Comparison of Ambisonic Loudspeaker Decoders for Channel-Based Material

Project Thesis

Markus Radke, B.Mus.

Supervisor: DI Matthias Frank, Ph.D.

Graz, WS 21/22

kunst uni graz

institut für elektronische musik und akustik

iem

Abstract

Encoding and decoding to loudspeakers in Ambisonics introduces interchannel crosstalk to channel-based material. This study investigates the perceptual impact of this phenomenon for different signals, Ambisonic orders, decoders, and listening positions with respect to spatial and timbral quality aspects. Open-access recordings made with different microphone arrays, as well as noise signals are encoded in Ambisonics with different orders. Then they are decoded to a loudspeaker setup matching the channel-based layout using either the AllRAD approach or a sampling decoder. In a MUSHRA-like listening experiment for center and off-center listening positions, perceptual differences to a direct loudspeaker playback of the material are investigated. Results show less perceptual differences for higher orders and musical material. In the particular environment of the listening experiment, basic weighting outperforms max-r_E weighting.

Kurzfassung

Das Kodieren und Dekodieren in Ambisonics verursacht Übersprechen zwischen den Kanälen bei kanalbasiertem Material. Diese Studie untersucht den wahrnehmbaren Einfluss dieses Phänomens für unterschiedliche Signale, Ambisonics-Ordnungen, -Decoder und Hörpositionen hinsichtlich klanglicher und räumlicher Qualitäten. Frei verfügbare Aufnahmen mit verschiedenen Mikrofon-Arrays sowie Rauschsignale werden in Ambisonics mit verschiedenen Ordnungen kodiert. Dann werden sie entweder mit dem AllRad-Ansatz oder einem Sampling Decoder auf ein Lautsprechersetup dekodiert, dass dem kanalbasierten Layout entspricht. In einem MUSHRA-ähnlichen Hörversuch für zentrale und dezentrale Hörpositionen werden wahrnehmbare Unterschiede zu einer direkten Lautsprecherwiedergabe des Materials untersucht. Die Ergebnisse zeigen, dass für höhere Ordnungen und musikalisches Material weniger Unterschiede wahrgenommen werden. In der Versuchsumgebung übertrifft die *basic*-Gewichtung die *max*-r_E-Gewichtung.

Contents

List of Figures

List of Tables

1 Introduction

Ambisonics is a format for recording and playing back sound on a full-sphere. Recently, Ambisonics has become a popular format for virtual reality applications [1]. Also, recent ITU, MPEG-H and ETSI standards include the format ([2, p. v]) and it is used by Google and Facebook for parts of their media services.[1][2] One advantage of Ambisonic playback is, that the sound scene can be rotated efficiently with a simple matrix multiplication. This is especially useful for binaural playback with headtracking. In addition, there are many other computationally efficient spatial effects available. Also, compact Higher-Order Ambisonics (HOA) microphone arrays, such as the Eigenmike[3] or Zylia[4] provide high usability.

Nevertheless, sound engineers still prefer spaced or near coincident microphone techniques for classical music recordings or film scoring over coincident Higher-Order Ambisonics microphone arrays. A wider spaced array yields a higher sense of spaciousness. This is due to a greater interchannel time difference (ITD) and therefore higher interchannel decorrelation, especially for low frequencies. Furthermore, spaced arrays with individual microphones offer more customization possibilities to the sound engineers [1, pp. 1 sq.].

However, Ambisonics is a possible delivery format for channel-based recordings, e.g. as specified in the MPEG-H standards [3]. In order to use Ambisonics as delivery format for channel-based recordings, the microphone array signals have to be encoded to Ambisonics. Then, depending on the Ambisonic order N, they can be delivered as $(N+1)^2$ spherical harmonics coefficient signals and finally be decoded to either loudspeakers or headphones (see fig. 1). Even when the angles for encoding and decoding match, interchannel crosstalk is introduced and the interchannel correlation increases. The special case of matching encoding and decoding positions is subject to this research project.

This study investigates the perceptual effect of interchannel crosstalk and increased correlation introduced to channel-based material by Ambisonic encoding and decoding. First of all, the theoretical background of the phenomenon will be discussed in more detail and measurments of crosstalk and correlation are presented. Then the method for the conducted listening experiment is presented. Finally, experimental results are evaluated and compared to the measurements described in the first part.

[1] https://vr.youtube.com/create/360 (visited on 01/14/2022).
[2] https://facebook360.fb.com/spatial-workstation (visited on 01/14/2022).
[3] https://mhacoustics.com/products (visited on 01/14/2022).
[4] https://www.zylia.co (visited on 01/14/2022).

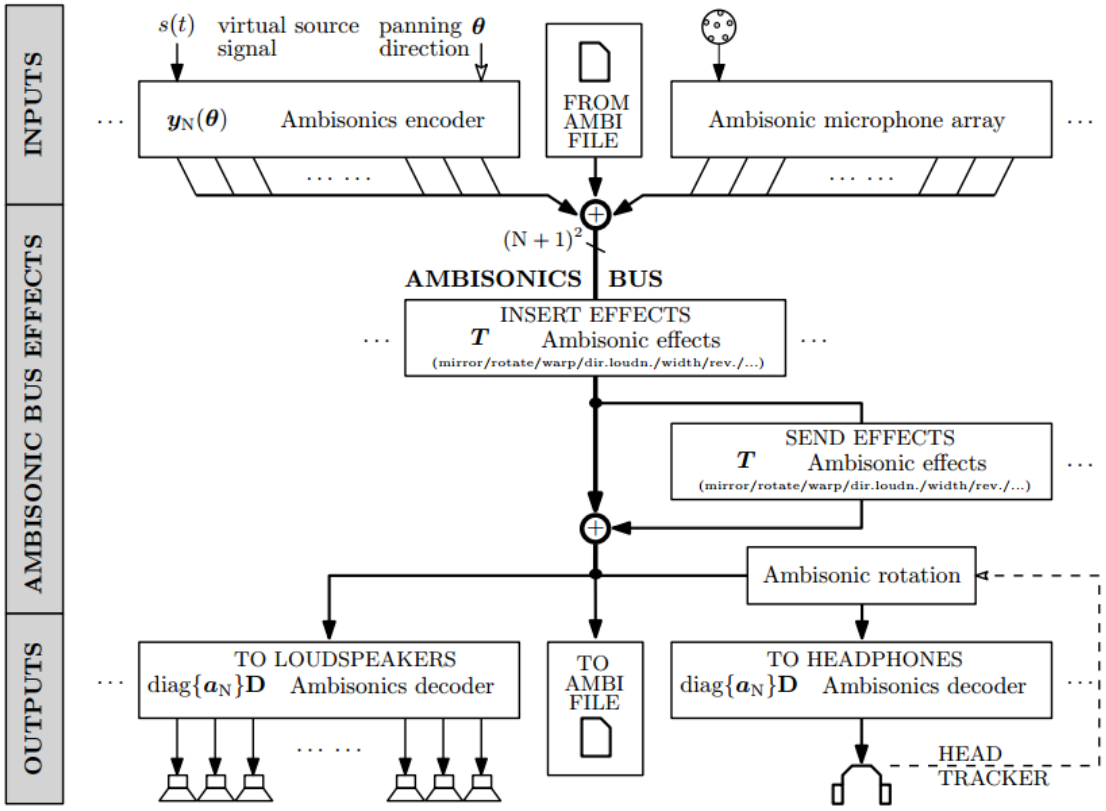


Figure 1: Typical Ambisonic signal flow, from [4].

2 Ambisonics and Interchannel Crosstalk and Correlation

2.1 Encoding and Decoding

Encoding Encoding is done by multiplying the signal s with an encoder representing the direction $\theta_\mathbf{s}$. The encoder is expressed as a vector with coefficients of spherical harmonics $Y_n^m(\theta_\mathbf{s})$, where $n <= N$ is the Ambisonic order and $m = -n...n$ is the degree of the spherical harmonics [2, p. 71]. In other words, the encoder matrix provides information on how much of a spherical harmonic function of order n and degree m is needed to pan a signal to direction $\theta_\mathbf{s}$ (see fig. 2). To achieve perfect panning, the signal would have to be encoded with infinitely many orders and spherical harmonics components. In practical applications typically up to seven order are used. A signal with $(N+1)^2$ channels is the result of the encoding process and contains the coefficients for the spherical harmonics to reproduce the panned sound on the full-sphere. Signals in the Ambisonics domain can be added and scaled. Consequently, to encode a microphone array, all microphone signals get encoded with a separate encoding vector and are added together subsequently.

