PERCEPTUAL EVALUATION OF MITIGATION APPROACHES OF IMPAIRMENTS DUE TO SPATIAL UNDERSAMPLING IN BINAURAL RENDERING OF SPHERICAL MICROPHONE ARRAY DATA: DRY ACOUSTIC ENVIRONMENTS

Tim Lübeck, Johannes M. Arend, Christoph Pörschmann [*]

Institute of Communications Engineering
TH Köln - University of Applied Sciences,
50678 Cologne, Germany
tim.luebeck@th-koeln.de

Hannes Helmholz, Jens Ahrens [†]

Division of Applied Acoustics
Chalmers University of Technology
412 96 Gothenburg, Sweden
hannes.helmholz@chalmers.se

ABSTRACT

Employing a finite number of discrete microphones, instead of a continuous distribution according to theory, reduces the physical accuracy of sound field representations captured by a spherical microphone array. For a binaural reproduction of the sound field, a number of approaches have been proposed in the literature to mitigate the perceptual impairment when the captured sound fields are reproduced binaurally. We recently presented a perceptual evaluation of a representative set of approaches in conjunction with reverberant acoustic environments. This paper presents a similar study but with acoustically dry environments with reverberation times of less than 0.25 s. We examined the Magnitude Least-Squares algorithm, the Bandwidth Extraction Algorithm for Microphone Arrays, Spherical Head Filters, spherical harmonics Tapering, and Spatial Subsampling, all up to a spherical harmonics order of 7. Although dry environments violate some of the assumptions underlying some of the approaches, we can confirm the results of our previous study: Most approaches achieve an improvement whereby the magnitude of the improvement is comparable across approaches and acoustic environments.

1. INTRODUCTION

Spherical microphone arrays (SMAs) allow for capturing sound fields including spatial information. The captured sound fields can be rendered binaurally if the head-related transfer functions (HRTFs) are available on a sufficiently dense grid. Mathematically, this is performed by means of spherical harmonics (SH) expansion of the sound field and the HRTFs [1, 2]. Conceptually, it is equivalent to bringing the listener's head virtually into the sound field captured with the array. Rotation of the HRTFs relative to the sound field according to the instantaneous head orientation of the listener allows for dynamic presentation.

The physical accuracy that can be achieved with SMAs is limited, mainly due to the employment of a finite number of microphones as opposed to the continuous distribution that the theory assumes. This leads to spatial undersampling of the captured sound field, which 1) induces spatial aliasing and 2) limits the maximum obtainable SH order representation. The order of the SH presentation directly corresponds to the spatial resolution of the captured sound field. Both phenomenons can lead to audible artifacts. Another practical impairment is caused by self-noise of the microphones in the array. Studying this aspect is beyond the scope of the present paper. We refer the reader to [3, 4].

In recent years, several approaches to mitigate such impairments in binaural rendering of undersampled SMA data have been proposed. We recently conducted a listening experiment to study the perceptual effects of the mitigation approaches [5]. The study employed the acoustic data of two rooms with a reverberation time of more than 1 s. In this contribution we present the results for a similar study, whereby the employed acoustic environments exhibit shorter reverberation times of less than 0.25 s.

2. SPATIAL UNDERSAMPLING

To outline the phenomenon of spatial undersampling, we briefly summarize the fundamental concept of binaural rendering of SMA data. For a more detailed explanation please refer to [2, 6]. The sound pressure $S(r, \phi, \theta, \omega)$ captured by the microphones on the array surface Ω is represented in the SH domain using the spherical Fourier transform (SFT)

$$S_{nm}(r, \omega) = \int_{\Omega} S(r, \phi, \theta, \omega)\, Y_n^m(\theta, \phi)^*\, \mathrm{d}A_\Omega\,, \tag{1}$$

whereby r denotes the array radius, ϕ and θ the azimuth and colatitude of a point on the array surface, and $\omega = 2\pi f$ the angular frequency. $Y_n^m(\theta, \phi)$ denotes the orthogonal SH basis functions for certain orders n and modes m and $(\cdot)^*$ the complex conjugate.

Based on knowledge of the sound field SH coefficients S_{nm}, the sound field on the array surface can be decomposed into a continuum of plane waves impinging from all possible directions

$$D(\phi, \theta, \omega) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} d_n\, S_{nm}(r, \omega)\, Y_n^m(\phi, \theta)\,, \tag{2}$$

with a set of radial filters d_n. Note that $S(r, \phi, \theta, \omega)$ and $D(\phi, \theta, \omega)$ do not necessarily represent the same sound fields. A SMA can incorporate a scattering body whose effect is contained in $S(r, \phi, \theta, \omega)$ but not in $D(\phi, \theta, \omega)$ where it is removed by the radial filters.

A HRTF $H(\phi, \theta, \omega)$ can be interpreted as the spatio-temporal transfer function of a plane wave to the listeners' ears. The binaural signals $B(\omega)$ for the left or right ear due to the plane wave components $D(\phi, \theta, \omega)$ impinging on the listener's head can therefore

[*] This work was partly supported by ERDF (European Regional Development Fund).

