L. McCormack, N. Meyer-Kahlen, D. L. Alon, Z. Ben-Hur,
S. V. Amengual Garí and P. Robinson, "Six-Degrees-of-Freedom Binaural
Reproduction of Head-Worn Microphone Array Capture"
J. Audio Eng. Soc., vol. 71, no. 10, pp. 638–649, (2023 October).
DOI: https://doi.org/10.17743/jaes.2022.0104

Six-Degrees-of-Freedom Binaural Reproduction of Head-Worn Microphone Array Capture

LEO MCCORMACK, *AES Student Member*, **NILS MEYER-KAHLEN,** *AES Student Member*,
(leo.t.mccormack@gmail.com) (nilsmk@meta.com)

DAVID LOU ALON, *AES Member*, **ZAMIR BEN-HUR,** *AES Associate Member*,
(davidalon@meta.com) (zamirbh@meta.com)

SEBASTIÀ V. AMENGUAL GARÍ, *AES Member* **AND PHILIP ROBINSON,** *AES Member*
(samengual@meta.com) (philrob22@meta.com)

Reality Labs Research, Meta, Redmond, WA

This article formulates and evaluates four different methods for six-degrees-of-freedom binaural reproduction of head-worn microphone array recordings, which may find application within future augmented reality contexts. Three of the explored methods are signal-independent, utilizing least-squares, magnitude least-squares, or plane wave decomposition–based solutions. Rotations and translations are realized by applying directional transformations to the employed spherical rendering or optimization grid. The fourth considered approach is a parametric signal-dependent alternative, which decomposes the array signals into directional and ambient components using beamformers. The directional components are then spatialized by applying binaural filters corresponding to the transformed directions, whereas the ambient sounds are reproduced using the magnitude least-squares solution. Formal perceptual studies were conducted, whereby test participants rated the perceived relative quality of the four binaural rendering methods being evaluated. Of the three signal-independent approaches, the magnitude least-squares solution was rated the highest. The parametric approach was then rated higher than the magnitude least-square solution when the listeners were permitted to move away from the recording point.

0 INTRODUCTION

The integration of multiple microphones into commercially available head-worn devices has invoked renewed interest into the development of suitable audio processing algorithms for application within augmented reality (AR) contexts. The objective is often to process the microphone array signals with low-latency and deliver them immediately to the user via headphones/ear pieces or, alternatively, to store them for future processing and playback. In the former use case, the intention is usually to enhance or augment the auditory experience of the listener. This may involve a reduction in background noise and improved speech intelligibility [1, 2], the preservation or modification of the perceived spatial properties of the scene [3, 4], or an extension of the listeners' hearing abilities beyond the audible range [5]. Whereas, for the other use case, these enhancements may also find application, but with less stringent latency constraints, while also permitting additional spatial modifications [6, 7] prior to reproducing the captured scene over the target playback setup. The topic of this article falls within this latter reproduction task for headphones, with the added goal of accounting for both the listener's head orientation and their position relative to the recording point, which is often collectively referred to as six-degrees-of-freedom (6DoF) rendering based on sound-field extrapolation [8].

Previous studies investigating the 6DoF binaural reproduction of microphone array recordings have, however, focused predominantly on the use of spherical microphone arrays (SMAs). Many existing rendering pipelines require the SMA signals to first be transformed into the spherical harmonic domain [9], which is often referred to as Ambisonic encoding [10, 11], and the subsequent mapping of the spherical harmonic/Ambisonic signals to the target playback setup is referred to as Ambisonic decoding [12, 13]. This Ambisonics framework is especially popular for the task of delivering binaural audio with three-degrees-of-freedom (3DoF) capability. This is because well-defined and efficient sound-field rotations may be achieved via a single broad-band matrix operation [14]. Subsequent binaural decoding solutions include those based upon the application of a plane wave decomposition (PWD) followed by binauralizing the plane wave signals [15, 13] (often referred to as virtual-loudspeaker–based decoding) or through a direct least-squares (LS) fitting of the spherical harmonic patterns to the binaural directivities [16]. The perceptually motivated magnitude LS (MagLS) optimization [17–19] is widely considered to represent the current state-of-the-art, signal-independent binaural decoding solution. 6DoF rendering support may then be incorporated into these decoding methods by treating the plane wave grid (or the optimization directional grid) as objects in space and applying the appropriate directional transformations and distance-dependent gain factors to account for a translated listener position [20, 21, 8, 22].

Directly adopting the aforementioned Ambisonics-based rendering solutions for head-worn microphone array recordings, however, may lead to two main problems. The first relates to the well-known perceptual limitations, which are inherently incurred when decoding lower-order Ambisonic signals; i.e., as one would likely expect to acquire when encoding head-worn arrays comprising relatively few microphones. These perceptual issues are largely a product of the spatial overlap of the signals delivered over the playback setup, with the resulting coherent spreading of directional and diffuse sounds leading to source localization ambiguities, timbral colorations, and a loss of perceived envelopment [23, 24]. These drawbacks have motivated the introduction of signal-dependent Ambisonic decoding alternatives, which have been shown to mitigate many of these issues [25–29]. Such decoders typically adopt a parametric sound-field model, conduct an acoustical analysis of the scene, and subsequently use this information to adaptively synthesize the target playback signals. 6DoF extensions to these parametric methods have also been explored and typically operate in a similar manner to their signal-independent counterparts. The exception being that rotations and translations are usually only applied to sound components that are analyzed as being directional, with other ambient/diffuse sounds left unchanged [30–33].

The second issue of concern relates to the Ambisonic encoding of irregularly shaped and physically larger arrays; such as those incorporated into the eyeglasses form-factor in particular. Although the parametric decoding methods described above may indeed improve the perceived spatial accuracy of lower-order Ambisonics material, they are unable to do so if the Ambisonic encoding scheme does not produce the correct Ambisonic patterns over a sufficiently wide frequency bandwidth. For example, a recent study involving a seven-sensor head-worn microphone array demonstrated poor encoding performance above 1.5 kHz, when using a conventional signal-independent encoding approach [34]. This has therefore motivated recent alternative encoding proposals, such as seeking to exploit the properties of an equatorial arrangement of sensors [35], or employing parametric/signal-dependent processing [36, 34], in order to help alleviate the encoding limitations of such arrays. However, it may be argued that circumventing any conversion into the intermediate Ambisonics format, and instead mapping the input array signals directly to the binaural channels, should represent the more optimal rendering strategy.

