Ambisonic Decoder Test Methodologies based on Binaural Reproduction

Enda Bates[1], William David[1], and Daniel Dempsey[1]

[1]*ADAPT Centre, School of Engineering, Trinity College Dublin, Ireland*

Correspondence should be addressed to Enda Bates (ebates@tcd.ie)

This is the Author's Accepted Manuscript (AAM) version of a paper originally published by the Audio Engineering Society at the 150th European Convention, Online, May 25–28th, 2021. The published version of the paper [1] is available in the AES E-library at https://www.aes.org/e-lib/browse.cfm?elib=21050

ABSTRACT

The comparative evaluation of the quality of different Ambisonic decoding strategies presents a number of challenges, most notably the lack of a suitable reference signal other than the original, real-world audio scene. In this paper a new test methodology for the evaluation of Ambisonic decoders is presented, using a virtual loudspeaker, binaural rendering approach. A sample study using a MUSHRA test paradigm and three different types of Ambisonic decoders was conducted and the results analyzed using a variety of different statistical approaches. The results indicate significant differences between decoders for some attributes and virtual loudspeaker layouts.

1 Introduction

The development of spatial audio reproduction techniques and technologies is an ever expanding area due to the growing demand for virtual reality and 360 video content. Ambisonics is a commonly used technique in these new immersive media, and involves two separate stages of encoding and decoding [2]. Therefore, the development of a robust decoder which can provide a good quality of experience for many different sound attributes and for a large range of environments is extremely important when reproducing the recorded or encoded scene. Traditional First Order Ambisonic (FOA) decoders utilize a passive decoding scheme, such as dual-band decoding via a velocity-based and energy-based models to optimize the reproduction in different frequency ranges. Other passive decoding schemes may also be used, as summarized in [3]. More recent parametric decoders (subsequently referred to as active decoders) attempt to improve on this passive method by extracting meaningful parameters from the Ambisonic signal and using different rendering techniques for each parameter [4]. However, the effectiveness of such active, parametric decoding methods based on a time-frequency domain analysis is highly dependent on the precise nature of the audio signal. Hybrid decoding strategies are an even more recent development, and attempt to utilize parametric decoding when it is beneficial to do so, but revert to traditional passive decoding if not [**?**].

Given the range of decoding strategies available, a test methodology for the comparative evaluation of the perceptual quality of such decoders is required. The assessment of audio rendering quality generally involves the comparison of the chosen renderer to some known, quality reference. The evaluation of Ambisonic decoding quality therefore poses a challenge in this regard due to the lack of a suitable reference, other than the real, physical audio scene. In addition, the specific subjective spatial attributes to be evaluated need to be defined, alongside their relationship to the chosen test methodology and reference material.

In this paper, such a test methodology is described and implemented in an initial study involving three, FOA decoders, rendered to binaural using a virtual-loudspeaker approach [5]. The three decoders consist of DAW-based plugins using three distinct decoding strategies based on traditional passive (classic, mode-matching weighting [3]), parametric, and hybrid decoding (subsequently referred to here as passive, active, and hybrid decoders). Although this initial study concentrates on FOA, the proposed methodology could also be utilized for the evaluation of higher order Ambisonic decoders.

2 Background

In many respects, the issue of how to subjectively evaluate different Ambisonic Decoders when no known quality reference is available, is quite similar to the comparison of different spatial audio recording techniques. In such tests, the only reference available as a ground truth is the original sound scene, and therefore are usually conducted without the use of a reference, and based on the comparative rating of different attributes [6, 7]. The evaluation of Ambisonics Decoders

to date either make reference to an internal, recalled judgement of some spatial attribute or scene [8], or perform a comparison between different decoders purely in terms of their relative performance for specific attributes, such as Apparent Source Width (ASW) [9]. In the former case, previous tests have relied on an internal reference based on either sound-scenes which are highly familiar to all listeners (such as traffic and other soundscapes), or scenes which are familiar for a specific set of expert listeners (such as concert hall acoustics).

The preceding discussion illustrates some of the challenges in designing a test methodology for the evaluation of Ambisonic decoders. While comparative studies may be performed without a reference, particularly for specific contexts such as classical music recordings, this approach relies on expert listeners with a strong familiarity with the test material. For the quantitative evaluation of Ambisonic decoders, which can and are used to decode a wide range of spatial audio signals in a variety of different contexts, a different test methodology involving the use of an appropriate reference signal for comparison is therefore required. Certain attributes such as localization accuracy and locatedness can potentially be assessed through the direct comparison with a simple, real audio object, such as a single loudspeaker. However, for other important attributes such as timbre, spaciousness, naturalness, or presence, determining a suitable reference is more challenging. This problem is the principle challenge when designing a test methodology for spatial audio decoders. Fundamentally, how can one assess which spatial audio decoder best represents the original, real sound-scene, if that original scene is not immediately available for comparison? Put simply, what is the available ground truth in any such comparative test?

While loudspeaker reproduction is optimal for the specific decoders being evaluated here, and will be evaluated in a later paper, it raises some significant challenges in terms of the test methodology, most notably the choice of a suitable reference. In addition, such an approach was problematic at this time due to the restrictions associated with the Covid pandemic. In this initial study, the use of binaural reproduction supported the delivery of an online listening test, using headphones, but also simplified the choice of reference signal. In addition, this paper contains an investigation of other relevant factors, such as the specific loudspeaker layout to be used for playback, and the choice of perceptual attributes to be assessed.

3 Test Methodology

The design of this test comprises of 2 groupings of test signals; object-based signals and recording-based signals. As binaural reproduction was used for playback, references could be created through direct binaural rendering of monophonic samples using Head Related Impulse Responses (HRIRs) or through direct binaural recording using a binaural microphone. Similarly, the test stimuli could be created through direct encoding to FOA or by recording using an FOA microphone. Once encoded or recorded, Ambisonic signals were decoded and rendered to binaural using a virtual loudspeaker approach, in which each loudspeaker signal was rendered to binaural via direct convolution with a HRIR [5]. These HRIR's were taken from the SADIE II database [10], which were captured using a Neumann KU-100 binaural microphone in an anechoic chamber, and included post-processing in the form of low frequency compensation and diffuse field equalization. Two loudspeaker layouts were selected for this study, namely a cube and a 11.1 (7.1.4) configuration. Although the commonly used Cube loudspeaker layout is highly efficient in terms of the low number of virtual loudspeakers involved, this along with the lack of speaker positions on the horizontal plane potentially represents a challenging layout for certain attributes, particularly azimuth localization accuracy. For the Cube layout, loudspeakers were placed at +/- 45 degrees elevation, and 90 degree intervals on the horizontal plane starting at 45 degrees. The second configuration chosen was a more cinematic style layout, namely 11.1 (7.1.4). In contrast to the Cube, this layout includes 7 loudspeakers on the horizontal plane (at 0, 30, -30, 90, -90, 150, & -150 degrees azimuth), and 4 overhead loudspeakers (at 45, -45, 135, & -135 degrees azimuth, and 45 degrees elevation).

