End-to-End Magnitude Least Squares Binaural Rendering of Spherical Microphone Array Signals

End-to-End Magnitude Least Squares Binaural Rendering of Spherical Microphone Array Signals

Thomas Deppisch, Hannes Helmholz, and Jens Ahrens
Division of Applied Acoustics
Chalmers University of Technology
412 96 Gothenburg, Sweden
{thomas.deppisch, hannes.helmholz, jens.ahrens}@chalmers.se

***Abstract*—Spherical microphone array (SMA) recordings are particularly suited for dynamic binaural rendering as the microphone signals can be decomposed into a spherical harmonic (SH) representation that can be freely rotated to match the head orientation of the listener. The rendering of such SMA recordings is a non-trivial task as the SH signals are impaired due to truncation of the SH decomposition order, spatial aliasing and the gain limitation of the employed radial filters. The perceptually most relevant consequence of this is an alteration of the magnitude transfer function at high frequencies. Previously, the magnitude least squares (MagLS) renderer for binaural rendering of SH signals was proposed to mitigate these effects under the assumption of ideal order-truncated plane waves, i.e., disregarding the influence of spatial aliasing as well as of non-ideal radial filters. Based on the MagLS renderer, we present a binaural rendering method for SMA recordings that integrates a comprehensive SMA model into the magnitude least squares objective. We evaluate the proposed end-to-end renderer by analyzing the reproduced binaural magnitude response. Our results suggest that the method significantly improves the high-frequency rendering mainly due to the inherent binaural diffuse-field equalization, while it achieves a slight improvement in the low and mid frequency range, where the error of the conventional method is already small. A reference implementation of the method accompanies this paper.**

***Index Terms*—Spherical Microphone Arrays, Ambisonics, Spherical Harmonics, Binaural Rendering**

I. Introduction

Ambisonics [1]–[3] provides a flexible framework for spatial audio capture, processing and rendering by utilizing a decomposition of the signals into spherical harmonics (SHs). Although Ambisonics neither dictates a specific capturing nor rendering technology, the sound capture with a spherical microphone array (SMA) and subsequent binaural rendering to headphones constitute the most common processing pipeline.

Sound capture with SMAs and binaural rendering using head-related transfer functions (HRTFs) are both non-trivial tasks. SMAs distort the captured sound field in several ways: i) They theoretically require a continuous distribution of microphones and the restriction to a finite set of discrete microphones in practice causes SH order truncation and spatial aliasing [4]. ii) Their limited size limits the ability to capture spatial information at low frequencies where the wavelength is much longer than the dimensions of the array. This manifests as the requirement of regularizing the incorporated so-called radial filters [5]. iii) To avoid ambiguities, the microphones have to be mounted on a rigid scattering object that further distorts the captured sound field [6].

The regularization of the radial filters, i.e., the limitation of the amplification gains, has to be performed at low frequencies only. Its impact on the binaural output signals is small as the SH coefficients of an HRTF set do not contain large amounts of energy in high SH orders at low frequencies [7]. The effects of spatial aliasing and order truncation become apparent above the spatial aliasing frequency as spatial ambiguities and a limitation of the spatial resolution, respectively. The spatial aliasing frequency of typical arrays is in the order of a few thousand Hz.

Binaural rendering typically involves the SH decomposition of an HRTF set. The circumstance that the ear position does not coincide with the center of the SH expansion, which is usually performed about the mid-point of the interaural axis, causes the higher SH orders of the HRTFs to contain significant energy at high frequencies [7]. This energy is missing in the output of renderers that use a reduced SH order, resulting in an attenuation of the high frequencies. Additionally, the rendering of high frequencies is impaired by a secondary effect of spatial aliasing: the energy at frequencies above the aliasing frequency, where the SH order truncation of the HRTF set causes an attenuation, is increased. However, these two effects on the magnitude transfer function do not cancel out and the impact of the spatial aliasing often dominates.

Following the idea of their independence, the challenges of SMA sound capture and binaural rendering have previously been tackled separately. The design of SMA radial filters [5], [8], [9], [3, Sec. 6.8] aims at recovering the directionality of a captured sound field and at removing the influence of the scattering off the rigid spherical microphone body. The SH order truncation can be equalized on average using a spherical head filter [10], while diffuse-field equalization [9, Sec. 3.8.2], [11] may be applied to compensate for the average influence of spatial aliasing. It was shown in [12] that the equalization of the magnitude transfer function to compensate for the influence of aliasing and truncation provides a significant perceptual improvement, and the resulting output is close to the ground truth ear signals if the rendering is performed at an SH order of 7 or higher.

We thank Facebook Reality Labs Research for funding this project.

Almost all methods proposed in the literature perform a direction-independent global equalization of the magnitude transfer function of the entire rendering pipeline. The magnitude least squares (MagLS) binaural rendering technique [13] is fundamentally different as it achieves an equalization that is dependent on the incidence direction of the sound. It follows the idea from [7], where it was shown that large amounts of energy in high orders of the SH representation of the HRTF set can be shifted to lower SH orders by time-alignment of the corresponding HRIRs. Based on Rayleigh's duplex theory, both approaches neglect the interaural time differences of the HRIR set at high frequencies, and MagLS finds a suitable phase response to optimally recreate the directional magnitude response of the HRTF set in a least-squares sense. Although the objective of the MagLS renderer does not minimize a perceptually motivated error function, user studies have shown its perceptual benefits [12], [13].

The original MagLS formulation [13] performs the equalization of the HRTFs in isolation. In other words, it produces an order-limited HRTF representation whose magnitude exhibits minimal deviations from the original non-truncated HRTF representation without considering signal impairments due to the employed sound capture and processing methods. Based on the MagLS renderer, we propose a novel binaural rendering algorithm with the aim of jointly solving the challenges of both, SMA sound capture and binaural rendering, from the capture-end to the rendering-end. This is achieved by integrating a comprehensive SMA model into the objective function of the MagLS renderer.

