Ambisonic Decoder Test Methodologies based on Loudspeaker 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 152nd European Convention, Online, May 2nd, 2022. The published version of the paper [1] is available in the AES E-library at https://www.aes.org/e-lib/browse.cfm?elib=21665

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 a previous paper, a new test methodology for such evaluations was presented via a listening test conducted using binaural reproduction. In this paper, this methodology is further refined and the results of a new listening test using loudspeaker reproduction over a 7.0.4 array is presented. The results again indicate some significant differences between the decoders for certain attributes.

1 Introduction

Spatial audio is a critical component of new immersive media such as Virtual Reality (VR) and 360 video. The Ambisonics format is commonly used in such media, and involves two separate stages of encoding and decoding [2]. The design of the decoder is particularly important in terms of ensuring quality of experience for a range of different attributes when reproducing the recorded or encoded scene. A number of different decoding strategies are available, therefore, a suitable test methodology to compare their respective performance is needed. In a previous paper the authors proposed such a subjective test methodology using binaural reproduction [1]. A set of specific subjective spatial attributes to be evaluated was defined, alongside their relationship to the chosen test methodology and test stimuli. In addition, results were presented of a subjective listening test based on this methodology and the evaluation of three distinct decoder types. Passive decoders are the traditional First Order Ambisonic (FOA) decoding approach and attempt to optimize the reproduction in different frequency ranges via velocity-based and energy-based models, as summarized in [3]. More recent active decoders attempt to improve the passive method by extracting and using different rendering techniques for meaningful parameters from the Ambisonic signal [4]. However, the effectiveness of this approach is highly dependent on the nature of the signal. Even more recently, hybrid decoders attempt to combine previous approaches, using parametric decoding for directional soundfield components, and reverting to traditional passive decoding for diffuse components[5].

The choice of a known quality reference signal, to which the evaluated decoders is compared, is a critical component of such a test methodology. However, the development of such a reference signal, other than the real, physical audio scene, represents a significant challenge. While a real source such as a single loudspeaker may be used to evaluate attributes such as localization accuracy, other attributes such as spaciousness are more challenging. While the use of binaural reproduction in this previous study was beneficial in this regard, the use of generic, non-personalized Head Related Impulse Responses (HRIRs) was potentially problematic. In this subsequent paper, the previously described test methodology was further refined and implemented in a new study using a 7.0.4 loudspeaker array. The same passive, active, and hybrid decoders were again evaluated, and further refinements were also made to the subjective attributes evaluated.

2 Background

Prior evaluations of Ambisonic decoders generally performed a relative comparison of specific attributes such as Apparent Source Width (ASW) [6], or made reference to an internal, recalled judgement of some spatial attribute [7]. In the latter case, this internal reference may be based on highly familiar soundscapes, or rely on expert listeners with a strong familiarity with the test material (such as concert hall acoustics). However, Ambisonic decoders can and are used to decode a wide range of signal types, in various contexts such as cinema, gaming, or VR. This raises the question, as to 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?

The proposed test methodology for the quantitative evaluation of Ambisonic decoders outlines the development of just such a reference signal, using either binaural reproduction [1], or in this case, loudspeaker reproduction. While potentially problematic in other respects, the use of binaural reproduction simplifies this issue to an extent, as the same HRIR datasets used for the binaural rendering of the Ambisonic decoders, may also be used for the generation of suitable reference signals. In this prior study, two categories of signals were evaluated, object-based signals, and recorded scenes. For the former, suitable references were created via the direct rendering of monophonic test signals with HRIRs, while for the latter, Binaural Room Impulse Responses (BRIRs) or direct binaural recordings were used. The test signals were created by direct encoding to First Order Ambisonics (FOA) using either FOA encoder plugins, or FOA microphones. This was then followed by decoding with the three decoders under evaluation, and binaural rendering using a virtual loudspeaker approach, in which each loudspeaker signal was rendered to binaural via direct convolution with a HRIR. As the same HRIR dataset and binaural microphone were used for both the creation of the reference signal, and for the binaural rendering of the decoded loudspeaker signals, this therefore significantly reduced the influence of the binaural rendering as a factor in the comparison. The results of the subsequent listening test revealed a number of differences between decoders for certain attributes [1]. The results indicated that the hybrid and active decoders outperformed the passive decoder for localization accuracy, with single or multiple sources, and locatedness, but only when the virtual loudspeaker layout was well matched to these decoding methods. No significant differences were revealed between decoders for ensemble width or source distance, for either virtual loudspeaker layout. In addition, no differences were indicated between decoders for timbre, spatial impression/envelopment, or naturalness/presence, however, some differences in performance for each of the virtual loudspeaker layouts was uncovered. For these three attributes, the Cube virtual loudspeaker layout was rated higher than 7.0.4, with the hybrid decoder suffering a marginally greater reduction in performance than the active decoder.

The use of binaural reproduction greatly simplified the creation of suitable reference signals, and the reported results suggested that the hybrid and active decoders outperformed the passive decoder for attributes such as localization and locatedness. However, other, subtle differences, particularly for more complex attributes, may have been masked by the use of generic HRIRs, which was necessary due to the experiment taking place during the Covid pandemic and associated lockdown.

