OPEN ACCESS Freely available online

B. Stahl and S. Riedel, "Perceptual Comparison of Dynamic Binaural Reproduction Methods for Sparse Head-Mounted Microphone Arrays," *J. Audio Eng. Soc.*, vol. 72, no. 7/8, pp. 442–454 (2024 Jul./Aug.). https://doi.org/10.17743/jaes.2022.0140.



# Perceptual Comparison of Dynamic Binaural Reproduction Methods for Sparse Head-Mounted Microphone Arrays

**BENJAMIN STAHL,*** *AES Student Member,* **AND STEFAN RIEDEL,** *AES Student Member*
(stahl@iem.at) (riedel@iem.at)

*Institute of Electronic Music and Acoustics, University of Music and Performing Arts, Graz, Austria*

This paper presents results of a listening experiment evaluating three-degrees-of-freedom binaural reproduction of head-mounted microphone array signals. The methods are applied to an array of five microphones whose signals were simulated for static and dynamic array orientations. Methods under test involve scene-agnostic binaural reproduction methods as well as methods that have knowledge of (a subset of) source directions. The results of an instrumental evaluation reveal errors in the reproduction of interaural level and time differences for all scene-agnostic methods, which are smallest for the end-to-end magnitude-least-squares method. Additionally, the inherent localization robustness of the array under test and different simulated microphone arrays is investigated and discussed, which is of interest for a parametric reproduction method included in the study. In the listening experiment, the end-to-end magnitude-least-squares reproduction method outperforms other scene-agnostic approaches. Above all, linearly constrained beamformers using known source directions in combination with the end-to-end magnitude-least-squares method outcompete the scene-agnostic methods in perceived quality, especially for a rotating microphone array under anechoic conditions.

## 0 INTRODUCTION

Head-mounted displays enable augmented reality (AR) and virtual reality (VR) applications. Typically, cameras and microphones are placed on the head-mounted device to capture the surrounding scene [2]. Spatial recording and reproduction of a scene is a viable use-case of such head-mounted devices [3, 4] (cf. [5] for a review). In AR/VR applications, binaural reproduction is targeted, and a dynamic mapping of the recorded signals is required to account for movements during recording and during reproduction (cf. Fig. 1).

A straightforward technique to map microphone array signals onto binaural signals is least-squares filtering [6], where a multichannel filter is matched to reproduce a set of head-related transfer functions (HRTFs), which describe directional characteristics of the ears:

$$\mathbf{w}_{\mathrm{LS},f} = \underset{\mathbf{w}_f}{\operatorname{argmin}} \|\mathbf{w}_f \mathbf{A}_f - \mathbf{h}_f\|^2, \tag{1}$$

where $\mathbf{h}_f \in \mathbb{C}^{1\times G}$ is a vector of HRTFs at frequency (bin) $f$ for a set of $G$ directions, $\mathbf{A}_f \in \mathbb{C}^{M\times G}$ are transfer functions of an array of $M$ microphones for the same directions, $\mathbf{w}_{\mathrm{LS},f} \in \mathbb{C}^{1\times M}$ is the multiple-input-single-output filter applied to the array signals, and $\mathbf{w}_{\mathrm{LS},f}$ is the filter minimizing the squared distance between the resulting HRTF modeled from the array transfer functions and the target HRTF. Chen et al. [6] proposed a least-squares approximation of cat HRTFs, and Poletti and Svensson [7] modeled individual human HRTFs with a 640-sensor spherical array.

The "Virtual Artificial Head" [8–10] used planar and open three-dimensional arrays to binaurally reproduce sound scenes with individual HRTFs. Fallahi et al. [11] evaluated these arrays and reproduction techniques in a listening experiment with head tracking. Madmoni et al. [12] conducted a simulation study on least-squares approximation of HRTFs using a semi-circular array of six microphones. However, only one static, frontal array orientation was considered.

Deppisch et al. [13] proposed an end-to-end magnitude-least-squares (MLS) approach for binaural rendering of microphone array signals, evaluated in a simulation study incorporating spherical and equatorial microphone arrays. Helmholz et al. [14] proposed improving this MLS approach for equatorial microphones by projecting ambisonically encoded signals onto the horizontal plane and evalu-

*To whom correspondence should be addressed, email: stahl@iem.at, Last updated: April 25, 2024

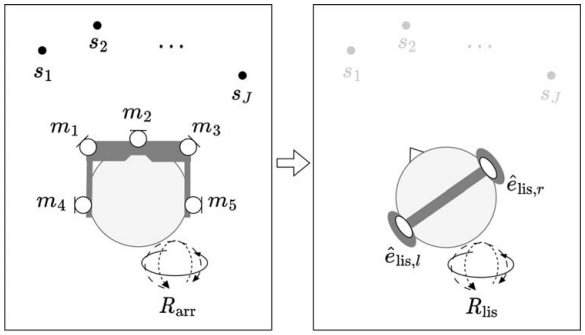



Fig. 1. Visualization of recording (a) and three-degrees-of-freedom binaural reproduction (b).

ated their approach in an instrumental evaluation for both horizontal and vertical head rotations. The MLS approach is based on Schörkhuber et al. [15], who proposed an MLS solution to better represent the magnitude of HRTFs in low Ambisonic orders. This representation can also be leveraged as a rendering method by encoding microphone signals into first-order Ambisonics (FOA) and dynamically decoding using Ambisonic rotation operations and the MLS FOA HRTFs [16].

Ifergan and Rafaely [17] used beamforming-based binaural rendering (BFBR), where a set of distortionless-response beamformers extracts direction signals to be convolved with head-related impulse responses (HRIRs). The optimal number of beamformers with BFBR was found to approximately correspond to the number of microphones, suggested by the results of a simulation study and a listening experiment for planar and spherical arrays with nine and 12 microphones, respectively. Lübeck et al. [4] compared MLS, BFBR, and other binaural reproduction methods in a static listening experiment including a head-mounted microphone array.

The methods named so far do not make use of direction-of-arrival (DOA) estimation techniques. Parametric encoding and upmixing algorithms use DOA estimation techniques to extract narrowband sound components and encode them into higher-order Ambisonics. For Ambisonic input, various estimation techniques have been proposed: intensity-based [18, 19], compressed sensing-based [20], and decomposition-based [21–24]. From spherical harmonic (SH)–domain input signals, the COMPASS algorithm [24] estimates the number of sources and source directions in each frequency bin using the SORTE algorithm [25] and MUSIC algorithm [26], respectively. It subsequently encodes the dominant directional components into the SH domain at a high spatial resolution while encoding the diffuse sound using a scene-agnostic beamforming-based approach. For microphone input, a recently proposed algorithm by McCormack et al. [3, 27] employs the same techniques directly on the microphone signals, which can be advantageous for arbitrary irregular microphone arrays. In the remainder of the paper, this method is referred to as the narrowband parametric (NBP) method.

Previous perceptual evaluation of binaural reproduction methods for microphone array signals used either static rendering [4] for head-mounted arrays or dynamic rendering, however, for planar and open arrays [11]. In this study, dynamic rendering for head-mounted arrays is targeted, where the time-variant orientation during signal acquisition is taken into account. Two classes of algorithms are considered: first, algorithms that do not have prior knowledge about source directions (scene-agnostic), such as FOA encoding and binaural decoding, BFBR, MLS, and NBP Ambisonic encoding, and second, an algorithm that makes use of prior knowledge of the source directions by extracting signals from sound sources with known directions using linearly constrained minimum-variance (LCMV) beamformers and rendering the residual ambient sound using the MLS approach. Furthermore, it is investigated whether knowledge of a subset of source directions already yields improvement. The algorithms are evaluated both in an instrumental evaluation and a listening experiment.

## 1 PROBLEM FORMULATION

The goal of the three-degrees-of-freedom binaural reproduction is to estimate ear signals $\hat{e}_\zeta[t]$ from microphone array signals $\mathbf{x}[t]$, taking into account the array and listener orientations. Let $\mathbf{A}[t] \in \mathbb{R}^{M\times G}$ be the impulse responses of an array with $M$ microphones measured for a dense grid of $G$ source directions, and its SH transform shall be denoted as $\mathbf{D}[t] = \mathbf{A}[t]\mathbf{Y}^\dagger \in \mathbb{R}^{M\times(\hat{N}+1)^2}$. The operation $(\cdot)\dagger$ denotes pseudo inverse (weighted by area weights). To avoid spatial aliasing $\hat{N} > kr$ is necessary, where $k$ is the wave number and $r$ is the array radius. Given a scenario with $J$ sources, the Ambisonic sound field $\boldsymbol{\chi}[t] \in \mathbb{R}^{(\hat{N}+1)^2\times 1}$ maps to the signals $\mathbf{x}[t]$ of a variably rotated microphone array as

$$\mathbf{x}[t] = \left(\mathbf{D} * \left(\boldsymbol{R}^{-1}_{\hat{N},\text{arr}}\,\boldsymbol{\chi}\right)\right)[t] \text{ with } \boldsymbol{\chi}[t] = \sum_{j=1}^{J}(\mathbf{r}_j * s_j)[t], \tag{2}$$

where * denotes (matrix) convolution along the time dimension. The $j$th anechoic source signal is denoted by $s_j[t]$, and the corresponding Ambisonic room impulse response by the matrix $\boldsymbol{R}_{\hat{N},\text{arr}} \in \mathbb{R}^{(\hat{N}+1)^2\times(\hat{N}+1)^2}$ is an Ambisonic (Wigner-$D$) rotation matrix to account for the array orientation [16]. The array orientation is generally time-varying.