II. Signal Model

The sound pressure at radius r from the origin due to a plane wave arriving from the spherical direction $\Omega = (\varphi, \vartheta)$, composed of azimuth angle φ and zenith angle ϑ, can be described in terms of SH pressure coefficients of order n and degree m [14, Eq. (2.3.6)],

$$p_n^m(kr, \Omega) = b_n(kr)\ Y_n^m(\Omega)\,, \tag{1}$$

where $Y_n^m(\Omega)$ are the SHs, and $k = \omega/c$ is the wave number depending on the angular frequency ω and the speed of sound c. The factor $b_n(kr)$ accounts for the radial dependency of the pressure and modifies the SH mode strengths accordingly. In a free-field, it is given by $b_n(kr) = 4\pi \mathrm{i}^n\, j_n(kr)$, with the imaginary unit $\mathrm{i} = \sqrt{-1}$ and the spherical Bessel function $j_n(kr)$. In the presence of a rigid sphere with radius $r_0 \leq r$ that is centered at the coordinate origin, $b_n(kr)$ fully represents the effect of the scattering of the incident sound field off the sphere and is given by [14, Eq. (4.2.10)]

$$b_n(kr) = 4\pi \mathrm{i}^n \left(j_n(kr) - \frac{j_n'(kr_0)}{h_n'(kr_0)}\ h_n(kr) \right)\,, \tag{2}$$

where $(\cdot)'$ denotes the derivative and $h_n(kr)$ is the spherical Hankel function. Without limiting the generality of the results, we assume such a rigid-sphere array in the remainder as this is the most common configuration. Expressions for $b_n(kr)$ for further configurations are provided in [15].

From the SH pressure coefficients (1), the corresponding space-domain sound pressure $p(kr, \Omega)$ at a location defined by the coordinates r and Ω_0 is obtained after SH expansion by [14, Eq. (2.1.65)]

$$p(kr, \Omega) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} Y_n^m(\Omega_0)^*\ p_n^m(kr, \Omega)\ . \tag{3}$$

The $(\cdot)^*$ operator denotes complex conjugation. An SMA samples the pressure given by (3) at several locations on the surface of the rigid sphere where $r = r_0$, such that $b_n(kr)$ loses the radial dependency and is denoted as $b_n(\omega)$ in the following. We express the vector of all pressure samples collected by the SMA in matrix-vector notation by limiting the maximum SH order to a finite $\tilde{N}$ as

$$\boldsymbol{p}_{\mathrm{M}}(\omega, \Omega) = \boldsymbol{Y}_{\tilde{N},\mathrm{M}}^{\mathrm{H}}\ \boldsymbol{B}_{\tilde{N}}(\omega)\ \boldsymbol{y}_{\tilde{N}}(\Omega)\ , \tag{4}$$

where the vector $\boldsymbol{y}_{\tilde{N}}(\Omega)$ contains the $(\tilde{N}+1)^2$ SH basis functions up to order $\tilde{N}$ evaluated at Ω, the matrix $\boldsymbol{Y}_{\tilde{N},\mathrm{M}} = [\boldsymbol{y}_{\tilde{N}}(\Omega_1), \ldots, \boldsymbol{y}_{\tilde{N}}(\Omega_M)]$ contains such SH vectors for all M microphone locations, and $\boldsymbol{B}_{\tilde{N}}(\omega)$ is a $(\tilde{N}+1)^2$ diagonal matrix with $b_n(\omega)$ replicated for all orders $n \in [0, \tilde{N}]$ and degrees $m \in [-n, n]$ on its diagonal. The superscript $(\cdot)^{\mathrm{H}}$ denotes the Hermitian (conjugate) transpose. Note that at this point, the limitation of the SH order to $\tilde{N}$ creates an error when compared to the infinite-order expansion of (3) that can be kept arbitrary small over the frequency range of interest by choosing a correspondingly high evaluation order $\tilde{N}$ [14, Ch. 9].

The processing of SMA pressure signals is typically performed in the SH domain by employing the discrete spherical harmonics transform (DSHT) matrix $\boldsymbol{E}_N$ [16, Sec. 3.6],

$$\boldsymbol{p}_N(\omega, \Omega) = \boldsymbol{E}_{N,\mathrm{M}}\ \boldsymbol{Y}_{\tilde{N},\mathrm{M}}^{\mathrm{H}}\ \boldsymbol{B}_{\tilde{N}}(\omega)\ \boldsymbol{y}_{\tilde{N}}(\Omega)\ . \tag{5}$$

The DSHT matrix $\boldsymbol{E}_{N,\mathrm{M}}$ is also referred to as microphone encoder in Ambisonics literature and is either obtained as the pseudoinverse $\boldsymbol{E}_{N,\mathrm{M}} = (\boldsymbol{Y}_{\tilde{N},\mathrm{M}}^{\mathrm{H}})^{\dagger}$ or by employing quadrature weights $\boldsymbol{\alpha}$ of the respective spherical sampling grid, $\boldsymbol{E}_N = \boldsymbol{Y}_{\tilde{N},\mathrm{M}}\ \mathrm{diag}(\boldsymbol{\alpha})$. In an ideal situation, i.e., in the absence of spatial aliasing and when $\tilde{N} \to \infty$, $\boldsymbol{p}_N(\omega, \Omega)$ is a vector of the coefficients $p_n^m(kr_0, \Omega)$ of the captured plane wave that are defined by (1). With a discrete microphone array, $\boldsymbol{p}_N(\omega, \Omega)$ contains a subset of the coefficients of the original plane wave for $n \leq \tilde{N}$ that may deviate to some extent from the true coefficients.