For the object-based tests, reference signals were created through the direct convolution of anechoic samples with SADIE HRIRs for a given angle(s), shown in fig 1. This therefore eliminated the influence of the binaural rendering as a factor in the comparison of the test signals and reference by using the same HRIR set for both the creation of the reference signal, and for the binaural rendering of the decoded loudspeaker signals. For the recording-based tests, the reference consisted of either a direct binaural recording of a real scene, or an ane-

choic test signal convolved with a BRIR. Direct recordings were taken using TCD's KU-100 binaural mic, while the anechoic samples were convolved with KU-100 BRIRs taken from the MAIR Huddersfield open access library [11]. The test stimuli signals comprised of a direct FOA recording or FOA encoded samples convolved with MAIR RIRs, which were then decoded, and rendered to binaural as before, as shown in figure 2. The difference in HRIRs/BRIRs sets used in the reference creation process and binauralization process resulted in unwanted timbral differences between the 2 sets which had to be compensated for to ensure a fair test was conducted.

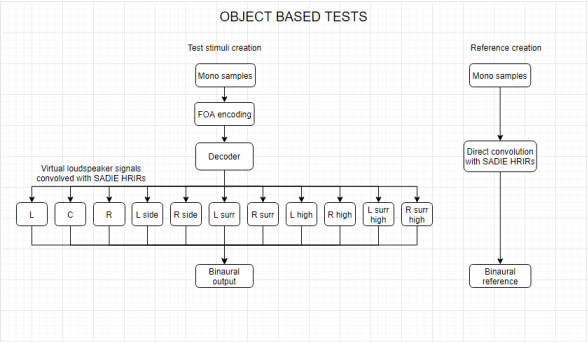

Fig. 1: Test stimuli generation for object-based tests

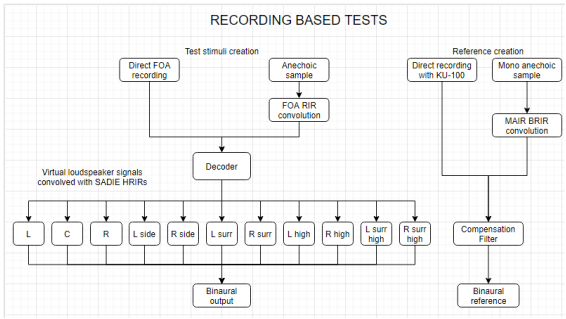

Fig. 2: Test stimuli generation for recording-based tests

3.1 HRIR Post-processing Compensation

The timbral difference between the reference signals and test signals for the recording based tests was attributed to the low frequency compensation and diffuse field equalization present in the SADIE HRIRs but not present in the MAIR HRIRs/BRIRs. A compensation filter was therefore designed and applied directly to the reference recordings to eliminate any timbral differences between the two signal sets, apart from those which resulted due to the decoders themselves. This was achieved through a diffuse-equalisation approach, previously used by McKenzie et al. as a method for improving high frequency reproduction in Ambisonics without resorting to the use of higher orders [12].

A set of SADIE HRIRs and newly recorded, unprocessed KU100 HRIRs (of identical angles) were used to generate a diffuse field for each set. This was calculated using the "computeDiffuseFieldIR" matlab function provided in the 3D3A toolbox developed by Princeton University [13]. This function applies a FFT to a set of HRIRs, calculates the root-mean-square of the HRTFs and applies an inverse FFT to give the final diffuse field response in the time domain, as seen in figure 3. Before compensation, the Genelec speaker response (contained in the new HRIRs through recording) had to be removed from the unprocessed diffuse field. This was achieved by applying an inverse linear-phase FIR filter built from the Genelec response recorded by a GRAS omnidirectional mic. This inverse filter was built using the "invFIR" Matlab function developed by Matthes [14]. This function follows the Nelson-Kirkeby approach which attempts to build a direct inverse of the input through a least-squares with regularisation method. A second inverse linear-phase FIR filter was designed to finally compensate for the difference between the SADIE, and unprocessed diffuse responses. This FIR filter was designed as the length of the signal, 0-20,000Hz range, with a Hanning window, 1/5th octave smoothing and regularisation of 25dB in the pass band and -5dB outside the pass-band. This filter was applied directly to the binaural reference recordings to, in essence, substitute the non-directional timbral information of the SADIE HRIRs to the reference recordings such as to mimic as if the reference recordings were generated with the SADIE HRIRs.

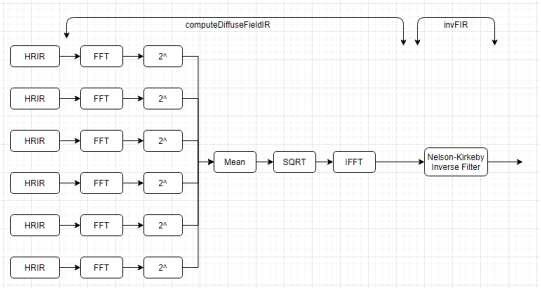

Fig. 3: Diffuse field equalisation signal flow diagram

4 Assessed Sonic Attributes

A number of studies have attempted to collate, define and condense various spatial attributes into a set of relevant and consistently defined attributes [15, 16]. Simple scenes containing a single source in an anechoic environment are highly controllable and describable, which simplifies the evaluation, but limits the number of variables which can be tested. Single sources in a reverberant environment allows for more attributes to be evaluated, while more typical program material containing multiple sources in various locations is far more complex to control, but also more ecologically valid. Certain attributes may become ambiguous or difficult to clearly define for the listener, however, the real-world nature of this material is still important to examine, despite these issues. Clearly there are a large number of spatial attributes which could potentially be evaluated in terms of the perceptual performance of Ambisonic decoders. However, given the need to ensure a reasonable test duration and the avoidance of listener fatigue, some rationalization of these many attributes is required. In this initial study therefore, eight specific attributes common to a number of past studies were evaluated, namely;

- Localization Accuracy - single static source
- Localization Accuracy - multiple static sources
- Locatedness
- Source Distance
- Ensemble Width
- Spatial Impression/Envelopment
- Timbre
- Naturalness/Presence

These sound characteristics were evaluated on the two aforementioned virtual loudspeaker layouts using object-based tests (localization accuracy and locatedness) and recording-based tests (all the remaining attributes).