[†] This work was partly supported by Facebook Reality Labs.

be computed by weighting all HRTFs $H(\omega)$ with the plane wave coefficients of $D(\phi_d, \theta_d, \omega)$ and integrating over all propagation directions

$$B(\omega) = \frac{1}{4\pi} \int_{\Omega} H(\phi, \theta, \omega)\, D(\phi, \theta, \omega)\, \mathrm{d}A_{\Omega}\,. \tag{3}$$

Transforming the HRTFs into the SH domain as well and exploiting the orthogonality property of the SH basis functions allows to resolve the integral and compute the binaural signals for either ear as [1]

$$B(\omega) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} d_n\, S_{nm}(\omega, r)\, H_{nm}(\omega)\,. \tag{4}$$

The exact formulation of Eq. (4) depends on the particular definition of the employed SH basis functions [7, p. 7].

So far, we have assumed a continuously and ideally sampled sound pressure distribution on the array surface. In this case, the computation of the ear signals is perfect i.e., $B(\omega)$ in (4) are the signals that arise if the listener with HRTFs $H(\phi, \theta, \omega)$ is exposed to the sound field that the microphone array captures. Real-world SMAs employ only a finite number of discrete microphones. As a result, spatial aliasing and truncation of the SH order n occur, which makes the ear signals that are computed by the processing pipeline differ from the true ones. This can significantly affect the perceptual quality of binaural reproduction, as shown by numerous research [2, 8, 9, 10]. These impairments due to spatial undersampling are briefly discussed in the following.

2.1. Spatial Aliasing

Similar to time-frequency sampling, where frequency components above the Nyquist-frequency are aliased to lower frequency regions, sampling the space with a limited number of sensors introduces spatial aliasing. Note that this applies for both, sampling of the sound field $S(\cdot)$ as well as for the sampling of the HRTFs $H(\cdot)$. In case aliasing occurs, higher spatial modes cannot be reliably resolved and leak into lower modes. Generally, higher modes are required for resolving high frequency components with smaller wavelengths. Spatial aliasing therefore limits the upper bound of the time-frequency bandwidth that can be deduced reliably from the array signals. While theoretically being apparent at all temporal frequencies f, spatial aliasing artifacts are considerable only above the temporal-frequency [6]

$$f_{\mathrm{A}} = \frac{N_{\mathrm{sg}}\, c}{2\pi r}\,. \tag{5}$$

Thereby, c denotes the speed of sound and N_{sg} the maximum resolvable SH order n of the sampling scheme. The leakage of higher spatial modes into lower spatial modes results in an increase of the magnitudes at temporal-frequencies above f_{A}. Although spatial aliasing primarily impairs spatial properties, it therefore also affects the time-frequency spectrum of the binaural signals.

2.2. Spherical Harmonic Truncation

Orthogonality of the SH basis functions $Y_n^m(\cdot)$ is given only up to the order $n = N_{\mathrm{sg}}$ (Eq. (5)) due to the discrete sampling of the SMA surface. Spatial modes for $n > N_{\mathrm{sg}}$ are spatially distorted and are ordinarily not computed. This order truncation results in a loss of spatial information. The sampling of the SMA is usually sparser than that of the HRTFs so that the SMA is the limiting factor.

Also the spatial order truncation affects the time-frequency representation by discarding components with mostly high frequency content. In addition, hard truncation of the SH coefficients at a certain order n results in side-lobes in the plane wave spectrum in Eq. (2) [11], which can further impair the binaural signals.

3. MITIGATION APPROACHES

In the last years, a number of different approaches to improve binaural rendering of SMA captures have been presented in the literature. In the following, a selection of approaches is summarized. These are the approaches that we evaluated in the experiment presented in Sec. 6.

3.1. Pre-Processing of Head-Related Transfer Functions

Since in practice, the SH order truncation of high-resolution HRTFs cannot be avoided, a promising approach to mitigate the truncation artifacts is to pre-process the HRTFs in such a way that the major energy is shifted to lower orders without notably decreasing the perceptual quality. Several approaches to achieve this have been introduced. A summary of a selection of pre-processing techniques is presented in [12]. In this paper, we investigate two concepts.

3.1.1. Spatial Subsampling

For the spatial subsampling method [2] (SubS), the HRTFs are transformed into the SH domain up to the highest SH order N_{sg} that the sampling grid supports. Based on this representation, the HRTFs are spatially resampled with a reduced maximum SH order N'_{sg} to the grid on which the sound field is sampled, which is usually more coarse.

This process modifies the spatial aliasing in the signals in a favorable way [2]. Fig. 1 depicts the energy distribution of dummy head HRTFs [13] with respect to SH order (y-axis) and frequency (x-axis). The left-hand diagram illustrates the untreated HRTFs with a significant portion of energy at high SH orders. The middle diagram shows the same HRTF set being subsampled to a 5th-order Lebedev grid. Evidently, the information can be reliably obtained only up to the 5th order.

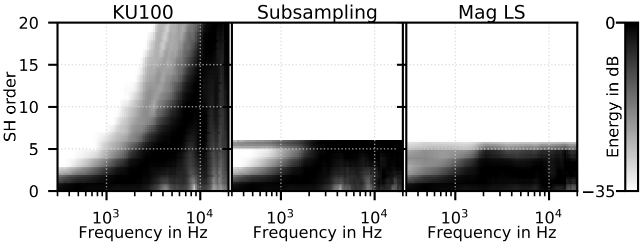


Figure 1: Energy distribution in dB with respect to order and frequency of the HRTFs of a Neumann KU100 dummy head. Untreated (left), subsampled (center), MagLS pre-processed (right).