3 Test Methodology

In this paper, an additional listening test was performed using broadly the same test methodology as before, but with loudspeaker reproduction instead of binaural. The listening room used for both the preparation of the test stimuli, and the listening test itself, contained a 7.0.4 loudspeaker array. The array was housed within six acoustic panels in a hexagonal arrangement, using an open design constructed under the concept that acoustic reflections can be minimised by building a treated structure where sound can freely travel out of. Each individual panel is approximately 2m wide and 2.25m tall, arranged in a hexagonal shape with 2 entrances on either side and an open top. The hexagonal panels house all speakers on the azimuth plane at the standard angles of 30,-30,0,150,-150,90,-90 degrees. Four elevated loudspeakers were positioned at 45 degrees elevation, and +/-45 and +/-135 degrees azimuth. Each loudspeaker was placed at 1.6m distance from the central listening position (as shown in Appendix A, Figure 1. The loudspeaker array consisted of 11 Equator D5 loudspeakers based on a coaxially designed transducer with a frequency response of 53Hz to 20kHz. The room had a background noise level at the listening position measured at 35.6 dBA, and a reverberation time at 1kHz of 0.419 seconds.

Similar to the previous experiment [1], two broad categories of signals were evaluated, object-based signals, and recorded scenes. For the object-based tests, the reference consisted of a single audio source, primarily single, additional Equator D5 loudspeakers (separate from the main 7.0.4 array). The monophonic, anechoic audio signals reproduced as the reference in this way, were similarly directly encoded into FOA and routed to the various decoders under evaluation, for playback and comparison to the reference.

The recording-based tests were significantly more challenging to implement in a number of ways. While

recordings of real audio scenes could be captured concurrently using FOA microphones and other types of spatial microphone arrays, any resulting comparison would in effect be comparing the Ambisonic signal path to that particular microphone technique, rather than the real audio scene itself. Therefore for the recording-based tests in this experiment, the playback of 7.0.4 spatial recordings (made with some non-Ambisonic microphone array) using the reproduction system described above, was used as the ground truth reference, to which the decoded FOA signals were compared. To generate the FOA test signals, a FOA microphone was placed in the listening position and the reference signal recorded, before then being decoded as required. The advantage of this approach is that the reference signal was always and immediately available for comparison during the test. It should be noted that in optimal conditions using an entirely anechoic space, then the only differences between the reference and test signals, would be the FOA recording and decoding process itself, which is the primary focus of the evaluation. In the more realistic listening room used here however, the FOA recording process also introduces an additional doubling of the room acoustic, one in the original FOA recording itself, and another during the final playback of the test stimuli withing the same listening room. As such, additional room compensation processing needed to be applied to the FOA test signal recordings, to reduce this effect.

3.1 Room Response Compensation

Various methods exist for both offline and active room compensation, as discussed in [8], however, in this study, the frequency domain deconvolution, Nelson-Kirkeby method was adopted. An inverse filter was generated to compensate for the early reflections only via the inversion of the windowed RIR in the Discrete Fourier Transform (DFT) domain. This method derived the inverse filter by dividing the complex spectrum of the target response by the real RIR response. As the method is sensitive to notches, a regularisation parameter was added to bias the method towards peak compensation only. This helped to avoid peaks in the inverse filters which can damage equipment and provide over-compensation. To generate the inverse filters, the speaker-room impulse response for each channel was first recorded using an omnidirectional measurement microphone, and then trimmed to remove pre-silence and late reflections. Inverse filters were then individually made from each speaker-room IR using the invFIR function developed by Matthes [9] which is based on the frequency deconvolution method. The filters were designed with minimum phase, with a filter length of 16384 samples, and with 5dB and -25dB regularisation with 1/8th octave smoothing pre-applied. The decoded uncompensated 7.0.4 test signals were finally compensated using the generated inverse filters, where each individual channel was convolved with its corresponding filter, resulting in compensated 7.0.4 renders for each sample, and for each decoder. Note that this room compensation was only necessary for the recording-based tests used to evaluate attributes such as timbre and spatial impression. For the object-based tests used to evaluate attributes such as localization, including at off-centre listener positions, no room compensation was necessary as the reference signals consisted of real, physical sources, namely individual loudspeakers.

4 Assessed Sonic Attributes

In this experiment, a number of changes to the attributes assessed in the previous study [1] were implemented. Specifically, given the lack of significant differences reported for source distance and ensemble width, these attributes were not evaluated here. However, the change from binaural to loudspeaker reproduction supported the evaluation of additional attributes, namely localization accuracy with elevated source directions, and at off-centre listener positions. The test signals were divided into two broad groups: object based (for localization and locatedness tests) and recording/scene based (for all other attributes), comprising of;

- Localization Accuracy - single static source
- Localization Accuracy - multiple static sources
- Localization Accuracy - single source, off-center listener position
- Locatedness
- Timbre
- Spatial Impression/Envelopment
- Naturalness/Presence

The reference and test signals for all of the tests were equalized in terms of loudness using a target level of -32LUFS (Loudness Unit Full Scale), as specified in the ITU-R BS.1770-3 standard [10]. The same three decoders (passive, active, and hybrid) and the same or highly similar test stimuli as the previous study were again used here. The MUSHRA test methodology includes a low quality hidden anchor, specified in ITU-R BS.1534-1 [11] as a low-pass filtered version of the

reference signal, in order to reduce the overall signal quality and timbre. However, as this process only represents one type of impairment, and in order to evaluate other attributes other types of impairments must be added. The choice of anchor for each attribute evaluated is discussed in detail below, and follows the same approach used in the previous experiment.