Examples of studies involving the direct binaural reproduction of head-worn arrays in the eyeglasses form-factor include the use of LS [37] or MagLS [38] solutions, which aim to fit the array directivities directly to the target binaural directivities. A general parametric spatial enhancement solution was also explored recently in [39], which demonstrated improved spatial accuracy over alternative signal-independent algorithms. However, as far as the present authors are aware, the only previous study involving the direct binaural rendering of head-worn microphone array recordings, while also accommodating for listener head movements, was conducted in [40]. Here, the study involved accounting for listener head-rotations around the z axis; i.e., with one degree of freedom. There currently exists no report of a study investigating the 6DoF (or 3DoF) direct rendering of head-worn array recordings.

Therefore, in this study, four different 6DoF binaural rendering methods were formulated and evaluated, which specifically target the use of a head-worn microphone array as input. Three of the methods are inspired by related Ambisonics literature and are signal-independent; achieving rotations and translations by applying directional transformations to the employed rendering/optimization grids. The fourth method is a parametric signal-dependent approach, which adopts the spatial analysis techniques employed recently in [34] and uses spatial filters to separate the input recording into directional and ambient components. The directional components are then rotated, translated, and spatialized as point-sources, whereas the ambient components are reproduced using one of the three signal-independent solutions. A seven-sensor head-worn microphone array is then described, which was used to record three sound scenes comprising different source stimuli. A multiple-stimulus listening test then followed, whereby test subjects compared the relative perceived rendering quality obtained using the four rendering methods under test.

This article is organized as follows. In SEC. 1, the three signal-independent approaches are formulated. Their 6DoF extensions are then presented in SEC. 2. The parametric alternative approach is described in SEC. 3. The test apparatus and methodology employed for the perceptual study is then detailed in SEC. 4, with the results and discussions provided in SEC. 5. The article is concluded in SEC. 6.

1 SIGNAL-INDEPENDENT BINAURAL RENDERING APPROACHES

It is first assumed that a sound-field $\mathbf{x}(t, f) \in \mathbb{C}^{Q \times 1}$ has been recorded using an array of Q microphones, which have been transformed into the time-frequency domain through either the application of a short-time Fourier transform (STFT) or a (near) perfect reconstruction filter-bank [41], in which t and f denote the time and frequency indices, respectively. The sound-field may then be modeled as a superposition of many plane waves, which are incident from a spherical grid of V directions as

$$\mathbf{x}(t, f) = \mathbf{A}(\mathbf{\Gamma}_V, f)\mathbf{z}(t, f), \tag{1}$$

where $\mathbf{z} \in \mathbb{C}^{V\times 1}$ are the plane wave signals and $\mathbf{A} \in \mathbb{C}^{Q\times V}$ are the respective array transfer functions (ATFs). Note that it is henceforth assumed that these ATFs are available for a dense spherical grid of directions, from which one may look up those corresponding to these V directions, $\mathbf{\Gamma}_V = [\boldsymbol{\gamma}_1, ..., \boldsymbol{\gamma}_V]$; where $\boldsymbol{\gamma}_v \in S^2$ is a unit-length Cartesian vector describing the direction of the vth plane wave. In practice, ATFs may be obtained via free-field measurements of the array in question or through simulations. Additionally, although the primary focus of the present study concerns the use of head-worn microphone arrays, it is noted that other arbitrary microphone array configurations may also be employed. This includes SMAs; in which case, the ATFs may also be obtained analytically [42, 43].

The microphone array signals may be linearly mapped to the binaural channels $\mathbf{y} \in \mathbb{C}^{2\times 1}$ as

$$\mathbf{y}(t, f) = \mathbf{M}(f)\mathbf{x}(t, f), \tag{2}$$

where $\mathbf{M} \in \mathbb{C}^{2\times Q}$ is an appropriate binaural mixing matrix.

1.1 Plane Wave Decomposition–Based Rendering

One option for computing the above binaural mixing matrix is to conduct a PWD of the microphone array signals and subsequently convolve these decomposed plane wave signals with the respective head-related transfer functions (HRTFs). This is often referred to as virtual-loudspeaker decoding in Ambisonics literature [13, 15] or beamforming-based reproduction when operating on the microphone array signals directly [44–46]. The approach may be formulated as

$$\mathbf{M}^{(\mathrm{PWD})}(f) = \mathbf{H}(\mathbf{\Gamma}_L, f)\mathbf{B}(\mathbf{\Gamma}_L, f), \tag{3}$$

where $\mathbf{B} \in \mathbb{C}^{L\times Q}$ is a beamforming matrix for $L \geq Q$ directions, $\mathbf{\Gamma}_L = [\boldsymbol{\gamma}_1, ..., \boldsymbol{\gamma}_L]$ and $\mathbf{H} \in \mathbb{C}^{2\times L}$ are HRTFs corresponding to those same directions.

In this work, an energy-preserving beamforming matrix was employed, which was originally formulated for Ambisonic decoding in [47], but has also been used in the space domain more recently in [34]. It is obtained by first applying a singular value decomposition to a matrix of ATFs, which correspond to these L directions, as

$$\mathbf{A}(\mathbf{\Gamma}_L, f) = \mathbf{U}(f)\mathbf{\Sigma}(f)\mathbf{V}^{\mathrm{H}}(f), \tag{4}$$

where $\mathbf{U} \in \mathbb{C}^{Q\times Q}$ and $\mathbf{V} \in \mathbb{C}^{L\times L}$ are matrices containing the left and right singular vectors, respectively, and $\mathbf{\Sigma} \in \mathbb{R}^{Q\times L}$ is a matrix containing the singular values along the main diagonal and zeros elsewhere.

The desired energy-preserving trait may then be obtained through the construction of the following unitary matrix

$$\mathbf{B}(\mathbf{\Gamma}_L, f) = \frac{1}{\sqrt{L}}\mathbf{V}(f)\mathbf{I}_{Q\times L}\mathbf{U}^{\mathrm{H}}(f), \tag{5}$$

where $\mathbf{I}_{M\times N}$ denotes constructing a $M \times N$ matrix containing ones on the main diagonal and zeros elsewhere.

1.2 Least-Squares–Based Approach

Alternatively, the intermediate plane wave representation used in the above approach may be bypassed, with a direct LS fitting of the microphone array directivities to the binaural directivities. The solution to this problem may be found as [37]

$$\mathbf{M}^{(\mathrm{LS})}(f) = \mathbf{H}(\mathbf{\Gamma}_V, f)\mathbf{W}\mathbf{A}^{\mathrm{H}}(\mathbf{\Gamma}_V, f)\Big(\mathbf{D}(f) + \lambda\mathbf{I}_{Q\times Q}\Big)^{-1}, \tag{6}$$

where $\mathbf{D} = \mathbf{A}\mathbf{W}\mathbf{A}^{\mathrm{H}} \in \mathbb{C}^{Q\times Q}$ is the diffuse coherence matrix for the array, $\lambda \geq 0$ is a regularization parameter, and $\mathbf{W} \in \mathbb{R}^{V\times V}$ is a diagonal matrix of integration weights to account for cases in which the common ATF and HRTF measurement grid is not uniform.

