Comparison of Mitigation Approaches of Spatial Undersampling Artifacts in Spherical Microphone Array Data Auralizations

Tim Lübeck[1,2], Johannes M. Arend[1,2], Hannes Helmholz[3], Jens Ahrens[3], Christoph Pörschmann[1]
[1] *TH Köln - University of Applied Sciences, Cologne, Germany*
[2] *Technical University of Berlin, Berlin, Germany*
[3] *Chalmers University of Technology, Gothenburg, Sweden*
Email: tim.luebeck@th-koeln.de

Introduction

In the last years, the auralization of spatial sound scenes based on spherical microphone array (SMA) captures has become very popular. SMA data can be reproduced in arbitrary virtual acoustic environment (VAE) formats, for example loudspeaker-based wave field synthesis [1], Ambisonics reproductions [17], or headphone-based binaural synthesis, which is the focus of this work. Employing SMAs for sound field analysis allows capturing the surrounding sound field at once including all dynamic spatial alternations. Instead of e.g. generating auditory scenes based on dummy head (DH) impulse response measurements, SMA captures are advantageous in terms of variety of VAE reproduction methods and in particular for realizing dynamic signal stream based applications.
The fundamental theory of SMA sound field capture and subsequent reproduction as binaural VAE has been discussed extensively, see e.g. [2, 6]. Furthermore, several real-time implementations have been introduced recently [8, 11, 17].
Capturing a sound field using real-world SMAs with a limited number of microphones leads to spatial undersampling. As a result, the undersampled sound field and the associated processing with a limited number of recorded channels introduce spatial aliasing and spherical harmonics (SH) order truncation. Both yield audible artifacts in the binaural reproduction.
To mitigate those artifacts, a number of approaches have been presented in the literature. Most of them have been evaluated independently. This contribution presents an overview of a selection of approaches, studies their influence on binaural synthesis, and compares them based on an instrumental evaluation.

Spatial Undersampling

To outline the phenomenon of spatial undersampling, we briefly summarize the fundamental concept of binaural reproduction of SMA captures. For a more detailed explanation please refer to [6, 12]. Binaural reproduction means virtually exposing the listeners' head to the sound field that is captured by the SMA. Therefore, the sound pressure S captured by the microphones on the array surface Ω is represented in the SH domain using the spherical Fourier transform (SFT)

$$S_{nm}(r,\omega) = \int_{\Omega} S(r,\phi,\theta,\omega)\, Y_n^m(\theta,\phi)^*\, dA_\Omega\,. \qquad (1)$$

Thereby, r denotes the array radius, ϕ the azimuth angle ranging from 0 to 2π, θ the colatitude ranging from 0 to π, and the angular frequency $\omega = 2\pi f$, with the temporal frequency f. $Y_n^m(\theta,\phi)$ denotes the orthogonal SH basis functions for certain order n and modes m and $(\cdot)^*$ the complex conjugate.
The surrounding sound field can then be decomposed into a continuum of plane waves impinging from all possible directions

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

with a set of radial filters d_n. Since a head-related transfer function (HRTF) $H^{l,r}$ describes the sound incident of a broadband plane wave to the listeners' ears in anechoic conditions, weighting the plane wave components D with the corresponding HRTF from that same direction, and integrating all possible incident directions yields the binaural signals, see e.g. [6, Eq. 2.147, p.64].
So far, a continuously sampled sound pressure distribution was assumed. However, real-world SMAs employ just a limited number of microphones. As a result, spatial aliasing and SH order truncation occur that significantly affect the perceptual quality of binaural reproduction. These impairments due to spatial undersampling are briefly discussed in the following.

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 artifacts. Higher spatial modes cannot be reliably resolved anymore and appear in lower modes. Generally, higher modes are required for resolving high frequency components with smaller wavelengths, thus the spatial aliasing limits the upper bound of the stably obtainable time-frequency bandwidth. Above the temporal frequency

$$f_\mathrm{A} = \frac{N_\mathrm{sg}\, c}{2\pi r}\,, \qquad (3)$$

the spatial aliasing artifacts increase significantly [12]. Thereby, c denotes the speed of sound and N_sg the maximum resolvable SH order of the sampling scheme. The appearance of higher modal components in the lower modal components results in increased magnitudes at higher temporal-frequencies. Hence, considering the

time-frequency spectrum of the binaural signals, spatial aliasing leads to a boosting high-shelf characteristic.