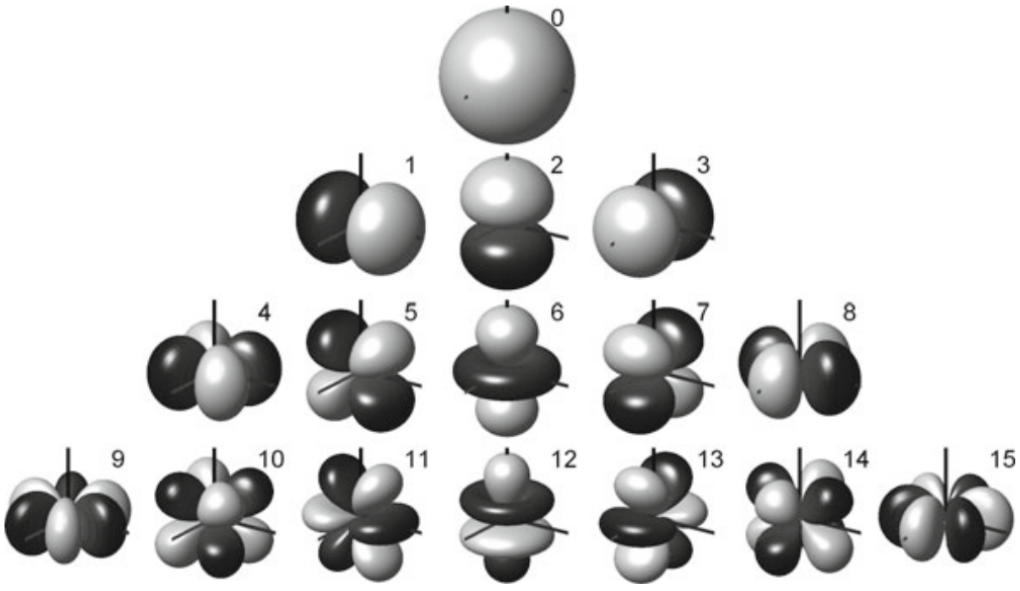
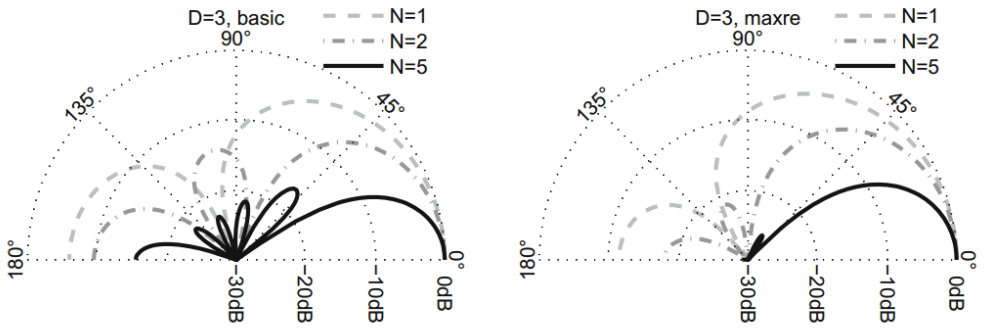


Figure 2: Spherical harmonics up to order 3, from [2, p. 68].

Weighting By limiting the number of orders used for encoding a signal at direction $\theta_\mathbf{s}$) side lobes are introduced to the main direction. The higher the order, the more side lobes



(a) basic $a_n = 1$ (b) max-$\boldsymbol{r}_\mathrm{E}$

Figure 3: Side lobes for basic and max r_e weighting, from [2, p. 70].

are present. At the same time their level is reduced (see fig. 3 (a)). The encoded Ambisonic signal can be weighted to suppress the side lobes. Max-r_E weighting maximizes the r_E vector, which is pointing into the panning direction. The weighting coefficients are described as:

$$a_n = \mathrm{P}_n \left[cos \left(\frac{137.9^\circ}{N + 1.51} \right) \right],$$

where P are n-order Legendre polynomials. [2, p. 69] Max-r_E weighting fades out higher Ambisonic orders. Apart from side-lobe suppression, this also leads to a wider main lobe and therefore possibly to an increased perceived apparent source width. In addition, zero crossings of the panning function are shifted by the weighting (see fig. 3 (b)).

In the course of this study, encoding and decoding were always done at the same position on the sphere for every channel. Please refer to A for channel naming conventions and corresponding positions. Two different types of decoders were investigated.

Sampling Decoder The sampling decoder (SAD) is the simplest decoder. It uses the spherical harmonics sampled at the loudspeaker position. The decoder matrix therefore is given as:

$$\mathrm{D} = \sqrt{\frac{S_{D-1}}{L}} \mathrm{Y}_N^T$$

Here, Y_N^T are spherical harmonics of order N, and the term $\sqrt{\frac{S_{D-1}}{L}}$ is a normalization term with $S_2 = 4\pi$ being the surface of the unit sphere and L being the number of loudspeakers. To yield constant loudness and width of decoded signals for every possible panning direction, the SAD requires an optimal loudspeaker layout (i.e. a t-design). If not provided, levels will be reduced at positions with a smaller loudspeaker density [2, p. 72].

This is not the case for this investigation: All signals will be mapped to a loudspeaker direction. This way, the effect of encoding and decoding can be shown at loudspeaker positions exclusively with a SAD.

ALL-Round Ambisonic Decoding All-Round Ambisonic Decoding (AllRAD) uses a combination of Vector-Base Amplitude Panning (VBAP) and SAD: First, the Ambisonic signal is rendered to 5200 virtual loudspeakers arranged in a t-design with SAD. After that the resulting multiple sources are synthesized to the loudspeaker positions using VBAP. Accordingly, the decoder matrix can be expressed as:

$$\mathrm{D} = \frac{S_{D-1}}{\hat{L}} \sum_{l=0}^{\hat{L}} g_{\mathrm{VBAP}} \left(\hat{\theta}_l\right) y_N^T \left(\hat{\theta}_l\right) = \frac{S_{D-1}}{\hat{L}} \hat{\mathrm{G}} \hat{\mathrm{Y}}_N^T$$

Again, the term $\sqrt{\frac{S_{D-1}}{\hat{L}}}$ is a normalization term with $S_2 = 4\pi$. As can be seen, the decoder can be implemented as a matrix that only depends on the number of loudspeakers and the ambisonic order, not the number of virtual loudspeakers [2, p. 76]. AllRAD decoding was chosen for this investigation since it is more related to practice. When including imaginary loudspeakers in the decoder design (see [5, p. 809]) it proves to be robust for every loudspeaker layout, including irregular hemispherical ones (e.g. ITU 7.0.4).

2.2 Interchannel Crosstalk and Correlation

The side lobes introduced by Ambisonic encoding cause interchannel crosstalk and increased interchannel correlation. In order to measure the crosstalk, an impulse played back at FC was encoded to 7th-order Ambisonics and then decoded with the different decoders used also in the listening experiment described below: AllRAD decoder with basic weighting for 1st, 3rst, 5th and 7th order (AB1, AB3, AB5, AB7); AllRAD decoder with max-r_E weighting for 3rd and 5th order (AM3, AM5); SAD with basic weighting for 3rd and 5th order (S3, S5). Then, level differences to FC for FL, FR, SL, and SR were measured and averaged for front and side channels. Results can be seen in fig. 4.

The level difference for AB1 decoder is very small. This can be explained by the low spatial resolution 1st-order spherical harmonics offer. Overall, larger level differences can be observed for higher orders. Looking at level differences for FL and FR, basic weighting performs better than max-r_E weighting. The reason is the widening of the main lobe with max-r_E weighting. SAD performs slightly better than AllRAD max-r_E but not as good as AllRAD basic when comparing the same Ambisonic orders. Taking a look at the results

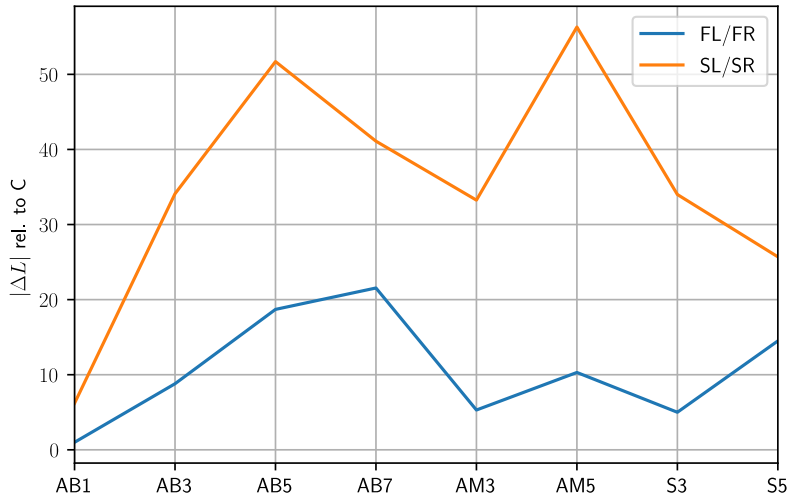


Figure 4: Averaged level difference to FC measured for different decoders and samples used in listening experiment.

for SL and SR, due to a higher angular distance to C there is less crosstalk in general. From Ambisonic order three on, the crosstalk with more than 30 dB is neglectable in practice. Crosstalk increases from AB5 to AB7 and from S3 to S5. Also, here AllRAD with max-r_E weighting performs better than AllRAD with basic weighting. This can be explained with the suppression of side lobes caused by the max-r_E weighting. Another explanation could be, that the weighting shifts zero crossings and therefore changes the angular position of side lobes.