4.1 Localizational Accuracy - single source

Four localization accuracy tests were used to evaluate the decoders for single sources at between-speaker positions on the horizontal plane for both loudspeaker layouts and two different signal types; namely anechoic speech and drums.These samples were chosen to provide good frequency coverage as well as representing common sound categories. References were created by directly convolving the anechoic samples with a HRIR for the chosen angle(s). Low quality anchors were created by applying a low pass filter at 2kHz to the anechoic samples and rendering to mono, to drastically reduce both timbral and spatial quality [17].

4.2 Localizational Accuracy - Multiple Sources

As two of the decoders evaluated utilize an active, parametric decoding paradigm involving the analysis of the input audio signal, it was important to also evaluate localization accuracy when more than one signal is present. This is particularly relevant when both source signals share similar time-frequency content, and when the source directions are either in close proximity, or in directly opposite directions. To evaluate localization accuracy in these scenarios, four tests were again devised, two for each loudspeaker layout, and for two differing source directions. The first two tests utilized two pseudo-anechoic viola and violin samples from the Mixing Secrets library [18] in close proximity on the horizontal plane (separated by 10 degrees) and in between speaker positions. The last two tests consisted of anechoic oboe and clarinet samples at diametrically opposite directions. References were created by directly convolving the anechoic samples with HRIRs for the chosen angle(s). Anchors were created by downmixing the sample pairs to a monophonic signal and applying a 2kHz low pass filter.

4.3 Locatedness

Locatedness describes the change in the focus or spreading of the source signal in either azimuth or elevation. Two tests were derived to test this attribute (one for each loudspeaker layout), where the test stimuli were compared to the binaural reference for the height and width, or overall spread of the sound source for constant pink noise positioned directly in front of the listener. The anchors consisted of two decorrelated streams of pink noise, encoded into FOA at 90 and -90 degrees azimuth, decoded and binaurally rendered, and then low pass filtered at 2kHz to degrade the frequency content and increase the spatial spread to a maximum.

4.4 Source Distance

For the four distance tests, the same drum and speech samples used in the object-based tests were used, for both loudspeaker layouts, and two different distances (3 and 5m). The test stimuli were created by convolving speech and drums samples with FOA RIRs from the Huddersfield MAIR library [11]) to create a FOA signal which was then decoded and binauralised as

before. Reference were created by convolving the anechoic samples with the MAIR KU100 BRIRs. In this way, the reference signals consisted of a direct binaural rendering of the anechoic test signals, but containing the early reflections, reverberation, and associated distance cues captured in the BRIR's. The anchor signals consisted of the anechoic samples, low pass filtered at 2kHz as before. In this way, the anchor contained no distance cues and are therefore internalized, thereby serving as a suitable zero distance anchor.

4.5 Ensemble Width

Ensemble width describes the horizontal spatial extent of the entire scene, containing multiple sources. While Localization and ASW describe changes in direction or spatial extent of individual sources, the ensemble width describes the overall effect of these potential changes to the entire scene. To evaluate this attribute, short samples of direct binaural and FOA recordings of an *a capella* vocal group were extracted from the 3D-MARCo open-access database [19]. As with the MAIR library, the recordings were made in the St. Paul's concert hall in Huddersfield, using a KU-100 and Sennheiser Ambeo microphone respectively. The anchor consisted of a monophonic, omnidirectional microphone recording taken from the same database, with a low pass filter applied at 2kHz as before.

4.6 Spatial Impression/Envelopment

A number of different attributes can be used to describe the overall spatial impression, particularly as this relates to the perception of enveloping reverberation. Therefore, this attribute was considered here as a single attribute, referred to here as Spatial Impression/Envelopment. The test signals used to evaluate this attribute consisted of direct binaural and FOA recordings from the 3D-MARCo database [19]. Samples of a recording of a classical trio (piano, violin, and cello) were used to ensure a degree of variety in frequency content. As with the Ensemble Width tests, the anchor consisted of a monophonic, omnidirectional microphone recording taken from the same database, with a low pass filter applied at 2kHz.

4.7 Timbre

Four tests were designed to investigate timbral quality, using two differing signal types, and for both loudspeaker layouts. Given the high familiarity of the timbre of human voices, a different sample from the *a capella* vocal group recording from the 3D-MARCo database [19] was used for the first two timbre tests. The anchor was created by taking the centre channel of the Decca tree arrangement from the recording as a monophonic sample, and low pass filtering it at 2kHz. The second pair of timbre tests consisted of a synthetic scene created using anechoic, multitrack recordings by the Anna Blanton Ensemble, taken from the mixing Secrets library[18]. Anechoic samples of a lead vocal, double bass, ukelele, violin, and congas were combined with a synthesized reverb and directly encoded to binaural to create the reference signal by convolution with HRIR's [10] for a range of azimuth angles (0, 45, 75, 285, 315 degrees) and 0 degrees elevation. The same anechoic samples were also directly encoded to FOA and subsequently decoded using the three decoders under evaluation and both loudspeaker layouts in order to generate the three test signals.

4.8 Naturalness/Presence

Attributes such as naturalness, realism, and presence were also considered here as a single attribute, referred to as Naturalness/Presence. To ensure ecological validity it was decided to create a new recording of a exterior scene containing a highly familiar soundscape, namely passing traffic, pedestrians, and other city street sounds. The recording was made concurrently using KU-100 and Sennheiser Ambeo binaural and FOA microphones. The anchor signal was derived from the omnidirectional, monophonic W channel of the FOA recording, with a low-pass filter at 2kHz applied.