III. Conventional Binaural Rendering

The computation of the binaural signals from microphone signals, or a SH decomposition thereof, is termed rendering or also decoding in the Ambisonics context. In a nutshell, the conventional rendering requires multiplying $p_n^m(kr_0, \Omega)$ with the inverse of $b_n(kr_0)$, which are referred to as radial filters in the SMA literature and in practice require regularization. The result is multiplied with the corresponding SH coefficients $H_n^m(\omega, \Omega)$ of the employed HRTF set. We refer the reader to [7] or [17] for further details.

Due to the circumstance that the ear positions do not coincide with the center of the SH expansion, the SH coefficients $H_n^m(\omega, \Omega)$ exhibit significant energy at high SH orders for high frequencies. Their truncation causes an error in the reconstructed magnitude response of the binaural signals when the higher orders are not used in the rendering. By neglecting the time information in the HRTF set at high frequencies, the energy in high SH orders can be shifted towards lower orders to improve the spectral balance of the order-limited binaural rendering. Binaural rendering with a high-frequency time-aligned HRTF set was proposed in [7]. Further improvement of the reconstructed binaural magnitude response can be achieved using the magnitude least squares (MagLS) approach [13] that calculates an optimal phase for an improved reconstruction of the magnitude response at high frequencies.

Although these methods significantly improve the rendering at high frequencies, they neglect the impact of the discrete sampling of the sound pressure performed by the SMA which leads to spatial aliasing and SH order truncation of the captured signal. Further, they assume ideal radial filtering of the renderer input signals. In other words, the underlying signal model of the original MagLS approach from [13] is given by

$$\boldsymbol{p}_N(\Omega) = \boldsymbol{y}_N(\Omega) \ . \tag{6}$$

Our contribution lies in incorporating the extended signal model into the MagLS objective function, i.e., we formulate the problem taking into account that the SH coefficients of the captured sound field deviate from the correct ones by some extent. This can be done in two different ways: i) by using the SH domain signal model from (5) or ii) by working directly with the microphone signals from (4) without SH decomposition. The expressions for the corresponding three renderers, conventional MagLS, the proposed end-to-end MagLS (eMagLS) based on the SH pressure coefficients (5), and the alternative end-to-end MagLS (eMagLS2) based on the microphone pressure signals (4), only differ in the employed pressure model. Hence, the derivations below are formulated using the generic pressure variable $\boldsymbol{p}(\omega, \Omega)$, that represents (4), (5), or (6), depending on the specific renderer.

IV. End-to-End Magnitude Least Squares Binaural Rendering

In this section, we integrate the signal model for the sound pressure $\boldsymbol{p}(\omega, \Omega)$ into the least squares and magnitude least squares objectives to find the corresponding optimal rendering filters $\boldsymbol{w}(\omega)$. Given that the entire processing pipeline from the microphones to the binaural output is composed of linear time-invariant (LTI) operations, it is possible to represent the pipeline – or arbitrary segments of it – by means of multiple-input-multiple-output (MIMO) LTI filters. In other words, M microphone signals are turned into two output signals. As the processing is independent for the left and right ear and differs only with respect to the employed HRTFs, we formulate the problem in the following as an M-to-1 multiple-input-single-output (MISO) problem for a selected ear.

The binaural signal for the left or right ear $s_e(\omega)$, with $e \in \{\mathrm{l}, \mathrm{r}\}$, is obtained by filtering the multichannel microphone signal $\boldsymbol{s}(\omega)$, due to an arbitrary sound field that was captured by the array, with the rendering filter for the ear $\boldsymbol{w}_e(\omega)$, that comprises the radial filtering and the HRTFs of that ear, and summing all filtered channels

$$s_e(\omega) = \boldsymbol{s}^{\mathrm{H}}(\omega)\ \boldsymbol{w}_e(\omega) \ . \tag{7}$$

In case of the MagLS and the eMagLS renderers, which are both based on an SH domain signal model, also the microphone signal $\boldsymbol{s}(\omega)$ is expressed in the SH domain, i.e., both $\boldsymbol{s}(\omega)$ and $\boldsymbol{w}_e(\omega)$ are vectors of length $(N{+}1)^2$, where N is the SH order. In case of the eMagLS2 renderer, which is based directly on the microphone signals (4), both $\boldsymbol{s}(\omega)$ and $\boldsymbol{w}_e(\omega)$ have length M, where M is the number of microphones.

For notational brevity, all derivations below are expressed for a single ear only, and the subscript $e \in \{\mathrm{l}, \mathrm{r}\}$ is omitted.

A. Least Squares Binaural Rendering

The MagLS renderer is typically derived as an extension of the least squares (LS) renderer [13]. The LS rendering filters minimize the squared difference between the HRTF $H(\omega, \Omega)$ and the filtered plane wave $\boldsymbol{p}(\omega, \Omega)$ for all incidence directions Ω on the unit sphere $\mathcal{S}^2$,

$$\boldsymbol{w}_{\mathrm{LS}}(\omega) = \arg\min_{\boldsymbol{w}} \int_{\Omega \in \mathcal{S}^2} \left| \boldsymbol{p}^{\mathrm{H}}(\omega, \Omega)\, \boldsymbol{w} - H(\omega, \Omega) \right|^2 \mathrm{d}\Omega \ . \tag{8}$$

For a discretely measured HRTF set, the objective function is discretized as

$$\boldsymbol{w}_{\mathrm{LS}}(\omega) = \arg\min_{\boldsymbol{w}} \left\| \boldsymbol{P}^{\mathrm{H}}(\omega)\, \boldsymbol{w} - \boldsymbol{h}(\omega) \right\|^2 \ , \tag{9}$$

where $\boldsymbol{h}(\omega)$ contains the HRTFs measured at K directions $\boldsymbol{h}(\omega) = [H(\omega, \Omega_1),\ \ldots,\ H(\omega, \Omega_K)]^{\mathrm{T}}$ and $\boldsymbol{P}(\omega) = [\boldsymbol{p}(\omega, \Omega_1),\ \ldots,\ \boldsymbol{p}(\omega, \Omega_K)]$ contains the sound pressure $\boldsymbol{p}(\omega, \Omega)$ described by one of the models from (4), (5) or (6), evaluated at the same directions. The least-squares solution is found analytically as the pseudoinverse of the pressure matrix $\boldsymbol{P}(\omega)$ times the HRTF vector,

$$\boldsymbol{w}_{\mathrm{LS}}(\omega) = \left(\boldsymbol{P}^{\mathrm{H}}(\omega)\right)^{\dagger}\ \boldsymbol{h}(\omega) \ . \tag{10}$$