The increase of correlation between channels was measured for noise and music samples that were used in the listening experiment described below. Correlation was measured between FL, FR, and FC. In case of decorrelated noise for FL and FR, non defined values for correlation evaluation of the reference were replaced with zero (i.e. the correlation between FC and FL as well as between FC and FR). The correlation between the three pairs was averaged. Also, the correlation measurements of all musical examples were averaged (see fig. 5).

All measured correlations are higher than the reference correlation. Clearly visible is a strong peak for the correlation for AB1. For the noise input and all other Ambisonic orders measurements stay within a range of 0.1. For AllRAD with basic weighting correlation slightly increases from 3rd order to 7th order. For AllRAD with max-r_E weighting and SAD it decreases a little from 3rd to 5th order. For the musical examples the correlation can be seen to decrease the higher the order.

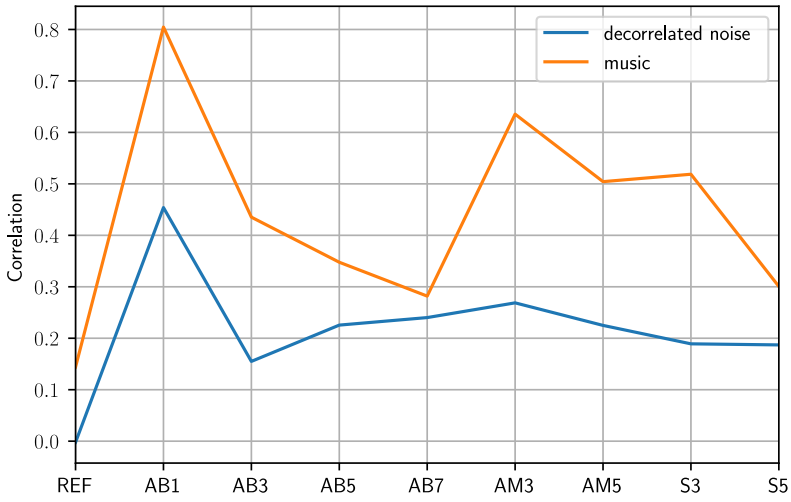


Figure 5: Averaged correlation measured for different decoders and samples used in the listening experiment.

High crosstalk might alter spatial perception and cause a blurred spatial image. Also, timbral perception might change, e.g. with comb filters added by the crosstalk. Perceptual differences are difficult to assess just based on the measurements presented above. Therefore, a listening experiment was conducted to investigate the perceptual effects of Ambisonic encoding and decoding with different decoders for channel-based material.

3 Listening Experiment

In order to investigate the influence of Ambisonic encoding and decoding with different decoder alternatives on perception of spatial and timbral qualities, a listening test was conducted. The experiment took place in the production studio at the *Institute of Electronic Music and Acousics* (IEM) in Graz. Effects were studied for different stimuli such as noise and musical examples as well as different listening positions. The test subjects were asked to rate the similarity of to a reference for different decoders with respect to spatial and timbral quality aspects. The reference was a direct mapping of the channel-based material to the loudspeakers.

3.1 Listening Environment

The experiment was conducted at the production studio at the IEM. The place was chosen to be a suitable typical studio environment for the listening test due to its acoustical properties and loudspeaker setup. The size of the room is about 41 m^2 and the volume can be estimated as 123m^2[6]. With the method presented in [7], the reverberation time RT_{60} could be estimated from measured impulse responses to be 0.3 s.

The loudspeaker setup used consisted of twelve loudspeakers of the model *KH 310*. The *KH 310* is a tri-amplified near-field monitor with no more than 3 dB deviation from a completely free-field impulse response between 34 Hz and 21 kHz [8]. Fig. 6 shows the loudspeaker arrangement at the studio. All loudspeakers were placed on a hemisphere above the listener. Seven loudspeakers formed the bottom layer. There were two height layers. The first at an elevation of 40° consisted of four speakers, the second height layer consisted of a speaker at an elevation of 90°, i.e. right above the listening position. 30° was the minimum angular distance between two speakers for the whole setup.

3.2 Scenes

As audio scenes five stimuli were chosen to be used for the experiment (see table 1). All of them were cut in a way so that they could be looped without distracting the subjects from rating the similarity. Loope lengths ranged between 11.8 and 16.1 seconds. As mentioned above, two scenes were based on pink noise. In one case, pink noise was played back on only channel FC. The other case was decorrelated pink noise on channels FL and FR. The decorrelation was generated by delaying one of the channels by one second. The correlation between FL and FR over the whole loop could be calculated as $\approx 0,01$.

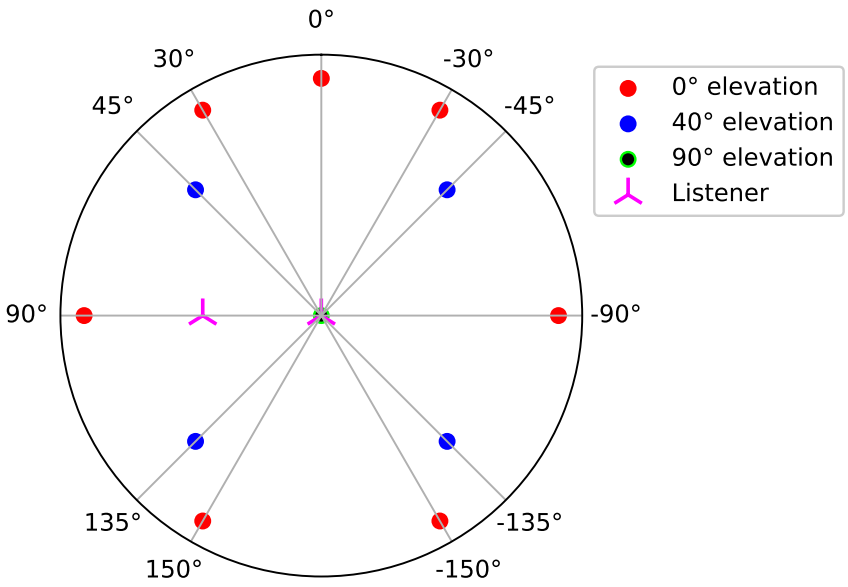


Figure 6: Loudspeaker setup at the production studio.

Noise stimuli were chosen because it supposedly was more easy to detect dissimilarities in spatial, but especially in timbral quality aspects than with music examples. As could be seen above, encoding and decoding material on the center channel only showed the crosstalk and leakage to other channels drastically. In the other case, the decorreleation of FL and FR is reduced critically by the encoding and decoding process. This was the motivation to investigate the perceptual effects for these scenarios.

All music examples were taken from the open data base of 3D microphone array recordings published by Hyunkook Lee and Dale Johnson in 2019 [9]. This way the listening experiment can be recreated (also, all Plug-ins needed for recreation are freely available). Lee and Johnson recorded various ensembles using 3D microphone arrays with varying conceptional approaches, layouts, and sizes. For this listening experiment, recordings made with three different 9-channel microphone setups were used.

PCMA-3D This setup is based on Hynkook Lee's "Perceptive Control Microphone Array" (PCMA) surround setup. A crucial concept for the PCMA is the idea to have control of the perspective during the post production process [10, p. 1]. The PCMA gets extended with a height layer for 3D audio recordings. The height microphones comprise four supercardioid microphones pointing to the ceiling, making the PCMA-3D a "horizontally spaced but vertically coincident array" [9, p. 2]. As Lee and Christopher Gibben showed

in 2015, the vertical distance between height and bottom layer has no significant impact on the spatial impression [11]. Therefore, the recording for the open database bottom and height layer used microphones placed on the same height. All microphones for the two layers were placed on a square with edge length 1 m, respectively. Five cardioid microphones pointing outward of the square were used for the bottom layer [9, p. 2].