4.1 Localizational Accuracy

Localizational accuracy with single, static sound sources were investigated using four different source directions, two horizontal, and two elevated. These directions were repeated for two different source signals. The reference was reproduced using one of four additional loudspeakers, positioned between the main loudspeakers of the 7.0.4 array at the angles shown in Appendix C, Table 1. The chosen angles concentrated on frontal and lateral directions in order to maintain a reasonable overall test duration, given the large number of tests involved. Anechoic speech and drum samples were chosen to provide good frequency content coverage as well as representing two common sound categories. The low quality anchors had reduced timbral quality via a low-pass filter at 2kHz. Directionality was also impaired by reproducing the anchor from all six main loudspeakers (excluding the center), or all eleven loudspeakers in the 7.0.4 array for the horizontal and elevated tests respectively.

As active decoding was being evaluated, localization accuracy was also tested using multiple sources with similar frequency content, and when the source directions are either in close proximity, or in directly opposite positions. In the former case, the references consisted of two static sources in close proximity on the horizontal plane separated by either 10 or 15 degrees, reproduced individually using one loudspeaker in the main 7.0.4 array (0 and 90 degrees azimuth), and one additional loudspeaker (-15 and 80 degrees respectively). The test stimuli were created by direct encoding into FOA, and subsequent decoding, and consisted of close-miked, pseudo-anechoic viola and violin samples from the Mixing Secrets library by Cambridge Music Technology [12].

Two tests were also performed using two sources at diametrically opposite directions, with the references reproduced using two loudspeakers in the main 7.0.4 array (-30 and 150 degrees, and +/-90 degrees). The test stimuli were created through straightforward FOA encoding and subsequent decoding. Test signals with strongly overlapping spectra were used in order to ensure these tests were a rigorous, difficult challenge for the decoders and consisted of anechoic oboe and clarinet samples from the Odeon Room Acoustics library of anechoic symphonic recordings [13]. The anchors for both types of multiple source localization tests consisted of monophonic downmixes of the sample pairs, low-pass filtered at 2kHz, and reproduced equally by both loudspeakers.

In order to evaluate the effective sweet spot of the different decoders, four additional localization accuracy tests were conducted with the listener re-seated at half the distance (0.75m) between the center and edge of the loudspeaker array, and at an angle of -115 degrees. The same voice and drum signals and horizontal source directions used in the previous localization tests were again used, as shown in Appendix C, Table 1.

4.2 Locatedness

The locatedness or potential change in the focus or broadening of the source signal in either horizontal or elevated planes (as opposed to Apparent Source Width which solely concerns horizontal broadening) was also evaluated. For this test, the reference consisted of a constant stream of pink noise, reproduced solely by the center loudspeaker at 0 degrees azimuth. The anchor consisted of decorrelated pink noise reproduced equally from five loudspeakers in the 7.0.4 array at +/-90, +/-30, and 0 degrees, producing an extremely wide, diffuse, and unfocused signal with no clearly defined direction. The test stimuli were again constructed via direct FOA encoding and decoding.

4.3 Timbre

Timbral quality was investigated using two tests and signals taken from the 3D-MARCo open-access database [14]. As in the previous study, one test signal consisted of an acapella vocal sample, recorded using a 5.0.4 PCMA-3D microphone array [15], supplemented with two additional side-fill cardioid microphones to complete the 7.0.4 signal. In a change to the previous study, the second timbre test replaced the synthesized scene with a church organ sample, also from the 3D-MARCo database. This was motivated by relatively poor performance of this synthesized scene in the previous timbre tests for all three decoders [1]. The organ test signal was also beneficial here, as it contained significant height information, and broad frequency content. In both timbre tests, the low quality anchor was created

using the standard process of low pass filtering at 2kHz.

4.4 Spatial Impression/Envelopment

The same test signals used to evaluate spatial impression/envelopment in the previous study [1] were again used here. These comprised of a classical trio (piano, violin, and cello) recording from St. Paul's concert hall in Huddersfield, from the the 3D-MARCo database, and with the 7.0.4 signal generated using the same procedure as in the timbre tests. In addition, a choral recording was also evaluated, namely an excerpt of a commercial choral recording recorded using a 7.0.4 PCMA-3D microphone array [15]. The anchors consisted of the monophonic center channels alone, with a low pass filter applied at 2kHz, thereby significantly degrading the overall spatial impression quality.

4.5 Naturalness/Presence

Attributes such as naturalness, realism, and presence were evaluated here as a single attribute, referred to as naturalness/presence, using two distinct signal types. The first test signal evaluated consisted of an exterior city scene, taken with the permission of the authors, from a dataset of city ambience recordings created by Felix Andriessens and Moritz Hoffmeister [16]. In this way, this signal replicated many of the characteristics of the stimulus used in the previous naturalness/presence test [1], and contained bird song from various directions, mostly elevated, distant vehicular and pedestrian traffic, and a single vehicle drive by. The recording was captured with a Williams Star 5-channel array, combined with an elevated IRT cross to produce the overall 5.0.4 signal [16], and therefore did not utilize the two side loudspeakers in the 7.0.4 array. The second test used a new recording consisting of approximately 12 individuals talking in small groups at various distances and directions in a highly reverberant basement. The microphone array used resembled that of Andriessens et al [16], and consisted of a combination of cardioid microphones arranged in a Williams star 6-channel array, and an IRT cross elevated by 1.3m above the main array. As with spatial impression/envelopment, the anchor for both tests consisted of the monophonic center channel signal, with a low pass filter applied at 2kHz. In the case of the new recording which did not contain a center-channel recording, a monophonic downmix of the left and right channels was used instead.