1.3 Magnitude Least-Squares–Based Approach

One popular perceptually motivated optimization to the above LS solution is to attempt to fit only to the HRTF magnitudes at higher frequencies, rather than fitting to both the magnitudes and phases [19]. This may be realized as

$$\mathbf{M}^{(\mathrm{MagLS})}(f) = \hat{\mathbf{H}}(\mathbf{\Gamma}_V, f)\mathbf{W}\mathbf{A}^{\mathrm{H}}(\mathbf{\Gamma}_V, f)\Big(\mathbf{D}(f) + \lambda\mathbf{I}_{Q\times Q}\Big)^{-1}, \tag{7}$$

which resembles the LS solution given by Eq. (6), except for an additional phase modification, which is applied above a certain frequency threshold f_{c}

$$\hat{\mathbf{H}}(\mathbf{\Gamma}_V, f) = \begin{cases} \mathbf{H}(\mathbf{\Gamma}_V, f), & f < f_{\mathrm{c}} \\ |\mathbf{H}(\mathbf{\Gamma}_V, f)|e^{i\ \mathbf{\Phi}(f)}, & f \geq f_{\mathrm{c}}, \end{cases} \tag{8}$$

where $|.|$ denotes obtaining the (element-wise) absolute values of the enclosed matrix values and $i = \sqrt{-1}$ is the imaginary unit.

Ideally, one would like to find an optimal phase $\mathbf{\Phi} \in \mathbb{R}^{2\times V}$ such that the squared errors between the magnitudes of the HRTF patterns and the reconstructed HRTF patterns are minimized. However, this magnitude least squares minimization is a non-convex problem and no closed-form solution exists. As an alternative to gradient-based optimization or approaches using semidefinite relaxation, in [13], a simple algorithm was proposed, which still yields an improved magnitude fit between the HRTFs and reconstructed HRTFs in the spherical harmonic domain. In this study, the same algorithm was applied, except in the space domain; i.e., instead using the ATFs as the basis. Here, $\mathbf{\Phi}$ is set to the phase response of the reconstructed HRTFs for the previous frequency index, which is obtained as [13]

$$\mathbf{\Phi}(f) = \angle[\mathbf{M}^{(\mathrm{MagLS})}(f-1)\mathbf{A}(\mathbf{\Gamma}_V, f-1)], \tag{9}$$

where $\angle[.]$ denotes obtaining the (element-wise) phase values of the enclosed matrix values. Note that the threshold, f_{c}, is typically set to around 1.5 kHz, which is inspired by the duplex theory [48]. Only above this threshold, the above solution aims to favor obtaining a better magnitude fit between the ATF and HRTF directivities, thereby yielding reduced interaural level difference (ILD) reconstruction errors (and higher interaural phase difference errors) at frequencies in which ILD cues are more important for source

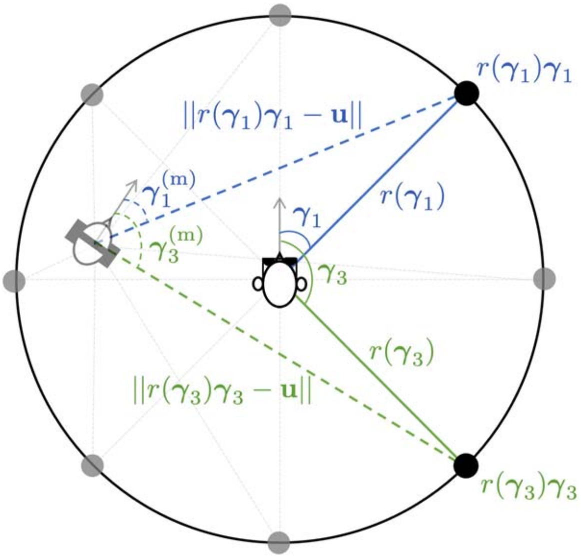

Fig. 1. An example of manipulating the plane wave directions (or grid optimization directions) based on a translated and rotated listener. The explicit mathematical notation describing two of the transformed directions are also included.

localization. An alternative to the above could also be to set $\boldsymbol{\Phi}$ to be linear phase [49], which is also expected to yield a better perceptual result compared to no modification.

2 SIX-DEGREES-OF-FREEDOM EXTENSIONS

The three signal-independent reproduction methods described in the previous section are now adapted for 6DoF rendering. This is realized by manipulating the plane wave directions, $\boldsymbol{\Gamma}_L^{(m)} = [\boldsymbol{\gamma}_1^{(m)}, ..., \boldsymbol{\gamma}_L^{(m)}]$, at each time frame, in order to account for the current position and head orientation of the listener. By projecting all of the plane waves onto a suitable distance map, which describes the source distances for all possible directions from the perspective of the recording point, these modified directions may be found as [30]

$$\boldsymbol{\gamma}_l^{(m)} = \mathbf{R}(\mathbf{o}) \frac{r(\boldsymbol{\gamma}_l)\boldsymbol{\gamma}_l - \mathbf{u}}{||r(\boldsymbol{\gamma}_l)\boldsymbol{\gamma}_l - \mathbf{u}||}, \quad \text{for } l = 1, ..., L, \tag{10}$$

where $||.||$ denotes taking the Euclidean norm of the enclosed vector, $\mathbf{u} \in \mathbb{R}^{3\times1}$ are Cartesian coordinates describing the current position of the listener, r denotes the source distance for the given plane wave direction, and $\mathbf{R} \in \mathbb{R}^{3\times3}$ is a rotation matrix to account for the orientation $\mathbf{o} = [\alpha, \beta, \omega]$ of the listener's head, given the yaw (α), pitch (β), and roll (ω) Euler angles. These trigonometric operations are depicted in Fig. 1. For the present study, the source distances are assumed to be the same for all possible source directions; i.e., employing a spherical distance map. Note that Eq. (10) also assumes that the listeners' head is static during the recording. Provided that the listeners' head orientation and position are tracked during the recording, anti-rotation and anti-translation operations may also be included at this stage; however, investigating the perceptual implications of this was beyond the scope of the present study. The time and frequency indices are also henceforth omitted for the brevity of notation.