5 Listening Test Implementation

While various test paradigms may be used for blind listening tests, a Multiple Stimuli with Hidden Reference and Anchor (MUSHRA) style test (ITU-R BS.1534-1) was utilized here due to its ability to allow multiple test signals to be rated per test instance. The AudioLabs webMUSHRA interface (shown in figure 4) [20] was used for the test, and supported standard features such as synchronous playback of the reference and test signals, and user controlled loop points. As remote testing was necessitated due to the Covid pandemic restrictions, the webMUSHRA test interface was installed on a virtual machine hosted in Trinity College Dublin and made available to test subjects via a URL. Test subjects were instructed to rate the stimuli with respect to the reference for the given characteristic under test using the vertical sliders. Apart from the given reference, five stimuli are presented to the listener per test, consisting

Test No	Test	Source	Layout	Angle (Azi,Ele)
1	Localization, Single Source	drums	7.1.4	60:0
2	Localization, Single Source	speech	7.1.4	-120:0
3	Localization, Single Source	drums	Cube	90:0
4	Localization, Single Source	speech	Cube	-60:0
5	Localization, two sources, close proximity	violin viola	7.1.4	40,0 ; 50,0
6	Localization, two sources, close proximity	violin viola	Cube	-80,0 ; -90,0
7	Localization, two sources, diametric positions	oboe clarinet	7.1.4	-20,0 ; 160,0
8	Localization, two sources, diametric positions	oboe clarinet	Cube	90,0 ; -90, 0
9	Locatedness	pink noise	7.1.4	0,0
10	Locatedness	pink noise	Cube	0,0

Table 1: Sonic attributes for object-based tests

Test No	Test	Source	Layout	Angle (Azi,Ele)
1	Source Distance	drums	7.1.4	KU-100 and Ambeo RIRs 5m / 10 degrees azi
2	Source Distance	speech	7.1.4	KU-100 and Ambeo RIRs 5m / 10 degrees azi
3	Source Distance	drums	Cube	KU-100 and Ambeo RIRs 3m / 0 degrees azi
4	Source Distance	speech	Cube	KU-100 and Ambeo RIRs. 3m / 0 degrees azi
5	Ensemble Width	vocal group	7.14	KU-100 and AmbiX recordings
6	Ensemble Width	vocal group	Cube	KU-100 and AmbiX recordings
7	Spatial Impression, Envelopment	classical trio	7.14	KU-100 and AmbiX recordings
8	Spatial Impression, Envelopment	classical trio	Cube	KU-100 and AmbiX recordings
9	Timbre	vocal group	7.14	KU-100 and AmbiX recordings
10	Timbre	vocal group	Cube	KU-100 and AmbiX recordings
11	Timbre	ensemble	7.14	KU-100 and Ambeo RIRs. various on-stage angles - 3m
12	Timbre	ensemble	Cube	KU-100 and Ambeo RIRs. various on-stage angles - 3m
13	Naturalness/Presence	exterior scene	7.14	KU-100 AMBEO recording
14	Naturalness/Presence	exterior scene	Cube	KU-100 AMBEO recording

Table 2: Sonic attributes for recording-based tests

of the passive, active, and hybrid decoder signals, a hidden reference and a hidden anchor. After completion of the test items, a short questionnaire was presented to gather data on gender, age, listening test experience, and professional working experience in spatial audio.

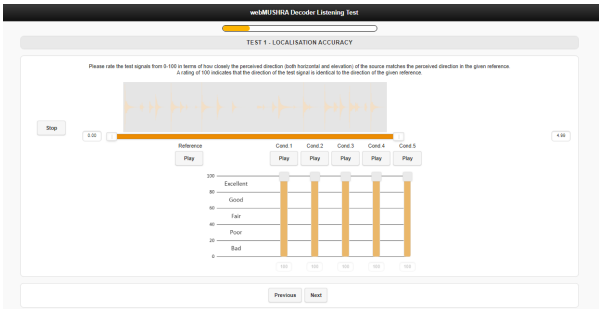

Fig. 4: WEBMUSHRA interface from Listening Test

5.1 Listening Test Procedure

Subjects were asked to rate all stimuli in terms of similarity to the available reference, for the given attribute under evaluation. 25 sets of results were gathered, 22 male and 3 female. The subjects consisted of academics with past experience of similar tests, and professional audio engineers. 9 subjects fell in the age range of 20-30, 11 subjects from 30-40, 3 subjects from 40-50, 1 subject from 50-60 and there was 1 60+ participant. In terms of experience, 6 people stated they had little experience with critical listening tests, while 19 stated they were very experienced. 6 people stated they have not worked professionally in audio, with 4 people stating they had little experience (1-3yrs), 6 people stating they had a fair amount of experience (3-5 years) and 9 people stating they have extensive experience working in audio (5+ years).

The ITU standard for the MUSHRA test methodology [17] specifies a post-screening method which excludes subjects who assign a very high grade to a significantly impaired anchor signal, and who frequently grade the hidden reference as though it were significantly impaired. Specifically, a test subject should be excluded from the final analysis if they rate the hidden reference condition lower than a score of 90 for more than 15% of the test items. As a result, 4 subjects were excluded from the analysis as they rated the reference <90 for over 15% of the tests (which in this case represents 4 individual tests) [17, 21].

6 Results and Analysis

The data collected for each test includes the type of decoder, the virtual loudspeaker layout used, and signal type (i.e., drums, vocals, etc.), although only 3 tests contained more than one signal type (specifically Localization accuracy, Source Distance, and Timbre). As suggested by [22], boxplots are primarily used to visualize and analyse the data here, rather than confidence intervals. [1] Boxplots display the whole data, convey skewness, and easily identify outliers. Confidence intervals on the other hand only display the mean and a multiple c of the standard error. It displays the range of values for which a t-test of 0 difference would fail to reject the null hypothesis at confidence level corresponding to c (for example, for a large sample size, $c = 2$ roughly implies a 95% confidence level). A visual inspection of the medians and interquartile ranges clearly indicated that the reference performed better than all three decoders, for all attributes, apart from locatedness. Similarly, the plots also indicated that all three decoders outperformed the anchor for all attributes, with the exception of locatedness for the passive decoder and the 7.1.4 layout.

In addition, the data was also analysed using a Repeated Measures Analysis of Variance (rmANOVA) as recommended by the ITU [17]. A mixed model [23] with the subjects as random effects was also used, as recommended by [24], but since the data has no missing values and is perfectly balanced, both methods gave the exact same p-values. It should also be noted that the presence of both dependent and independent samples is here handled by the usage of a repeated measures ANOVA / mixed effects model, as discussed in [22].