Spherical Harmonic Truncation

Besides spatial aliasing, the truncation of the natural SH order series leads to audible artifacts. The limited number of microphones leads to the discretization of the integral in Eq. 1, resulting in the discrete SFT. Stable SH coefficients can only be obtained up to a certain SH order. Since higher SH modes directly correspond to higher spatial resolution, order truncation results in a loss of spatial detail. Usually, HRTF sets are measured on dense grids and therefore allow for SH processing on sufficiently high SH orders. However, when merging high order HRTF SH coefficients with order-limited sound field coefficients, the higher orders of the HRTFs are not triggered.
Besides the loss of high spatial information encoded in the higher modes of the HRFTs, considering the spectrum of binaural time-frequency responses, order-truncation decreases the magnitudes of higher frequencies and thus has an attenuating low-pass effect. In addition, hard truncation of the SH coefficients at a certain order results in side-lobes in the plane wave spectrum [9], which can further impair the binaural signals.

Mitigation Approaches

In the last years, a number of different approaches to improve binaural rendering of SMA captures has been presented in the literature. In the following, a selection of approaches is summarized.

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 [7]. In this paper, we investigate two concepts in more detail.

Spatial Subsampling

For the spatial subsampling method [6], the HRTFs are transformed into the SH domain on the highest stable SH order. Subsequently, employing the inverse SFT up to the same SH order, but for a reduced number of directions, allows resampling the HRTF set. These so-called HRTFs with reduced modal order (RHRTFs) are resampled to the same grid the sound field has been sampled on. Using the RHRTFs for binaural rendering of SMA data avoids SH order truncation, adversely, for the cost of significantly higher spatial aliasing artifacts. Fig. 1 depicts the energy distribution of *KU100* HRTFs [5] 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 subsampled to a 5th order Lebedev grid. Evidently, the information can be reliably obtained only up to the 5th order.

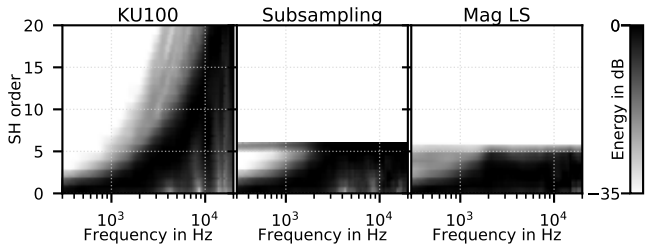

Figure 1: Energy distribution in dB with respect to order and frequency of *Neumann KU100* HRTFs; left: untreated, center: subsampled; right: MagLS pre-processed.

Magnitude Least-Squares

Another HRTF pre-processing approach is the Magnitude Least-Squares (MagLS) [14] algorithm, as an improvement of the Time Alignment (TA) proposed by the same authors. Both approaches are based on the duplex theory [13]. 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 order 5. The major part of the energy is shifted to SH coefficients of orders below 5. The major difference of both 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.

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) [4, 6] synthesizes the SH coefficients of the unstable higher frequency regions by extracting spatial and spectral information from components below f_A (Eq. (3)). To estimate the spatial information of the higher frequencies, the average spatial energy distribution of the lower components denoted as spatio-temporal image I_{nm} is calculated. The time-frequency spectral information is obtained by an additional omnidirectional microphone in the center of the microphone array. 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 impact 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 influence of BEMA. For

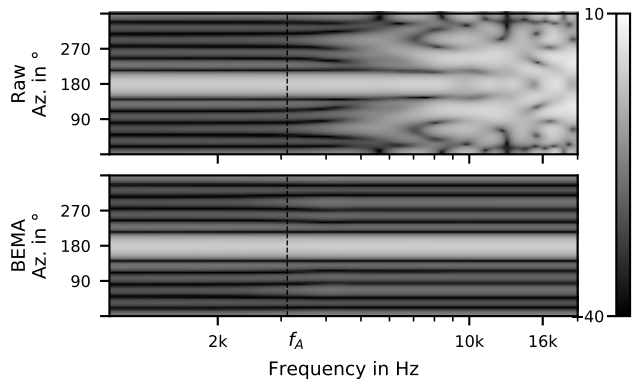

Figure 2: Plane wave magnitudes of a plane wave impact from $\phi = 180°$, $\theta = 90°$ on a 50 sampling point Lebedev grid SMA with a radius of 8.75 cm; The top diagram depicts the untreated magnitudes, the bottom diagram plane waves calculated after BEMA processing.

a single plane wave, the sound field is reconstructed perfectly over the entire audible bandwidth by projecting the spatial information below f_A to higher frequency regions.