5 Listening Test Implementation

A Multiple Stimuli with Hidden Reference and Anchor (MUSHRA) style test (ITU-R BS.1534-1) was utilized due to its ability to allow multiple test signals to be rated per test instance. The HULTI-GEN interface was used for the test (shown in Appendix A, Figure 2) and supported standard features such as synchronous playback of the reference and test signals, and user controlled loop points [17]. Test subjects were instructed to rate the stimuli using the vertical sliders with respect to the reference for the given characteristic under test. In addition to the given reference, five stimuli were presented to the listener per test, consisting of the passive, active, and hybrid decoder signals, a hidden reference and a hidden anchor. A short questionnaire was also used to gather data on gender, age-range, listening test experience, and professional working experience in spatial audio.

5.1 Listening Test Procedure

Following an initial training session, subjects were then asked to to rate all stimuli in terms of their similarity to the available reference for the specific attribute being evaluated. 19 sets of results were gathered, 17 male and 2 female. The test subjects consisted of academics with past experience of similar tests, and audio professionals. 6 subjects were between the age category of 20-30, 9 between 30-40, 1 between 40-50, 2 between 50-60 and there was 1 60+ subject. In terms of experience, 10 subjects stated they had considerable experience with critical listening tests, 8 subjects reported little experience (1 or 2 previous tests), and 1 subject had none. 8 subjects reported extensive professional experience in audio (5+ years), 3 with fair experience (3-5 years), 5 with little experience (1-3 years) and 3 with 0 years professional experience.

The ITU MUSHRA test methodology standard [11] specifies post-screening to exclude subjects who frequently grade the hidden reference as though it were significantly impaired, or assign a very high grade to a significantly impaired anchor signal. Test subjects should be excluded from the final analysis if they rate the hidden reference condition less than 90 for more than 15% of the test items, and/or if they rate the hidden anchor above 90 for more than 15% of tests. In this case, no subjects were excluded from the analysis as none exceeded this threshold of more than 15% of tests, which in this experiment represents 3 to 4 tests,

of the 23 in total.

6 Results and Analysis

For each test, the data collected included the type of decoder, azimuth or elevated source direction (for single source localization), and the signal type. Boxplots were used to visualize and analyse the data, rather than confidence intervals (as suggested by [18]), and shown in Appendix B. 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 the active decoder for locatedness, and possibly the hybrid decoder also. Similarly, the plots also indicated that all three decoders outperformed the anchor for all attributes, with the exception of the passive decoder again for locatedness, where the anchor scored unusually high on average. The anchor was also competitive against the passive decoder for off-centre localization.

The data was also analyzed using a Repeated Measures Analysis of Variance (rmANOVA) as recommended by the ITU [11]. The presence of both dependent and independent samples was handled using a repeated measures ANOVA/mixed effects model, as discussed in [18]. Locatedness was not included due to its violations of the ANOVA assumptions, however, the box plot (shown in Figure 8) clearly indicated a significant difference between decoders, with the passive decoder severely outperformed by the hybrid and active decoders. The multiple source localization tests consisted of two distinct questions involving two sources in either close proximity, or at diametrically opposite positions. For this reason, two, one-way rmANOVAs were performed for this test, one for each signal type and source directions. For the remaining tests, a two-way rmANOVA was performed. Also, the data for single-source localization was split into two parts to distinguish between horizontal and elevated sources. Since a number of tests were performed, the significance threshold was set to 0.01, somewhat more conservative than the usual 0.05, to account for the possibility of inflated Type-I error due to multiple testing. Given this significance threshold, the results displayed some significant differences between decoder types for all localization tests (all p-values <0.001), and naturalness/presence (p-value <0.003). The signal type differences were significant for spatial impression/envelopment, and naturalness/presence (all p-values <0.001). The interaction between the decoder and signal type was also significant for timbre, and localization off-centre (p-values 0.001, and <0.001 respectively). This means that while the decoders and signal types were not that different overall, the signal types did have different effects on each decoder. However, the ANOVA only indicated that a significant difference existed somewhere in the groups; it does not reveal which pairwise differences are significantly different. This was determined with Tukey's Honest Significant Difference test, as recommended by the ITU [11]. This calculates a p-value in such a way that it takes into account the multiple testing issue, and is thus far more reliable than simple t-tests. Again, a significance threshold of 0.01 was used.

6.1 Discussion

For localization accuracy, significant differences were indicated by the rmANOVA model for the decoder for both source directions (horizontal or elevated). The Tukey HSD test results indicated that the active and hybrid decoders could not be distinguished for either signal type for horizontal sources, however, both clearly outperformed the passive decoder for either signal type, as shown in Figure 3.

Similar results were apparent for elevated sources, as shown in Figure 4 with no significant differences apparent between the hybrid and active decoders. However, the active speech and drums combinations outperformed passive drums and passive speech (all p-values <0.001). The hybrid speech combination also outperformed the passive drums and speech (p-values <0.01 and <0.001), however, hybrid drums only outperformed the passive speech (p-value <0.001) but not passive drums. These results largely replicated the findings of the previous study [1] for single source localization in that both active and hybrid decoders outperformed the passive decoder for the 7.0.4 layout for both signal types, and for both horizontal and elevated sources. However, this improvement was slightly less evident for elevated sources, and in particular the hybrid decoder failed to significantly outperform the passive decoder for the drum signal.