3.1.2. Magnitude Least-Squares

Another HRTF pre-processing approach is the Magnitude Least-Squares (MagLS) [14] algorithm, which is an improvement of the Time Alignment (TA) proposed by the same authors. Both approaches are based on the duplex theory [15]. At high frequencies, the interaural level differences (ILDs) become perceptually more relevant than the interaural time differences (ITDs). However, at high frequencies, the less relevant phase information constitutes a major part of the energy. Thus, removing the linear phase at high frequencies decreases the energy in high modes, without losing relevant perceptual information. MagLS aims to find an optimum phase by solving a least-squares problem that minimizes the differences in magnitude to a reference HRTF set, resulting in minimal phase in favor of optimal ILDs. Fig. 1 (right) illustrates the energy distribution of MagLS pre-processed HRTFs for SH order 5. The major part of the energy is shifted to SH coefficients of orders below 5.

The major difference between both HRTF pre-processing approaches is that subsampling results in a HRTF set defined for a reduced number of directions and thus allowing only for a limited SH representation. In contrast, MagLS does not change the HRTF sampling grid and thus, theoretically, allows expansion up to the original SH order.

3.2. Bandwidth Extension Algorithm for Microphone Arrays

Besides pre-processing of the HRTFs, there are algorithms that are applied to the sound field SH coefficients. The Bandwidth Extension Algorithm for Microphone Arrays (BEMA) [16, 2] synthesizes the SH coefficients at $f \geq f_{\mathrm{A}}$ by extracting spatial and spectral information from components $f < f_{\mathrm{A}}$. The time-frequency spectral information is obtained by an additional omnidirectional microphone in the center of the microphone array (which is evidently not feasible in practice if a scattering object is employed). The BEMA coefficients can then be estimated as the combination of spatial and spectral information.

Fig. 2 depicts the magnitudes of plane wave components calculated for a broadband plane wave impinging from $\phi = 180°$, $\theta = 90°$ on a 50 sampling point Lebedev grid SMA with respect to azimuth angle (x-axis) and frequency (y-axis). The top diagram is based on untreated SH coefficients, the bottom diagram illustrates the effect of BEMA. For the example of a single plane wave, the sound field is perfectly reconstructed over the entire audible bandwidth.

3.3. Spherical Harmonic Tapering

SH order truncation induces side-lobes in the plane wave spectrum, which can be reduced by tapering high orders n [11]. In other words, an order-dependent scaling factor is applied to all SH modes and coefficients of that order. Different windows have been discussed, and a cosine-shaped fade-out was found to be the optimal choice. Additionally, the authors recommend to equalize the binaural signals with the so-called Spherical Head Filter, as discussed in the subsequent section. The combination of SH tapering and spherical head filters is referred to as Tap+SHF in the remainder.

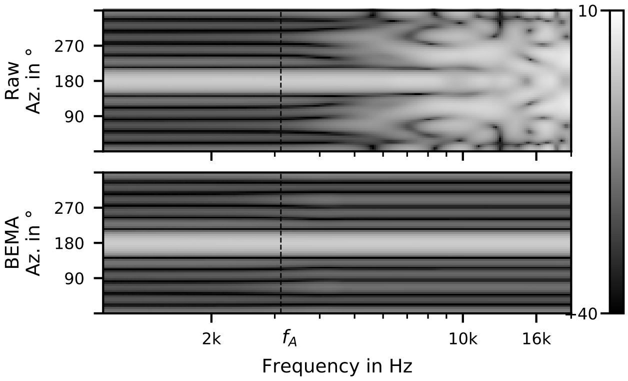


Figure 2: Plane wave magnitudes of a plane wave impact from $\phi = 180°$, $\theta = 90°$ on a 50 sampling point Lebedev grid SMA with a radius of $8.75\,\mathrm{cm}$. The top diagram depicts the untreated magnitudes, the bottom diagram the plane wave calculated after BEMA processing.

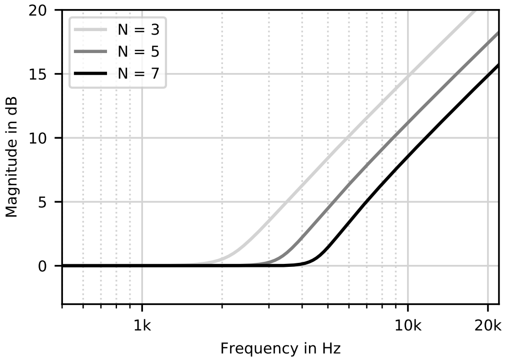


Figure 3: Spherical Head Filter (SHF) for orders $N = (3, 5, 7)$.

3.4. Spectral Equalization

The modification of the time-frequency response due to spatial undersampling is a perceptually distinctive impairment, as shown e.g. in [10]. Therefore, a third category of mitigation approaches is global equalization of the binaural signals. Different approaches have been introduced in the literature to design such equalization filters. The Spherical Head Filter (SHF) [8] compensates for the low-pass behavior of SH order truncation. The authors disregard spatial aliasing effects and proposed a filter based on the plane wave density function of a diffuse sound field. The resulting filters for different SH orders are depicted in Fig. 3. A similar approach to equalize this low-pass effect has been discussed in [17]. In the following we investigate the SHFs.