The least-squares solution $\boldsymbol{w}_{\mathrm{LS}}(\omega)$ implicitly comprises the inverse SH transform and the HRTF filtering. If the conventional definition of $\boldsymbol{p}_N(\Omega)$ from (6) is employed, the same solution is obtained if the HRTF set is SH transformed by using a pseudoinverse as DSHT matrix [18]. If the definition of the sound pressure $\boldsymbol{p}(\omega, \Omega)$ from (4) or (5) is used, the pseudoinverse needs to be regularized as the model does not include radial filtering and the scattering term $b_n(\omega)$ attenuates high-order modes for low frequencies. We propose regularization by limitation of the singular values on the diagonal of $\boldsymbol{\Sigma}$, that are obtained via the singular value decomposition $\boldsymbol{P}^{\mathrm{H}} = \boldsymbol{U}\, \boldsymbol{\Sigma}\, \boldsymbol{V}^{\mathrm{H}}$, to a small percentage β of the maximum singular value $\sigma_{\max}$,

$$\mathrm{diag}(\tilde{\boldsymbol{\Sigma}}) = \max\left(\mathrm{diag}(\boldsymbol{\Sigma}),\ \beta\, \sigma_{\max}\right) \ . \tag{11}$$

The regularized inverse is then obtained as $\boldsymbol{V}\, \tilde{\boldsymbol{\Sigma}}^{-1}\, \boldsymbol{U}^{\mathrm{H}}$.

B. Magnitude Least Squares Binaural Rendering

For low frequencies, magnitude and phase of the HRTFs are reproduced with little error by the LS renderer because the underlying physical problem is solvable [7]. At high frequencies, the ambiguities and information loss introduced by spatial aliasing and SH order truncation prevent a precise solution. As shown in [7], it is not necessary to preserve the timing information at high frequencies and optimizing only the magnitude by fitting a suitable phase reduces the error of the solution compared to the LS renderer [13].

Hence, the MagLS renderer utilizes the LS approach for low frequencies but transitions to the magnitude least squares objective for frequencies higher than the transition frequency ω_t [13],

$$\boldsymbol{w}_\mathrm{MLS}(\omega) = \arg\min_{\boldsymbol{w}} \Big[\lambda(\omega) \left\| \boldsymbol{P}^\mathrm{H}(\omega)\, \boldsymbol{w} - \boldsymbol{h}(\omega) \right\|^2 + (1-\lambda(\omega)) \left\| \left| \boldsymbol{P}^\mathrm{H}(\omega)\, \boldsymbol{w} \right| - |\boldsymbol{h}(\omega)| \right\|^2 \Big] \;, \quad (12)$$

where the function

$$\lambda(\omega) = \begin{cases} 1 & \text{for} \quad \omega < \omega_t \\ 0 & \text{else} \;, \end{cases} \quad (13)$$

defines the transition between the least squares and the magnitude least squares objective. The corresponding optimal filters $\boldsymbol{w}_\mathrm{MLS}(\omega)$ represent the MagLS, eMagLS or eMagLS2 renderer, depending on the employed pressure model $\boldsymbol{p}(\omega, \Omega)$ from (6), (5), or (4), respectively.

The magnitude least squares optimization problem is generally difficult to solve; an approximate solution can be obtained by reformulation as a complex two-way partitioning problem and subsequent relaxation of the rank-1 constraint [19]. In practice, however, it was found that the solution $\boldsymbol{w}_\mathrm{MLS}(\omega_l)$ for the current frequency bin ω_l is well approximated by iteratively reconstructing the $K \times 1$ phase vector $\boldsymbol{\Phi}(\omega_l)$ for all K directions from the solution for the previous bin ω_{l-1} [3, Sec. 4.11.2],

$$\boldsymbol{\Phi}(\omega_l) = \angle \left(\boldsymbol{P}^\mathrm{H}(\omega_l)\; \boldsymbol{w}_\mathrm{MLS}(\omega_{l-1}) \right) \;, \quad (14)$$

$$\boldsymbol{w}_\mathrm{MLS}(\omega_l) = \left(\boldsymbol{P}^\mathrm{H}(\omega_l) \right)^\dagger |\boldsymbol{h}(\omega_l)|\; \mathrm{e}^{\mathrm{i}\boldsymbol{\Phi}(\omega_l)} \;, \quad (15)$$

where element-wise definitions of the angle function $\angle(\cdot)$ and the exponential $\mathrm{e}^{(\cdot)}$ are employed for vector-valued arguments. The pseudoinverse is again regularized as described in Sec. IV-A.

As noted in [13], a global time delay can be applied above the transition frequency to avoid a discontinuity in the group delay of the decoding filters at the transition frequency.

The diffuseness constraint from [7] matches the interaural coherence of the output of the time-alignment renderer to the coherence of the HRTF set and can also be integrated into the MagLS renderer [3, Sec. 4.11.3], as well as into the eMagLS and eMagLS2 renderers, by replacing the conventional model of the plane wave from (6) that was used ibidem with one of the extended models from (4) or (5).