OCT-3D The OCT-3D was suggested by Günther Teile and Helmut Wittek. For the front triplet of the bottom layer, a cardioid microphone for the center channel is combined with two supercardioid microphones pointing to +- 90° to minimize interchannel crosstalk. This is supposed to guarantee a precise frontal image localization. For this recording, the FL and FR microphone were placed 70 cm apart and the C microphone was placed 8 cm in front of the array base point. For the rear channels two cardiod microphones were arranged 40 cm behind the FL and FR microphones facing backwards [9, p. 2]. The height layer is meant to capture no direct sound from the stage. Accordingly, four additional supercardiod microphones were put 1 m above the bottom layer facing upwards. This way they are presumed to record early reflections and diffuseness and add to a "natural music recording" [12].

Decca Cuboid This microphone array consists of eight omni microphones placed at the edges of a cube with an edge length of 2 m. In addition, a microphone for the center channel was set 0.25 m in front of the edge between the FL and FR microphone. As the spacing is larger as in the two other microphone systems presented above, the decorrelation between the channels increases. This provides a more spacious sound image. At the same time localization relies heavily on a strong precedence effect. This reduces the number of effective image localization points to the panning directions of the microphone channels [9, pp. 2 sq.]. Also, the interchannel crosstalk induced by the use of omni microphones might cause horizontal localization blur and vertical image shift [13].

Noise	**Music**
C	string quartet (PCMA-3D)
FL & FR decorrelated noise	piano trio (OCT-3D)
	a capella chorus (Decca Cuboid)

Table 1: Overview of used scenes.

Three recordings from the database were selected for the listening experiment: The first was an excerpt from Antonin Dvorak's string quartet in G major op. 106, recorded with the PCMA-3D setup. The second recording was a passage from Ludwig van Beethoven's

piano trio op. 1 no. 1, recorded with the OCT-3D array. Finally, the last sample was an a capella arrangement of the Amber song "I found", recorded with the Decca Cuboid. The recordings were chosen due to musical diversity, ensemble size, and different spatial qualities.

3.3 Signal Flow

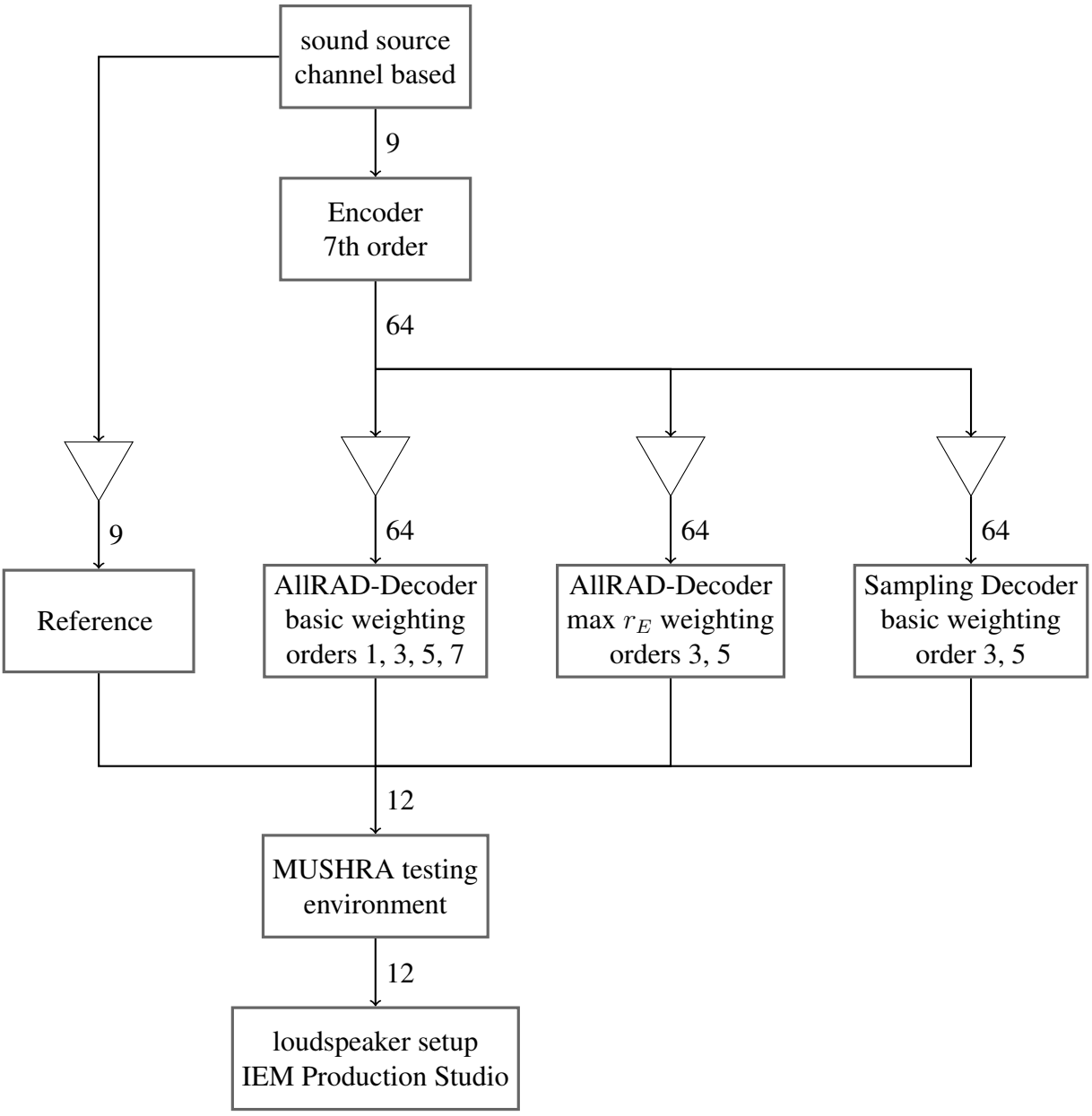


Figure 7: Signal flow in *REAPER*.

All signal processing was done in *REAPER 6* [5]. The signal flow can be seen in fig. 7. In case of the music recordings, the nine microphone signals were mapped directly to the loudspeaker channels FL, FR, FC, BL, BR, TFL, TFR, TBL, and TBR for the reference. For Ambisonic encoding and decoding the microphone signals were first encoded to 7th-order Ambisonics using the MulitEncoder from the *IEM* Plug-in suite [6]. In order to do that, the signals were panned to the corresponding loudspeaker positions. For decoding the *IEM* Simple Decoder Plug-in was used. Loudspeaker directions were slightly adjusted for encoding as well as for decoding: Instead of 30° elevation, the loudspeakers in the first height layer were mapped to 45° elevation. This yields better results in practice [2, p. 82]. Three decoder variants were used for the listening experiment: An AllRAD decoder with basic weighting was calculated with the *IEM* AllRAD Decoder Plug-in and tested for Ambisonic orders 1, 3, 5, and 7. Another AllRAD decoder with max-r_E weighting was also calculated with this Plug-in and tested for orders 3 and 5. Lastly, a sampling decoder was calculated using MATLAB [7] and also tested for orders 3 and 5.

Since the decoded signals varied greatly in perceived loudness, their levels were adjusted to match the reference. For a first impression of the necessary adjustments, an impulse was convolved with measured impulse responses (IRs) for each loudspeaker of the production studio. This was done using Matthias Kronlachner's *mcfx* Plug-in suite [8]. Impulses emitted from FC speaker only, FL and FR speaker as well as all loudspeakers playing at the same time were simulated. From the result, level differences in the subband between 200 Hz and 4 kHz were used for a first level adjustment. After this, the simulations were done again to take a look at the magnitude responses. In fig. 8 the differences to the reference in the magnitude response for an impulse emitted from FC can be seen in third-octave bands for all different decoders. It is visible, that in case only the FC speaker is playing, sampling decoder and AllRad decoder with basic weighting are closer to the magnitude response of the reference. The main differences are to be found in higher subbands. The lower the order, the more deviation is evident. Max-r_E weighting causes more attenuation of higher frequencies than basic weighting as can be seen when comparing the different decoders with same Ambisonic order.

It was tried to equalize the differences towards high frequencies. This worked only to a certain extend and the success was sometimes hard to asses due to interference with other spatial and timbral artifacts caused by the encoding and decoding process. Therefore, for

[5] https://www.reaper.fm (visited on 01/13/2022).
[6] https://plugins.iem.at (visited on 01/13/2022).
[7] https://www.mathworks.com/products/matlab.html (visited on 01/13/2022).
[8] http://www.matthiaskronlachner.com/?p=1910 (visited on 01/12/2022).