For the repeated measures ANOVA a three-way fit for the spatial attributes which have more than one signal was applied, and two-way fits were applied for the remaining attributes. Locatedness was not included due to its violations of the ANOVA assumptions. However, the box plot visualization for locatedness clearly indicates that there is a strong interaction effect between the loudspeaker layout and decoders, and that the 7.1.4 layout massively outperforms the Cube layout for locatedness when using the active and hybrid decoders, to the point where a hypothesis test of 0 difference was not necessary.

[1] Data visualization in the form of boxplots of the scores separated by the decoder, are available at https://www.endabates.net/AmbiDecoderTestMethodPlots.html

In order to offset the compounding Type-I error the Bonferroni correction was used to adjust the significance threshold, with the aim of an experiment wide Type-I error rate of 5%, as is standard. Starting with this commonly accepted value of 5%, and then dividing by the number of concurrent hypothesis tests running $(= (3 \times 5) + (7 \times 3) = 36)$, this results in a significance threshold (rounding up slightly) of 0.0014.

The three-way ANOVAs indicate that Localisation accuracy has the most significant discrepancy in scores arising, with the decoder, signal, and interaction between decoder and speaker all showing significance (all p-values <0.0001). The interaction between speaker and signal is also very close to the threshold (p-value 0.0024). Timbre also displays significant speaker and signal effects (p-values <0.0001). The two-way ANOVAs indicate that for localisation accuracy, two sources, close proximity, the interaction effect of the loudspeaker layout and the decoders is significant (p-value <0.0001). The independent loudspeaker layout effect was also very close to significance in this case (p-value 0.0016). The independent loudspeaker layout effect does manage to breach the significance threshold for the spatial impression tests and naturalness/presence tests (p-values 0.0005 and <0.0001).

As the ITU recommends [17], post-hoc analyses for significant ANOVA results, such as Tukey's Honest Significant Difference test, should also be applied to MUSHRA test data. This enables combinations of factors to be contrasted to further explore which differences are truly significant. Here this was achieved using pairwise p-value plots, which use the Tukey adjustment to compute p-values for differences. As such, smaller p-values imply a more significant difference. As with the ANOVA, the significance threshold was adjusted from the standard value of 0.05. Naive multiple comparison tests (such as multiple paired t-tests) without adjusting the significance threshold can result in an inflated error rate since every comparison has a 5% chance of type-I error [22]. Tukey tests circumvent this problem by computing a p-value in such a way that it accounts for the inflated error. However, as multiple Tukey tests are conducted here, each with an error rate equal to the significance threshold, a slightly reduced value of 0.01 was used.

6.1 Discussion

For Localization accuracy, the box plot reveals that the active and hybrid decoders perform well for the 7.1.4 layout achieving "good" quality for the drum signal, and "good/excellent" for the voice signal. This distinction is less evident for the passive decoder, as can be seen from the increased variability for both signal types, with the drum signal achieving a rating of "poor/fair" and the voice signal "fair/good". The results for both signal types indicate little difference in performance for the Cube layout for all three decoders, with all three rated as "fair/good". The results of the Tukey HSD test for localization accuracy indicate that for the 7.1.4 layout, the active and hybrid decoders could not be distinguished for either signal type (p-values 1.000 and 0.9995), but both clearly outperform the passive decoder for the drum signal (p-value <0.0001) and the voice signal (p-values 0.0005 and <0.0001) and this layout.

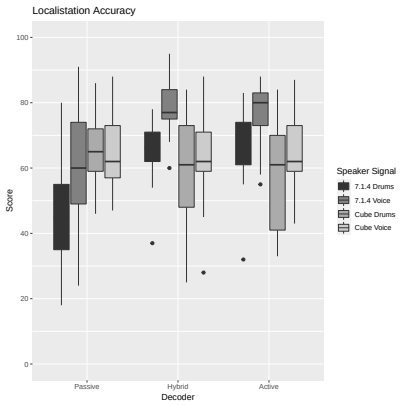

Fig. 5: Box plot - single source localization accuracy

The hybrid-7.1.4-voice combination performed significantly better than hybrid-7.1.4-drums (p-value 0.0067), hybrid-cube-voice (p-value 0.0008), hybrid-cube-drums (p-value <0.0001), active-cube-voice (p-value 0.0013), hybrid-cube-voice (p-value 0.0008), and various combinations of the passive decoder. The active-7.1.4-voice combination performed similarly well, but to a slightly lesser extent, namely outperforming the active-cube-drums (p-value <0.0001), hybrid-cube-drums (p-value 0.0003), and again various combinations of the passive decoder. For the passive decoder, significant differences are also apparent for the drum signal and the 7.1.4 layout, with most other layouts and decoders outperforming this configuration, suggesting that the passive decoder struggled with the 7.1.4 layout and particularly the drums signal, but also the voice signal to an extent. These results suggest that the hybrid

decoder marginally outperformed the active decoder for localization accuracy. However, the general trend is that both active and hybrid decoders outperformed the passive decoder for the 7.1.4 layout for both signal types, with a general improvement in performance for voice over drums for all decoder types and this layout. In contrast, for the Cube layout, little difference was apparent for either signal or decoder type, with the performance generally worse than the 7.1.4 layout for the voice signal, most notably with the hybrid and active decoders. Overall these results suggest that the active and hybrid decoding methods provide some improvements in localization accuracy for the 7.1.4 loudspeaker layout, but this improvement is not evident for the less ideal, yet still commonly used Cube layout.

The analysis revealed no significant differences or interaction between decoders, or loudspeaker layout for localization accuracy for two sources at diametrically opposite positions, with all three decoders rated as "good/excellent". In contrast, for localization accuracy with two sources in close proximity, the 7.1.4 layout obtained a score of "Good/Excellent" for the active decoder, and "Good" for the hybrid decoder, and with a reduction in quality ("Fair/Good") for the Cube layout. For the passive decoder, the opposite is the case, with the Cube layout rated as "Good" while the 7.1.4 layout is only rated as "Fair/Good". The Tukey HSD test also indicates that both the active and hybrid decoders significantly outperformed the passive decoder for the 7.1.4 layout (p-values 0.001 and <0.0001 respectively), and also hybrid-cube (p-values 0.0068 and <0.0001 respectively). However, in addition the hybrid-7.1.4 combination also outperformed active-cube (p-value 0.0001). Much as with localization accuracy, this suggests again that the hybrid and active decoding methods perform better with 7.1.4 layouts, compared to the Cube, with a marginally greater improvement in performance by the hybrid decoder for this layout, compared to the Cube.