From the Ambisonic sound field $\boldsymbol{\chi}[t]$, a set of reference ear signals are obtained by

$$e_\zeta[t] = (\mathbf{h}^{\mathsf{T}}_{\text{MLS},\zeta} * \boldsymbol{R}^{-1}_{N,\text{lis}}\boldsymbol{\chi}_N)[t], \tag{3}$$

where $\mathbf{h}_{\text{MLS},\zeta}[t] \in \mathbb{R}^{(N+1)^2\times 1}$ is an $N$th-order Ambisonic MLS HRIR of the ear $\zeta \in \{\text{L}, \text{R}\}$, $\boldsymbol{\chi}_N[t]$ denotes the Ambisonic reference limited to order $N$, and $\boldsymbol{R}_{N,\text{lis}}$ is the $N$th-order Ambisonic rotation matrix corresponding to the listener's orientation. MLS binaural decoding is described in [15], where HRTFs of Ambisonic orders as low as $N = 3$ have resulted in high perceived binaural reproduction quality.

Table 1. Investigated rendering methods and their abbreviations. Scene-agnostic algorithms do not assume knowledge about source directions.

| | Scene-agnostic | Description |
|---|---|---|
| FOA | Yes | First-order Ambisonic rendering |
| BFBR | Yes | Beamforming-based binaural rendering |
| MLS | Yes | Magnitude-least-squares rendering |
| NBP | Yes | Narrowband parametric rendering [3] |
| BF+MLS | No | Beamformers on all sources |
| BF+MLS* | No | Beamformers on subset of sources |

## 2 BINAURAL REPRODUCTION METHODS

### 2.1 End-To-End MLS

Table 1 gives an overview of the investigated binaural reproduction methods. End-to-end MLS has been described in [13]. With this approach, the ear signals are directly estimated from the array signals. The MLS optimization problem in the frequency domain is

$$\begin{aligned}&\mathbf{w}_{\mathrm{MLS},\zeta,f}(\boldsymbol{R}_{\mathrm{rel}})\\&=\underset{\mathbf{w}_f}{\operatorname{argmin}}\begin{cases}\left\|\mathbf{w}_f\mathbf{A}_f-\mathbf{h}_{\zeta,f}(\boldsymbol{R}_{\mathrm{rel}}^{-1})\right\|^2 & ,\text{for } f\le f_c\\ \left\|\left|\mathbf{w}_f\mathbf{A}_f\right|-\left|\mathbf{h}_{\zeta,f}(\boldsymbol{R}_{\mathrm{rel}}^{-1})\right|\right\|^2 & ,\text{for } f> f_c\end{cases},\\&\text{s.t. } ||\mathbf{w}_f\mathbf{A}_f||=||\mathbf{h}_{\zeta,f}||,\end{aligned}\tag{4}$$

where $\boldsymbol{R}_{\mathrm{rel}}=\boldsymbol{R}_{\mathrm{lis}}\boldsymbol{R}_{\mathrm{arr}}^{-1}$ is the relative rotation between microphone array and listener, $\mathbf{A}_f\in\mathbb{C}^{M\times G}$ is a matrix of array transfer functions sampled on a fine grid of directions that uniformly sample the sphere, $\mathbf{h}_{\zeta,f}(\boldsymbol{R})\in\mathbb{C}^{1\times G}$ denotes a set of HRTFs sampled on the same grid rotated by $\boldsymbol{R}$, $f_c$ is the cutoff frequency above which MLS optimality is targeted, and $|\cdot|$ denotes element-wise absolute value.

The MLS solution is approximated by continuing the phase of reconstructed HRTFs over frequency, which has been proposed in [16]:

$$\boldsymbol{\phi}_f=\angle(\tilde{\mathbf{w}}_{\zeta,f-1}\mathbf{A}_f)\tag{5}$$

$$\tilde{\mathbf{w}}_{\zeta,f}=\begin{cases}\mathbf{h}_{\zeta,f}\mathbf{A}_f^{\dagger} & ,\text{for } f\le f_c\\ |\mathbf{h}_{\zeta,f}|\,e^{j\boldsymbol{\phi}_f}\mathbf{A}_f^{\dagger} & ,\text{for } f> f_c\end{cases},\tag{6}$$

where $\tilde{\mathbf{w}}_{\zeta,f}$ is the solution to the unconstrained MLS problem. The operations $e^{(\cdot)}$ and $\angle(\cdot)$ are applied element-wise on their vector-valued arguments. Note that the dependency on $\boldsymbol{R}_{\mathrm{rel}}$ has been omitted here for better readability. Note that, while [13] used a covariance constraint to preserve the interaural diffuse-field correlation, only a monaural diffuse-field equalization for each ear is applied. This side constraint can be fulfilled by scaling the solution to the unconstrained problem:

$$\mathbf{w}_{\mathrm{MLS},\zeta,f}=\tilde{\mathbf{w}}_{\zeta,f}\frac{||\mathbf{h}_{\zeta,f}||}{||\tilde{\mathbf{w}}_{\zeta,f}\mathbf{A}_f||}.\tag{7}$$

The time-domain ear signal estimates are then

$$\hat{e}_\zeta[t]=\left(\mathbf{w}_{\mathrm{MLS},\zeta}(\boldsymbol{R}_{\mathrm{rel}})*\mathbf{x}\right)[t],\tag{8}$$

where $\mathbf{w}_{\mathrm{MLS},\zeta}(\boldsymbol{R}_{\mathrm{rel}})[t]$ is the inverse discrete Fourier transform (DFT) of $\mathbf{w}_{\mathrm{MLS},\zeta,f}(\boldsymbol{R}_{\mathrm{rel}})$. In practice, the MLS filter weights $\mathbf{w}_{\mathrm{MLS},\zeta}(\boldsymbol{R}_{\mathrm{rel}})[t]$ can be precomputed for a set of possible relative rotations, and the filter weights can then be dynamically selected.

### 2.2 FOA Rendering

The FOA rendering is based on estimating the FOA sound field from the available array signals, which is done in the frequency domain:

$$\begin{aligned}\mathbf{W}_{\mathrm{FOA},f}=\underset{\mathbf{W}_f}{\operatorname{argmin}}&\left\{||\mathbf{W}_f\,\mathbf{D}_{1,f}-\mathbf{I}_4||_{\mathrm{fro}}^2\right\}\\&\text{s.t. } ||\mathbf{W}\,\mathbf{D}_f||_{\mathrm{fro}}^2=4,\end{aligned}\tag{9}$$

where $\mathbf{D}_f$ denotes the DFT of the directional array impulse responses $\mathbf{D}[t]\in\mathbb{R}^{M\times(\hat{N}+1)^2}$ at frequency bin $f$; $\mathbf{D}_{1,f}$ denotes the first-order sound field components, i.e., the first four columns of $\mathbf{D}_f$; and $\mathbf{W}_f$ is a multiple-input-multiple-output filter applied to the array signals. The filter $\mathbf{W}_{\mathrm{FOA},f}$ projects the first-order components of the array signals onto an ideal first-order SH representation of the sound field in a least-squares sense. Like in Sec. 2.1, this optimization problem can be solved by first computing the least-squares solution and then scaling the solution by the factor that is required to fulfill the constraint[1]:

$$\tilde{\mathbf{W}}_{\mathrm{FOA},f}=\mathbf{D}_{1,f}^{\dagger},\tag{10}$$

$$\mathbf{W}_{\mathrm{FOA},f}=\tilde{\mathbf{W}}_f\frac{2}{\|\tilde{\mathbf{W}}_f\mathbf{D}_f\|_{\mathrm{fro}}}.\tag{11}$$

The Ambisonic soundfield in the time domain is then

$$\hat{\boldsymbol{z}}=\boldsymbol{R}_{1,\mathrm{arr}}(\mathbf{W}_{\mathrm{FOA}}*\boldsymbol{x})[t],\tag{12}$$

where $\mathbf{W}_{\mathrm{FOA}}[t]$ is the inverse DFT of $\mathbf{W}_{\mathrm{FOA},f}$, and $\boldsymbol{R}_{1,\mathrm{arr}}$ denotes the FOA rotation matrix accounting for the array rotation. The time-domain ear signals are computed as

$$\hat{e}_\zeta[t]=(\mathbf{h}_{\mathrm{MLS}-1,\zeta}^{\mathsf{T}}*\boldsymbol{R}_{1,\mathrm{lis}}^{-1}\hat{\mathbf{z}})[t],\tag{13}$$

where $\mathbf{h}_{\mathrm{MLS}-1,\zeta}^{\mathsf{T}}$ denotes the MLS FOA HRIR of ear $\zeta$, and $\boldsymbol{R}_{1,\mathrm{lis}}$ is the FOA rotation matrix corresponding to the listener's orientation.