V. Instrumental Evaluation

A. Evaluation Measures

We evaluate the proposed renderers using the summed magnitude response (SMR) error

$$\epsilon_\mathrm{SMR}(\omega, \Omega) = \mathrm{SMR}_\mathrm{MLS}(\omega, \Omega) - \mathrm{SMR}_\mathrm{ref}(\omega, \Omega) \;, \quad (16)$$

i.e., the difference between the SMR from a magnitude least squares renderer and the reference SMR obtained from a least-squares renderer with SH order $N = 32$. The SMR was also used in [13],

$$\mathrm{SMR}(\omega, \Omega) = 10 \log \Big(\left| \boldsymbol{p}^\mathrm{H}(\omega, \Omega)\, \boldsymbol{w}_\mathrm{l}(\omega) \right|^2 + \left| \boldsymbol{p}^\mathrm{H}(\omega, \Omega)\, \boldsymbol{w}_\mathrm{r}(\omega) \right|^2 \Big) \;, \quad (17)$$

and is inspired by the binaural composite loudness level [20], [21]. In the evaluation, the pressure signal $\boldsymbol{p}(\omega, \Omega)$ is modeled by the comprehensive SMA signal model. More precisely, the SH-domain model from (5) is applied to evaluate the MagLS and the eMagLS renderers, and the model without the SH decomposition from (4) is applied to evaluate the eMagLS2 renderer.

As a global metric, we define the average SMR error

$$\bar{\epsilon}_\mathrm{SMR} = \frac{1}{B} \sum_{k=1}^{K} \sum_{b=1}^{B} \alpha_k\, \mathcal{B}_b\{ |\epsilon_\mathrm{SMR}(\omega, \Omega_k)| \} \;, \quad (18)$$

by third-octave-band averaging $\mathcal{B}_b\{\cdot\}$ of the absolute SMR error $|\epsilon_\mathrm{SMR}(\omega, \Omega)|$ for each band with index b, and subsequently averaging over all B third-octave bands and K evaluation directions using the quadrature weights α_k.

The binaural diffuse-field magnitude response of an HRTF set is approximated by the average magnitude response

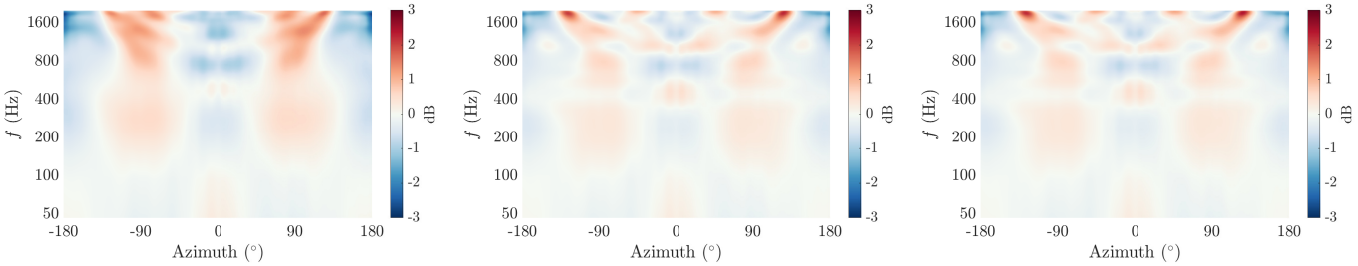
$$|H_\mathrm{df}(\omega)| = \frac{1}{2} \sum_{k=1}^{K} \alpha_k \left(|H_\mathrm{l}(\omega, \Omega_k)| + |H_\mathrm{r}(\omega, \Omega_k)| \right) \;, \quad (19)$$

i.e., by averaging the magnitude of the HRTF over K directions on the surface of the unit sphere $\mathcal{S}^2$ across both ears.

Similarly, in case of the MagLS renderers, we average the rendered binaural ear signals due to plane-waves impinging from K directions

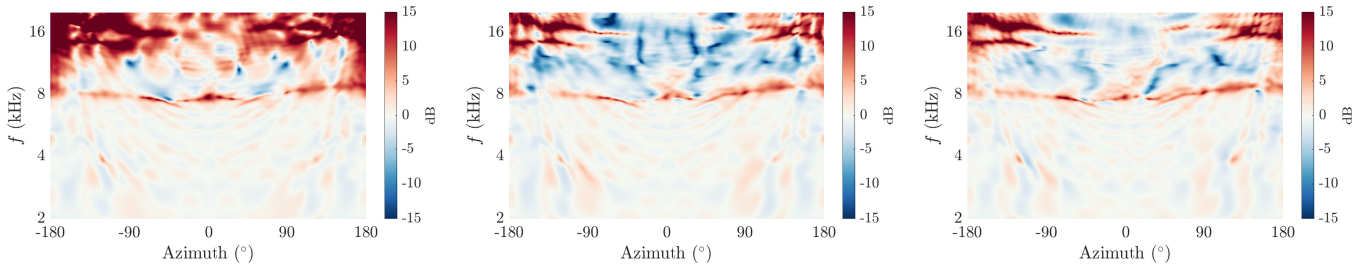
$$|H_\mathrm{df,MLS}(\omega)| = \frac{1}{2} \sum_{k=1}^{K} \alpha_k \left(|\boldsymbol{p}^\mathrm{H}(\omega, \Omega_k)\, \boldsymbol{w}_\mathrm{MLS,l}(\omega)| + |\boldsymbol{p}^\mathrm{H}(\omega, \Omega_k)\, \boldsymbol{w}_\mathrm{MLS,r}(\omega)| \right) \;. \quad (20)$$

All of the following simulations are based on the HRIR set of the *Neumann KU100* dummy head from [22] that includes 2702 measurement directions on a Lebedev grid. For the renderers, we use a transition frequency $f_\mathrm{t} = \omega_\mathrm{t} \,/\, (2\pi)$ of 2 kHz as proposed in [13]. In case of the MagLS renderer that uses the conventional pressure model from (6), Tikhonov-regularized radial filters $b_n(\omega)^{-1}$ (cf. (2)) with a regularization weight of 0.01 were used to remove the spherical scattering [5]. In case of the eMagLS and eMagLS2 renderers based on (5) and (4), respectively, a regularization weight $\beta = 0.01$ is used (cf. (11)).



