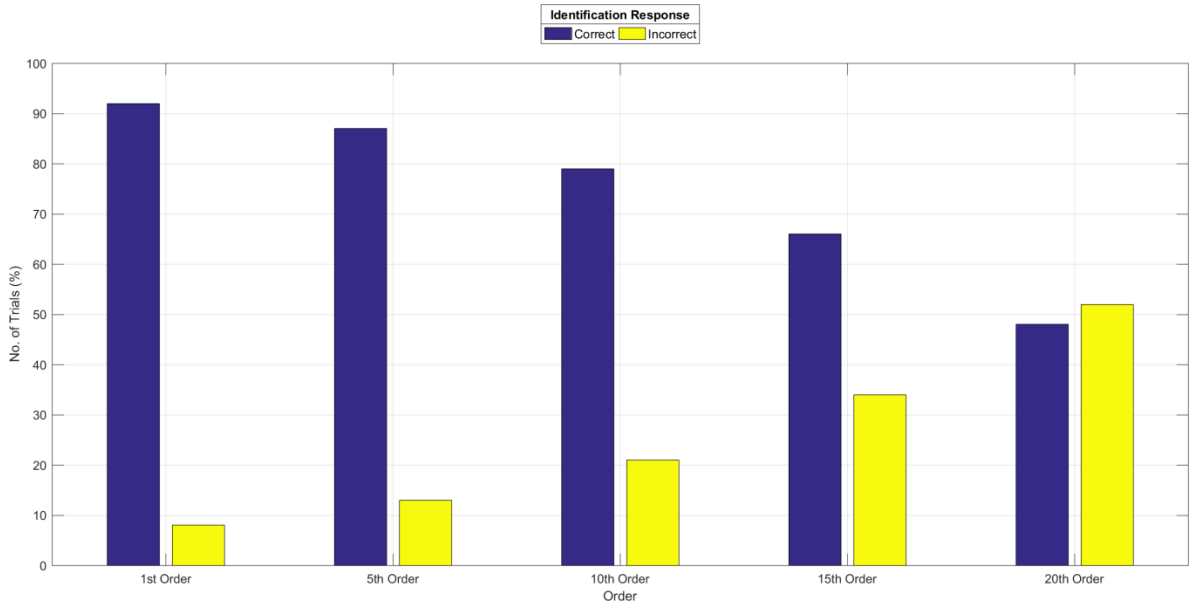


Figure 6: Combined Participant Responses (in %) for each Order

The evidence strongly shows that with each incremental change in circular harmonic order presented to the cohort of participants, the total number of correct responses falls to the point where at 20th order a greater number of responses are incorrectly identified. 92% of the responses were correctly observed to be different when comparing 1st order against 31st order, however collectively the participants only correctly identified 48% of the 20th order stimuli against the 31st order (with a probability of 50% for a forced, chance response). Table 1 presents the same data in non-graphical format with the inclusion of the p-value in the final row for statistical validation.

Table 1: All Trial Responses separated by Order

	1st Order	5th Order	10th Order	15th Order	20th Order
Correct (No.)	46	87	79	66	48
Incorrect (No.)	4	13	21	34	52
Correct (%)	92	87	79	66	48
Incorrect (%)	8	13	21	34	52
P-Value	0.0000	0.0000	0.0000	0.0005	0.0735

It is observed from the data in the above table that only the responses to differentiate the binaural presentation of the spatial attributes for the chosen audio stimuli of 20th and 31st order above the 0.05 significance level. From the results we reject the hypothesis that a difference between 31st order and 20th order can be determined (i.e. a difference is perceived). Therefore, the null hypothesis is rejected for orders 1st, 5th, 10th and 15th as statistically they are shown to have p-values less than 0.05 but is retained for 20th order.

3.2 Results by Participant

Observation of the choices made by each individual is shown in Table 2, where ‘correct responses’ are reported (as a percentage) per order per participant (participant 5 excluded due to save failure).

Table 2: Correct Reponses (in %) by Participant for each Order

Participant	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
1st Order	100	100	100	100	0	0	100	100	100	50	100	100	100	100	100	100	100	100	100	100	100	100	50	100	100	100
5th Order	75	100	100	75	0	75	100	75	100	75	100	100	75	100	50	100	50	100	100	75	75	100	75	100	100	100
10th Order	100	100	75	75	0	25	100	50	100	50	100	75	75	100	50	50	50	100	75	75	100	50	100	100	100	100
15th Order	50	25	75	50	0	50	75	100	100	100	75	75	75	50	25	50	100	50	25	50	75	50	50	100	75	100
20th Order	25	25	50	25	0	25	50	50	75	50	25	50	75	100	50	25	75	75	50	25	0	75	50	25	75	50

Only 3 out of 25 (12%) of the listeners failed to distinguish the difference between 1st order and 31st order, in comparison to 24 out 25 (96%) when listening to the 20th order reproduction. Participant 14 is the only one to correctly identify the 4 trials of 20th order stimuli as being different to the 31st order, however all other participants have been unable to observe this for at least 1 out of 4 trials. It is evident from the table that a reduction in correct responses occurs with an increase in order, as such the percentage of participants that achieve 100% correct responses for a particular order are 88% at 1st order, 56% at 5th order, 48% at 10th order, 24% at 15th order and 4% at 20th order.