### 2.3 BFBR

BFBR has been proposed in [17]. With BFBR, the microphone array is used to extract a set of directional signals, from directions equally distributed to sample the sound scene. Typically, maximum-directivity beamformers are used: Let $\{\mathbf{u}_1,\ldots,\mathbf{u}_K\}$ be a set of directions (cartesian unit

[1] **Note:** The employed first-order encoder neglects higher-order SH components of the array impulse responses but is commonly used with arrays that equally sample the sphere [16], which is why the authors used such an encoder in this study. The authors later found that an encoder that takes higher-order SHs into account by $\operatorname{argmin}_{\tilde{\mathbf{W}}}\{||\tilde{\mathbf{W}}\,\mathbf{D}_f-[\mathbf{I}_4\ \mathbf{0}_{4\times(\hat{N}+1)^2-4}]||_{\mathrm{fro}}^2\}$ with solution $\tilde{\mathbf{W}}=[\mathbf{I}_4\ \mathbf{0}_{4\times(\hat{N}+1)^2-4}]\mathbf{D}_f^{\dagger}$ is more suitable for irregular arrays, because it suffers from less systematic spatial distortions due to aliasing and, therefore, would have been a better candidate.

vectors), which sample the soundfield. The corresponding steering vectors $\mathbf{a}_k \in \mathbb{C}^{M\times 1}$ can be computed as

$$\mathbf{a}_{k,f} = \mathbf{D}_f \, \boldsymbol{R}^{-1}_{\hat{N},\mathrm{arr}} \, \boldsymbol{y}(\mathbf{u}_k), \tag{14}$$

where $\boldsymbol{y}(\mathbf{u})$ is an order-$\hat{N}$ vector of SHs evaluated at the direction $\mathbf{u}$. The diffuse-field covariance matrix of the array is computed as

$$\mathbf{C}_f = \frac{1}{4\pi}\mathbf{D}_f\mathbf{D}_f^{\mathsf{H}}. \tag{15}$$

Taking the diffuse-field covariance into account, maximum-directivity beamformers [28] are computed as

$$\mathbf{w}_{\mathrm{MD},k,f} = \frac{\mathbf{a}_{k,f}^{\mathsf{H}}\mathbf{C}_f^{-1}}{\mathbf{a}_{k,f}^{\mathsf{H}}\mathbf{C}_f^{-1}\mathbf{a}_{k,f}}. \tag{16}$$

The corresponding beamformer output signals in the time domain become

$$y_k[t] = (\mathbf{w}_{\mathrm{MD},k} * \mathbf{x})[t], \tag{17}$$

where $\mathbf{w}_{\mathrm{MD},\,k}[t]$ is the inverse DFT of $\mathbf{w}_{\mathrm{MD},\,k,f}$. Finally, the ear signals are obtained by convolution with HRIRs corresponding to the beam directions.

$$\hat{e}_\zeta[t] = \sum_{k=1}^{K}(h_\zeta(\boldsymbol{R}^{-1}_{\mathrm{lis}}\mathbf{u}_k) * y_k)[t], \tag{18}$$

where $h_\zeta(\mathbf{u})$ denotes the HRIR of the ear $\zeta$ evaluated at the direction $\mathbf{u}$, and $\boldsymbol{R}_{\mathrm{lis}}$ is a cartesian rotation matrix describing the listener's rotation.

### 2.4 NBP Method

The NBP method [3] estimates the number of sources and source directions directly from the short-time Fourier transforms of the microphone signals $\mathbf{x}[t]$ using SORTE [25] and MUSIC [26], respectively. To extract the dominant directional signals, it uses an LCMV beamformer. These signals are encoded into the SH domain, e.g., at fifth order. The ambient sound components are extracted from the residual microphone signals using an energy-preserving plane-wave decomposition followed by decorrelation (optional) and encoding of these ambience signals.

### 2.5 LCMV Beamformers With Known Source Directions and MLS Residual Rendering

If the directions of the sound sources are known, which can be retrieved from visual and audiovisual tracking [29] or from user input, it is reasonable to extract their direct sound signals using beamformers. Inspired by the NBP method proposed in [3], the rendering is split into a beamforming-based processing path for the direct sound and an additional residual processing path. However, the method proposed here differs from the NBP method in that the beamformer directions are not estimated but assumed to be known and do not differ across frequency.

Let $\{\Omega_1, \ldots, \Omega_J\}$ be the set of known source directions and $\{\mathbf{a}_1, \ldots, \mathbf{a}_J\}$ be the corresponding steering vectors obtained from the rotated array as in Eq. (14). Let $\mathbf{A} = [\mathbf{a}_1 \ldots \mathbf{a}_J]$ be a matrix whose columns are the source steering vectors. The LCMV beamformer extracting and separating these source signals is formulated in the frequency domain as

$$\mathbf{W}_{\mathrm{LCMV},f} = \mathbf{A}_f^{\mathsf{H}}\mathbf{C}_f^{-1}\left(\mathbf{A}_f^{\mathsf{H}}\mathbf{C}_f^{-1}\mathbf{A}_f\right)^{-1}. \tag{19}$$

The corresponding time-domain output signals are

$$\mathbf{y}[t] = (\mathbf{W}_{\mathrm{LCMV}} * \mathbf{x})[t], \tag{20}$$

where $\mathbf{y}[t] = [y_1[t] \ \ldots \ y_J[t]]^{\mathsf{T}}$ contains the individual beamformer output signals, and $\mathbf{W}_{\mathrm{LCMV}}[t]$ is the inverse DFT of $\mathbf{W}_{\mathrm{LCMV},\,f}$. The output signals are convolved with HRIRs as in Eq. (18). The residual microphone signals are obtained by

$$\mathbf{W}_{\mathrm{res},\,f} = \mathbf{I}_M - \mathbf{A}_f\mathbf{W}_{\mathrm{LCMV},\,f}, \tag{21}$$

$$\mathbf{x}_{\mathrm{res}}[t] = (\mathbf{W}_{\mathrm{res}} * \mathbf{x})[t], \tag{22}$$

where $\mathbf{I}_M$ is a $M \times M$ identity matrix, and $\mathbf{W}_{\mathrm{res}}[t]$ is the inverse DFT of $\mathbf{W}_{\mathrm{res},\,f}$. The residual microphone signals are convolved with the MLS filter as in Eq. (8). From here, this method is denoted by BF+MLS, if all source directions are known, and by BF+MLS* if only a subset of source directions is known.

### 2.6 Algorithm Settings

In the instrumental evaluation and experiments described below, the following algorithm settings were used:

- For the FOA rendering method, the vertical component ($z$-component) was not considered in the pseudo-inverse in Eq. (10) and was filled with zeros in the first-order encoding matrix.
- For the BFBR method, $K = 5$ beams were steered at five directions equally distributed on the equator.
- The pseudo-inversion in Eqs. (6) and (10) and the inversion of the covariance matrix in Eqs. (16) and (19) were regularized by adding $0.001 \cdot \lambda_f$ to the matrix diagonal, where $\lambda_f$ is the covariance matrix's largest eigenvalue.
- For the inversion of the constraint covariance $\left(\mathbf{A}_f^{\mathsf{H}}\mathbf{C}_f^{-1}\mathbf{A}_f\right)$ in Eq. (19), diagonal loading with 0.1 $\lambda_f$ was applied.

## 3 ARRAY TRANSFER FUNCTION DATA

This study employs the EasyCom dataset [2] in all evaluations. The head-mounted array transfer functions provided with the EasyCom dataset had been measured on a head and torso simulator for a rectangular grid of 60 (azimuth) × 17 (elevation) = 1,020 directions. While the original array is composed of six microphones, only five of them were used for symmetry. Three microphones are located on a pair of glasses, and two additional microphones are located at the ears, as indicated in Fig. 1. All microphones have an omnidirectional free-field polar pattern.

This *irregular* microphone array differs from regular microphone arrays, such as tetrahedral or higher-order uniform (e.g., *Eigenmike em32* or *em64*) designs on a spherical scatterer or tetrahedral arrays of cardioid microphones in

free field (*Soundfield microphone*), in that its microphones do not sample the sphere in a uniform way. The array used in this study particularly poses a challenge to encoding and rendering methods, as there are no microphones at the back of the head, making it hard to accurately reproduce sources from this direction.

Array transfer functions were resampled to $f_s = 32\,\text{kHz}$. As described above, the array measurements were encoded into the SH domain at order $\hat{N} = 25$. The SH-encoding pseudo inverse was regularized by diagonal loading with $0.1\ \lambda_f$, where $\lambda_f$ is the largest eigenvalue of $\boldsymbol{Y}\boldsymbol{Y}^{\mathsf{T}}$.

## 4 INSTRUMENTAL EVALUATION

### 4.1 Evaluation of Resulting Binaural Cues

From the SH-domain array transfer functions $\mathbf{D}_f$, transfer functions $\mathbf{a}_f$ were obtained for plane waves impinging from directions on the equator. This was done for azimuth angles spaced by 5° both for a forward array orientation and for the array oriented 90° to the right. By applying the methods in Sec. 2 to $\mathbf{a}_f$ in the frequency domain, resulting HRTFs $\hat{h}_{\zeta,f}$ were obtained that can be compared to reference HRTFs. The reference HRTFs are fifth-order MLS Ambisonic representations of the KU100 mannequin head HRTFs [30]. To show the effect of a beamformer on residual rendering, which is important to the BF+MLS* method, a beamformer was applied that was not steered toward the source direction but fixed at 45° azimuth angle.

The interaural level difference (ILD) is a primary localization cue. Therefore, ILDs are computed for the reference HRTFs and for the HRTFs resulting from the processing methods:

$$\text{ILD}_f = 20\log_{10}|\hat{h}_{\text{L},f}| - 20\log_{10}|\hat{h}_{\text{R},f}|. \tag{23}$$

For frequencies $\leq$1,500 Hz, the interaural time difference (ITD) is a primary localization cue (along with the ILD). To evaluate ITDs, the resulting HRTFs were bandpass-filtered using a bank of Gammatone filters $w_{f_c,f}$ around center frequencies $f_c$ spaced by one equivalent rectangular bandwidth. The ITDs are computed as

$$\text{ITD}_{f_c} = \underset{t}{\text{argmax}} \sum_f \text{Re}\left\{\hat{h}_{\text{L},f}\hat{h}^*_{\text{R},f}\ w^2_{f_c,f}\ e^{j2\pi f t}\right\}. \tag{24}$$

To evaluate spectral distortions associated with the rendering methods, the summed power response (SPR) [13] was calculated:

$$\text{SPR} = 10\log_{10}\left(|\hat{h}_{\text{L},f}|^2 + |\hat{h}_{\text{R},f}|^2\right). \tag{25}$$

The SPR level error for a rendering method is defined as $\Delta\text{SPR} = \text{SPR} - \text{SPR}_{\text{ref}}$, where $\text{SPR}_{\text{ref}}$ denotes the SPR level of the reference HRTF.