The results for off-center single source localization accuracy were highly similar to that of single source localization in general. Significant differences were found for the decoder and for the interaction between the decoder and signal type. Both the active and hybrid decoders performed well with the passive decoder rated far lower, for both signal types, as shown in Fig-

ure 5. Notably, the anchor results were competitive against the passive decoder here, indicating the relatively poor performance of this decoder in this case. The Tukey HSD tests indicated that the active and hybrid decoders could again not be distinguished, however, both clearly outperformed the passive decoder for either signal type when the listener is seated off-center (all p-values <0.001). Overall these results suggest that the hybrid and active decoders achieved a localization accuracy that was largely comparable to their performance at the centre position. For the passive decoder, despite the usage of a Max-Re weighting scheme, localization accuracy was degraded significantly at an off-centre listener position. Overall this suggests that when multiple listeners are present, active and hybrid decoding schemes provide significant advantages over more traditional FOA decoders in terms of localization accuracy.

The 2-way ANOVA results for localization accuracy for two sources indicated some significant differences between decoders, and source direction. For two sources in close proximity, the hybrid and active decoders could not be distinguished but both significantly outperformed the passive decoder (p-values = 0.00) much as in the previous experiment [1], as shown in Figure 6. However, for two sources at diametrically opposite positions, the hybrid and passive decoders could not be distinguished, but both were significantly outperformed by the active decoder (p-values = 0.00), as shown in Figure 7.

As mentioned previously, the locatedness results did not support ANOVA analysis. However, the active and hybrid decoders displayed a clearly apparent improvement in locatedness compared to the passive decoder, as shown in Figure 8, with both the hybrid and active decoders centered very close to the highest possible score of 100. The passive decoder was rated significantly worse and comparable to the anchor which scored unusually high on average. This result is quite similar to the prior binaural study [1], where the passive decoder was rated similarly and also showed some overlap of its interquartile ranges with the anchor. Overall, the results suggest that both the active and hybrid decoders are largely indistinguishable from the reference in terms of locatedness, while the performance of the passive decoder was poor in both experiments.

The results for timbre, shown in Figure 9, revealed no significant differences between decoders but some significant differences for the decoder/signal interaction with the hybrid-organ combination significantly different and outperformed by both the passive-choir (p-value = 0.004), and the active-organ (p-value = 0.00). This result tentatively suggests some degradation in timbre for the hybrid decoder and organ sample, which may only have been apparent in the broader frequency range of the organ stimulus, compared to the choir. However, given the use of repeated significance tests here, and the limited results, this finding should perhaps be interpreted with some degree of caution.

The results for spatial impression/envelopment revealed no significant differences between decoders, or decoder/signal interaction, but some significant differences in signal type. The classical trio was rated similarly for all three decoders, but with a quite large spread in the interquartile ranges as shown in Figure 10. In contrast for the vocal group sample, the active decoder was rated higher than both the hybrid and passive decoders and with a narrower interquartile range. The Tukey HSD test indicated some significant differences between signal types, with the active and passive choir samples both significantly different and rated higher than the classical trio and all three decoders. However, in contrast the hybrid choir sample showed no significant differences with all other classical trio samples. The generally higher and narrower scores obtained for the vocal group compared to the classical trio, perhaps indicate a greater degree of difficulty in the evaluation of spatial impression with this stimulus. It was noticeable that in the previous binaural tests, an increased variance in data for all three decoders for this attribute also perhaps suggested the difficulty in the judgement of this attribute [1]. However, the similar results for both do also perhaps indicate that the decoding method is not in of itself a significant factor for spatial impression/envelopment.

The results for naturalness/presence revealed some significant differences between decoders and signal type, but not in the decoder/signal interaction. The active decoder outperformed both the hybrid and passive decoders for the outdoor scene, while for the indoor scene, both the active and passive decoders outperformed the hybrid decoder, as shown in Figure 11. The Tukey HSD test indicated some significant differences between factors with the passive and active indoor scene combinations significantly different and rated higher than both the hybrid and passive outdoor scene combinations. The active outdoor scene combination was

also significantly different and rated higher than the hybrid outdoor scene. Overall these results suggest a generally higher rating for the indoor scene, compared to the outdoor scene, at least for the active and passive decoders. In addition, the results tentatively suggest a slightly worse performance by the hybrid decoder for this attribute, and the outdoor scene specifically. Although firm conclusions are again difficult to determine for this attribute, the results for spatial impression/envelopment, and naturalness/presence both suggest that at least for certain stimuli and loudspeaker layouts, the hybrid decoder performed marginally worse than the other decoders to some extent over the two experiments.

7 Conclusion

The results of this second listening test overall largely replicated the findings of the first, but with some additional findings. The most significant results again suggested that the hybrid and active decoders performed similarly for certain attributes, most notably in terms of localization accuracy and locatedness. Some differences were uncovered between the hybrid and other two decoders for other attributes, and while these findings are somewhat tentative, they did seem to be replicated in both the binaural and loudspeaker experiments.

For localization accuracy, very similar results were revealed in both experiments, and suggested that the active and hybrid decoders significantly outperformed the passive decoder in this regard. This appeared to hold in the case of single and multiple source localization, for different signal types, for both horizontal and elevated source directions, and for an off-center listener. The one potential exception was in the case of two sources with highly similar spectral content in diametrically opposite directions. For this test, the loudspeaker experiment did tentatively suggest a slightly better performance by the active decoder for this attribute. The results for locatedness were also highly similar across both experiments, and again suggested that the active and hybrid decoders significantly outperformed the passive decoder in this regard.