4. EMPLOYED DATA

The stimuli in our study were created from measured array room impulse responses using the `sound_field_analysis-py` Python toolbox [18] and the impulse response data set from [19]. This data set contains both binaural room impulse responses (BRIRs) measured with a Neumann KU100 dummy head as well as array room impulse responses (ARIRs) captured on various Lebedev grids under identical conditions. This allows for a direct

Figure 4: The Control Room 1 (left) and 7 (right) (CR1, CR7) with reverberation times of less than 0.25 s (measured at 500 Hz and 1 kHz) at the WDR Broadcast studios, that were auralized in the listening experiment.

comparison of binaural auralization of SMA data to the ground truth dummy head data. The ARIR measurements were performed with the VariSphear device [20], which is a fully automated robotic measurement system that sequentially captures directional impulse responses on a spherical grid for emulating a SMA. To obtain impulse responses of a rigid sphere array, the Earthworks M30 microphone was flush-mounted in a wooden spherical scattering body (see [19, Fig. 12]). All measurements were performed in four different rooms at the WDR broadcast studios in Cologne, Germany. In this study we employ the measurement data of the rooms Control Room 1 (CR1) and Control Room 7 (CR7) (Fig. 4), which both have short reverberation times of less than 0.25 s. Recall that we conducted a similar study with the rooms Small Broadcast Studio (SBS) and Large Broadcast Studio (LBS) with approximate reverberation times of 1 s and 1.8 s in [5].

The Neumann KU100 HRIR set, measured on a 2702 sampling point Lebedev grid [13], is used to synthesize binaural signals $B(\omega)$ for a pure horizontal grid of head orientations with 1° resolution based on ARIRs according to Eq. (4). We denote this data "ARIR renderings" in the following. Likewise, the BRIRs of the dummy head are available for the same head orientations so that a direct comparison of both auralizations is possible.

In order to restrict the gain of the radial filters $d_n(\omega)$ in (4), we employ a soft-limiting approach [2, pp. 90-118]. Fig. 5 illustrates the influence of the soft-limiting for the left-ear binaural room transfer functions (BRTFs) resulting from a broadband plane wave impinging from ($\phi = 0°$, $\theta = 90°$) on a simulated 2702 sampling point Lebedev SMA. The BRTFs were calculated up to the 35th-order using the different radial filter limits 0, 10, 20, and 40 dB. It can be seen that a limit of 0 dB leads to a significant attenuation of the high frequency components, but provides an advantageous signal-to-noise ratio in the resulting ear signals nevertheless [2, 4]. Although this is not required for the ideal rendering conditions in this study, we chose 0 dB soft-limiting for this contribution in order to produce comparable results to previous studies [2, 10].

All mitigation algorithms were implemented with `sound_field_analysis-py` [18]. Solely the MagLS HRIRs were pre-processed with MATLAB code provided by the authors of [14]. Every ARIR parameter set was processed with each of the mitigation algorithms MagLS, Tapering+SHF, SHF, and SubS (Spatial Subsampling), as well as an untreated (Raw) ARIR rendering was produced.

Previous studies showed that SH representations of an order of less than 8 exhibit audible undersampling artifacts, i.e., a clear perceptual difference to the reference dummy head data [10]. Since this work investigates the effectiveness of mitigation approaches for undersampled sound fields, we chose to focus on SH orders below 8 for the subsequent instrumental and perceptual evaluation. Significant beneficial effects of the mitigation approaches for higher orders are not expected.

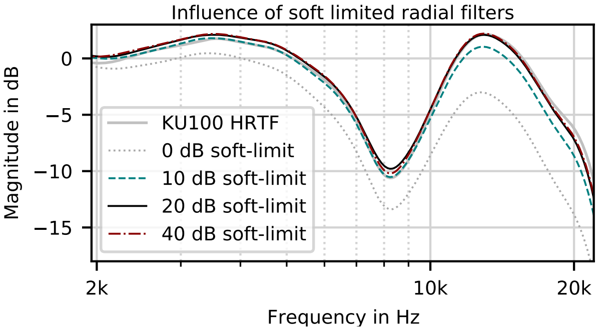


Figure 5: Left ear magnitude responses of the frontal KU100 HRTF, and ARIR binaural renderings up to order 35 involving radial filters with different soft-limits. The ARIR renderings are based on a simulated broadband plane wave impinging virtual 2702 Lebedev SMAs from ($\phi = 0°$, $\theta = 90°$). The deviation to the magnitudes of the HRTF illustrates the influence of the soft limit. All magnitude responses are 1/3-octave-smoothed.

5. INSTRUMENTAL EVALUATION

In this section, we compare the mitigation approaches based on 3rd SH order array data of CR7, which has a reverberation time of about 0.25 s. We used ARIRs from a 50-point Lebedev grid. We calculated the BRIRs for 360 azimuth directions in the horizontal plane in steps of 1° and compare them to the measured ground truth dummy head BRIRs for the same head orientations.

Absolute spectral differences between dummy head and array BRIRs in dB are illustrated in Fig. 6. The top diagram depicts the deviations averaged over all 360 directions with respect to frequency (x-axis). The bottom diagram shows the differences averaged over 40 directions contralateral to the source position. It is evident that the spectral differences tend to be larger on this contralateral side.