Fig. 2 shows the resulting ILDs and ITDs along with the reference ILD and ITD as well as SPR level errors for the forward and right-rotated arrays. The FOA and BFBR methods only coarsely reconstruct the ILDs for the frontal array orientation and suffer from severe distortions for the rotated array. The FOA method reveals a strong shift toward positive ILDs (associated with left localization) for frequencies above 1,000 Hz. The MLS method reconstructs the ILDs reasonably well for the frontal array orientation but suffers from distortions for the rotated array. As expected, the BF+MLS method reconstructs the ILD (and ITD) perfectly when a beam is steered toward the source direction, regardless of array orientation. If the beam is steered at 45°, minor distortions arise for other directions due to the sidelobes of the beamformer compared to the pure MLS rendering. While this distortion becomes evident for the frontal array orientation, the additional ILD distortion from the beamformer sidelobes is comparable to the distortions caused by the pure MLS method for a rotated array.

Regarding ITD cues, all but the BFBR and BF+MLS* conditions reproduce the time difference accurately up until 1 kHz for the frontal array orientation. The BFBR method overestimates ITDs for low frequencies, which can be explained by the interaction between maximum-directivity beampatterns. The BF+MLS with a beam steered at 45° yields distortions for directions where beamformer sidelobes interact with the residual rendering (negative azimuth range). The FOA and BFBR methods, as well as the MLS method with a beam steered at 45°, show severe narrowband ITD errors (inverted sign/lateralization) for frequencies around 1 kHz. These errors become more frequent for the rotated microphone array.

The SPR level error analysis reveals that all methods except for the the BF+MLS method result in clear spectral distortions up to around $\pm 10\,\text{dB}$. The residual error of the BF+MLS method is due to the regularization of the beamformer, which impedes a perfectly distortionless response. Among the scene-agnostic methods, MLS rendering leads to significantly lower errors than FOA and BFBR for frequencies below approximately 6 kHz for the frontal orientation.

### 4.2 Steering Vector Analysis for NBP Method

For the NBP method, the inherent localization robustness of the microphone array is expected to impact the reproduction accuracy. Therefore, the MUSIC [26] localization errors are evaluated for a varying direct-to-diffuse energy ratio (DDR), comparing the array used in this study (EasyCom) and other simulated arrays.

The array covariance matrix of a partially diffuse sound field is expressed as

$$\mathbf{C}_{\text{mix},f}(\mathbf{u}) = \mathbf{a}_f(\mathbf{u})\,\mathbf{a}_f^{\mathsf{H}}(\mathbf{u}) + 10^{-\frac{\text{DDR}}{10}}\mathbf{C}_f, \tag{26}$$

where $\mathbf{a}_f(\mathbf{u})$ is the steering vector for direction $\mathbf{u}$, and $\mathbf{C}_f$ is the diffuse-field covariance matrix. The MUSIC DOA estimation method first carries out an eigenvalue decomposition on $\mathbf{C}_{\text{mix},f}(\mathbf{u})$ in order to obtain the noise subspace $\mathbf{N}_f(\mathbf{u}) \in \mathbb{C}^{M\times(M-1)}$, whose columns are the eigenvectors associated with all but the largest eigenvalue. The estimated direction is

$$\hat{\mathbf{u}}_{\text{MUSIC},f} = \underset{\hat{\mathbf{u}}}{\text{argmax}}\ \frac{1}{\mathbf{a}_f^{\mathsf{H}}(\hat{\mathbf{u}})\mathbf{N}_f(\mathbf{u})\mathbf{N}_f^{\mathsf{H}}(\mathbf{u})\mathbf{a}_f(\hat{\mathbf{u}})}. \tag{27}$$

Next to the array used in this study (*Easycom-5MicEncDec*), whose steering vectors and diffuse-field covari-

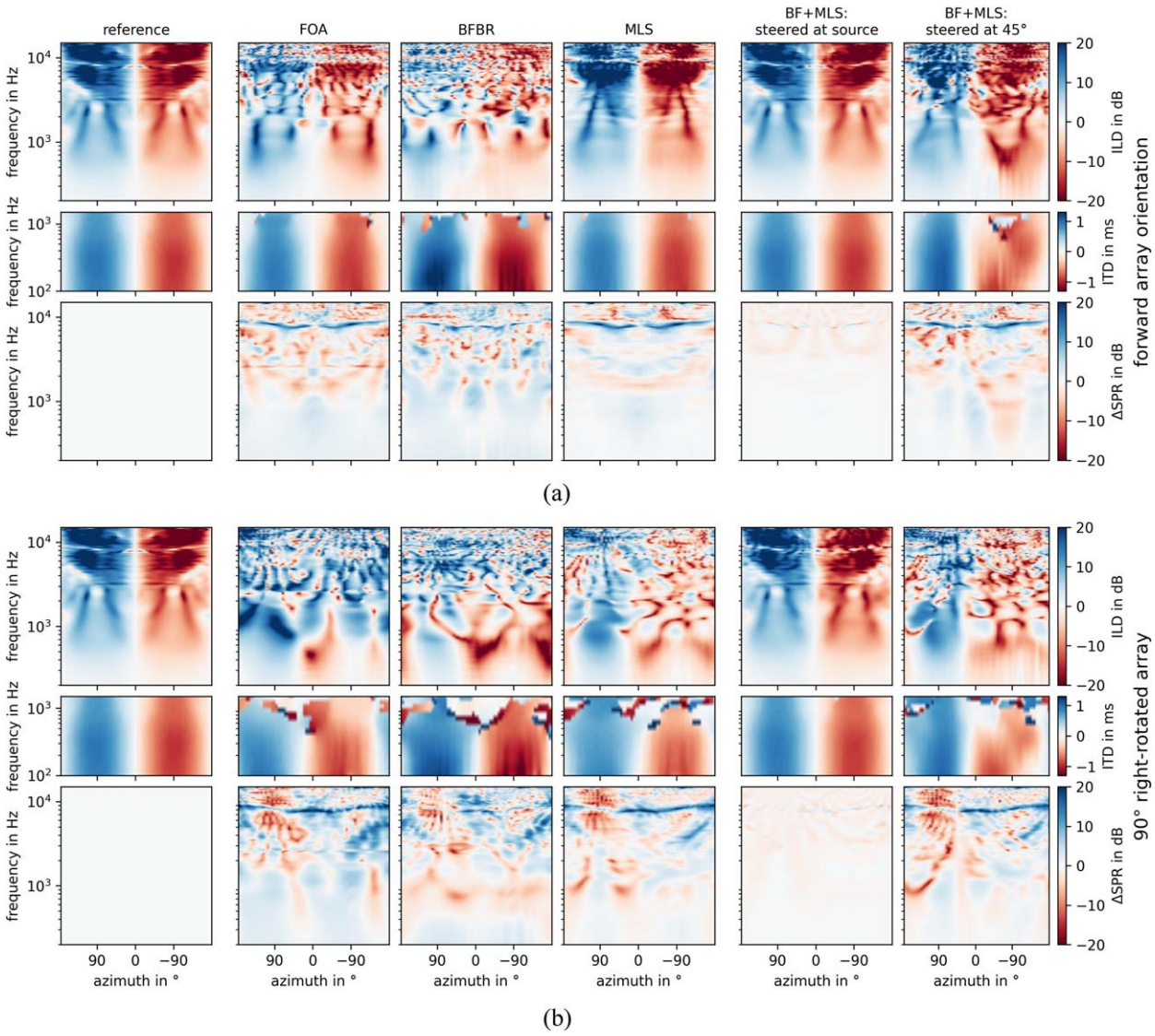



Fig. 2. ILD, ITD, and SPR error (Δ SPR) for a plane wave impinging from directions along the azimuth (zero elevation). (a) array oriented toward the front; (b) array rotated by 90° to the right. As the NBP method [3] is signal-dependent, it could not be evaluated in this way and is not shown here.

ance matrix are obtained from its 25th-order Ambisonic representation, the original EasyCom array was also investigated, denoted by *Easycom-5Mic*. For *Easycom-5Mic*, the diffuse-field covariance was computed directly from the space-domain steering vectors weighted by area weights accounting for the nonuniform measurement grid.

Next, a spherical array with five (omnidirectional) pressure sensors equally spaced along the frontal semi-equator (*Sphere-5Mic*) was simulated using a 25th-order spherical scatterer model (sphere radius $r = 8$ cm) [16]. Additionally an array with seven sensors, adding a rear and top microphone, was simulated (*Sphere-7Mic*).

The head-mounted array used by McCormack et al. [3] was also included. This array of (omnidirectional) pressure sensors had been simulated using the boundary element method (*BEM-7Mic*). To resample to the grid used in this analysis, the array transfer functions were encoded into 25th-order SH domain. Like *Sphere-7Mic*, the BEM-simulated array has five sensors along the frontal semi-equator, a rear microphone, and a top microphone. To allow for a comparison to the EasyCom and spherical arrays, a reduced version was obtained by omitting the top and rear sensors (*BEM-5Mic*).