As mentioned previously, the distribution of data for the locatedness test does not support ANOVA analysis. However, the box plots illustrate the clearly apparent interaction between the loudspeaker layout and decoder type. For locatedness and the 7.1.4 layout, both the hybrid and active decoders were rated as "excellent", with both decoders centered very close to the highest possible score of 100. For the passive decoder, 7.1.4 layout, locatedness was rated significantly worse as only "Poor/Fair". In contrast, for the Cube layout, little difference is evident between decoders, with all three rated as "Fair", but a great deal of variability for both the hybrid and active decoders. This result suggests the active and hybrid decoding methods do indeed provide the expected improvement in locatedness and the focusing of sources when compared to traditional, passive FOA decoding. However, this improvement is only evident in the 7.1.4 loudspeaker layout, and not the Cube. The results suggest that both the active and hybrid decoders are largely indistinguishable from the reference in this regard, at least for this particular layout.

The analysis revealed no significant differences between factors for source distance, with all being rated as "good" or "good/excellent".

Similarly, no significant differences were indicated for ensemble width, with the Cube layout rated as "good" for all decoder types. For the 7.1.4 layout the hybrid decoder was also rated as "good", and the passive and active decoders as "good/excellent", although this difference was not statistically significant.

For spatial impression/envelopment the Cube layout was rated as "good/excellent" for all decoder types. In contrast, the 7.1.4 layout was rated as "good" for the passive decoder, but only "fair/good" for the hybrid and active decoders, although the analysis suggests that there was no significant differences between decoder types. However, in the Tukey HSD test, all three decoders and the Cube layout are quite close to the significance threshold when compared to the hybrid-7.1.4 combination. In particular, the passive-cube (p-value 0.0104) and active-cube (p-value 0.0109) are extremely close to the significance threshold, which perhaps tentatively suggests a slight decrease in performance for the hybrid-7.1.4 layout for this attribute, compared to the other combinations.

Similarly for Naturalness/Presence, the Cube layout was rated mostly as "good" for all decoder types, while in contrast, the 7.1.4 layout was rated as only "fair/good" for all three decoders. Notably, the hybrid-7.1.4 combination was significantly different and outperformed by all three decoders for the Cube layout, specifically passive-cube (p-value 0.0025), active-cube (p-value <0.0001), and hybrid-cube (p-value 0.0001). The results for the active decoder were somewhat similar but to a lesser extent, with the active-7.1.4 revealed to be significantly different and performed worse than the active-cube combination (p-value 0.006). This again suggests an improved performance with the Cube