It can be observed that just 1 participant accurately identified the difference between 20th order and 31st order for each trial, although in a forced choice test, this could happen by chance. Although not statistically valid due to the small number of trials per order per individual; this observation is a good indicator that when compared with 31st order, the 20th order stimuli is presented with a high enough approximation that makes it perceptually indistinguishable in a high majority of cases.

A known flaw when using ABX testing is the favoring of negative results by participants that may be disinterested or frustrated when determining small differences. An alternative result from that in section 3.1, is observed by screening the three participants (6, 10, 23) who were unable to identify the differences between the 1st and 31st order, shown in Table 3. This action is justified by acknowledging that beyond these listeners all others were able to perceive this change. With the contrast in the performance of the system between these orders being noticeably stark, a lack of motivation is present in these three results and considered acceptable to be discounted.

Table 3: Screened Trial Responses separated by Order

	1st Order	5th Order	10th Order	15th Order	20th Order
Correct (No.)	44	78	72	58	43
Incorrect (No.)	0	10	16	30	45
Correct (%)	100	89	82	66	49
Incorrect (%)	0	11	18	34	51
P-Value	0.0000	0.0000	0.0000	0.0010	0.0829

Although the screening of participants does not affect the order at which the null hypothesis is rejected, it does reinforce previous observations seen in 3.1, as the p-value at 20th order increases from 0.0735 to 0.0829.

4 CONCLUSIONS

The method used to simulate and binaurally present a modelled space correctly responding to listener head rotations using very high order circular harmonics is evaluated using an ABX test; listeners are tasked to evaluate the performance of the system responding to three specific spatial attributes.

The standout observation in the series of results is observed in section 3.1, combining the trials for each order. We show that listeners are unable to perceive the difference between the spatial presentation of the audio stimuli presented at 20th and 31st order, verified statistically with a p-value greater than 0.05. The null hypothesis, that a difference between 31st order and 'x' order cannot be determined, is rejected for orders 1st, 5th, 10th and 15th as statistically they are shown to have p-values less than 0.05.

This paper presents a pragmatic approach to the auralisation of real spaces using very high order circular harmonics to binaural techniques. The impact of these findings on our future work is that we can save on processing time through the reduction in the number of BRIRs (head rotation approximately every 8.5°) required to model the space to a perceptually equivalent standard.

4.1 Future Work

The authors except that some limitations exist in the process outlined in this paper. These include the potential for inaccuracies and losses in the methods used by acoustic modelling software and generic HRTF's, however some research has found that implementing low-latency head-tracking can overcome issues with non-individualised HRTFs[7,12].

The conclusions also assume the participants used were sufficiently aware of the attributes under analysis during the test. In addition, the combination of three spatial attributes into a single judgement limits our ability to observe if a change in order influences one attribute more than another and which of these are providing stronger 'clues' during the listening test. Future work would benefit from the inclusion of participant screening and a training phase to better assure appropriate sensitivity towards the test attributes[13].

Further subjective testing will be required to validate if circular harmonic auralistion of the simulated space over headphones can be used as a viable alternative when seeking to represent acoustic interactions of sound reproduction systems in real spaces.

5 REFERENCES

1. Gerzon, M. A. (1974a) Sound Reproduction Systems. Patent No. 1494751.
2. McKeag, A., McGrath, D. (1996) Sound Field Format to Binaural Decoder with Head-Tracking. 6th Austrailian Regional Convention of the AES, Melbourne, Australia. 10 – 12 September. Preprint 4302. URL: http://www.aes.org/e-lib/browse.cfm?elib=7477
3. Politis, A. and Poirier-Quinot, D. (2016) JSAmbisonics: A Web Audio library for interactive spatial sound processing on the web. Interactive Audio Systems Symposium, York, UK.
4. Wiggins, B. (2017) Analysis of Binaural Cue Matching using Ambisonics to Binaural Decoding Techniques. 4th International Conference on Spatial Audio, 7-10 Sept., Graz, Austria
5. Madonna, Pettibone, S. (1990). Vogue. On *I'm Breathless* [CD]. Sire Records.
6. Farina, A.; *X-volver*. 2017. Available online: http://www.angelofarina.it/Public/Xvolver/
7. Rumsey, F.; Binaural Audio and Virtual Acoustics. *J. Audio Eng. Soc.* **2017**, Vol. 65, pp. 524-528, June.
8. Romanov, M.; Berghold, P., Implementation and Evaluation of a Low-cost Head-tracker for Binaural Synthesis. *AES 142nd Convention*, Berlin, Germany, May 20-23, 2017.
9. Boley, J.; Lester, M., Statistical Analysis of ABX Results Using Signal Detection Theory. *AES 127th Convention*, New York, New York, USA, October 9-12, 2009.
10. Rumsey, F., Spatial Quality Evaluation for Reproduced Sound: Terminology, Meaning, and a Scene-Based Paradigm. *J. Audio Eng. Soc.* **2002**, Vol. 50, pp. 651-666, September.
11. Wiggins, B. (2004) An Investigation into the Real-time Manipulation and Control of Three-dimensional Sound Fields. (2903 downloads) PhD thesis, University of Derby, Derby, UK.
12. Hendrickx, E.; Stitt, P., Influence of head tracking on the externalization of speech stimuli for non-individualized binaural synthesis. *J. Acoustic. Soc. Am.* **2017**, Vol. 141[3], pp. 2011-2023, March.
13. Klein, F.; Neidhardt, A., Training on the acoustical identification of the listening position in a virtual environment. *AES 143rd Convention*, New York, New York, USA, October 18-21, 2017.