Finally, a spherical array with seven sensors distributed on the equator was also evaluated (*Sphere-7Mic-circ*). This distribution of microphones was included as [27] used a head-mounted array with a similar distribution of microphones.

As a search grid of vectors $\hat{\mathbf{u}}$ for the MUSIC algorithm, the upper-hemisphere 60(azimuth) × 10(elevation) = 600 measurement positions of the grid described in Sec. 3 were used in order to exclude trivial misdetections by the *Sphere-5Mic* array, for which the steering vectors are symmetric around the horizontal plane. For the simulated directions $\mathbf{u}$, the same grid was used.

Per frequency bin, the median angular error over all source directions $\mathbf{u}$ was evaluated for 11 DDR values uniformly spaced in the range [–20, 10] dB. As a point of

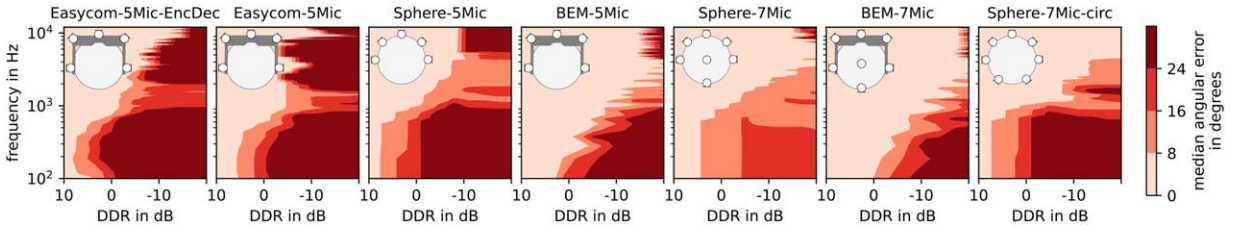



Fig. 3. MUSIC median angular errors for a single source plus isotropic diffuse sound evaluated for various DDRs. The SH-represented (EncDec) EasyCom array is used in this study.

reference, the reverberant conditions in the experiment described below in Sec. 5 used a direct-to-reverberant energy ratio (DRR) of –6 dB. Note that DDR and DRR cannot be directly compared, as DDR describes the ratio of direct sound to fully diffuse sound, cf. Eq. (26), while DRR is used to describe room impulse responses including early reflections, which had been simulated for the listening experiment.

The median angular errors are visualized in Fig. 3. Generally, prominent errors are expected at low frequencies due to wavelengths larger than the array dimensions ($\lambda \gg r$) and high frequencies due to spatial aliasing for $\lambda/2 < d$, where $d$ is the sensor spacing.

The SH-representation of the EasyCom array shows larger errors in the low-frequency range compared to the original EasyCom transfer functions, likely due to the reduction in directional differences caused by the regularization described in Sec. 3. Both the SH-represented and original EasyCom transfer functions lead to significantly larger localization errors in the high-frequency range above 2 kHz compared to the simulated microphone arrays.

The BEM-simulated array shows smaller errors compared to the spherical array, which can be explained by more unique steering vectors due to the more complex shape of the scatterer. As expected, the seven-sensor variants show lower errors than their five-sensor counterparts. This can be seen both for the spherical scatterer arrays and for the BEM arrays. Compared to the *Sphere-5Mic* array, the *Sphere-7Mic* array reduces the error for all frequencies, while the *Sphere-7Mic-circ* array mostly reduces the error at frequencies above 1 kHz. Overall, the simulated seven-sensor BEM array used in [3] shows the lowest localization errors, significantly lower than the measured five-sensor EasyCom array used in this study. This analysis indicates that the EasyCom array poses a challenge for a parametric rendering technique such as the NBP method considered in this study [3].

## 5 LISTENING EXPERIMENT

### 5.1 Experiment Design and Setup

The algorithms were compared in a dynamic listening experiment. To this end, microphone signals were simulated for three base scenarios listed in Table 2. The scenarios were chosen to cover possible use cases for recordings with head-mounted microphone arrays. From each of these base scenarios, four scenarios resulting from the combination of two reverberation conditions (anechoic and reverberant) and array orientations (forward-facing array, dynamically rotated array) were derived. Audio samples were obtained from the EBU Sound Quality Assessment Material CD (speech) and an anechoic string quartet recording dataset [31]. The resulting simulated microphone signals and stimuli had a length of 2.5 s. This short duration was chosen to enable participants to form a global (integrated over time) impression of similarity/quality for each rated condition.

Table 2. Spatial configurations of the scenarios.

| Scenario | Sources, directions given as {(azimuth$_1$, zenith$_1$), (azimuth$_2$, zenith$_2$), ...} in the order of naming |
|---|---|
| *string_quartet* | Violins, viola, and cello {(− 90 °, 90 °), (− 30 °, 90 °), (30 °, 90 °), (90 °, 90 °)} |
| *speech_front* | Male speaker speaker and female speaker {(− 30 °, 90 °), (0 °, 90 °)} |
| *speech_opposite* | Male speaker speaker and female speaker {(− 135 °, 90 °), (45 °, 70 °)} |

For the reverberant condition, a simple image source model with frequency-independent reverberation was parameterized according to a room with dimensions (7 × 6 × 3.5 m), with the listener positioned at (3.5, 4, 1.5 m). The distance of all sources and the wall reflection coefficient was selected such that the DRR was –6 dB and reverberation time was 0.6 s. For the dynamic array rotation, the array linearly rotated from 90° to −90° yaw angle within the 2.5 s of the sound. This was simulated by dividing the 25th-order Ambisonic reference into Hann-windowed 50% overlapping blocks of length 20 ms, multiplying each block with a corresponding Ambisonic rotation matrix, and recomposing the signal.

The simulated microphone signals were processed using the methods described Sec. 2. To include a case in which not all the source directions are known, the BF+MLS method was also applied with beams steered only at the two central instruments known in the *string_quartet* scenario, a beam steered only at the male speaker in the *speech_front* scenario, and a beam steered only at the female speaker in the *speech_opposite* scenario. This method is denoted by BF+MLS*.

All filters had a length of 16 ms (512 samples at $f_s$ = 32 kHz) and were obtained from the inverse DFT of a frequency-domain filter with DFT size 1,024 by shifting the inverse DFT by half the DFT size, taking the central 512 samples, and applying a Tukey window with shape parameter 0.1 to assure the filters start and end at zero.

To account for time-dependency, filters were applied to 50% overlapping, 20-ms Hann-windowed blocks of the microphone signals. To facilitate real-time three-degrees-of-freedom rendering, the BFBR and BF+MLS beamformer outputs were encoded into fifth-order Ambisonics, and a fifth-order Ambisonic listening-experiment reference was derived from the 25th-order reference.

For the MLS method, multichannel filters were computed on a grid of relative orientations $R_{\text{rel}}$ with yaw angle $\alpha \in [-180^\circ, 180^\circ)$, pitch angle $\beta \in [-60^\circ, 60^\circ]$, and roll angle $\gamma \in [-45^\circ, 45^\circ]$, each spaced by $6^\circ$, which was found to be enough for filter switches to be inaudible in informal listening tests. The MLS filters were compiled into a custom Virtual Studio Technology (VST) plug-in allowing for real-time filter switching based on the array and listener orientations. As with other methods, the MLS plug-in also applied filters to 50 % overlapping Hann windows. To also evaluate the NBP method [3], microphone signals were processed using this method, resulting in fifth-order Ambisonic signals. The NBP method processing was done without any parameter optimization regarding the investigated array and scenarios. Additionally, a mono anchor was obtained from the omnidirectional channel of the Ambisonic reference. The loudness of the stimuli was manually matched by ear.

Stimuli were played back to the participants using AKG K702 headphones. The listener orientation was tracked in real time using an OptiTrack system and a reflecting rigid-body mounted onto the headphones. The REAPER digital audio workstation and the IEM Plug-in Suite[2] were utilized to process the fifth-order Ambisonic signals according to the listener orientation, to re-rotate the NBP Ambisonic signals according to the simulated dynamic array orientation, and to decode Ambisonic signals to binaural signals using KU100 MLS HRTFs.

Participants rated the stimuli in a Multiple Stimulus with Hidden Reference and Anchor experiment in which they were advised to dynamically rotate their head. In one experiment part, participants were asked to rate the similarity to the reference with respect to *perceived direction and distance of sources*. In a second part, participants were asked to rate the similarity with respect to *timbre and freedom from artifacts*. Similarity was rated on a scale ranging from 0 (very different) to 100 (identical). Each part consisted of 12 trials presenting the 12 scenarios. In each trial, eight stimuli were presented for comparison (hidden reference, FOA, BFBR, MLS, BF+MLS*, BF+MLS, NBP, mono anchor). The order of parts and trials was randomized.

Ratings were obtained from 16 participants, all of which had a background in working with sound (audio engineering or computer music). Participants were acquired from university staff and students.

## 5.2 Results

Medians and interquartile ranges of the ratings to all scenarios are shown in Fig. 4. The following sections proceed with analyses of variance and a detailed comparison of rendering methods in terms of subjective ratings.