The untreated (Raw) rendering indicated by the dashed line is clearly affected by undersampling artifacts above f_A. Around the

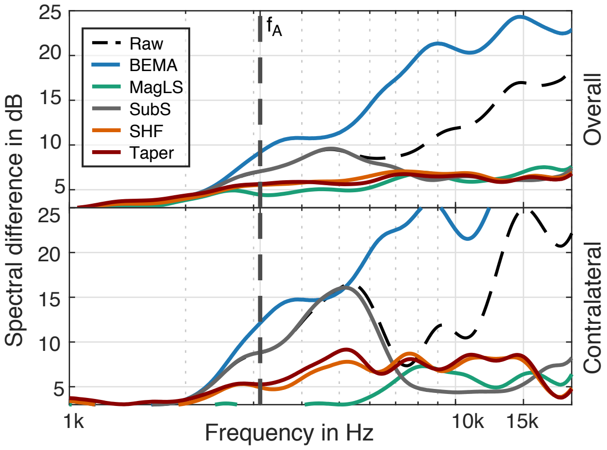


Figure 6: Absolute spectral differences of dummy head and SMA binaural signals in dB. Top: averaged over 360 horizontal directions. Bottom: averaged over 40 directions around the contralateral side.

contralateral side, these differences increase rapidly. Both HRTF pre-processing algorithms (SubS (gray) and MagLS (green)) significantly decrease the difference to the reference whereby MagLS tends to produce the lowest deviations.

Although BEMA (blue) was shown to be effective for very simple sound fields like a single plane wave, it produces significantly larger deviations from the reference than Raw. As noted by the authors of BEMA [2], even for a simple sound field composed of three plane waves from different directions and arbitrary phase, BEMA introduces audible comb filtering artifacts. Additionally, the averaging of the SH coefficients from lower modes to extract the spatial information for higher modes, leads to a perceivable low-pass effect, which produces the large differences towards higher frequencies.

The SHFs and Tapering perform comparably. Both methods employ global filtering to the binaural signals. The differences at the contralateral side are larger than for frontal directions.

6. PERCEPTUAL EVALUATION

Some of the approaches considered here have already been perceptually evaluated in listening experiments. Subsampling showed to significantly improve the perceptual quality [2], although it provokes stronger spatial aliasing. Time Alignment, Subsampling and SHFs were compared in [9]. The results showed that mostly Time Alignment, which is a predecessor of MagLS, yields better results than Subsampling. The SHFs were rated worst of the three tested methods, matching the instrumental results depicted in Fig. 6. This may be due to the fact that global equalization shifts the error in binaural time-frequency spectra to lateral directions. The perceptual evaluation of BEMA showed improvements when auralizing simulated sound fields with a limited number of sound sources [2]. However, for measured diffuse sound fields, BEMA introduces significant artifacts and thus is no promising algorithm for real-world applications. To our knowledge, Tapering has not been evaluated perceptually in a formal manner.

6.1. Methods

6.1.1. Stimuli

The stimuli were calculated as described in Sec. 4 for the SH orders 3, 5 and 7 for 360 directions along the horizontal plane with steps of 1° for the room CR7 and CR1. The 3rd and 5th-order renderings are based on impulse response measurements on the 50 sampling point Lebedev grid while for order 7 the 86 sampling point Lebedev grid was used. Previous studies showed strong perceptual differences between ARIR and dummy head auralizations in particular for lateral sound sources [9, 10]. Therefore, each ARIR rendering was generated for a virtual source in the front ($\phi = 0°$, $\theta = 90°$) and at the side ($\phi = 90°$, $\theta = 90°$). To support transparency, static stimuli for both tested sound source positions are publicly available [1]. Anechoic drum recordings were used as the test signal in particular because drums have a wide spectrum and strong transients making them a critical test signal. Previous studies showed that certain aspects are only induced with critical signals [2, 10].

6.1.2. Setup

The experiment was conducted in a quiet acoustically damped audio laboratory at Chalmers University of Technology. The SoundScape Renderer (SSR) [21] in binaural room synthesis (BRS) mode was used for dynamic auralization. It convolves arbitrary input test signals with a pair of BRIRs corresponding to the instantaneous head orientation of the listener, which was tracked along the azimuth with a Polhemus Patriot tracker. The binaural renderings were presented to the participants using AKG K702 headphones with a Lake People G109 headphone amplifier at a playback level of about 66 dBA. The output signals of the SSR were routed to an Antelope Audio Orion 32 DA converter at 48 kHz sampling frequency and a buffer length of 512 samples. Equalization according to [19] was applied to the headphones and the dummy head. The entire rendering and performance of the listening experiment were done on an iMac Pro 1.1.

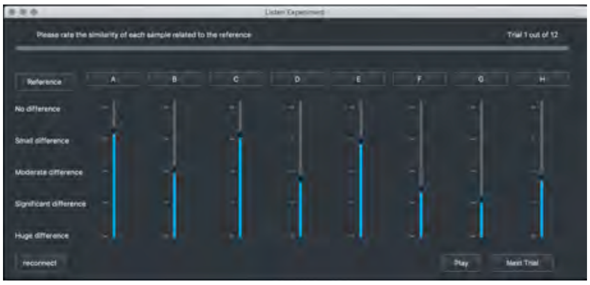
6.1.3. Paradigm and Procedure

The test design was based on the Multiple Stimulus with Hidden Reference and Anchor (MUSHRA) methodology proposed by the International Telecommunication Union (ITU) [22]. The participants were asked to compare the ARIR renderings to the dummy head reference in terms of overall perceived difference. The anchor consists of diotic non-head-tracked BRIRs, low-pass filtered at a cutoff at 3 kHz. Each trial, i.e., a MUSHRA page, comprised 8 stimuli to be rated by the subjects (BEMA, MagLS, SHF, Tapering+SHF, SubS, Raw, hidden reference (Ref), Anchor). The experiment was composed of 12 trials: 3 SH orders $(3, 5, 7) \times 2$ nominal source positions $(0°, 90°) \times 2$ rooms (CR1, CR7).