Spherical Harmonic Tapering

SH order truncation induces side-lobes in the plane wave spectrum. In time-frequency processing, side-lobes in the frequency domain are suppressed by using smoother roll-offs in the time domain instead of rectangular windowing. Similarly, applying smooth window functions to the SH coefficients instead of harsh truncation reduces the magnitudes of the plane wave side-lobes [9]. In fact, these window functions constitute an order-dependent scaling factor equally applied to all SH coefficients for the same order. Different windows have been discussed and half-sided Hanning windows were found to be the optimal choice. Additionally, the authors equalized the binaural signals with the so-called Spherical Head Filters which are discussed in the subsequent section.

Spectral Equalization

The modification of the time-frequency response due to spatial undersampling is a perceptual distinctive impairment, as shown e.g. in [2]. Therefore, a third category of mitigation approaches is global equalization of the binaural signals. Two approaches have been introduced in the literature to design such equalization filters. The Spherical Head Filters (SHFs) [3] compensate for the low-pass behavior of SH order truncation. The authors neglect spatial aliasing effects and deployed filters based on the plane wave density function of a diffuse sound field.
The second equalization approach compensates for the spatial aliasing high-shelf boost in the binaural time-frequency spectrum [6, pp. 83]. These filters were designed with negligible truncation artifacts by using subsampled HRTFs (RHRTFs). The average deviation of binaural signals directly measured with a DH and binaural signals based on SMA renderings in a diffuse sound fields follows a +6 dB per octave slope starting above f_A. Both equalization filters, depicted in Fig. 3, are directly applied to the binaural signals, for every direction equally.
Since informal listening showed that the truncation

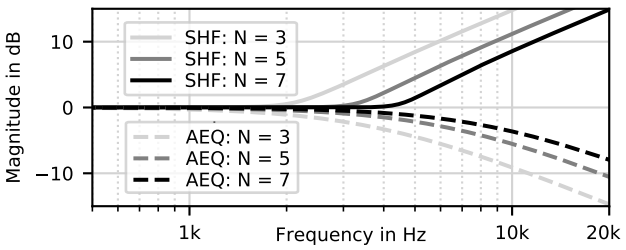

Figure 3: Spherical Head Filters (SHFs) and spatial aliasing compensation filters (AEQs) for orders $N = (3, 5, 7)$

low-pass is more prominent than the high-shelf due to spatial aliasing, applying solely the low-pass filters has no perceptual benefits. We thus exclusively considered the SHFs for instrumental evaluation.

Evaluation and Discussion

This section focuses on the influence of the mitigation approaches on binaural renderings. We used an impulse response database containing array impulse responses measured on different Lebedev grids, as well as binaural room impulse responses (BRIRs) measured with a DH under exact the same conditions in the *WDR Broadcast Studios* [15]. This allows for a direct comparison of binaural auralization of SMA and DH data, which are the ground truth. In the following, we compare 3rd SH order array renderings based on a 50 sampling point Lebedev grid measurement of *Control Room 7*, that has an RT_{60} of about 0.9 s (0.5 kHz to 1 kHz). We calculated the BRIRs for 360 azimuth directions in the horizontal plane in 1°-steps to be compared against the DH BRIRs measured on the same grid. As a measure for the effectiveness of the mitigation approaches, we calculated the absolute spectral differences between DH and array BRIRs in dB as illustrated in Fig. 4. The top diagram depicts the deviations averaged over all 360 directions with respect to frequency (x-axis). To demonstrate that the differences are very high in magnitude in particular at the contralateral side, the bottom diagram shows the differences averaged over 40 directions around the contralateral side.
The untreated (Raw) rendering is clearly affected by undersampling artifacts above f_A as indicated by the vertical dashed line. Around the contralateral side, these differences increase rapidly. Comparing the HRTF preprocessing techniques reveals that both algorithms significantly decrease the difference to the reference. However, MagLS leads to better results, especially for frequencies between 2 kHz and 6 kHz. At the contralateral side, the Subsampling performs slightly better around 9 kHz.
BEMA yields more impairments than improvements. As already found by the authors of BEMA, it is able to perfectly reconstruct the sound field for a single plane wave. However, even for 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 the Tapering approach perform rather similar. Both methods employ

global filtering to the binaural signals. The differences at the contralateral side are still higher than for frontal directions. Overall, the instrumental evaluation reveals that MagLS performs best, at least for the cases tested.