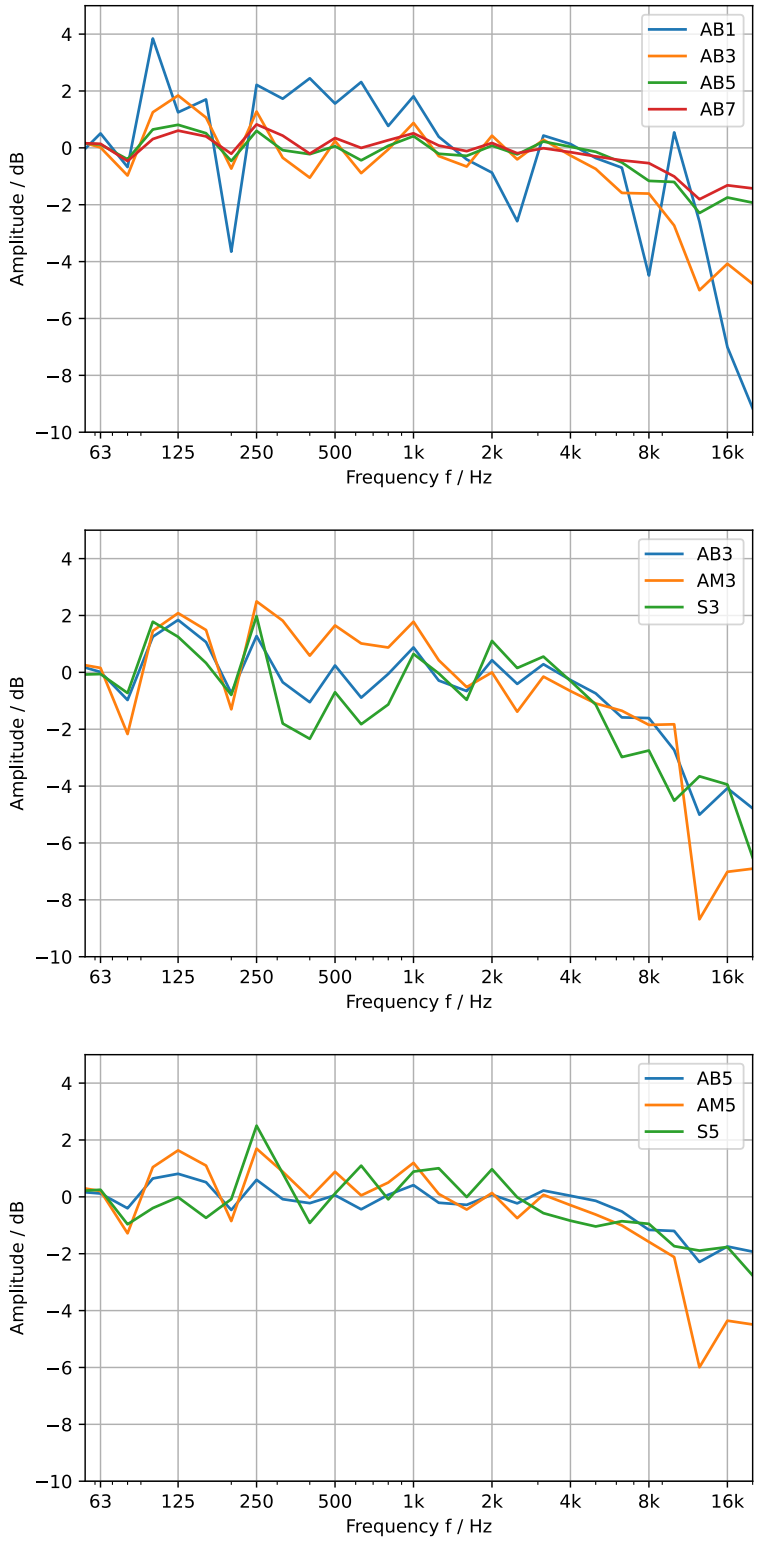


Figure 8: Difference to reference for smoothed frequency responses for all decoders; impulse emitted from FC.

the experiment the simulation could only be an orientation. Manual balancing was done to provide the same perceived loudness for every decoder version compared against the reference. No timbral changes to the output of the decoders were done for the listening test.

3.4 Test Design

MUSHRA-test Originally the "Multiple Stimuli with Hidden Reference and Anchor" test (MUSHRA) is a method for testing intermediate audio quality codecs. That means that it is typically used to compare audio encoding with noticeable impairments [14, p. 3]. The International Telecommunication Union (ITU) specified the design for MUSHRA tests in recommendation *BS.1534: Method for the subjective assessment of intermediate quality level of audio systems* [14, pp. 6 sqq.]: 20 subjects are often sufficient. Scene lengths are proposed to be ten to twelve seconds and no more than twelve signals (here: decoder alternatives) are recommended for the test. Hidden reference and anchor signals for low and medium quality are used and subjects are trained before the listening test. During the test, test subjects can switch freely between the reference and any other signals under test. They rate the difference to the reference on a scale from 0 to 100.

The test setup used for the investigation of different Ambisonic decoders is similar to MUHSRA, but not exactly the same. There were no hidden anchors, just the hidden reference. In two parts subjects were asked to rate:

1. the similarity of the signal to the reference with respect to spatial quality aspects, such as localization and apparent source width,
2. the similarity of the signal to the reference with respect to timbral aspects.

For both attributes ratings had to be done for all scenes. Ratings were done using sliders on a scale with 100 steps, ranging from "very different" to "identical". For each scene, all sliders were reset to 50 as a starting point. Including the hidden reference the subjects had to rate ten signals for each scene. After finishing both parts subjects were asked to repeat the experiment for another listening position: The experiment was done for a center and off-center listening position (see fig. 6). The order of scenes, attributes, and starting listening positions was randomized in order to avoid any bias. Until now twelve subjects took part in the listening experiment.

From a technical point of view, the experimental environment was set up using a software coded at *IEM* that controls *REAPER*'s mixer via OSC. Please refer to fig. 9 for an

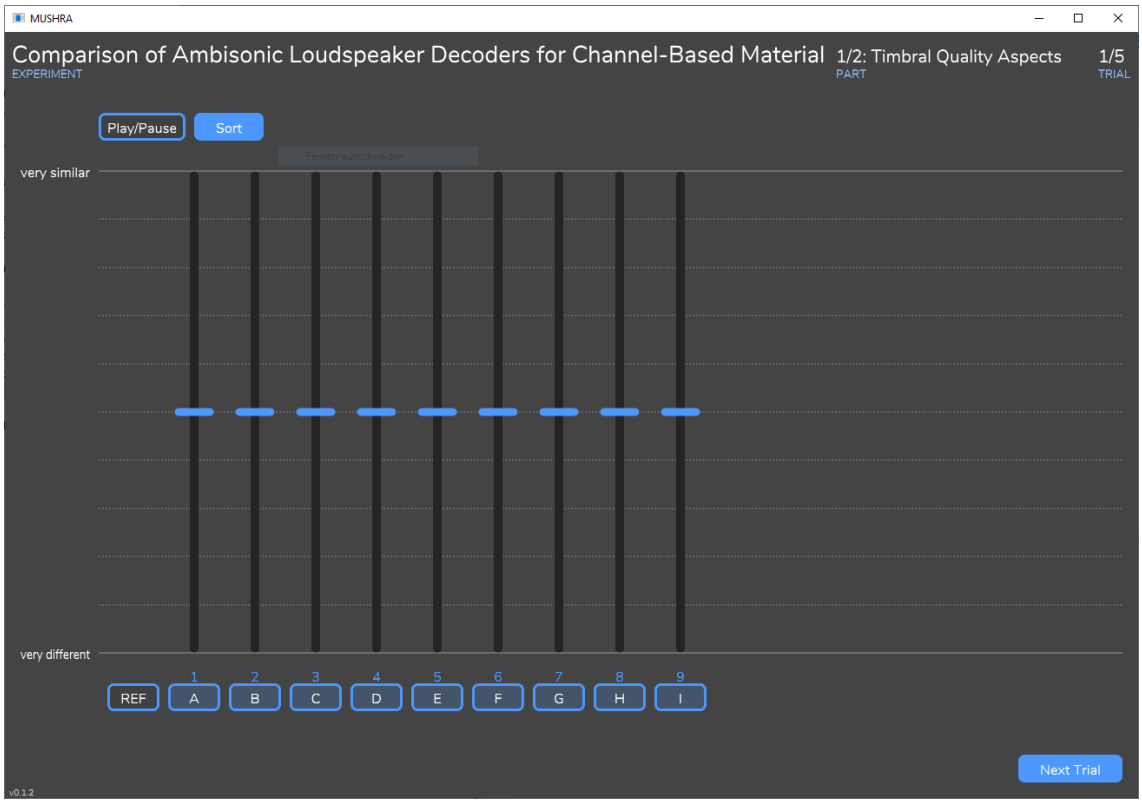


Figure 9: example window of MUSHRA-like application

example window of the application. By choosing any of the signals a fade was done for the corresponding faders in the mixer. Sometimes this fade was audible. However, the audibility seemed not to correlate with the chosen signals.