The PWD-based binaural rendering method may now be rewritten for 6DoF playback as

$$\tilde{\mathbf{M}}^{(\text{PWD})} = \mathbf{H}(\boldsymbol{\Gamma}_L^{(m)})\mathbf{G}_L\mathbf{B}(\boldsymbol{\Gamma}_L), \tag{11}$$

where $\mathbf{G}_L = \text{diag}[g_1, ..., g_L] \in \mathbb{R}^{L\times L}$ is a diagonal matrix of gains accounting for the broad-band inverse-distance law and are calculated as

$$g_l = \min\left(g_{\max}, \frac{||r(\boldsymbol{\gamma}_l)\boldsymbol{\gamma}_l||}{||r(\boldsymbol{\gamma}_l)\boldsymbol{\gamma}_l - \mathbf{u}||}\right), \quad \text{for } l = 1, ..., L, \tag{12}$$

where $g_{\max}$ is a hard threshold on the maximum gain amplification to allow. Note that, because of the employed frequency-dependent nature of the rendering framework, near-field/proximity effects [50, 51] and source directivity may also be incorporated into the method at this point. However, only the inverse-distance law was adopted for the present study.

In a similar manner, but rather using the (typically) denser optimization grid of V directions, Eqs. (10) and (12) may be used to obtain $\boldsymbol{\Gamma}_V^{(m)} = [\boldsymbol{\gamma}_1^{(m)}, ..., \boldsymbol{\gamma}_V^{(m)}]$ and $\mathbf{G}_V \in \mathbb{R}^{V\times V}$, respectively. This then facilitates the incorporation of 6DoF capabilities into the LS-based approach,

$$\tilde{\mathbf{M}}^{(\text{LS})} = \mathbf{H}(\boldsymbol{\Gamma}_V^{(m)})\mathbf{G}_V\mathbf{W}\mathbf{A}^{\text{H}}(\boldsymbol{\Gamma}_V)\Big(\mathbf{D} + \lambda\mathbf{I}_{Q\times Q}\Big)^{-1}, \tag{13}$$

and also for the perceptually motivated MagLS variant as

$$\tilde{\mathbf{M}}^{(\text{MagLS})} = \hat{\mathbf{H}}(\boldsymbol{\Gamma}_V^{(m)})\mathbf{G}_V\mathbf{W}\mathbf{A}^{\text{H}}(\boldsymbol{\Gamma}_V)\Big(\mathbf{D} + \lambda\mathbf{I}_{Q\times Q}\Big)^{-1}, \tag{14}$$

which uses the same modified HRTFs as described by Eq. (8), except propagating $\tilde{\mathbf{M}}^{(\text{MagLS})}(f-1)$ for frequencies above f_c in Eq. (9).

3 PARAMETRIC BINAURAL RENDERING APPROACH

The signal-dependent rendering approach considered in the present study relies on a parametric sound-field model. In this instance, the array signals are modeled as a superposition of $K \ll Q$ source signals $\mathbf{s} \in \mathbb{C}^{K\times1}$ and (potentially anisotropic) diffuse sounds $\mathbf{d} \in \mathbb{C}^{Q\times1}$, which may be expressed as

$$\mathbf{x} = \mathbf{A}_{\text{s}}\big(\boldsymbol{\Gamma}_K\big)\mathbf{s} + \mathbf{d}, \tag{15}$$

where $\mathbf{A}_{\text{s}} \in \mathbb{C}^{Q\times K}$ are the array steering vectors corresponding to the source directions $\boldsymbol{\Gamma}_K = [\boldsymbol{\gamma}_1, ..., \boldsymbol{\gamma}_K]$.

The array spatial covariance matrices (SCMs) may be modeled as

$$\mathbf{C}_{\text{x}} = \mathbb{E}[\mathbf{x}\mathbf{x}^{\text{H}}] = \mathbf{A}_{\text{s}}\big(\boldsymbol{\Gamma}_K\big)\mathbf{C}_{\text{s}}\mathbf{A}_{\text{s}}^{\text{H}}\big(\boldsymbol{\Gamma}_K\big) + \mathbf{C}_{\text{d}}, \tag{16}$$

where $\mathbb{E}[.]$ denotes the expectation operator, which is typically achieved via temporal averaging in the range of tens of milliseconds; $\mathbf{C}_{\text{s}} = \mathbb{E}[\mathbf{s}\mathbf{s}^{\text{H}}] \in \mathbb{C}^{K\times K}$ is the SCM of the source signal(s); and $\mathbf{C}_{\text{d}} \in \mathbb{C}^{Q\times Q}$ is the SCM of the array, given the capture of diffuse sounds.

In order to account for potential anisotropic energy distributions of diffuse sounds in the captured scene, the diffuse array signal vector is modeled as the superposition of many plane waves $\mathbf{z} \in \mathbb{C}^{V \times 1}$, which are incident from V directions, as $\mathbf{d} = \mathbf{A}\mathbf{z}$, in a similar manner as Eq. (1). The array SCM, given the capture of only diffuse sounds, is therefore expressed as

$$\mathbf{C}_{\mathrm{d}} = \mathbb{E}[\mathbf{d}\mathbf{d}^{\mathrm{H}}] = \mathbf{A}\mathbf{C}_{\mathrm{z}}\mathbf{A}^{\mathrm{H}}, \tag{17}$$

where $\mathbf{C}_{\mathrm{z}} = \mathbb{E}[\mathbf{z}\mathbf{z}^{\mathrm{H}}] \in \mathbb{C}^{V \times V}$ is the SCM for these plane waves. Note that if the array captures an isotropic diffuse field, then this SCM becomes $\mathbf{C}_{\mathrm{d}}^{(\mathrm{isotropic})} = P_{\mathrm{d}}\mathbf{D}$, in which $P_{\mathrm{d}} = \mathrm{trace}[\mathbf{C}_{\mathrm{z}}]$ is the total energy of diffuse components.

3.1 Rendering Source Components

It is henceforth assumed that appropriate spatial analysis techniques have been applied to the input head-worn microphone array signals, in order to obtain estimates of the source number K and their respective directions of arrival (DoAs) $\mathbf{\Gamma}_K$ over time and frequency. In the present study, the same techniques as described in [34] were used; i.e., Second ORder sTatistic of Eigenvalues estimator (SORTE) for the source number estimation and Multple-Signal Classiciation (MUSIC) for the DoA estimation.

With this information at hand, source signal estimates may be obtained with the following

$$\mathbf{s} = \mathbf{B}_{\mathrm{s}}(\mathbf{\Gamma}_K)\mathbf{x}, \tag{18}$$

where $\mathbf{B}_{\mathrm{s}} \in \mathbb{C}^{K \times Q}$ is a matrix of beamforming weights for the estimated DoAs.