Table 3. Results of analyses of variance. Main factors are rendering method (Method), reverberation condition (Rev.), rotated/static array (Rot.), and scenario (Scen.). Numerator degrees of freedom are denoted by DFn. Due to the repeated-measures design, factors have different error terms and denominator degrees of freedom (DFd). The $p$ values are reported for the tests on the spatial response ($p_{\text{Spat}}$), the timbral response ($p_{\text{Timb}}$), and the bivariate spatial-timbral analysis ($p_{\text{S.T.}}$). The $p$ values in bold indicate significance at the 95 % confidence level.

| Source | DFn | DFd | $p_{\text{Spat}}$ | $p_{\text{Timb}}$ | $p_{\text{S.T.}}$ |
|---|---|---|---|---|---|
| Method | 7 | 105 | $<$ **0.001** | $<$ **0.001** | $<$ **0.001** |
| Rev. | 1 | 15 | $<$ **0.001** | $<$ **0.001** | $<$ **0.001** |
| Rot. | 1 | 15 | **0.002** | **0.001** | $<$ **0.001** |
| Scen. | 2 | 30 | $<$ **0.001** | **0.019** | $<$ **0.001** |
| Method × Rev. | 7 | 105 | $<$ **0.001** | $<$ **0.001** | $<$ **0.001** |
| Method × Rot. | 7 | 105 | $<$ **0.001** | $<$ **0.001** | $<$ **0.001** |
| Method × Scen. | 14 | 210 | $<$ **0.001** | $<$ **0.001** | $<$ **0.001** |
| Rev × Rot. | 1 | 15 | 0.881 | 0.734 | 0.946 |
| Rev × Scen. | 2 | 30 | 0.608 | 0.371 | 0.676 |
| Rot × Scen. | 2 | 30 | 0.480 | 0.768 | 0.662 |

### 5.2.1 Analyses of Variance

To test for the effects of the experimental factors (rendering method, reverberation condition, static/rotated array, scenario), a four-way repeated-measures analysis of variance (RM-ANOVA) was conducted both for the spatial and timbral ratings. Additionally, a four-way repeated-measures multivariate analysis of variance (RM-MANOVA) for the combined two-dimensional rating was carried out. To this end, the *R* software was used. For the RM-MANOVA, $p$ values were computed using the Wilks' Lambda test statistic [32, pp. 81–82].

The results of the analyses of variance are shown in Table 3. For all three ANOVAs the same factors are significant. All main effects are significant at the 95% confidence level. Only two-way interaction effects are reported, because higher-order interactions are hard to interpret and are of little relevance to this study. All interactions involving the method factor are significant, which indicates that the relative performance of the different rendering methods is impacted by different experimental conditions (reverberation, static/rotated array and spatial configuration/choice of sound sources). Other two-way interactions are all non-significant.

### 5.2.2 Comparison of Rendering Methods

In nearly all scenarios, the FOA, BFBR, and NBP methods were rated lower than MLS, BF+MLS, and BF+MLS* with regard to both spatial and timbral similarity (cf. Fig. 4). In the *string_quartet* scenario, the BF+MLS method performs best in the anechoic scenario, while the MLS and BF-MLS* approaches respectively perform best in the reverberant scenario without and with dynamic array rotation both for spatial and timbral similarity. The MLS method performed clearly better in the reverberant condition than in the anechoic condition. The BF+MLS method shows

[2]http://plugins.iem.at/.

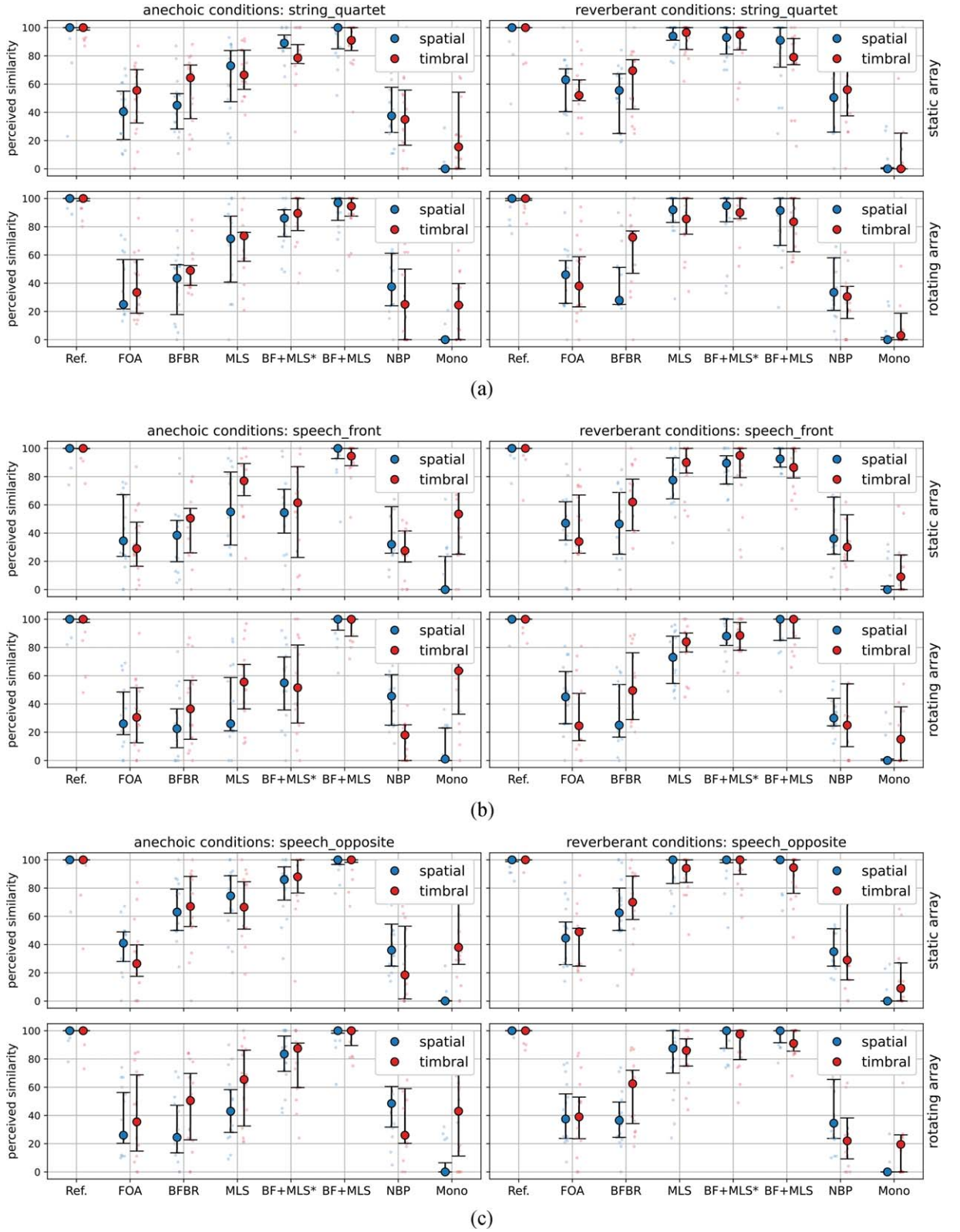



Fig. 4. Median and interquartile range for ratings of all spatial scenarios: *string_quartet* (a), *speech_front* (b), *speech_opposite* (c). The (hidden) reference is a fifth-order Ambisonic binaural rendering.

a small degradation in the reverberant conditions. In the *speech_front* scenario, the BF+MLS method performed best for both spatial and timbral similarity. The MLS method shows a large degradation with regard to spatial similarity in the anechoic dynamically rotating condition. Again, the MLS method performs better in the reverberated conditions. In the *speech_opposite* scenario, ratings are similar to the *speech_front* scenario. However, the degradation of the MLS method in the anechoic, dynamically rotated condition is smaller than in the *speech_front* scenario.

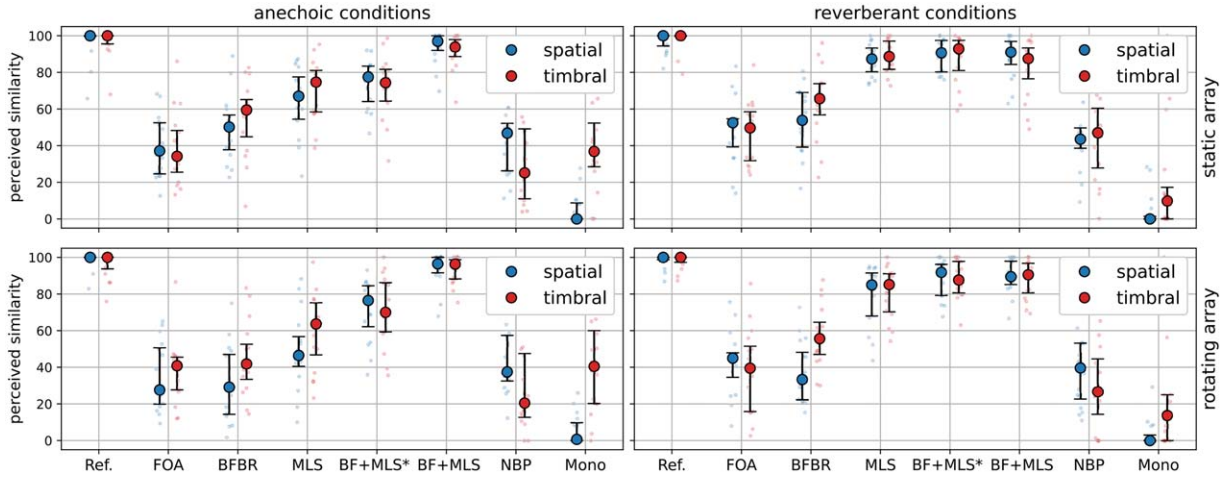



Fig. 5. Median and interquartile range for ratings pooled across spatial scenarios (*string_quartet*, *speech_front*, and *speech_opposite*).