(a) MagLS (b) eMagLS (c) eMagLS2

Fig. 1. SMR errors for incidence directions in the horizontal plane as a function of incidence angle and frequency for frequencies below the transition frequency $f < f_\mathrm{t}$. The simulated array is the *Eigenmike em32* microphone array.



(a) MagLS (b) eMagLS (c) eMagLS2

Fig. 2. Same as Fig. 1 but for frequencies above the transition frequency $f > f_\mathrm{t}$. Note the different color scale compared to Fig. 1.

B. Results

Fig. 1 and 2 show the SMR error $\epsilon_\mathrm{SMR}(\omega, \Omega)$ of the MagLS, the eMagLS and the eMagLS2 renderers for plane-wave incidence directions in the horizontal plane and a simulation of the *Eigenmike em32* SMA. The SMA configuration contains $M = 32$ microphones distributed on a rigid sphere of radius $4.2\,\mathrm{cm}$. It supports a SH decomposition of order $N = 4$, and the spatial aliasing frequency is approximately $5\,\mathrm{kHz}$.

Fig. 1 shows $\epsilon_\mathrm{SMR}(\omega, \Omega)$ for frequencies below the transition frequency $f_\mathrm{t} = 2\,\mathrm{kHz}$. The SMR errors stay below $3\,\mathrm{dB}$ in magnitude for all evaluated directions. The SMR errors of the eMagLS and eMagLS2 renderers are very similar to each other in this frequency range and are only slightly lower than the SMR errors generated by the MagLS renderer. This confirms that the assumption of ideal radial filters in the MagLS renderer creates only a small error.

As shown in Fig. 2, all three renderers create much larger errors at high frequencies above $6\,\mathrm{kHz}$. (Note the different color scales in Fig. 1 and 2.) However, in comparison to the conventional binaural renderer in Fig. 3, that uses the SH transformed HRTF set or, equivalently, the LS renderer (10) with the signal model from (6), all three MagLS-based renderers achieve a significant improvement in the high-frequency range. The low-frequency performance of the conventional binaural renderer is not shown explicitly as the conventional renderer is equivalent to the MagLS renderer for low frequencies, cf. Fig. 1a and (12). Spatial aliasing prevents the accurate rendering of the binaural magnitude response with a low

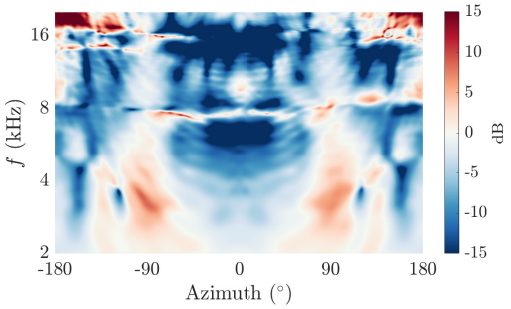


Fig. 3. SMR errors for incidence directions above the transition frequency $f > f_\mathrm{t}$ (same as Fig. 2) but for the conventional binaural renderer using the SH transformed HRTF set. For frequencies below the transition frequency $f < f_\mathrm{t}$, the conventional renderer and the MagLS renderer are equivalent, cf. Fig. 1a.

SH decomposition order of $N = 4$. However, by including the full SMA description into the definition of the plane-wave model, the eMagLS and eMagLS2 renderers improve the rendered magnitude response. In comparison to the MagLS renderer, Fig. 2a, which overemphasizes high frequencies in all directions, the eMagLS and eMagLS2 renderers, Figs. 2b and 2c, equalize the high-frequency magnitude response on average, which is similar to what a diffuse-field equalization filter does. The eMagLS2 renderer, Fig. 2c, creates a slightly lower SMR error than the SH-domain-based eMagLS renderer, Fig. 2b, especially for frontal incidence directions.

The inherent diffuse-field equalization of the proposed

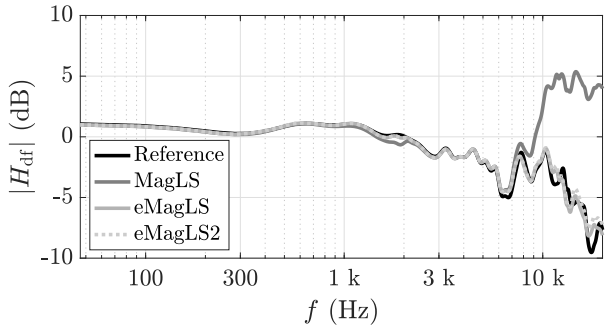


Fig. 4. The binaural diffuse-field magnitude response $|H_{\mathrm{df}}|$ of the employed HRTF set (Reference) and for different rendering approaches. $|H_{\mathrm{df}}|$ is reproduced accurately by the proposed eMagLS and eMagLS2 renderers, while the MagLS renderer overemphasizes the diffuse-field magnitude for high frequencies above 8 kHz.

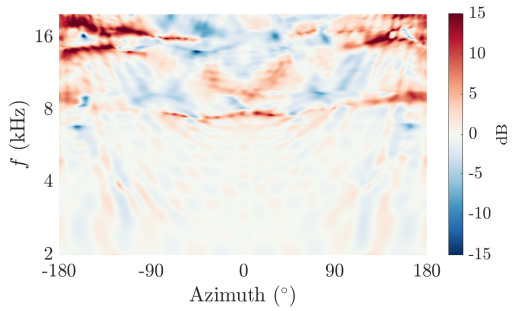


Fig. 5. SMR errors for incidence directions in the horizontal plane, frequencies above the transition frequency $f > f_{\mathrm{t}}$ and the simulated equatorial microphone array (EMA) using the eMagLS renderer.