The results across both tests for timbre, spatial impression/envelopment, and naturalness/presence were more complex, and indicative of the generally more complicated nature of such attributes. As in the first experiment, no significant differences were reported between decoders in terms of timbre, however, some significant differences were uncovered for the decoder/signal interaction. No significant differences were revealed between decoders for the vocal group stimulus, which was also the case in the original study. The organ sample used in this second test contained a much broader frequency spectrum than the vocal sample used previously, and this perhaps revealed some differences between decoders in terms of timbre. Most notably, the hybrid decoder was outperformed by both the passive and active decoders for this stimulus. This result tentatively suggested some degradation in timbre for the hybrid decoder, which may only have become apparent with the broader frequency range of this stimulus. However, given the use of repeated significance tests here, and the limited results, this result should perhaps be interpreted with some caution.

Much as with timbre, the results for spatial impression/envelopment were somewhat tentative, and the similar results in both experiments perhaps indicated that the decoding method was not in of itself a significant factor for spatial impression/envelopment. The results again indicated some degree of difficulty in the evaluation of this attribute, and the generally higher and narrower scores obtained for the vocal group for spatial impression/envelopment suggested that this stimulus type was perhaps not particularly revealing for this attribute. The other stimulus, namely a classical trio, was rated higher than the vocal group in terms of spatial impression for the active and passive decoders, but not for the hybrid decoder.

A similar overall trend was finally also evident for naturalness/presence. Once again, the signal containing largely voice signals and reverberation (the indoor scene) was generally rated higher than the other stimulus which contained a broader frequency spectrum and range of content. The results for naturalness/presence in this second test again tentatively suggested a marginally worse performance by the hybrid decoder for this attribute, but only for certain signal types. Although firm conclusions were again difficult to determine for this attribute, it was noticeable that both experiments tentatively suggested a slightly greater decrease in performance for the hybrid decoder with the 7.0.4 layout for this attribute, compared to the other decoders. However, this was only evident for certain stimuli in the second experiment, and for certain loudspeaker layouts in the first. Nonetheless, the trend across both experiments for timbre, spatial impression/envelopment, and naturalness/presence did

seem to suggest that at least for certain stimuli and loudspeaker layouts, the hybrid decoder performed marginally worse than the other decoders, at least to some extent.

Overall, the results of both experiments suggest that the proposed listening test methodology is a viable approach, and that some consistent differences between decoders are apparent. While the use of binaural reproduction offers some advantages, it also precludes the evaluation of localization with elevated sources, and at off-centre listener positions. The results for more complex attributes such as spatial impression, naturalness or timbre were more inclusive with either reproduction method, and suggest that vocal stimuli may not be particularly revealing for such attributes, perhaps due to the limited frequency range of this signal type. In addition, for timbre, the use of loudspeaker reproduction and object-based stimuli reproduced using a single loudspeaker may be more revealing, as the use of generic HRTFs or room compensation methods may mask subtle differences in this attribute.

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," *Journal of the Audio Engineering Society*, 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," 2019. https://doi.org/10.25836/sasp.2019.26.

[5] A. Politis, S. Tervo, and V. Pulkki, "Compass: Coding and multidirectional parameterization of ambisonic sound scenes," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6802–6806, 2018.

[6] 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. Preprint 4996.

[7] 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. https://doi.org/10.1121/1.1763973.

[8] S. Cecchi, A. Carini, and S. Spors, "Room response equalization—a review," *Applied Sciences*, 2017.

[9] Matthes, "inverse fir filter." https://uk.mathworks.com/matlabcentral/fileexchange/19294-inverse-fir-filter, 2008.

[10] ITU-R, "Algorithms to measure audio programme loudness and true-peak audio level," Tech. Rep. BS.1770-3, International Telecommunication Union, Aug. 2012. https://www.itu.int/dms_pubrec/itu-r/rec/bs/R-REC-BS.1770-3-201208-S!!PDF-E.pdf.

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

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

[13] T. Lokki, J. Pätynen, and V. Pulkki, "Anechoic recordings of symphonic music," 2002. https://odeon.dk/downloads/anechoic-recordings/.

[14] H. Lee and D. Johnson, "An open-access database of 3d microphone array recordings," *147th Audio*

Engineering Society Convention: AES 2019, October 2019.

[15] H. Lee, "A new multichannel microphone technique for effective perspective control," *130th Audio Engineering Society Convention: AES 2011*, 2011.

[16] F. Andriessens and M. Hoffmeister, "A new 3d microphone array for cinema, vr and games – a comparison," *International Conference on Spatial Audio (ICSA) 2017*, 2017. Graz, Austria.

[17] D. Johnson and H. Lee, "Huddersfield universal listening test interface generator (hulti-gen) version 2," *Journal of the Audio Engineering Society*, October 2020.