The subjects were provided a graphical user interface (GUI) with continuous sliders ranging from '*No difference*', '*Small difference*', '*Moderate difference*', '*Significant difference*' to '*Huge difference*' as depicted in Fig. 7.

14 participants in the age between 21 and 50 years took part in the experiment. Most of them were MSc students or staff at the Division of Applied Acoustics of Chalmers University of Technology. The subjects were sitting in front of a computer screen with a keyboard and a mouse. The drum signal was playing continuously,

[1] `http://doi.org/10.5281/zenodo.3931629`



Please rate the similarity of each sample related to the reference

Trial 1 out of 12

Figure 7: Employed graphical user interface of the listening experiment.

and it was possible to listen to each stimulus as often and long as desired. The participants were allowed and strongly encouraged to move their heads during the presentation of the stimuli. At the beginning of each experiment, the subjects rated four training stimuli that covered the entire range of perceptual differences of the presented stimuli in the main part of the experiment. These training stimuli consisted of a BEMA and MagLS rendering of CR1 data at order 3 for the lateral sound source position as well as the corresponding anchor and reference. The experiment took on average about 30 minutes per participant.

6.2. Results

As recommended by the ITU [22], we post-screened all reference and anchor ratings. Two participants rated the anchor higher than 30 (44, 36). We found no further inconsistencies so that we chose not to exclude these participants.

In the listening experiment, we solely presented one order and one direction per trial. We want to therefore highlight that the direct comparison of the ratings for different orders and different source positions as well as subsequent interpretation has to be performed with reservation. All stimuli were presented in randomized order and the corresponding references and anchors were always the same for each condition so that some amount of consistency in the subject's responses may be assumed. We therefore present a statistical analysis in the following that includes comparisons between orders and positions as it is commonly performed with MUSHRA data.

Fig. 8 presents the interindividual ratings in form of boxplots. The plots are divided for each room and sound source position and present the ratings with respect to the algorithm (x-axis) and order as indicated by the color. Two major observations can be made: 1) Considering the ratings of the Raw conditions shows that mostly higher-order renderings were perceived closer to the reference than lower-order renderings. 2) The algorithms MagLS, Tapering+SHF, SubS, and SHF all improve ARIR renderings compared to untreated renderings. This improvements seems to become weaker with increasing order.

For statistical analysis of the results, a repeated measures ANOVA was performed. We applied a Lilliefors test for normality to test the assumptions for the ANOVA. It failed to reject the null hypothesis in 4 of 72 conditions at a significance level of $p = 0.05$. However, parametric tests such as the ANOVA are generally robust to violations of normality assumption [25]. For further analysis Greenhouse-Geisser corrected p-values are considered, with the associated ϵ-values for correction of the degrees of freedom of the F-distribution being reported.

A four-way repeated measures ANOVA with the within-subject factors algorithm (BEMA, MagLS, Tapering+SHF, SHF, SubS, and Raw), order $(3, 5, 7)$, room (CR1, CR7), and nominal source position $(0°, 90°)$ was performed. The associated mean values with respect to algorithm (x-axis), and SH order (color) are depicted in Fig. 9. Each value was calculated as the mean value of the ratings of all participants for both directions and both rooms. The 95 % within-subject confidence intervals were determined as proposed by [23, 24] based on the main effect of algorithm. Similar to the boxplots, the mean values indicate that all algorithms except BEMA yield considerable improvements.

The ANOVA revealed the significant main effects algorithm ($F(5, 65) = 143.64,\ p < .001,\ \eta_p^2 = .917,\ \epsilon = .457$), and order ($F(2, 26) = 37.382$, $p < .001$, $\eta_p^2 = .742$, $\epsilon = .773$). These significant effects match the observations made so far. Mostly, higher-order renderings yielded smaller perceptual differences than lower-order ones. Further, the algorithm significantly influences the perceptual character of ARIR renderings. The ANOVA revealed the significant interaction of algorithm×order ($F(10, 130) = 4.756,\ p < .001,\ \eta_p^2 = .268,\ \epsilon = .556$). Thus, the algorithms seem to perform differently with respect to the rendering order. The significant effect of the interaction of algorithm×source position ($F(5, 65) = 7.176$, $p < .001$, $\eta_p^2 = .356$, $\epsilon = .774$) shows that the performance of the algorithm also depends on the sound source position.

The ANOVA also revealed two significant interactions involving the factor room: The interaction of algorithm×room ($F(5, 65) = 2.864,\ p < .040,\ \eta_p^2 = .181,\ \epsilon = .695$), as well as order×room ($F(2, 26) = 4.736$, $p < .024$, $\eta_p^2 = .267$, $\epsilon = .853$) were found to be significant. The results of the listening experiment and the ANOVA values, are available as well [1].

7. DISCUSSION AND CONCLUSIONS

We presented a perceptual evaluation of approaches for mitigating the perceptual impairment due to spatial aliasing and order truncation in binaural rendering of spherical microphone array data. The present results employing dry acoustic environments together with previous results on reverberant environments [5] suggest the following:

- Bandwidth Extension Algorithm for Microphone Arrays (BEMA) is the only method that causes larger perceptual differences to the ground truth signal than without mitigation.
- Depending on the condition, all other mitigation approaches produce either no improvement or an improvement that is comparable in magnitude.
- Mitigation is more effective at lower orders and is hardly detectable at order 7.
- We did not find a dependency on the room although some mitigation approaches are based on a diffuse field assumption, which fulfilled better in more reverberant rooms.
- In both experiments Tapering+SHF was sometimes rated closer to the reference when rendered at order 5, instead of order 7. This might be caused by the cosine-shaped windowing of the Tapering algorithm, which modifies higher rendering orders more than lower ones.