TABLE I
AVERAGE SMR ERRORS $\bar{\epsilon}_{\mathrm{SMR}}$ IN dB FOR THE MAGLS, EMAGLS AND EMAGLS2 RENDERERS, AND DIFFERENT SMA CONFIGURATIONS. THE SMA CONFIGURATIONS INCLUDE THE EIGENMIKE EM32 WITH A SH DECOMPOSITION ORDER OF $N = 4$ AND RADIUS $r = 4.2\,\mathrm{cm}$, AND FIVE DIFFERENT LEBEDEV GRIDS WITH N BETWEEN 1–5 AND $r = 8.5\,\mathrm{cm}$.

	em32	*Leb1*	*Leb2*	*Leb3*	*Leb4*	*Leb5*
MagLS	4.2	15.0	11.3	9.4	7.9	6.8
eMagLS	2.0	3.6	3.7	3.1	2.8	2.5
eMagLS2	1.8	3.0	3.0	2.6	2.6	2.3

renderers is also observable when considering the binaural diffuse-field magnitude responses in Fig. 4 that were calculated using 2702 plane-wave incidence directions on the entire unit-sphere, including incidence directions outside of the horizontal plane. The eMagLS and eMagLS2 renderers generate binaural diffuse-field magnitude responses that follow the reference diffuse-field magnitude response of the pure HRTFs closely, while the MagLS renderer overemphasizes frequencies above 8 kHz by up to +10 dB.

To further investigate the high-frequency performance of the renderers, average SMR errors $\bar{\epsilon}_{\mathrm{SMR}}$ according to (18) for frequencies above the transition frequency $f_{\mathrm{t}} = 2\,\mathrm{kHz}$ and different SMA configurations are provided in Tab. I. The average SMR errors $\bar{\epsilon}_{\mathrm{SMR}}$ were again calculated for 2702 incidence directions on the entire unit sphere. The array configurations include the *Eigenmike em32* with radius $4.2\,\mathrm{cm}$ and five other SMAs with radius $8.5\,\mathrm{cm}$ and microphones distributed according to Lebedev grids. The Lebedev grids of the arrays *Leb1* to *Leb5* support SH decomposition orders of $N = 1$ to $N = 5$, respectively. For all tested SMA configurations, the MagLS renderer creates the highest average SMR error $\bar{\epsilon}_{\mathrm{SMR}}$ between $4.2\,\mathrm{dB}$ for the *em32* and $15.0\,\mathrm{dB}$ for the *Leb1* grid, followed by the eMagLS and the eMagLS2 renderers. The eMagLS and eMagLS2 renderers reduce $\bar{\epsilon}_{\mathrm{SMR}}$ significantly. The eMagLS renderer generates errors between $2.0\,\mathrm{dB}$ and $3.7\,\mathrm{dB}$, and the eMagLS2 renderer further improves the rendering of the binaural magnitude response slightly, generating average SMR errors between $1.8\,\mathrm{dB}$ and $3.0\,\mathrm{dB}$. In contrast to the SMR errors in Figs. 1–3, the average SMR error $\bar{\epsilon}_{\mathrm{SMR}}$ considers the absolute value of the magnitude error. The eMagLS and eMagLS2 renderers reduce this error considerably for all array configurations under test, which proves that they do not only reduce the magnitude error on average but also in total.

VI. END-TO-END MAGLS RENDERING FOR THE EQUATORIAL MICROPHONE ARRAY

The proposed eMagLS binaural rendering approach is not limited to specific array geometries. In this section, we apply it to the recently proposed equatorial microphone array (EMA) [23] to demonstrate the flexibility of the renderer.

The EMA is essentially an SMA with microphones distributed exclusively around the equator of the spherical scatterer. The EMA supports a signal decomposition into $2N + 1$ circular harmonics (CHs) that are a subset of the SHs for the representation of functions only varying in azimuth. This has the advantage that only $2N+1$ instead of at least $(N+1)^2$ microphones are required for a decomposition order of N, if an EMA is employed instead of an SMA. EMAs provide similar accuracy as SMAs for sound fields from sources inside the horizontal plane at the same SH order N. Interaural time differences are conveyed correctly for all angles of sound incidence. Monaural elevation cues are impaired for non-horizontal incidence with deviations of the magnitude transfer function in the order of a few dB [24].

In the following, we employ an EMA with 9 microphones mounted on a spherical scatterer, supporting a signal decomposition into CHs of up to 4th order. The radius of the array and the maximum decomposition order are identical to the ones of the SMA employed in the previous section. As evident from Fig. 5, the accuracy of the 9-element EMA when using eMagLS rendering is similar to that of the 32-element SMA (depicted in Fig. 2c) for horizontal sound incidence. The binaural diffuse-field magnitude response of the EMA in Fig. 6 deviates by up to $5\,\mathrm{dB}$ from the binaural diffuse-field magnitude response of the HRTFs. This is because the binaural diffuse-field magnitude response comprises sound incidence distributed over all possible angles, including non-horizontal incidence for which the EMA is less accurate. Fig. 6 confirms that the impairment for non-horizontal incidence is moderate.

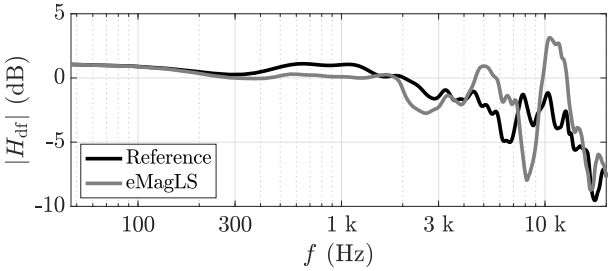


Fig. 6. The binaural diffuse-field magnitude response $|H_{\mathrm{df}}|$ of the employed HRTF set (Reference) and for the equatorial microphone array (EMA) after rendering with the eMagLS renderer.