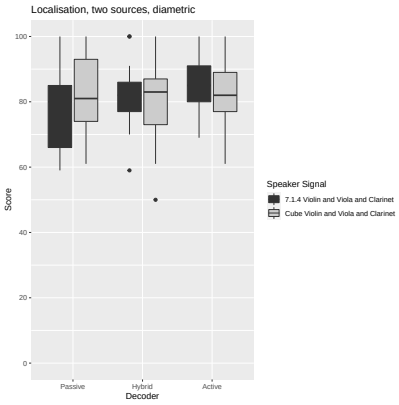

Fig. 6: Box plot for sound localization, 2 sources diametrically opposed

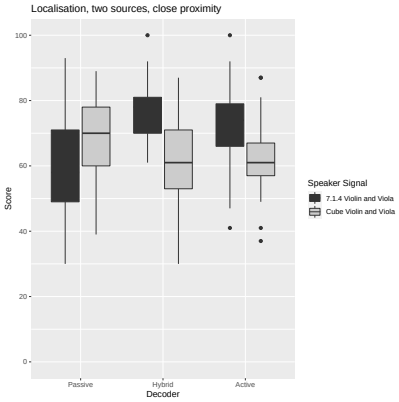

Fig. 7: Box plot for Sound localization, 2 sources close proximity

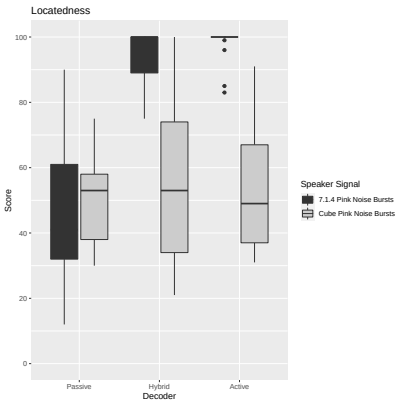

Fig. 8: Box plot for Locatedness

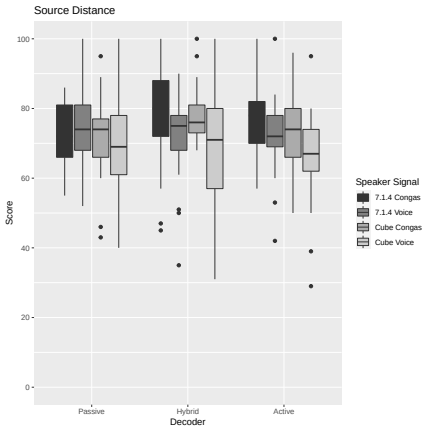

Fig. 9: Box plot for Source Distance

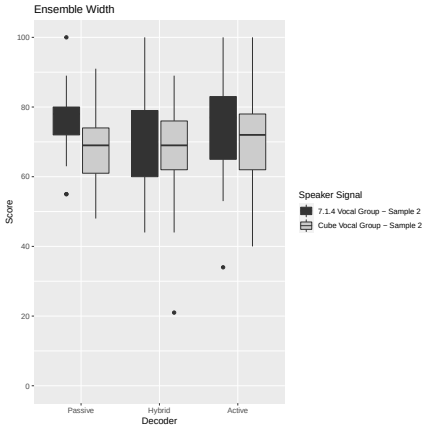

Fig. 10: Box plot for Ensemble width

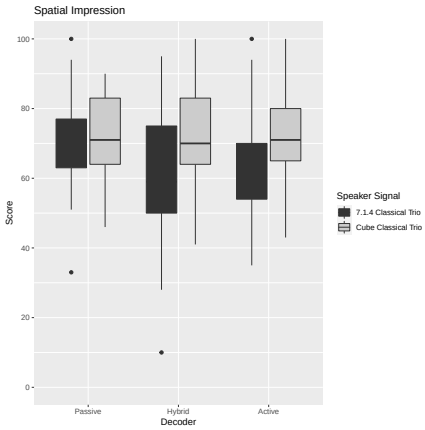

Fig. 11: Box plot for Spatial impression

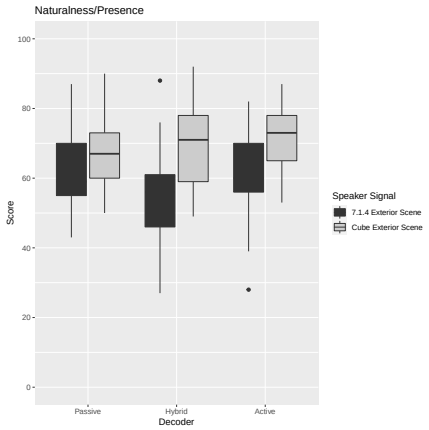

Fig. 12: Box plot for Naturalness/Presence

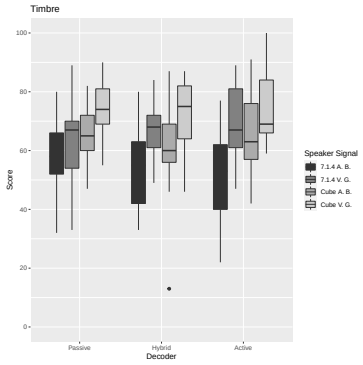

Fig. 13: Box plot for Timbre

layout, compared to 7.1.4 for all decoder types for this attribute, with a slightly greater decrease in performance for the hybrid decoder with the 7.1.4 layout, compared to the other decoders.

For Timbre and the Cube layout, the Vocal Group stimulus was rated as "good/excellent" for all decoder types, while the Anna Blanton Ensemble was rated as "fair/good" for all three decoders. For the 7.1.4 layout, the Vocal Group stimulus was rated as "good" for the active and hybrid decoders, and "fair/good" for the passive decoder, while the Anna Blanton Ensemble was rated as "fair/good" for all three decoders. The results of the TUKEY HSD tests for timbre are complex and so are only summarized here. However, the results indicate that the Cube layout and vocal group sample achieved significantly better results than most other combinations. In addition, the 7.1.4 layout and the Anna Blanton ensemble performed relatively poorly, particularly for the active and hybrid decoders. More specifically the active-7.1.4-ABensemble and hybrid-7.1.4-ABensemble combinations were significantly worse than the cube-ABensemble combination for both the active and passive decoders, and also the hybrid-7.1.4-vocal group combination.

Notably, the two test signals used for the timbre tests differed significantly. Specifically, the vocal group is a direct, live recording using binaural and FOA microphones in a reverberant church, while the Anna Blanton Ensemble was synthetically constructed using anechoic instrument and voice signals, encoded using HRIRs and FOA and combined with synthetic reverberation. Therefore, it may be that this reduction in timbral quality was due to the process used to synthesize the ensemble scene.

7 Conclusion

In summary, the overall results of the listening test indicate no significant differences between the decoders, or indeed the loudspeaker layout used for attributes such as ensemble width and source distance. The results clearly indicate that the hybrid and active decoding methods outperformed traditional passive decoding for certain attributes such as localization accuracy, with single or multiple sources, and locatedness, but only when the loudspeaker layout is well matched to this decoding method (i.e. 7.1.4). These improvements are not evident with the Cube layout however. In addition, there is some evidence that this improvement in performance for these attributes was marginally stronger for the hybrid decoder, than the active decoder.

While no significant differences were revealed between decoders for timbre, spatial impression/envelopment, or naturalness/presence, the Cube layout was rated higher than the 7.1.4 layout (although only for one signal type in the case of timbre) for these attributes. In addition, the hybrid decoder suffered a marginally greater reduction in performance than the active decoder for the 7.1.4 layout for naturalness/presence, and to lesser extent for spatial impression/envelopment. However, overall the result suggest a general trend for an improved performance for the Cube layout compared to the 7.1.4 layout for all three decoders. This perhaps suggests a slight trade-off between improvements in localization and focusing of sources, and other attributes, particularly those related to the impression of the room

or environment. In future tests, particularly those using real loudspeaker reproduction, it may be worth investigating different 2D and 3D loudspeaker layouts to determine if these results are also apparent in this context. It should also be noted that a negative influence of the generic HRTFs and lack of headphone equalization on attributes cannot be discounted.

In conclusion, the results of this initial phase of the project suggest a number of considerations in terms of the design of a test methodology for Ambisonic decoders. Firstly, while tests such as this can involve a very large number of different perceptual attributes, decoder configurations, and test signals, clearly some rationalizations of these many factors are needed to reduce the overall test duration. In this regard, the consistent results uncovered here for the different decoders between different sets of attributes, suggest that this reduced list of test attributes was a suitable choice, with a manageable duration and encompassing most of the relevant factors to a reasonable degree. However, the lack of significant differences reported for two attributes in particular, namely source distance and ensemble width, perhaps suggest that these attributes may be discounted in future tests, if necessary.

Overall, the results suggest there are both benefits and drawbacks to the use of headphone reproduction for the comparative assessment of the performance of different Ambisonic decoding methods. The clear benefit to this approach consists in the ability to create suitable reference signals for comparison, which is much simpler in this context than for loudspeaker reproduction. In addition, the use of the same HRIRs (whether using actual HRIRs or direct microphone recordings and virtual loudspeaker rendering) for the creation of both the reference and decoded test signals appears to be a valid approach, as evidenced by the reported and perhaps expected improvement in performance for the hybrid and active decoders for certain attributes such as localization and locatedness, when compared to traditional decoding methods. One notable result in this regard is the reported differences in performance between the two loudspeaker layouts. The better performance of the active and hybrid decoders for the 7.1.4 layout suggests that the choice of loudspeaker layout is highly significant for such decoding methods. In addition, the better performance for the Cube layout for all decoders for spatial impression/envelopment, naturalness/presence, and to a certain extent timbre was an interesting result.