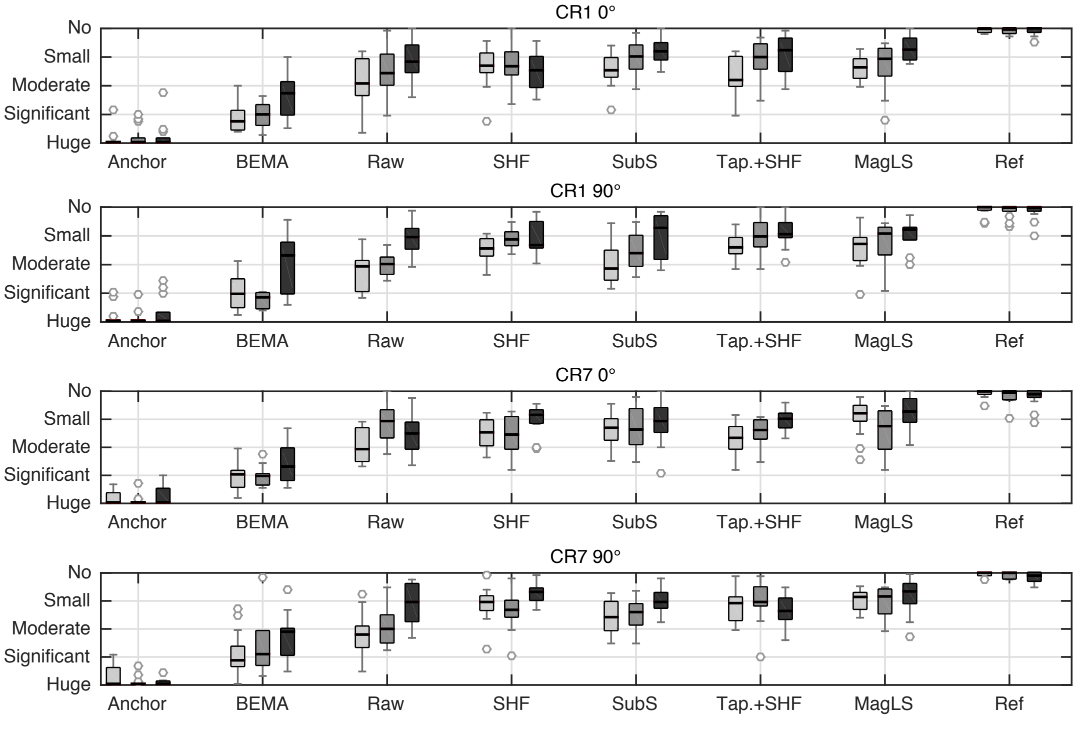


Figure 8: Interindividual variation in the ratings of perceptual difference between the stimulus and the dummy head reference with respect to the algorithm (x-axis), and SH order (color) for each room and virtual source position separately. Each box indicates the 25th and 75th percentiles, the median value (black line), the outliers (grey circles) and the minimum / maximum ratings not identified as outliers (black whiskers).

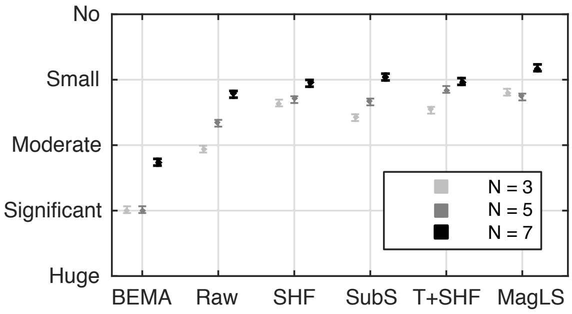


Figure 9: Mean values of the ratings pooled over both rooms with respect to the algorithm. The 95 % within-subject confidence intervals were calculated according to [23, 24]. The ratings for different SH orders are displayed separately as indicated by the color.

8. ACKNOWLEDGMENTS

We thank Christian Schörkhuber, Markus Zaunschirm, and Franz Zotter of IEM at the University of Music and Performing Arts in Graz for providing us with their code of MagLS, and all participants of the listening experiment for their support.

9. REFERENCES

[1] Boaz Rafaely and Amir Avni, “Interaural cross correlation in a sound field represented by spherical harmonics,” *The Journal of the Acoustical Society of America*, vol. 127, no. 2, pp. 823–828, 2010.

[2] Benjamin Bernschütz, *Microphone Arrays and Sound Field Decomposition for Dynamic Binaural Recording*, Ph.D. thesis, Technische Universität Berlin, 2016.

[3] Hannes Helmholz, Jens Ahrens, David Lou Alon, Sebastià V. Amengual Garí, and Ravish Mehra, “Evaluation of Sensor Self-Noise In Binaural Rendering of Spherical Microphone Array Signals,” in *Proc. of the IEEE Intern. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, May 2020, pp. 161–165, IEEE.