In this work, linearly constrained minimum-power beamformers were selected [52], with the beamforming weights computed as [34]

$$\begin{aligned}\mathbf{B}_{\mathrm{s}}(\mathbf{\Gamma}_K) = \big(&\mathbf{A}_{\mathrm{s}}^{\mathrm{H}}(\mathbf{\Gamma}_K)(\mathbf{C}_{\mathrm{x}} + \zeta\mathbf{I}_{Q \times Q})^{-1}\mathbf{A}_{\mathrm{s}}(\mathbf{\Gamma}_K)\big)^{-1} \\ &\mathbf{A}_{\mathrm{s}}^{\mathrm{H}}(\mathbf{\Gamma}_K)(\mathbf{C}_{\mathrm{x}} + \zeta\mathbf{I}_{Q \times Q})^{-1},\end{aligned} \tag{19}$$

where $\zeta \geq 0$ is a regularization parameter.

The binaural signals corresponding to these source components may then be obtained as

$$\mathbf{y}_{\mathrm{s}} = \mathbf{H}_{\mathrm{s}}\big(\mathbf{\Gamma}_K^{(\mathrm{m})}\big)\mathbf{G}_K\mathbf{s}, \tag{20}$$

where $\mathbf{H}_{\mathrm{s}} \in \mathbb{C}^{2 \times K}$ is a matrix of HRTFs for the potentially translated and/or rotated listener, given the directions $\mathbf{\Gamma}_K^{(\mathrm{m})} = [\boldsymbol{\gamma}_1^{(\mathrm{m})}, ..., \boldsymbol{\gamma}_K^{(\mathrm{m})}]$, which are obtained similarly as in Eq. (10), and $\mathbf{G}_K = \mathrm{diag}[g_1, ..., g_K] \in \mathbb{R}^{K \times K}$ is a diagonal matrix of source-dependent gains to account for the inverse-distance law, as described by Eq. (12).

3.2 Rendering Ambient Components

The estimated source components are then spatially subtracted from the input array recording, in order to obtain an estimate of the ambient array signals, $\mathbf{d}$, which should ideally encapsulate only diffuse reverberation and weakly directional sounds. This has been conducted previously based on Ambisonic signals in [27] and also more recently in the space domain in [34]. It may be formulated as

$$\mathbf{d} = [\mathbf{I}_{Q \times Q} - \mathbf{A}_{\mathrm{s}}(\mathbf{\Gamma}_K)\mathbf{B}_{\mathrm{s}}(\mathbf{\Gamma}_K)]\mathbf{x}. \tag{21}$$

The ambient binaural signals, $\mathbf{y}_{\mathrm{d}} \in \mathbb{C}^{2 \times 1}$ may then be obtained by reproducing $\mathbf{d}$ using one of the three signal-independent methods described in Sec. 1 or, for example, using the diffuse rendering strategy detailed in [34]. In this study, the MagLS approach was selected for this task,

$$\mathbf{y}_{\mathrm{d}} = \tilde{\mathbf{M}}^{(\mathrm{MagLS})}\mathbf{d}. \tag{22}$$

3.3 Overall Rendering

The final output signals may then be obtained as

$$\mathbf{y}_{\mathrm{par}} = \mathbf{y}_{\mathrm{s}} + \mathbf{y}_{\mathrm{d}}. \tag{23}$$

Note that if $K = 0$, then the proposed parametric rendering would revert to the signal-independent MagLS approach, since $\mathbf{s}$ would become undefined and $\mathbf{d} = \mathbf{x}$ during such cases. The presented parametric rendering framework, as configured for the current study, may therefore be viewed as a spatial sharpening method, whereby sounds that are analyzed as emanating from directional sound sources are isolated by the beamformers and subsequently collapsed into pin-point directions on the sphere. This type of rendering has previously been shown to improve the perceived spatial accuracy, when compared with signal-independent alternatives, in a number of perceptual studies [27, 39, 34, 53].

However, two aspects of the rendering may give rise to potentially audible artefacts in the present context, namely: 1) as the listener moves closer to a sound source, any time-varying and frequency-varying angular errors incurred during the DoA estimation will become exaggerated during the selection of the appropriate HRTFs, and 2) by addressing the inverse-distance law, any potentially unstable or misidentified sound source will now also become louder as the listener approaches it. Therefore, an additional balancing parameter may be included $\delta \in [0, .., 1]$, which allows for a trade-off to be made between rendering the signals using a less spatially articulate (but inherently stable) signal-independent approach, and the spatially sharper (but potentially less stable) parametric approach as

$$\mathbf{y}_{\mathrm{overall}} = \delta\mathbf{y}_{\mathrm{par}} + (1 - \delta)\mathbf{y}. \tag{24}$$

4 EVALUATION

The evaluation conducted in this study sought to assess the relative perceived quality of the four 6DoF binaural rendering methods under test; i.e., the three signal-independent approaches described in Secs. 1 and 2 and the parametric approach described in Sec. 3. The authors hypothesized that the MagLS approach would perform better than the other two signal-independent approaches. The rationale being that, by prioritizing the fitting to the binaural magnitudes at high frequencies (rather than to both magnitudes and phases), the method should result in a perceptually more accurate sound image and also incur reduced timbral colorations. The authors also postulated that the parametric approach would perform better than MagLS, since it should lead to the sharpest rendering of directional sounds in the scene.

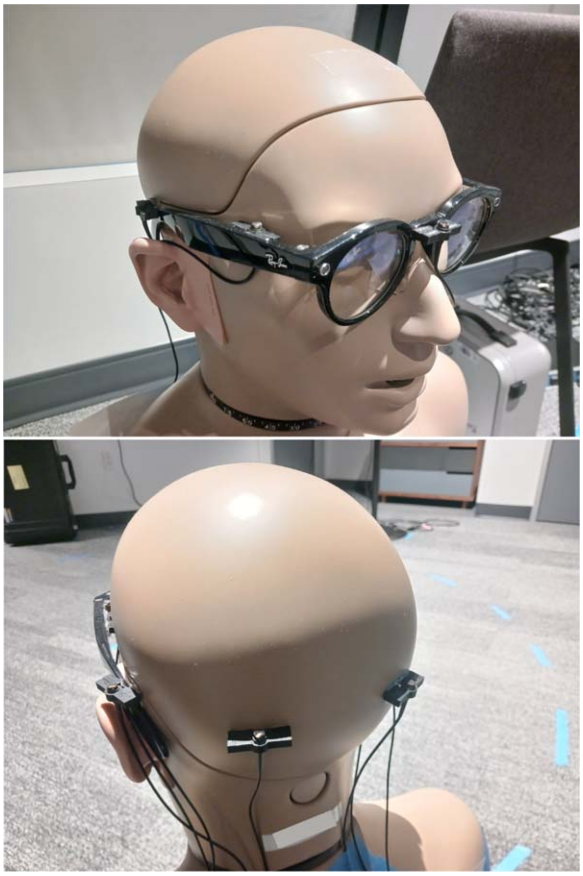

Fig. 2. Photos of the employed head-worn microphone array, which was affixed to a KEMAR head-and-torso simulator.