4 Results

There are basically two aspects that will be discussed when evaluating the data collected in the listening experiment: First, for the AllRAD decoder with basic weighting all orders are compared to each other and the hidden reference. This provides insight in how different orders perform for the loudspeaker setup used for the experiment. Secondly, differences between the decoders and weightings are compared. This is done for Ambisonic orders three and five. The number of pair comparisons of the ratings was reduced to 16 instead of 36 by limiting the investigation to the described aspects instead of comparing all decoders against each and every other decoder.

The significance of the differences in pairwise comparison was tested using the Wilcoxon signed-rank test with Bonferroni-Holm correction. Those methods are described briefly below.

4.1 Statistical Evaluation Design

Wilcoxon signed rank test The Wilcoxon signed-rank test is a non-parametric hypothesis test for two populations. It is non-parametric, because it also works for probability distributions that are not parametrized (e.g. a parametric distribution is the normal distribution with parameters mean and variance). One can argue, that the collected data is most certainly non-parametric for the listening experiment presented above, given perceptual identity as the upper boundary of the rating scale.

To get the test statistic T, first the differences between observations in the populations are calculated. Then ranks are assigned to the absolute differences. The ranks are then multiplied with the sign of the difference and summed afterwards. The result is T, which can be can be compared to its distribution under the null hypothesis to produce a p-value [15]. Here, a two-sided test was done. The significance level α was chosen to be 0.05.

Bonferroni-Holm correction A correction is necessary, because by comparing more than two populations (i.e. more than two decoder comparisons), more than one hypotheses are checked. By doing that, the chance of Type-I errors (false positives) rises. This can be corrected. n p-values for n comparisons can be treated the following way: First, all p-values are sorted, starting with the smallest value. If the first sorted p-value is larger than or equal to α/n, the procedure is stopped. Then none of the p-values is significant. Else, the process can be continued using the second p-value and $\alpha/(n-1)$ and so on. The

process is stopped, as soon as the p-value gets larger than or equal to α. This way, the more hypotheses are tested, the more penalty is caused. Results tend to be less significant after the correction when compared to results without correction [16].

4.2 Evaluation of listening test results

Subject validity For all scene ratings of each subject, the difference between the rating for hidden reference and maximum rating was checked. If any other decoder was rated more similar to the reference than the hidden reference, the difference was negative. The most similar rated decoder was found in that case. Fig. 10 shows which decoders were chosen as most similar if not the hidden reference was chosen. It could be seen that none of the subjects disqualified for the test since all recognized the hidden reference for most of the trials: With 240 data sets in total, confusion with the hidden reference happend in less than 10% of the sets for 7th-order basic-weighted AllRAD and in about 5% of the sets for 5th-order basic-weighted AllRAD and 5th-order SAD. For all other decorders, there was even less confusion with the hidden reference. Therefore, the 7th order AllRAD-decoder with basic weighting was rated more similar to the reference than the hidden reference most. This resembles the findings that will be looked at below.

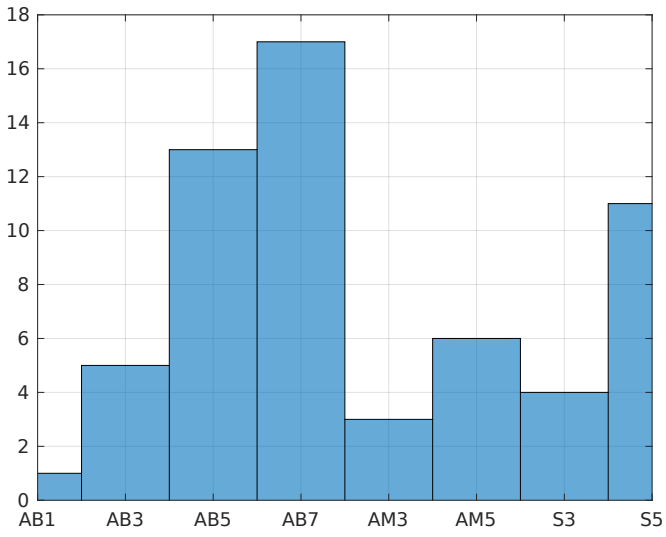


Figure 10: Maximum similarity ratings different from hidden reference for all scenes and subjects.

Data Evaluation Fig. 11 to fig. 18 show the median and 95% confidence intervals of the ratings for each attributes and scene. In fig. 11 to fig. 14 an increase of similarity to the reference towards higher orders for AllRAD basic weighted decoding is clearly visible. It can be observed, that all orders are significantly different to the hidden reference for the center listening position ($p \leq 0.0391$). Most orders are also significantly different to each other. For the off-center listening position, the 7th order is not significantly different to the reference ($p \geq 0.125$). An exception for this is found for decorrelated noise and timbral quality aspects, where the ratings are significantly different to the hidden reference ($p = 0.0117$). Also, no significant differences between 5th and 7th order can be found here ($p = 1.3809$).

When all 3rd order decoders are compared, there is a tendency that basic-weighted AllRAD and SAD are rated more similar to the reference than the AllRAD decoder with max-r_E weighting. However, differences are not always significant: For noise from FC and off-center listening position there are no significant differences between the decoders ($p \geq 0.3047$). For decorrelated noise at center listening position for both attributes, as well as for timbral quality at the off-center listening position, AllRAD with basic weighting and SAD perform significantly better than AllRAD with max-r_E weighting ($p \leq 0.0293$). For noise from FC, the center position and timbral quality aspects, AllRAD basic is rated more similar to the reference than the SAD ($p \leq 0.0249$).

The same trend can be seen for the 5th order decoders. Here, for center listening position, noise from FC and spatial aspects, for center listening position, decorrelated noise and timbral aspects, and for off-center listening position, decorrelated noise and timbral aspects AllRAD basic and SAD are rated more similar than AllRAD max-r_E ($p \leq 0.0391$). There is also the case that all decoders are rated significantly different. If so, AllRAD basic is rated more similar than SAD and SAD is rated more similar than AllRAD max-r_E ($p <= 0.0117$). This is the case for noise from FC and timbral aspects for both center and off-center listening position. For the off-center position, noise from FC and spatial aspects, AllRAD basic is significantly more similar than AllRAD max-r_E ($p = 0.049$). SAD here is significantly different to neither of the two other decoders.

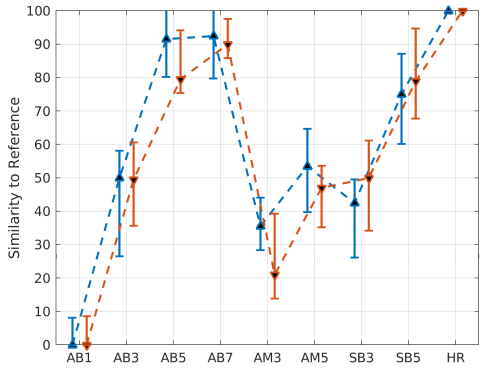


Figure 11: Center position spatial similarity noise scenes.

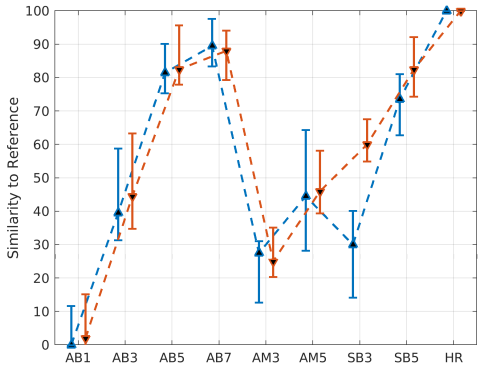


Figure 12: Center position timbral similarity noise scenes.

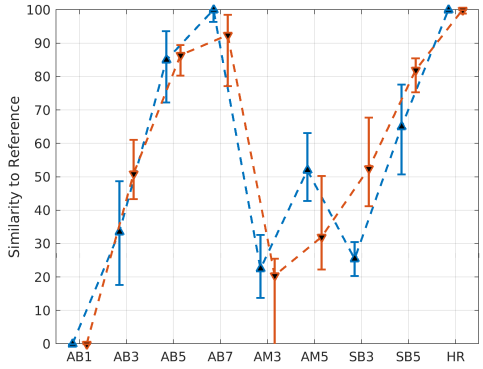


Figure 13: Off-center position spatial similarity noise scenes.