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

A Listening Room and Test Interface

Fig. 1: 7.0.4 Listening Room

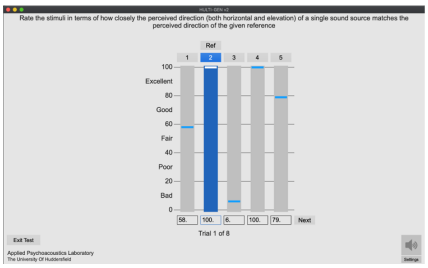

Fig. 2: HULTI-GEN Test Interface

B Box Plots

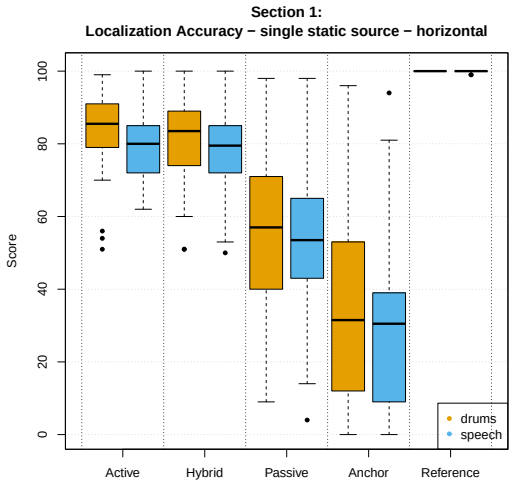

Fig. 3: Boxplot for localization accuracy, horizontal source

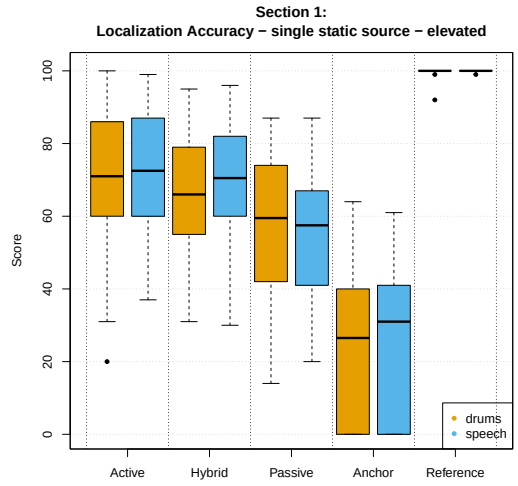

Fig. 4: Boxplot for localization accuracy, elevated source

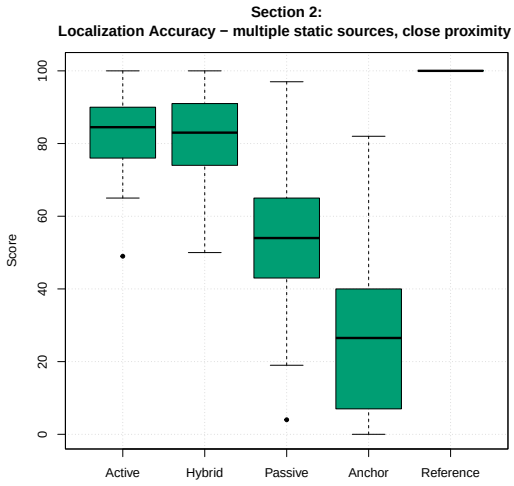

Fig. 6: Boxplot for localization accuracy, multiple sources in close proximity

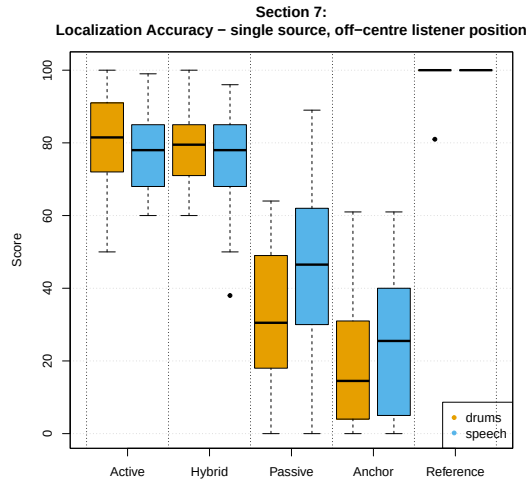

Fig. 5: Boxplot for localization accuracy, off-center listener

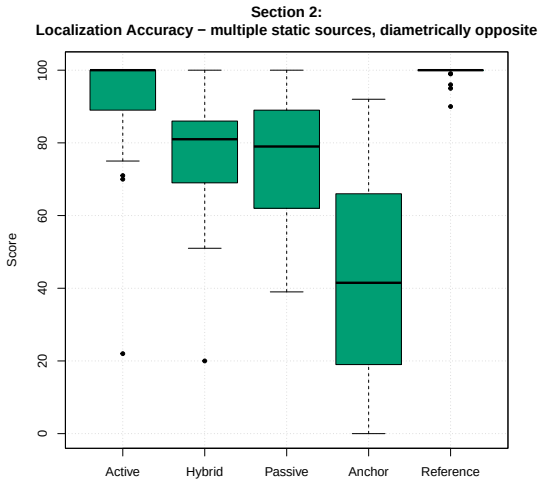

Fig. 7: Boxplot for localization accuracy, multiple sources, diametrically opposite positions