4.1 Test Apparatus

The microphone array selected for this study comprised seven DPA 4061-OC-C-B00 omni-directional microphones, which were distributed approximately uniformly around the circumference of a pair of eyeglasses worn by a KEMAR head-and-torso simulator. Such a configuration has previously been used, for example, in [35], and is depicted in Fig. 2. A total of 2,699 free-field ATF measurements were made using a grid approximately conforming to a Lebedev grid of order 47, obtained after quantizing the measurements to the nearest points in the Lebedev grid, and subsequently eliminating any duplicates and points below –69° elevation. Whereas, the HRTFs of the KEMAR were measured for a total of 1,625 directions, approximately corresponding to a Lebedev grid of order 35, obtained using the same elimination scheme as with the ATFs.

The KEMAR was then placed approximately in the middle of an apartment room, with the center of its head 1.80 m from the floor. The reverberation time (RT60) of the room was [0.56,0.72,0.80,0.71,0.54,0.42] s in octave bands from 125 Hz to 4 kHz, and the background noise level was $L_{\mathrm{A,\ eq,\ 30s}} = 34.4$ dB SPL(A); i.e., a room that has received minimal acoustic treatment. Two Genelec 8331A coaxial loudspeakers at a height of 1.65 m were then placed 1.83 m away from the KEMAR. These two loudspeakers were placed with 71.4° of angular separation on the horizontal plane, as depicted in Fig. 3. The test participants were then given Mysphere 3 headphones to wear for the full duration of the test. Note that these headphones have been shown to be among the most acoustically transparent, commercially available headphones [54]. The position and orientation of the headphones were tracked using an OptiTrack system comprising five PRIME 41/17W sensors (240 Hz), which were mounted to the ceiling.

4.2 Implementation of the Rendering Approaches

All four 6DoF rendering approaches were implemented into a real-time Virtual Studio Technology (VST) audio plugin.[1] The alias-free STFT [41] was employed as the time-frequency transform, with a window size of $5.\dot{3}$ ms (256 samples at 48 kHz), and a hopsize of $2.\dot{6}$ ms (128 samples at 48 kHz). The spatial analysis and rendering matrices were recomputed for every window. A one-pole filter, with a coefficient value of 0.3, was used to recursively average the array SCMs. The PWD approach employed a grid of $L = 24$ directions, corresponding to a minimum t-design of degree 6. Whereas, the LS and MagLS approaches employed a grid of $V = 240$ directions, corresponding to a minimum t-design of degree 21, and used $\lambda = 0.01$. At run-time, after applying the directional transformation described by Eq. (10), the transformed L and V grids were quantized to the nearest directions available in the employed HRTF grid. Note that the loudspeaker distances from the recording point were also informed to the plugin at this stage by specifying a spherical distance map of radius $r =$ 1.83 m. A maximum gain amplification of $g_{\mathrm{max}} = 8$ was also imposed (≈18 dB).

For the parametric approach, the spatial analysis was conducted as described in [34], and the linearly constrained minimum-power beamformers used $\zeta = 0.1$. In order to improve the perceived robustness of the reproduction, a rendering balance of $\delta = 0.85$ was empirically selected, since this was deemed by the authors to represent a reasonable trade-off between the spatial sharpening of directional sounds and perceived image stability. Furthermore, to reduce the computational complexity of the rendering system, while maintaining the low latency of the present STFT configuration, all rendering methods transitioned into the computationally efficient PWD approach for frequencies above 10 kHz. This was deemed by the authors to not significantly affect the rendered output, since spatial aliasing likely occurs well below this limit, and thus, all methods were spatially ambiguous above 10 kHz.

4.3 Test Design and Methodology

The developed real-time plugin was hosted by the MaxMSP (Cycling '74) software program. Three sets of contrasting source material were selected to be played out of the two loudspeakers (listed in the order of the left and

[1] Note that open-source software implementations of the four explored reproduction methods may be found here: https://github.com/facebookresearch/6DoF-Auraliser.

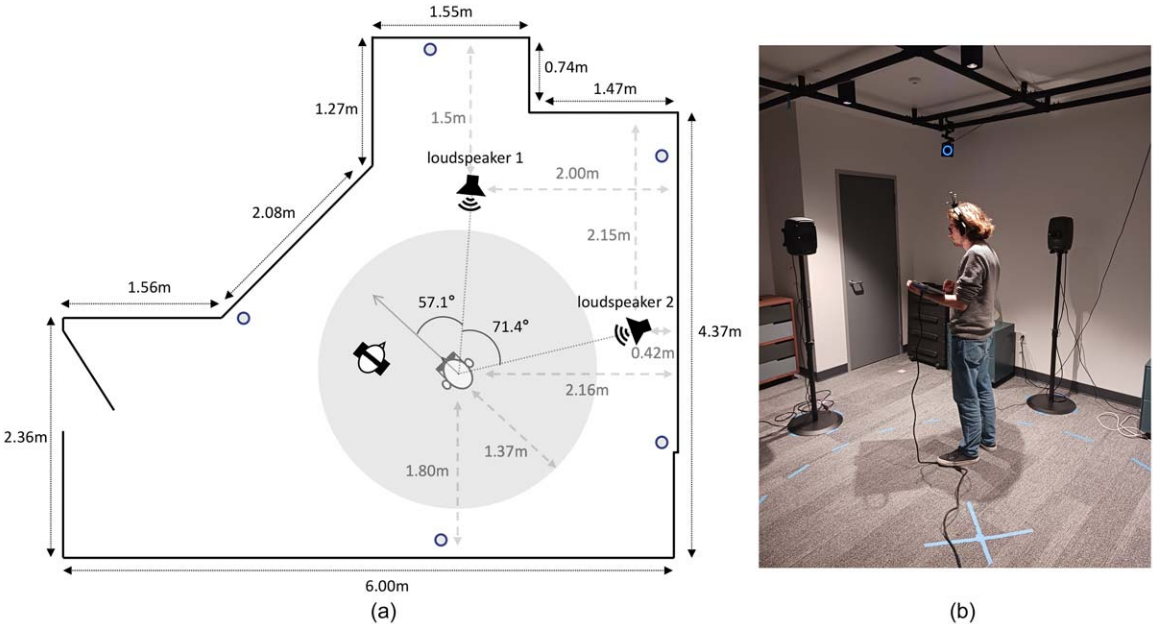

Fig. 3. (a) Illustration of the apartment room dimensions from the top-down (note that the five OptiTrack sensor positions are depicted as small circle icons). (b) A photo of the setup, with a listener located within the permitted navigable area marked on the floor, wearing the Mysphere 3 headphones and holding the tablet hosting the test interface.