It should also be noted that further improvements to the use of headphone reproduction in such test methodologies could potentially be applied, and may reveal further distinctions between the decoding methods. In particular, the use of headphone equalization, headtracking, and potentially personalized HRIRs may help to reveal subtle distinctions in performance, which were not revealed here. However, these features were not practically realizable at this time due to constraints associated with the Covid lockdown and the necessity of an online listening test. In addition, the use of personalized HRIRs would make the use of recorded test signals significantly more complex, and logistically highly demanding, perhaps unrealistically so. More generally, it cannot be discounted that other, subtle differences in performance, for example between the hybrid and active decoders, were masked to an extent by the headphone-based reproduction method. Therefore a similar test using loudspeaker reproduction is planned for the next part of this project. The most notable challenge in designing such a test is certainly the design and capture of a suitable reference signal, something which is far more challenging in this context when compared to headphone reproduction. Nonetheless, the results reported here represent a valuable first step, particularly in terms of the choice of perceptual attributes to be evaluated. In addition, the specific results presented here did reveal some important differences in performance between the different decoders, and suggest a number of avenues for future research and the development of a standardized test methodology for the assessment of Ambisonics rendering quality.

8 Acknowledgement

This research was conducted with the financial support of Science Foundation Ireland under Grant Agreement No.13/RC/2106 at the ADAPT SFI Research Centre at Trinity College Dublin.

References

[1] E. Bates, W. David, and D. Dempsey, "Ambisonic decoder test methodologies based on binaural reproduction," *Audio Engineering Society - 150th Audio Engineering Society Convention*, May 2021.

[2] M. A. Gerzon, "General metatheory of auditory

localisation," *Journal of the Audio Engineering Society*, March 1992.

[3] F. Zotter and M. Frank, *Ambisonics: A Practical 3D Audio Theory for Recording, Studio Production, Sound Reinforcement, and Virtual Reality*. Springer Topics in Signal Processing, Springer International Publishing, 1st ed., Jan 2019.

[4] L. McCormack and S. Delikaris-Manias, "Parametric first-order ambisonic decoding for headphones utilising the cross-pattern coherence algorithm," *EAA Spatial Audio Signal Processing Symposium*, 09 2019. https://doi.org/10.25836/sasp.2019.26.

[5] M. Noisternig, T. Musil, A. Sontacchi, and R. Holdrich, "3d binaural sound reproduction using a virtual ambisonic approach," in *IEEE International Symposium on Virtual Environments, Human-Computer Interfaces and Measurement Systems*, pp. 174–178, 2003.

[6] F. Camerer and C. Sodl, "Classical music in radio and TV - a multichannel challenge. the ORF surround listening test," ORF Austrian Broadcasting Corporation, Nov 2001.

[7] R. Jacques, M. Fleischer, S. Fuhrmann, B. Steglich, U. Reiter, and H. Kutschbach, "Empirical comparison of microphone setups for 5.0 surround," *22nd Tonmeistertagung of the VDT, Hannover, Germany*.

[8] C. Guastavino and B. F. G. Katz, "Perceptual evaluation of multi-dimensional spatial audio reproduction," *Journal of the Acoustical Society of America*, vol. 116(2), pp. 1105–15, 09 2004.

[9] G. Martin, W. Woszczyk, J. Corey, and R. Quesnel, "Controlling phantom image focus in a multichannel reproduction system," *107th Audio Engineering Society Convention*, September 1999.

[10] C. Armstrong, L. Thresh, and G. Kearney, "Sadie II database." https://www.york.ac.uk/sadie-project/database.html, 2018.

[11] H. Lee and C. Millns, "Microphone array impulse response (MAIR) library for spatial audio research," October 2017. https://github.com/APL-Huddersfield/MAIR-Library-and-Renderer.

[12] T. McKenzie, D. T. Murphy, and G. Kearney, "Diffuse-field equalisation of binaural ambisonic rendering," *Applied Sciences*, vol. 8, no. 10, 2018.

[13] R. Sridhar and J. Tylka, "3d3a matlab toolbox." https://github.com/PrincetonUniversity/3D3A-MATLAB-Toolbox.

[14] Matthes, "Inverse fir filter." https://uk.mathworks.com/matlabcentral/fileexchange/19294-inverse-fir-filter.

[15] J. Berg and F. Rumsey, "Spatial attribute identification and scaling by repertory grid technique and other methods," *Audio Engineering Society 16th International Conference: Spatial Sound Reproduction*, 1999.

[16] S. Le Bagousse, M. Paquier, and C. Colomes, "Families of sound attributes for assessment of spatial audio," *Audio Engineering Society Convention 129*, November 2010.

[17] International Telecommunications Union, *Method for the Subjective Assessment of Intermediate Quality Level of Audio Systems*, 2014. ITU-R Rec. BS.1534-2.

[18] Camebridge Music Technology, *The 'Mixing Secrets' Free Multitrack Download Library*, 2020. https://www.cambridge-mt.com/ms/mtk/#topAnchor.

[19] H. Lee and D. Johnson, "An open-access database of 3d microphone array recordings," *147th Audio Engineering Society Convention: AES 2019*, October 2019.

[20] International Audio Laboratories Erlangen, *webMUSHRA*, 2020. https://www.audiolabs-erlangen.de/resources/webMUSHRA.

[21] ITU-R, "Method for the subjective assessment of intermediate quality level of audio systems," Tech. Rep. BS.1534-3, International Telecommunication Union, Oct. 2015. https://www.itu.int/dms_pubrec/itu-r/rec/bs/R-REC-BS.1534-3-201510-I!!PDF-E.pdf.

[22] F. Nagel, T. Sporer, and P. Sedlmeier, "Toward a Statistically Well-Grounded Evaluation of Listening Tests—Avoiding Pitfalls, Misuse, and Misconceptions," in *Audio Engineering Society - 144th Audio Engineering Society Convention 2018*, 2010.

[23] R. H. Baayen, D. J. Davidson, and D. M. Bates, "Mixed-effects modeling with crossed random effects for subjects and items," *Journal of Memory and Language*, vol. 59, pp. 390–412, Nov. 2008.

[24] R. Gueorguieva and J. H. Krystal, "Move Over ANOVA: Progress in Analyzing Repeated-Measures Data andIts Reflection in Papers Published in the Archives of General Psychiatry," *Archives of General Psychiatry*, vol. 61, pp. 310–317, Mar. 2004.