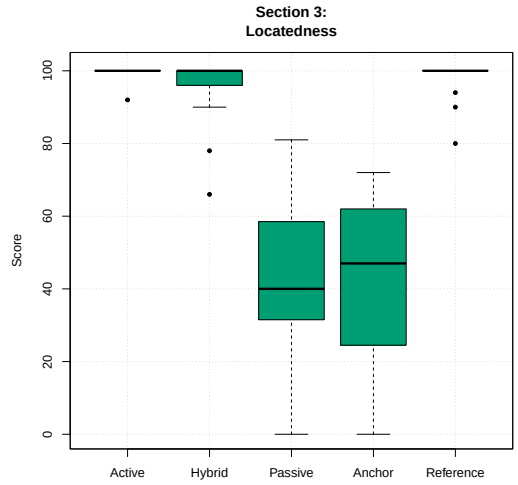

Fig. 8: Boxplot for locatedness

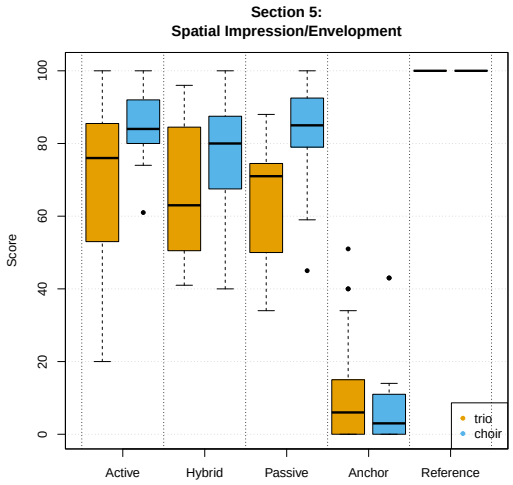

Fig. 10: Boxplot for spatial impression/envelopment

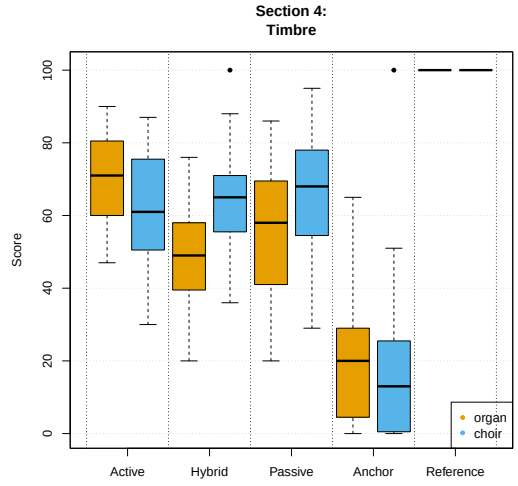

Fig. 9: Boxplot for timbre

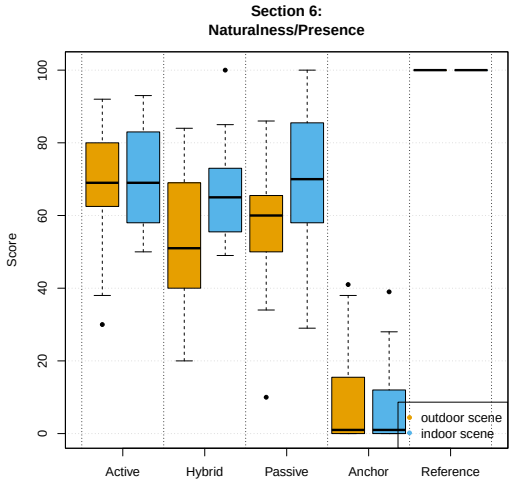

Fig. 11: Boxplot for naturalness/presence

C Listening Test Stimuli

Test No.	Test	Description	Angles (Azi,Ele)	Anchor
1	Localisation Accuracy - single source	Anechoic drums sample	80,0	All azimuth speakers excl. center
2	Localisation Accuracy - single source	Anechoic speech sample	-15,0	All azimuth speakers excl. center
3	Localisation Accuracy - single source	Anechoic speech sample	80,0	All azimuth speakers excl. center
4	Localisation Accuracy - single source	Anechoic drums sample	-15,0	All azimuth speakers excl. center
5	Localisation Accuracy - single source	Anechoic drums sample	60,30	All 11 speakers
6	Localisation Accuracy - single source	Anechoic speech sample	-35,35	All 11 speakers
7	Localisation Accuracy - single source	Anechoic speech sample	60,30	All 11 speakers
8	Localisation Accuracy - single source	Anechoic drums sample	-35,35	All 11 speakers
9	Localisation Accuracy - two sources, close proximity	Anna Blanton Voila & Voilin samples	0,0 & -15,0	Both signals at -15,0 & 120,0
10	Localisation Accuracy - two sources, close proximity	Anna Blanton Voila & Voilin samples	80,0 & 90,0	Both signals at -90,0 & 90,0
11	Localisation Accuracy - two sources, opposite	Odeon Room Acoustics Oboe & Clarinet samples	-30,0 & 150,0	Both signals in -30,0 & 150,0
12	Localisation Accuracy - two sources, opposite	Odeon Room Acoustics Oboe & Clarinet samples	-90,0 & 90,0	Both signals in -90,0 & 90,0
13	Locatedness	Pink noise	0,0	Decorrelated noise in front 5 azimuth speakers
14	Timbre	3D-MARCo organ sample	-	Low pass filtered to 2kHz
15	Timbre	3D-MARCo acapella sample	-	Low pass filtered to 2kHz
16	Spatial Impression/Envelopment	3D-MARCo classical trio sample	-	Mono, low pass filtered to 2kHz
17	Spatial Impression/Envelopment	Hyunkook Lee choral sample	-	Mono, low pass filtered to 2kHz
18	Naturalness/Presence	Felix Andriessens outdoor scene	-	Mono, low pass filtered to 2kHz
19	Naturalness/Presence	Recorded basement scene	-	Mono, low pass filtered to 2kHz
20	Localization - off center listener	Anechoic drums sample	80,0	All azimuth speakers excl. center
21	Localization - off center listener	Anechoic drums sample	-15,0	All azimuth speakers excl. center
22	Localization - off center listener	Anechoic speech sample	80,0	All azimuth speakers excl. center
23	Localization - off center listener	Anechoic speech sample	-15,0	All azimuth speakers excl. center

Table 1: Listening Test Content Summary