then right loudspeaker): a female singer and an acoustic guitar (*music1*), a broad-band percussive shaker and synthesized strings (*music2*), and a female English speaker and a male Danish speaker reciting nonsensical phrases simultaneously (*speech*). The three scenes were then recorded by the head-worn microphone array in question and passed onto the developed plugin, which was informed of the listener head orientation and position at run time.

During the perceptual study, direct comparisons between the real scene and the four binaural rendering methods under test were not made possible. Rather, the stimuli were first played through the two loudspeakers located in the room. The participants then waited for 15 s before being able to experience and judge the quality of the four binaural reproduction methods. Therefore, the test design aims to more faithfully recreate the primary scenario in which these methods would be applied in practice; whereby listeners would experience previously recorded sound scenes at a later date. By including this pause, the intention is for the listeners to be less able to discern between smaller differences between the rendering methods and the real-world reference, such as slight coloration changes and localization shifts, and instead focus more on the perceived reproduction accuracy and quality between the methods themselves. Theories of cognition suggest there exists a short-term auditory sensory memory, which rapidly decays once signals are no longer available to be perceived by the listener [55, 56]. This auditory sensory memory is thought to decay within a time span in the range of 10 s [57], hence, motivating the selection of the 15 s pause.

The test participants were provided with a computer tablet, which allowed them to play the stimuli through the two loudspeakers. After 15 s had elapsed, they were then able to select, listen, and then rate the four test conditions based on their perceived spatial and timbral quality. Note that the user interface included a slider for each of the four conditions, as is commonly the case in multiple stimulus comparison tests. The interface also displayed the verbal anchors, “Bad,” “Poor,” “Fair,” “Good,” and “Excellent,” in 20-point increments next to these four sliders. The listening experiment was then conducted twice. In the first round, the listeners were restricted to the center of the setup (i.e., as a 3DoF control case), and in the second round, the listeners were able to move 1.37 m from the center (i.e., the 6DoF case). The permitted navigable area was marked clearly on the floor of the room, and it is also illustrated in Fig. 3. The three sets of test stimuli were presented to the listeners twice (i.e., one repetition) in randomized order. The listeners were also naive as to the rendering methods under test. Finally, in order to reduce audible sensor noise, all output audio was high-passed filtered using a fourth-order IIR filter (24 dB/octave) with a cutoff frequency of 200 Hz.

5 RESULTS AND DISCUSSION

The results based on 16 participants are presented in Fig. 4. The results for the 3DoF test case are shown in Fig. 4(a), and the results for the 6DoF case are shown in Fig. 4(b). Note that the medians are depicted with white dots, and the 95% confidence intervals are shown with black lines. The 32 individual data points for each combination of stimuli and method are depicted as colored dots.

It can be observed that for both parts of the perceptual study, and for all three sets of stimuli, the median scores for the MagLS approach are higher than both the PWD and regular LS approaches. Additionally, the median scores for the parametric approach were all higher than the median scores for the MagLS approach; however, these results were

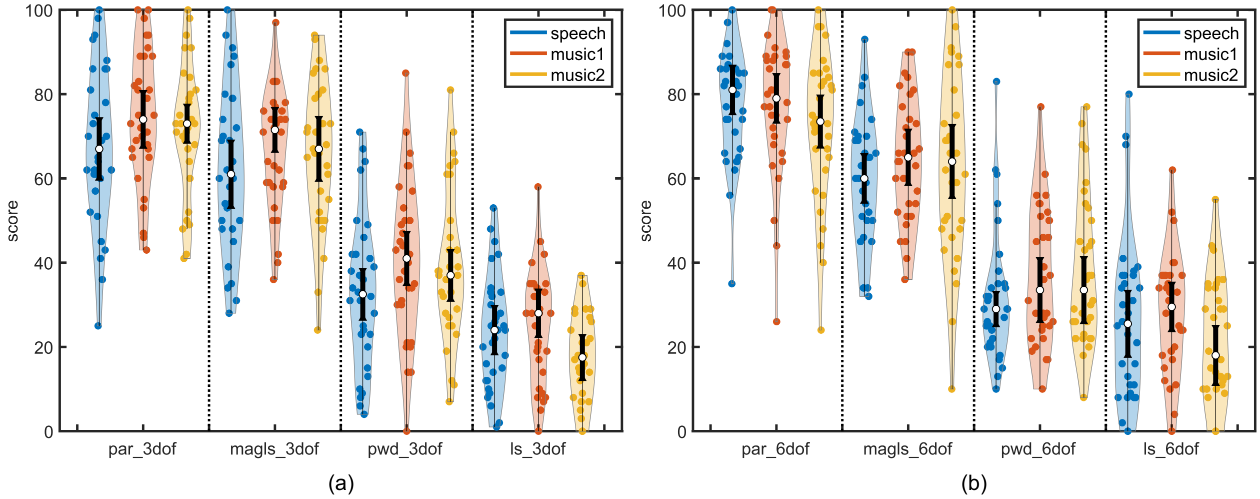

Fig. 4. Medians and 95% confidence intervals for the 3DoF perceptual study (a) and for the 6DoF perceptual study (b).

notably closer for the 3DoF case. Since not all the data conformed to a normal distribution, hypothesis testing was conducted using the non-parametric Wilcoxon signed rank test. The p values were corrected for multiple testing using the Bonferroni-Holm procedure. For the 3DoF part of the study, the differences between MagLS and LS, as well as MagLS and PWD, were found to be highly significant for all scenes ($p < 0.001$), whereas the differences between the parametric approach and MagLS were not found to be significant ($p > 0.1$).