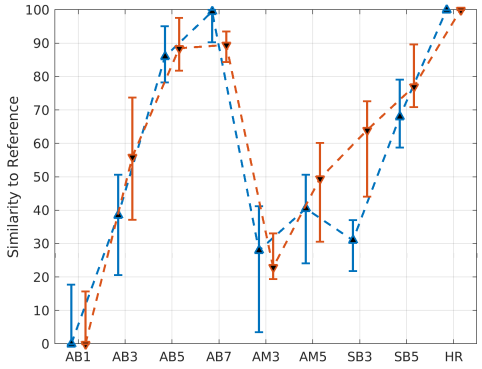


Figure 14: Off-center position timbral similarity noise scenes.

A similar behavior can be recognized for the three music scenes. In general, encoding and decoding seems to be more difficult to notice for music scenes than for noise scenes. For the AllRAD basic decoder 5th and 7th order are rated not significantly different to the hidden reference ($p \geq 0.0527$). The first order is always significantly different to all others as well as to the hidden reference ($p \leq 0.0156$). Two times the third order is significantly different to the hidden reference ($p \leq 0.023$), but not to higher orders. This is the case for timbral quality aspects for the center position, piano trio and off-center position, a capella ensemble.

Again, there is a trend for the differences between decoders for the third order similar to the trend for the noise: AllRAD basic and SAD are rated more similar than AllRAD max-r_E. Sometimes AllRAD basic is rated significantly more similar to the reference than SAD. Sometimes there are no significant differences. For off-center position, piano trio and timbral aspects, SAD was rated to be significantly more similar to the reference than AllRAD with basic weighting ($p = 0.0041$).

However, comparing all 5th order decoders, no significant differences can be found between the decoders for the center listening position and timbral aspects of the off-center position ratings ($p \geq 0.0591$). Apart from that, AllRAD basic and SAD perform better than AllRAD with max-r_E weighting. This shows that the difference between the decoders is more noticeable for lower orders.

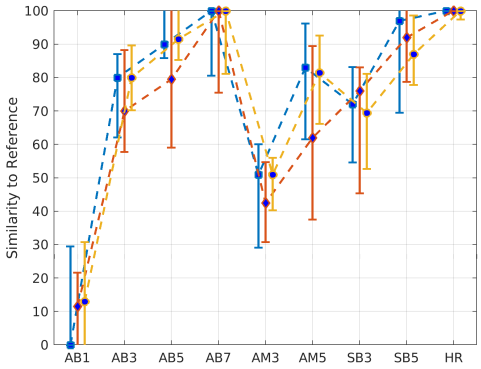


Figure 15: Center position spatial similarity music scenes.

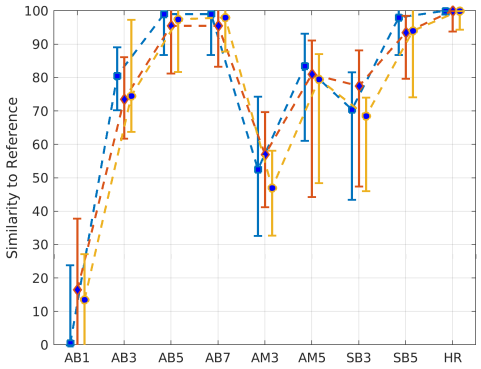


Figure 16: Center position timbral similarity music scenes.

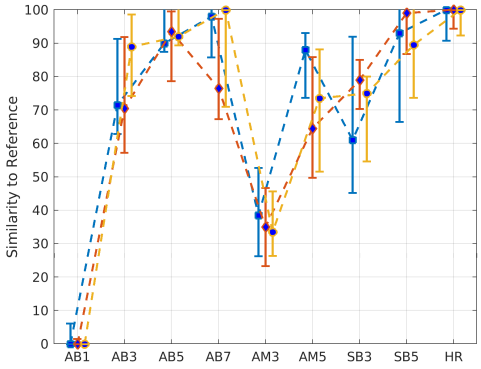


Figure 17: Off-center position spatial similarity music scenes.

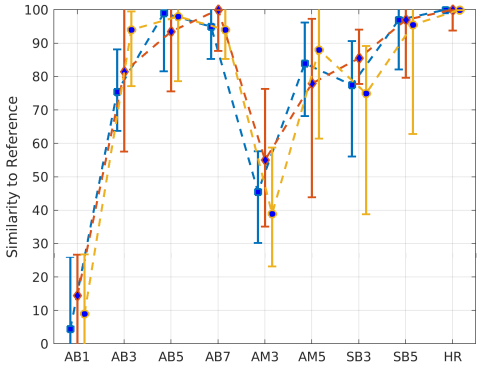


Figure 18: Off-center position timbral similarity music scenes.

Combinations of data Although differences between the ratings for noise scenes and music scenes are obvious, there are only little differences within the ratings for noise and music scenes, respectively. To show this, the correlation between the ratings for each listening position and attribute was calculated for the noise scenes and the music scenes. The minimum correlations between the scenes range from 0.92 to 0.99. Due to such high correlation, all noise scenes and all music scenes can be combined. This way, there are 24 ratings for the noise scenes and 36 ratings for the music scenes for the twelve subjects that took part in the listening experiment.

In fig. 19 and fig. 20 we see the medians and confidence intervals for spatial and timbral quality aspects for the combined noise scene ratings. The center listening position is plotted against the off-center listening position. For AllRAD basic weighting, all orders are significantly different to the reference, except for the off-center listening position and spatial aspects ($p = 0.0873$). For the center position and spatial aspects and the off-center position and timbral aspects the 5th order is not significantly different from the 7th order ($p \geq 0.0544$). Comparing all 3rd and 5th orders, AllRAD basic and SAD are always rated significantly more similar to the reference than AllRAD max-r_E. For the 5th order, off-center position and spatial quality aspects AllRAD basic is rated significantly more similar than SAD ($p = 0.007$).

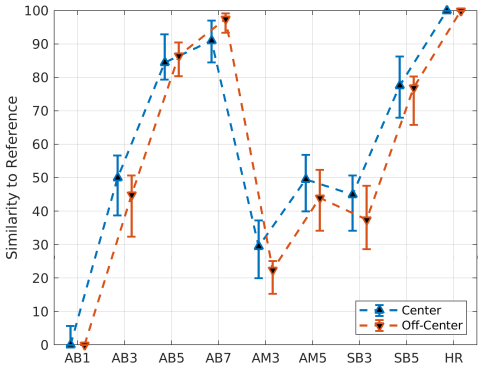


Figure 19: Spatial similarity combined noise scenes.

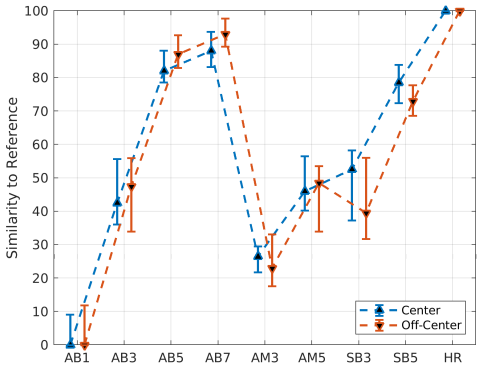


Figure 20: Timbral similarity combined noise scenes.

The combined ratings for the music examples are shown in fig. 21 and fig. 22. 5th and 7th order of basic-weighted AllRAD are both not significantly different to the hidden reference for the off-center listening position ($p \leq 0.1388$). For the center listening position there are differences between spatial and timbral quality aspects: For spatial aspects only the 5th order is not significantly different to the hidden reference ($p = 0.1675$), for timbral aspects only the 7th order ist ($p = 0.0975$). For both listening positions and spatial aspects only the 1st order is significantly different to all other orders ($p = 0$). For timbral aspects 1st and 3rd order are significant to each other and to 5th and 7th order ($p \leq 0.0077$). 5th and 7th order AllRAD with basic weighting are always not significantly different from each other ($p \geq 0.3157$). This implies that working in 7th order does not have significant advantages over 5th order for working in Ambisonics with this loudspeaker setup and similar material.

Comparing the different decoders, AllRAD with basic weighting and SAD are always rated more similar to the reference than AllRAD with max-r_E weighting. ($p \leq 0.0197$). An exception to that are the ratings for 5th order, center position and spatial aspects. Here,

only SAD is rated significantly better than AllRAD with max-r_E weighting ($p = 0.0156$). Additionally, for 3rd order, the center position and timbral aspects AllRAD with basic weighting was rated to be significantly more similar to the reference than SAD ($p = 0.0344$).