VII. Conclusion

We extended the MagLS binaural rendering technique by a comprehensive microphone model, including the impact of SH order truncation and spatial aliasing, in order to facilitate a complete end-to-end processing routine for SMA signals without the need of any further equalization filters. The formulation as magnitude least squares optimization problem allows for two different solutions: the eMagLS renderer processes SMA signals in the SH domain, while the eMagLS2 renderer directly operates on raw microphone signals. Both proposed approaches outperform the conventional MagLS renderer at high frequencies, mainly due to their inherent diffuse-field equalization. However, they not only reduce the average error over all directions, resulting in an improved diffuse-field response, but also the absolute error, which we showed for six different SMA configurations. The eMagLS2 renderer performs slightly better than the eMagLS renderer at high frequencies but requires M convolutions, where M is the number of microphones in the array. The eMagLS renderer operates in the SH domain and hence only requires $(N{+}1)^2$ convolutions, where N is the SH decomposition order. Typically, SMAs employ $M > (N{+}1)^2$ microphones such that the eMagLS renderer has a lower computational cost. Moreover, the SH domain operation of the eMagLS renderer has the advantage of facilitating dynamic head rotations by a single matrix multiplication.

We demonstrated the high flexibility of the eMagLS concept by applying it to the recently proposed equatorial microphone array. For horizontal sound incidence, the eMagLS renderer using signals from the equatorial microphone array performs similar to the one using SMA signals while requiring only $2N + 1$ microphones.

MATLAB reference implementations of the proposed eMagLS and eMagLS2 renderers are provided online[1].

References

[1] M. A. Gerzon, "Ambisonics in Multichannel Broadcasting and Video," *Journal of the Audio Engineering Society*, vol. 33, no. 11, p. 859–871, 1985.

[2] J. Ahrens, *Analytic Methods of Sound Field Synthesis*. Heidelberg, Germany: Springer, 2012.

[3] F. Zotter and M. Frank, *Ambisonics, A Practical 3D Audio Theory for Recording, Studio Production, Sound Reinforcement, and Virtual Reality*. Springer, 2019.

[4] B. Rafaely, B. Weiss, and E. Bachmat, "Spatial aliasing in spherical microphone arrays," *IEEE Transactions on Signal Processing*, vol. 55, no. 3, pp. 1003–1010, 2007.

[5] S. Moreau, J. Daniel, and S. Bertet, "3D Sound Field Recording with Higher Order Ambisonics - Objective Measurements and Validation of a 4th Order Spherical Microphone," in *120th Convention of the Audio Eng. Soc.*, 2006.

[6] B. Rafaely, "Analysis and design of spherical microphone arrays," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 1, pp. 135–143, 2005.

[7] M. Zaunschirm, C. Schörkhuber, and R. 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.

[8] C. T. Jin, N. Epain, and A. Parthy, "Design, optimization and evaluation of a dual-radius spherical microphone array," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 22, no. 1, pp. 193–204, 2014.

[9] B. Bernschütz, "Microphone Arrays and Sound Field Decomposition for Dynamic Binaural Recording," Ph.D. dissertation, Technical University of Berlin, 2016.

[10] Z. Ben-Hur, F. Brinkmann, J. Sheaffer, S. Weinzierl, and B. 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.

[11] T. McKenzie, D. T. Murphy, and G. Kearney, "Diffuse-Field Equalisation of binaural ambisonic rendering," *Applied Sciences (Switzerland)*, vol. 8, no. 10, 2018.

[12] T. Lübeck, H. Helmholz, J. M. Arend, C. Pörschmann, and J. 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*, vol. 68, no. 6, pp. 428–440, 2020.

[13] C. Schörkhuber, M. Zaunschirm, and R. Höldrich, "Binaural Rendering of Ambisonic Signals via Magnitude Least Squares," in *Fortschritte der Akustik – DAGA*, 2018, pp. 339–342.

[14] N. Gumerov and R. Duraiswami, *Fast Multipole Methods for the Helmholtz Equation in Three Dimensions*. Amsterdam: Elsevier, 2005.

[15] I. Balmages and B. Rafaely, "Open-Sphere Designs for Spherical Microphone Arrays," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 2, pp. 727–732, 2007.

[16] B. Rafaely, *Fundamentals of Spherical Array Processing*. Springer, 2015.

[17] J. Ahrens and C. 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*, vol. 145, no. April, pp. 2783–2794, 2019.

[18] B. Rafaely and A. 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.

[19] P. W. Kassakian, "Magnitude Least–Squares Fitting via Semidefinite Programming with Applications to Beamforming and Multidimensional Filter Design," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 3, 2005, pp. iii/53–iii/56.

[20] V. Pulkki, M. Karjalainen, and J. Huopaniemi, "Analyzing Virtual Sound Source Attributes Using a Binaural Auditory Model," in *Journal of the Audio Engineering Society*, 1999, pp. 203–217.

[21] K. Ono, V. Pulkki, and M. Karjalainen, "Binaural Modeling of Multiple Sound Source Perception: Coloration of Wideband Sound," in *112th Convention of the Audio Eng. Soc.*, Munich, 2002, pp. 1–8.

[22] B. Bernschütz, "A Spherical Far Field HRIR/HRTF Compilation of the Neumann KU 100," *Fortschritte der Akustik – AIA-DAGA 2013*, pp. 592–595, 2013.

[23] J. Ahrens, H. Helmholz, D. Alon, and S. Amengual Garí, "Spherical Harmonic Decomposition of a Sound Field Based on Observations Along the Equator of a Rigid Spherical Scatterer," *J. Acoust. Soc. Am.*, 2021, (submitted).

[24] J. Ahrens, H. Helmholz, D. L. Alon, and S. V. Amengual Garí, "A head-mounted microphone array for binaural rendering," in *Int. 3 D Audio Conference (I3DA)*, 2021.

[1] https://github.com/thomasdeppisch/eMagLS