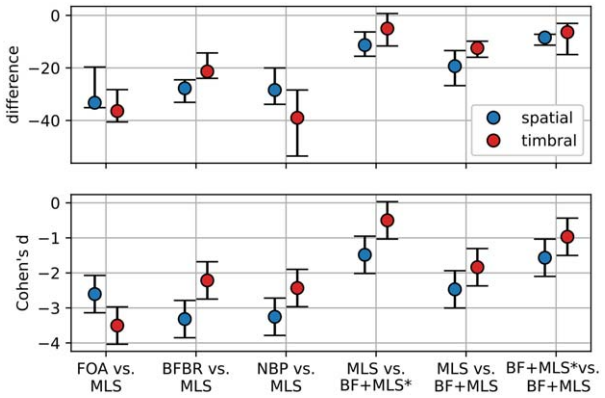



Fig. 6. Medians and interquartile ranges of rating differences averaged across spatial scenes and static/rotating array conditions (a). Cohen's $d$ along with 95% confidence intervals (b).

To highlight the general performance of the reproduction methods, ratings for each participant were averaged over the three spatial base scenarios (cf. Fig. 5). A performance ranking follows from the rated similarity regarding spatial and timbral aspects, which holds for all scenarios (except for some small deflections): The MLS method outperforms the FOA, BFBR, and NBP methods. The BF+MLS* method in turn outperforms the MLS and is itself outperformed by the BF+MLS method, which uses knowledge of all the source directions. The MLS performs worse for dynamic array orientation than for static array orientation and better for reverberant scenarios than for anechoic ones.

To statistically quantify the ranking differences, the participant ratings were further averaged over anechoic/reverberated and static/dynamic conditions, and per-participant differences for the pairs of methods that are of interest were computed. Fig. 6 shows difference distributions, Cohen's $d$ effect sizes, and their $t$-based 95% confidence intervals. If a confidence interval does not overlap with the zero line, a difference is significant at $P = 95\%$. All differences are significant except for the MLS vs. BF+MLS* with respect to timbre, verifying the ranking stated above.

## 6 DISCUSSION

The general perceptual ranking of methods described in Sec. 5.2 confirms what could be interpreted from the instrumental evaluation described in Sec. 4: The MLS method outperforms other scene-agnostic methods and performs worse with array rotation. This can be attributed to the unequal distribution of microphones (no microphones on the back of the head), which causes a distortion of horizontal localization cues for yaw rotations of the array. If this happens under time variation, disturbing "pseudo-movement" of sources can arise. The fact that the MLS shows more spatial degradation in the *speakers_front* than in the *speakers_opposite* scenario can be explained by the source directions: Participants were likely more sensitive to the "pseudo-movement" of sources directly in front of them, compared to sources that are further to the side or behind.

The MLS method performed better in the reverberant conditions than in the anechoic conditions in terms of both spatial and timbral similarity, also with rotated arrays. The good performance can be explained by the accurate reproduction of the diffuse field magnitude response, which is guaranteed by the constraint in Eq. (4). In low DDR conditions, the reverberation may make localization more difficult and conceal errors in spatial cues for the direct sound.

Steering beams at source directions and spatializing the beamformer output signals stabilizes and sharpens the spatial impression, especially (but not only) if all source directions are known. An exception to this general tendency is the *string_quartet* scenario, where the BF+MLS method performs slightly worse than MLS and BF+MLS* (cf. Fig. 4). This observation can be explained by the four-fold constraint on the beamformer with the BF+MLS method in this scenario, which might distort the reproduction of the diffuse reverberation residuum.

The NBP method aims to estimate source directions in each frequency bin and steers beamformers at these directions, which theoretically stabilizes and sharpens the spatial impression [33], but only if direction estimates are accu-

rate. However, for the semi-circular array investigated in this study, NBP outperform the other scene-agnostic methods. Informal listening tests showed that spatial and timbral degradations with the NBP method are noticeable mainly at frequencies above 2 kHz for the static array condition, while for the dynamically rotated array condition, errors are also noticeable at low frequencies. The degradations can potentially be attributed to errors in the DOA estimation, which is suggested by the analysis of the inherent localization robustness in Sec. 4.2 (columns 1 and 2).

In the study by McCormack et al. [27], the NBP method resulted in median overall quality ratings that were at least as high as for the MLS method, whereas in the present study, the MLS method significantly outperformed the NBP method in terms of both spatial and timbral scores (cf. Fig. 6). This discrepancy of results is possibly explained by the difference in the employed microphone array: The array in [27] was a circular seven-sensor array with microphones distributed on a mannequin head with glasses, whereas the array used in this study contains only five sensors and does not cover the back of the head. The beneficial effect of two additional microphones on the back of the head can be seen in Fig. 3 (column 3 vs. column 7).

It is important to note that the array transfer functions from the EasyCom dataset employed in this study seem to be inherently problematic regarding localization robustness, cf. Sec. 4.2. Future research should compare this array transfer function dataset to other measured head-mounted microphone arrays with respect to localization robustness.

Another possible explanation for the discrepancy between the results of the present study and the ones described in [27] is the diffuse-field equalization applied in the MLS and FOA methods in the present study. Such an equalization is not reported for the MLS method in [27]. Informal listening tests revealed that the diffuse-field equalization greatly improves the timbral similarity to the reference and left-right balance for the MLS method under reverberant conditions.

## 7 CONCLUSION

This study investigated dynamic binaural rendering of head-mounted microphone array signals in a three-degrees-of-freedom listening experiment. A sparse array of five sensors was considered, both for a static array orientation and a turning motion during signal recording. Two classes of algorithms were evaluated: 1) scene-agnostic algorithms, which operate without prior knowledge about source directions in the recorded scene, and 2) scene-aware algorithms, for which knowledge about sound source directions is assumed, e.g., due to user input in an AR/VR use case.

Among the scene-agnostic algorithms, the MLS rendering, which maps microphone array signals directly onto ear signals, clearly outperforms the other algorithms, namely the FOA, BFBR, and NBP upmixing. In the instrumental evaluation, the MLS rendering minimizes errors in the reproduction of ILD across direction and frequency.

Accurate reproduction is challenging for scenes recorded by a rotating head-mounted array, in which time-variant spatial and timbral mapping artifacts become audible. Results from this listening experiment revealed that the artifacts can be mitigated by steering of broadband distortionless-response beamformers (minimum variance distortionless response/LCMV) to obtain direct source signals of interest, which can then be encoded and reproduced with high spatial and timbral fidelity, while residual array signals are reproduced as faithfully as possible with a scene-agnostic approach like the MLS.

The code for rendering the stimuli, experiment result data, and analysis scripts are provided.[3] Static renders of the experiment stimuli can be listened to online.[4]

## 9 ACKNOWLEDGMENT

The authors thank Archontis Politis for processing the simulated microphone signals using the NBP method and Franz Zotter for fruitful discussions and input on the structuring of the paper.

## 10 REFERENCES

[1] B. Stahl and S. Riedel, "Perceptual Comparison of Dynamic Binaural Reproduction Methods for Head-Mounted Microphone Arrays," presented at the *155th Audio Engineering Society Convention* (2023 Oct.), paper 10672. http://aes.org/e-lib/browse.cfm?elib=22253.

[2] J. Donley, V. Tourbabin, J.-S. Lee, et al., "EasyCom: An Augmented Reality Dataset to Support Algorithms for Easy Communication in Noisy Environments," *arXiv preprint arXiv:2107.04174* (2021 Oct.). https://doi.org/10.48550/arXiv.2107.04174.

[3] L. McCormack, A. Politis, R. Gonzalez, T. Lokki, and V. Pulkki, "Parametric Ambisonic Encoding of Arbitrary Microphone Arrays," *IEEE Trans. Audio Speech Lang. Process.*, vol. 30, pp. 2062–2075 (2022 Jun.). https://doi.org/10.1109/TASLP.2022.3182857.

[4] T. Lübeck, S. V. Amengual Garí, P. Calamia, et al., "Perceptual Evaluation of Approaches for Binaural Reproduction of Non-Spherical Microphone Array Signals," *Front. Sig. Process.*, vol. 2, paper 883696 (2022 Aug.). https://doi.org/10.3389/frsip.2022.883696.

[5] Rafaely, Boaz, Tourbabin, Vladimir, Habets, Emanuel, et al., "Spatial Audio Signal Processing for Binaural Reproduction of Recorded Acoustic Scenes - Review and Challenges," *Acta Acust.*, vol. 6, paper 47 (2022 Oct.).https://doi.org/10.1051/aacus/2022040.

[6] J. Chen, B. D. Van Veen, and K. E. Hecox, "External Ear Transfer Function Modeling: A Beamforming Approach," *J. Acoust. Soc. Am.*, vol. 92, no. 4, pp. 1933–1944 (1992 Oct.). https://doi.org/10.1121/1.405241.

[7] M. A. Poletti and U. P. Svensson, "Beamforming Synthesis of Binaural Responses From Com-

[3] http://github.com/BenjSta/array2binaural.
[4] http://array2binaural_demo.iem.sh/.

puter Simulations of Acoustic Spaces," *J. Acoust. Soc. Am.*, vol. 124, no. 1, pp. 301–315 (2008 Jul.). https://doi.org/10.1121/1.2924206.