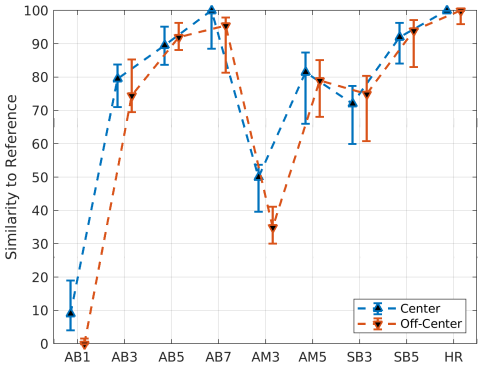


Figure 21: Spatial similarity combined music scenes.

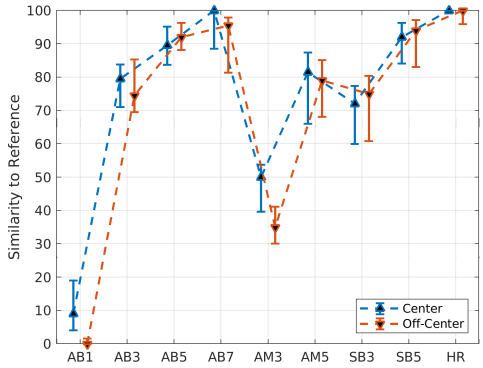


Figure 22: Timbral similarity combined music scenes.

Correlation with theoretical reasoning The results of the listening experiment were compared with the crosstalk and correlation measurements presented above in section 2.2. This was done by correlating the medians of the ratings with the measurements. The measured correlation was inverted by subtracting it from 1 beforehand: Increased correlation should match lower similarity to the reference whereas less absolute crosstalk level should match lower similarity to the reference. Results of the correlation can be seen in table 2. Looking at the ratings for the noise samples, the crosstalk seems to have a high influence on the perception of spatial as well as timbral quality aspects with correlation ranging from 0.82 to 0.98. For music material the increase in correlations appears to be more important than the introduced crosstalk for the perceived spatial and timbral impairments. Here, the correlation ranges from 0.90 to 0.97. However, it must be noted that the correlation between measured crosstalk and rating medians is nearly as high for musical examples, too. It is even equally high with 0.97 for the ratings for the center listening position and timbral quality aspects.

Generally, side lobe levels seem to be less important as the main lobe widening in this loudspeaker setup as can be inferred from the lower correlation with the crosstalk at SL/SR compared to the correlation with the crosstalk at L/R. However, increased correlation of the ratings with the measured crosstalk at SL/SR can be noticed for the off-center positions.

Ratings	Levels FL/FR	Levels SL/SR	Correlation FL/FR/FC
center spatial noise	0.96	0.61	0.75
center timbral noise	0.98	0.56	0.69
off-center spatial noise	0.84	0.73	0.81
off-center timbral noise	0.82	0.70	0.80
center spatial music	0.92	0.58	0.97
center timbral music	0.97	0.62	0.97
off-center spatial music	0.86	0.74	0.93
off-center timbral music	0.80	0.73	0.90

Table 2: Medians of ratings for different listening positions, attributes and combined scenes correlated with curves for crosstalk from section 2.2.

4.3 Conclusion and Outlook

Having considered all these findings, it is clear that significant perceptible differences are introduced to a signal by Ambisonic encoding and decoding. Especially the 1st order was perceived very different compared to the reference regarding spatial and timbral quality aspects. For the music scenes (which are therefore more related to practice) less differences were recognized. For this material, 5th order was sufficiently similar in most situations. For noise-based material, 7th order was sometimes not significantly different from the hidden reference.

In this special environment SAD and basic-weighted AllRAD decoder perform better than max-r_E-weighted AllRAD decoder. There is a tendency that the basic-weighted AllRAD decoder introduces less differences compared to SAD.

As an outlook on further investigations it could be very interesting to repeat the listening test on other loudspeaker layouts in other rooms. Especially larger layouts could behave different for spatial quality aspects. Here, max-r_E weighting might have a positive influence on similarity when the off-center position is further out of the sweet spot than it was the case in this studio environment. Another possible experiment is the placement of channel-based sources on other positions than the loudspeakers positions and comparing the encoded and decoded Ambisonic signal to a reference speaker setup at the actual positions. Further, equalization as a compensation for the timbral changes caused by the encoding and decoding process seems worth to be investigated.

This research project was extended to 15 test subjects and the results were then presented at the German Annual Conference on Acoustics (DAGA) in Stuttgart in March 2022.

A Channel Naming Conventions and Positions

Abbreviation	**Full Name**	**Azimut / °**	**Elevation / °**
FL	Front Left	-30	0
FL	Front Right	30	0
FC	Front Center	0	0
SL	Side Left	-90	0
SR	Side Right	90	0
BL	Back Left	-150	0
BR	Back Left	150	0
TFL	Top Front Left	-45	30
TFR	Top Front Right	45	30
TBL	Top Back Left	-135	30
TBR	Top Back Right	135	30
VOG	Voice of God	-	90

References

[1] Hyunkook Lee, Matthias Frank, and Franz Zotter. "Spatial and Timbral Fidelities of Binaural Ambisonics Decoders for Main Microphone Array Recordings". In: *Audio Engineering Society Conference: 2019 AES International Conference on Immersive and Interactive Audio*. Mar. 2019. URL: http://www.aes.org/e-lib/browse.cfm?elib=20392 (visited on 01/13/2022).

[2] Franz Zotter and Matthias Frank. *Ambisonics: A Practical 3D Audio Theory for Recording, Studio Production, Sound Reinforcement, and Virtual Reality*. Vol. 19. Springer Topics in Signal Processing. Cham: Springer International Publishing, 2019.

[3] Jürgen Herre et al. "MPEG-H Audio—The New Standard for Universal Spatial/3D Audio Coding". In: *J. Audio Eng. Soc* 62.12 (2015), pp. 821–830.

[4] Matthias Frank, Franz Zotter, and Alois Sontacchi. "Producing 3D Audio in Ambisonics". In: *Audio Engineering Society Conference: 57th International Conference: The Future of Audio Entertainment Technology – Cinema, Television and the Internet*. Mar. 2015. URL: http://www.aes.org/e-lib/browse.cfm?elib=17605 (visited on 01/14/2022).

[5] Franz Zotter and Matthias Frank. "All-Round Ambisonic Panning and Decoding". In: *J. Audio Eng. Soc* 60.10 (2012), pp. 807–820.

[6] Institut für Elektronische Musik und Akustik. *Räume*. URL: https://iem.kug.ac.at/services/raeume.html (visited on 01/11/2022).

[7] M. R. Schroeder. "New Method of Measuring Reverberation Time". In: *The Journal of the Acoustical Society of America* 37.3 (Mar. 1965), pp. 409–412.

[8] Georg Neumann GmbH. *KH 310*. URL: https://de-de.neumann.com/kh-310-a (visited on 01/11/2022).

[9] Hyunkook Lee and Dale Johnson. "An Open-Access Database of 3D Microphone Array Recordings". In: *Audio Engineering Society Convention 147*. Oct. 2019. URL: http://www.aes.org/e-lib/browse.cfm?elib=20566 (visited on 12/30/2021).

[10] Hyunkook Lee. "A New Multichannel Microphone Technique for Effective Perspective Control". In: *Audio Engineering Society Convention 130*. May 2011. URL: http://www.aes.org/e-lib/browse.cfm?elib=15804 (visited on 01/09/2022).

[11] Hyun Kook Lee and Christopher Gibben. "Effect of Vertical Microphone Layer Spacing for a 3D Microphone Array". In: *Journal of the Audio Engineering Society* 62.12 (Jan. 5, 2015), pp. 870–884.

[12] Günther Theile and Helmut Wittek. "3D Audio Natural Recording". In: 27th Tonmeistertagung. 2012.

[13] Rory Wallis and Hyunkook Lee. "The Reduction of Vertical Interchannel Crosstalk: The Analysis of Localisation Thresholds for Natural Sound Sources". In: *Applied Sciences* 7.3 (Mar. 14, 2017), p. 278.

[14] ITU. *Recommendation BS.1534: Method for the subjective assessment of intermediate quality level of audio systems*. 2015. URL: `https://www.itu.int/rec/R-REC-BS.1534-3-201510-I/en` (visited on 01/09/2022).

[15] Mathworks. *Wilcoxon signed rank test - MATLAB signrank*. URL: `https://www.mathworks.com/help/stats/signrank.html` (visited on 01/18/2022).

[16] Sture Holm. "A Simple Sequentially Rejective Multiple Test Procedure". In: *Scandinavian Journal of Statistics* 6.2 (1979), pp. 65–70.