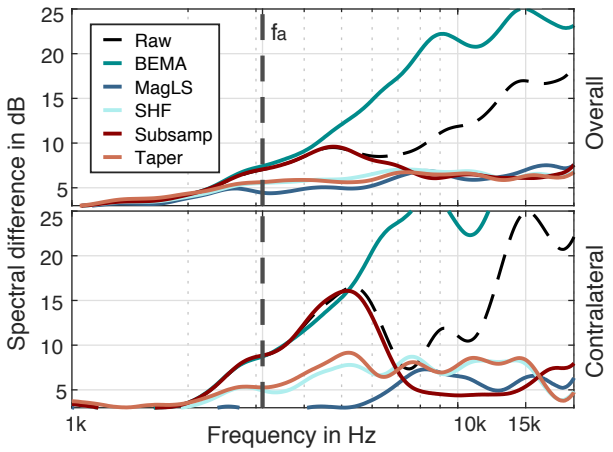

Figure 4: Absolute spectral differences of DH and SMA binaural signals in dB; top: averaged over 360 horizontal directions; bottom: averaged over 40 directions around the contralateral side.

Some of the approaches considered here have already been perceptually evaluated in listening experiments. Subsampling showed to significantly improve the perceptual quality [6], although it provokes higher spatial aliasing. The Time Alignment, Subsampling and SHFs were compared in [16]. The results showed that mostly Time Alignment, which is an early stage of MagLS, yields better results than Subsampling. It can thus be assumed that MagLS perceptually outperforms the Subsampling procedure, which corresponds to the larger deviations depicted in Fig. 4. The SHFs were rated worst of the three tested methods, matching the instrumental results in Fig. 4. 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 [6]. However, for measured diffuse sound fields, BEMA introduces significant artifacts and thus is no promising algorithm for real-world applications. To our knowledge, no listening experiment evaluated the Tapering approach so far. For a broad perceptual comparison of the approaches presented in this paper, we recently conducted a listening experiment. The study is already submitted for publication [10].

Conclusion

The instrumental evaluation revealed that MagLS, Tapering, and the global SHFs significantly improve undersampled SMA auralizations. Global equalization as the SHFs are applied equally for every direction. This however has the disadvantage that errors are shifted to lateral directions. Although Tapering and MagLS try to tackle this directional dependency, small lateral artifacts still persist. The instrumental analysis did not indicate that Tapering improves the directional dependency, however informal listening revealed slightly more stable auralizations with the tapered SH coefficients. BEMA might be an adequate approach to process direct sound components in impulse response based auralizations, but is not applicable in diffuse environments.

References

[1] Jens Ahrens. *Analytic Methods of Sound Field Synthesis*. Springer, 2012.

[2] Jens Ahrens and Carl Andersson. Perceptual evaluation of headphone auralization of rooms captured with spherical microphone arrays with respect to spaciousness and timbre. *Journal of the Acoustical Society of America*, 145(April):2783–2794, 2019.

[3] 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*, 141(6):4087–4096, 2017.

[4] Benjamin Bernschütz. Bandwidth Extension for Microphone Arrays. *AES 133th Convention*, pages 1–10, 2012.

[5] Benjamin Bernschütz. A Spherical Far Field HRIR/HRTF Compilation of the Neumann KU 100. In *39th DAGA*, pages 592—-595, 2013.

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

[7] Fabian Brinkmann and Stefan Weinzierl. Comparison of head-related transfer functions pre-processing techniques for spherical harmonics decomposition. In *AES Conference on Audio for Virtual and Augmented Reality*, pages 1–10, Redmond, WA, USA, 2018.

[8] Hannes Helmholz, Carl Andersson, and Jens Ahrens. Real-Time Implementation of Binaural Rendering of High-Order Spherical Microphone Array Signals. In *45th DAGA*, pages 2–5, 2019.

[9] 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 *International Conference on Acoustics, Speech and Signal Processing*, volume 2, pages 261–265, 2019.

[10] 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. *submitted for publication*, 2020.

[11] Leo McCormack and Archontis Politis. SPARTA & COMPASS: Real-time implementations of linear and parametric spatial audio reproduction and processing methods. In *AES Conference on Immersive and Interaktive Audio*, York, 2019.

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

[13] Lord Rayleigh. XII. On our perception of sound direction. *Philosophical Magazine Series 6*, 13(74):214–232, 1907.

[14] Christian Schörkhuber, Markus Zaunschirm, and Robert Holdrich. Binaural rendering of Ambisonic signals via magnitude least squares. In *44th DAGA*, pages 339–342, 2018.

[15] Philipp Stade, Benjamin Bernschütz, and Maximilian Rühl. A Spatial Audio Impulse Response Compilation Captured at the WDR Broadcast Studios. In *27th Tonmeistertagung - VDT International Convention*, pages 551—-567, 2012.

[16] 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*, 143(6):3616–3627, 2018.

[17] Franz Zotter and Matthias Frank. *Ambisonics A Practical 3D Audio Theory for Recording, Studio Production, Sound Reinforcement, and Virtual Reality*. Springer-Verlag, 2019.