[8] E. Rasumow, M. Blau, M. Hansen, et al., "The Impact of the White Noise Gain (WNG) of a Virtual Artificial Head on the Appraisal of Binaural Sound Reproduction," in *Proceedings of the EAA Joint Symposium on Auralization and Ambisonics*, pp. 174–180 (Berlin, Germany) (2014 Apr.). https://doi.org/10.14279/depositonce-29.

[9] E. Rasumow, M. Blau, S. Doclo, et al., "Individualized Binaural Reproduction Using a Virtual Artificial Head," in *Proceedings of the 40th DAGA*, pp. 26–27 (Oldenburg, Germany) (2014 Mar.). http://pub.dega-akustik.de/DAGA_2014/data/articles/000296.pdf.

[10] E. Rasumow, M. Blau, M. Hansen, et al., "Smoothing Individual Head-Related Transfer Functions in the Frequency and Spatial Domains," *J. Acoust. Soc. Am.*, vol. 135, no. 4, pp. 2012–2025 (2014 Apr.). https://doi.org/10.1121/1.4867372.

[11] M. Fallahi, M. Hansen, S. Doclo, et al., "Dynamic Binaural Rendering: The Advantage of Virtual Artificial Heads over Conventional Ones for Localization With Speech Signals," *Appl. Sci.*, vol. 11, no. 15, paper 6793 (2021 Jul.). https://doi.org/10.3390/app11156793.

[12] L. Madmoni, J. Donley, V. Tourbabin, and B. Rafaely, "Beamforming-Based Binaural Reproduction by Matching of Binaural Signals," in *Proceedings of the AES International Conference Audio for Virtual and Augmented Reality* (2020 Aug.), paper 3-1. http://aes.org/e-lib/browse.cfm?elib=20878.

[13] T. Deppisch, H. Helmholz, and J. Ahrens, "End-to-End Magnitude Least Squares Binaural Rendering of Spherical Microphone Array Signals," in *Proceedings of the Immersive and 3D Audio: from Architecture to Automotive (I3DA)*, pp. 1–7 (Bologna, Italy) (2021 Sep.). https://doi.org/10.1109/I3DA48870.2021.9610864.

[14] H. Helmholz, T. Deppisch, and J. Ahrens, "End-to-End Magnitude Least Squares Binaural Rendering for Equatorial Microphone Arrays," in *Proceedings of the 48th DAGA*, pp. 1679–1682 (Hamburg, Germany) (2023 Mar.). http://pub.dega-akustik.de/DAGA_2023/data/articles/000052.pdf.

[15] C. Schörkhuber, M. Zaunschirm, and R. Höldrich, "Binaural Rendering of Ambisonic Signals via Magnitude Least Squares," in *Proceedings of the 44th DAGA*, pp. 339–342 (Munich, Germany) (2018 Mar.). http://pub.dega-akustik.de/DAGA_2018/data/articles/000301.pdf.

[16] F. Zotter and M. Frank, *Ambisonics: A Practical 3D Audio Theory for Recording, Studio Production, Sound Reinforcement, and Virtual Reality*, Springer Topics in Signal Processing, vol. 19 (Springer, Cham, Switzerland, 2019). https://doi.org/10.1007/978-3-030-17207-7.

[17] I. Ifergan and B. Rafaely, "On the Selection of the Number of Beamformers in Beamforming-Based Binaural Reproduction," *EURASIP J. Audio Speech Music Process.*, vol. 2022, no. 1, paper 6 (2022 Mar.). https://doi.org/10.1186/s13636-022-00238-7.

[18] V. Pulkki, "Spatial Sound Reproduction With Directional Audio Coding," *J. Audio Eng. Soc.*, vol. 55, no. 6, pp. 503–516 (2007 Jun.). http://aes.org/e-lib/browse.cfm?elib=14170.

[19] A. Politis, J. Vilkamo, and V. Pulkki, "Sector-Based Parametric Sound Field Reproduction in the Spherical Harmonic Domain," *IEEE J. Sel. Top. Signal Process.*, vol. 9, no. 5, pp. 852–866 (2015 Mar.). https://doi.org/10.1109/JSTSP.2015.2415762.

[20] A. Wabnitz, N. Epain, and C. T. Jin, "A Frequency-Domain Algorithm to Upscale Ambisonic Sound Scenes," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 385–388 (Kyoto, Japan) (2012 Aug.). https://doi.org/10.1109/ICASSP.2012.6287897.

[21] S. Berge and N. Barrett, "A New Method for B-Format to Binaural Transcoding," in *Proceedings of the 40th AES International Conference: Spatial Audio: Sense the Sound of Space* (2010 Oct.), paper 6-5. http://aes.org/e-lib/browse.cfm?elib=15527.

[22] S. Berge and N. Barrett, "High Angular Resolution Planewave Expansion," in *Proceedings of the 2nd International Symposium on Ambisonics and Spherical Acoustics*, pp. 1–6 (Paris, France) (2010 May). http://ambisonics10.ircam.fr/drupal/files/proceedings/presentations/O15_18.pdf.

[23] B. Jo, F. Zotter, and J.-W. Choi, "Extended Vector-Based EB-ESPRIT Method," *IEEE Trans. Audio Speech Lang. Process.*, vol. 28, pp. 1692–1705 (2020 May). https://doi.org/10.1109/TASLP.2020.2996090.

[24] A. Politis, S. Tervo, and V. Pulkki, "COMPASS: Coding and Multidirectional Parameterization of Ambisonic Sound Scenes," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6802–6806 (Calgary, Canada) (2018 Apr.). https://doi.org/10.1109/ICASSP.2018.8462608.

[25] Z. He, A. Cichocki, S. Xie, and K. Choi, "Detecting the Number of Clusters in n-Way Probabilistic Clustering," *IEEE Trans. Pattern Analysis Mach. Intell.*, vol. 32, no. 11, pp. 2006–2021 (2010 Nov.). https://doi.org/10.1109/TPAMI.2010.15.

[26] R. Schmidt, "Multiple Emitter Location and Signal Parameter Estimation," *IEEE Trans. Antennas Propag.*, vol. 34, no. 3, pp. 276–280 (1986 Mar.). https://doi.org/10.1109/TAP.1986.1143830.

[27] L. McCormack, N. Meyer-Kahlen, D. Lou Alon, et al., "Six-Degrees-of-Freedom Binaural Reproduction of Head-Worn Microphone Array Capture," *J. Audio Eng. Soc.*, vol. 71, no. 10, pp. 638–649 (2023 Oct.). http://aes.org/e-lib/browse.cfm?elib=22237.

[28] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A Consolidated Perspective on Multimicrophone Speech Enhancement and Source Separation," *IEEE Trans. Audio Speech Lang. Process.*, vol. 25, no. 4, pp. 692–730 (2017 Apr.). https://doi.org/10.1109/TASLP.2016.2647702.

[29] C. Huang, Y. Tian, A. Kumar, and C. Xu, "Egocentric Audio-Visual Object Localization," in *Proceedings of the IEEE/CVF Conference on*

*Computer Vision and Pattern Recognition (CVPR)*, pp. 22910–22921 (Vancouver, Canada) (2023 Jun.). https://doi.org/10.1109/CVPR52729.2023.02194.

[30] B. Bernschütz, "A Spherical Far Field HRIR/HRTF Compilation of the Neumann KU 100," *in Proceedings of the AIA-DAGA*, pp. 592–595 (Deutsche Gesellschaft für Akustik e.V., Merano, Italy) (2013 Mar.).

[31] O. C. Gomes, W. Lachenmayr, J. Thilakan, and M. Kob, "Anechoic Multi-Channel Recordings of Individual String Quartet Musicians," in *Proceedings of the International Conference Immersive and 3D Audio (I3DA)*, pp. 1–7 (Bologna, Italy) (2021 Sep.). https://doi.org/10.1109/I3DA48870.2021.9610912.

[32] K. Mardia, J. Kent, and J. Bibby, *Multivariate Analysis* (Academic Press, London, UK, 1979).

[33] B. Stahl and S. Riedel, "Perceptual Comparison of Parametric Spatial Audio Codings for Binaural Rendering," in *Proceedings of the AES International Conference Spatial and Immersive Audio* (2023 Aug.), paper 13. http://aes.org/e-lib/browse.cfm?elib=22205.

## THE AUTHORS



Benjamin Stahl



Stefan Riedel

Benjamin Stahl is a researcher and Ph.D. student at the Institute of Electronic Music and Acoustics (IEM) of the University of Music and Performing Arts Graz (KUG). He received his M.Sc. in electrical and audio engineering from TU Graz and KUG in 2018. His Ph.D. research focuses on multichannel signal processing and Deep Learning–based speech enhancement. In 2021, he was a research intern at Microsoft Applied Sciences Group in Munich. In 2023, his submission to the Speech Enhancement For Augmented Reality challenge was ranked best among all challenge submissions. At the 2023 AES International Conference on Spatial and Immersive Audio in Huddersfield, his paper on spatial audio coding (co-authored by Stefan Riedel) won the best paper award.

•

Stefan Riedel is a researcher and Ph.D. student at the Institute of Electronic Music and Acoustics (IEM) of the University of Music and Performing Arts Graz (KUG). He graduated in electrical and audio engineering from TU Graz and KUG with a Master's thesis on mixed-order spherical beamforming in 2019. The corresponding publication won the best paper award at the 2019 International Computer Music Conference in New York City. His current research focuses on the perception of envelopment in spatial sound reproduction with recent publications in *JASA Express Letters* and the *Journal of the Audio Engineering Society*. Further research interests include binaural rendering for augmented and virtual reality, virtual acoustics, and psychoacoustics.