[4] Hannes Helmholz, David Lou Alon, Sebastià V. Amengual Garí, and Jens Ahrens, “Instrumental Evaluation of Sensor Self-Noise in Binaural Rendering of Spherical Microphone Array Signals,” in *Forum Acusticum*, Lyon, France, 2020, pp. 1–8, EAA.

[5] Tim Lübeck, Hannes Helmholz, Johannes M. Arend, Christoph Pörschmann, and Jens Ahrens, “Perceptual Evaluation of Mitigation Approaches of Impairments due to Spatial Undersampling in Binaural Rendering of Spherical Microphone Array Data,” *Journal of the Audio Engineering Society*, pp. 1–12, 2020.

[6] Boaz Rafaely, *Springer Topics in Signal Processing Springer Topics in Signal Processing*, Springer, 2015.

[7] Carl Andersson, "Headphone Auralization of Acoustic Spaces Recorded with Spherical Microphone Arrays," M.S. thesis, Chalmers University of Technology, 2017.

[8] Zamir Ben-Hur, Fabian Brinkmann, Jonathan Sheaffer, Stefan Weinzierl, and Boaz Rafaely, "Spectral equalization in binaural signals represented by order-truncated spherical harmonics," *The Journal of the Acoustical Society of America*, vol. 141, no. 6, pp. 4087–4096, 2017.

[9] Markus Zaunschirm, Christian Schörkhuber, and Robert Höldrich, "Binaural rendering of Ambisonic signals by head-related impulse response time alignment and a diffuseness constraint," *The Journal of the Acoustical Society of America*, vol. 143, no. 6, pp. 3616–3627, 2018.

[10] Jens Ahrens and Carl Andersson, "Perceptual evaluation of headphone auralization of rooms captured with spherical microphone arrays with respect to spaciousness and timbre," *The Journal of the Acoustical Society of America*, vol. 145, no. April, pp. 2783–2794, 2019.

[11] Christoph Hold, Hannes Gamper, Ville Pulkki, Nikunj Raghuvanshi, and Ivan J. Tashev, "Improving Binaural Ambisonics Decoding by Spherical Harmonics Domain Tapering and Coloration Compensation," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, 2019, pp. 261–265.

[12] Fabian Brinkmann and Stefan Weinzierl, "Comparison of head-related transfer functions pre-processing techniques for spherical harmonics decomposition," in *Proceedings of the AES Conference on Audio for Virtual and Augmented Reality*, Redmond, USA, 2018, pp. 1–10.

[13] Benjamin Bernschütz, "A Spherical Far Field HRIR/HRTF Compilation of the Neumann KU 100," in *Proceedings of the 39th DAGA*, Meran, Italy, 2013, pp. 592–595.

[14] Christian Schörkhuber, Markus Zaunschirm, and Robert Holdrich, "Binaural rendering of Ambisonic signals via magnitude least squares," in *Proceedings of 44th DAGA*, Munich, Germany, 2018, pp. 339–342.

[15] Lord Rayleigh, "XII. On our perception of sound direction," *Philosophical Magazine Series 6*, vol. 13, no. 74, pp. 214–232, 1907.

[16] Benjamin Bernschütz, "Bandwidth Extension for Microphone Arrays," in *Proceedings of the 133th AES Convention*, San Francisco, USA, 2012, pp. 1–10.

[17] Thomas McKenzie, Damian T. Murphy, and Gavin Kearney, "Diffuse-Field Equalisation of binaural ambisonic rendering," *Applied Sciences*, vol. 8, no. 10, 2018.

[18] Christoph Hohnerlein and Jens Ahrens, "Spherical Microphone Array Processing in Python with the sound field analysis-py Toolbox," in *Proceedings of the 43rd DAGA*, Kiel, Germany, 2017, pp. 1033–1036.

[19] Philipp Stade, Benjamin Bernschütz, and Maximilian Rühl, "A Spatial Audio Impulse Response Compilation Captured at the WDR Broadcast Studios," in *Proceedings of the 27th Tonmeistertagung - VDT International Convention*, Cologne, Germany, 2012, pp. 551–567.

[20] Benjamin Bernschütz, Christoph Pörschmann, Sascha Spors, and Stefan Weinzierl, "Entwurf und Aufbau eines variablen sphärischen Mikrofonarrays für Forschungsanwendungen in Raumakustik und Virtual Audio," in *Proceedings of 36th DAGA*, Berlin, Germany, 2010, pp. 717–718.

[21] Matthias Geier, Jens Ahrens, and Sascha Spors, "The soundscape renderer: A unified spatial audio reproduction framework for arbitrary rendering methods," in *Proceedings of the 124th AES Convention*, Amsterdam, Netherlands, 2008, pp. 179–184, Code publicly available at "http://spatialaudio.net/ssr/".

[22] ITU-R BS.1534-3, "Method for the subjective assessment of intermediate quality level of audio systems," 2015.

[23] Geoffrey R Loftus, "Using confidence intervals in within-subject designs," *Psychonomic Bulletin & Review*, vol. 1, no. 4, pp. 1–15, 1994.

[24] Jerzy Jarmasz and Justin G. Hollands, "Confidence Intervals in Repeated-Measures Designs: The Number of Observations Principle," *Canadian Journal of Experimental Psychology*, vol. 63, no. 2, pp. 124–138, 2009.

[25] Jürgen Bortz and Christof Schuster, *Statistik für Human- und Sozialwissenschaftler*, Springer-Verlag, Gießen, Germany, 7 edition, 2010.