For the 6DoF part of the study, the differences between MagLS and PWD, and between MagLS and LS, remained large and were found to be significant ($p < 0.001$). Notable differences in the median scores between the parametric and MagLS approaches were also found for this part of the study. For the speech stimuli, the difference of 21 points in the median rating was found to be significant ($p < 0.001$). For the first set of musical stimuli (*music1*), the difference of 14 points in the median rating was also found to be significant ($p = 0.002$). Whereas, for the second set of musical stimuli (*music2*), the difference in the medians was 9.5 points, but a 5% confidence level was not reached after correction for multiple comparisons ($p = 0.1$).

The results therefore clearly demonstrate an advantage when using the MagLS approach over the other two signal-independent approaches. These results also somewhat align with the results of previous perceptual studies, which were instead conducted within the context of binaural Ambisonics decoding [17, 58]. Due to the better fitting of the array directivities to the magnitudes of the HRTFs, the MagLS approach may be expected to inherently achieve reduced ILD reconstruction errors and thus may sound spatially more accurate. For the same reason, signal colorations may also be minimized compared to the other two signal-independent approaches under test. Therefore, the listener's scores for the MagLS approach may have also been influenced by a preference for improved timbral accuracy.

The parametric approach was found to receive the strongest advantage over the MagLS approach for the 6DoF part of the perceptual study. Whereas, for the 3DoF part of the study, the MagLS approach received similar scores to the parametric approach. The present authors postulate that this is likely due to both approaches being timbrally similar. Whereas, although improvements in spatial resolution were identifiable by the present authors also for the parametric 3DoF case, it is possible that such improvements only became apparent to the test participants when they were permitted to move closer to the sound sources. Nevertheless, both experiments suggest that the parametric approach was largely free of audible artefacts, when compared to signal-independent approaches, since such artefacts would have likely negatively impacted the ratings. In the 6DoF part of the study, the parametric method yielded measurable perceptual improvements over the signal-independent MagLS approach. The smallest improvement was observed for the second combination of musical stimuli. The present authors theorize that this may have been connected to the inclusion of the broad-band percussive shaker, since scenes comprising a small number of temporally and frequency overlapping sound sources generally pose the greatest challenge to parametric methods.

Avenues for future work include estimating additional parameters describing the composition of the sound scene, possibly through the application of other modalities, such as computer-vision solutions applied to the corresponding video captured by the head-worn device, or through the use of light detection and ranging systems. This additional information would remove the need for the user to inform the system of the source distances from the recording point (as required in the present study), since the estimated source signals could be projected onto these estimated source positions. Room reflections could also be projected onto the surrounding room geometry, leading to a further improved perception of the space. Sound source positions could also be ascertained by using a distributed arrangement of multiple head-worn microphone arrays, similarly to those explored recently in [59, 60], which instead used multiple Ambisonic receivers.

6 CONCLUSION

This article investigates the use of four different binaural rendering approaches, which are all based on a single head-worn microphone array recording as input and are able to account for both the listener's head orientation and position relative to the recording point. This is commonly referred to as 6DoF rendering via sound-field extrapolation. Three of the explored approaches are signal-independent, delivering binaural audio through either LS-based or MagLS-based optimizations or through a PWD-based approach. Rotations and translations are accounted for by applying simple directional transformations to the employed rendering or optimization grids. The fourth approach considered is instead signal-dependent and utilizes a sound-field model and parametric spatial analysis, in order to steer beamformers and divide the input array signals into directional and ambient components. These components are then reproduced separately using dedicated rendering strategies.

The four considered 6DoF binaural reproduction approaches were integrated into a real-time system. A formal perceptual study was then conducted, whereby test participants compared the relative perceived quality of the four binaural rendering methods. The study was conducted in two stages. For the first part of the perceptual study, the participants remained at the center/recording point, with only their head orientations taken into consideration by the rendering methods (i.e., a 3DoF control case). Whereas, for the second part, the participants were permitted to navigate around the room, with their positions (relative to the recording point) also tracked and relayed to the real-time rendering system. The results of the perceptual study indicated that the MagLS approach outperformed the other two signal-independent binaural rendering approaches in all test cases, and performed similarly to the parametric approach for the 3DoF part of the study. The signal-dependent parametric method was then shown to perform better than the MagLS approach, when the listeners were permitted to navigate away from the recording point.

7 ACKNOWLEDGMENT

The authors would like to thank Jens Ahrens for designing and assembling the microphone array used for the evaluations and Samuel Clapp for measuring the ATFs and HRTFs. Thanks are also extended to Lachlan Birnie, Jacob Donley, and Vladimir Tourbabin for the insightful discussions that transpired during the course of the project, and to Zachery Schramm and Jessica Liberman for managing logistics and recruiting the test participants. The authors would also like to thank the reviewers for their helpful comments and suggestions.

8 REFERENCES

[1] S. M. Kuo, S. Mitra, and W.-S. Gan, "Active Noise Control System for Headphone Applications," *IEEE Trans. Control Syst. Technol.*, vol. 14, no. 2, pp. 331–335 (2006 Feb.). https://doi.org/10.1109/TCST.2005.863667.

[2] V. Hamacher, J. Chalupper, J. Eggers, et al., "Signal Processing in High-End Hearing Aids: State of the Art, Challenges, and Future Trends," *EURASIP J. Adv. Signal Process.*, vol. 2005, no. 18, paper 152674 (2005 Nov.). https://doi.org/10.1155/ASP.2005.2915.

[3] D. Marquardt, E. Hadad, S. Gannot, and S. Doclo, "Theoretical Analysis of Linearly Constrained Multi-Channel Wiener Filtering Algorithms for Combined Noise Reduction and Binaural Cue Preservation in Binaural Hearing Aids," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 23, no. 12, pp. 2384–2397 (2015 Sep.). https://doi.org/10.1109/TASLP.2015.2479940.

[4] J. Fernandez, L. McCormack, P. Hyvärinen, A. Politis, and V. Pulkki, "A Spatial Enhancement Approach for Binaural Rendering of Head-Worn Microphone Arrays," in *Proceedings of the 24th International Congress on Acoustics (ICA)* (Gyeongju, South Korea) (2022 Oct.).

[5] V. Pulkki, L. McCormack, and R. Gonzalez, "Superhuman Spatial Hearing Technology for Ultrasonic Frequencies," *Sci. Rep.*, vol. 11, no. 1, paper 11608 (2021 Jun.). https://doi.org/10.1038/s41598-021-90829-9.

[6] A. Politis, T. Pihlajamäki, and V. Pulkki, "Parametric Spatial Audio Effects," in *Proceedings of the 15th International Conference on Digital Audio Effects (DAFx)* (York, UK) (2012 Sep.).

[7] M. Kronlachner and F. Zotter, "Spatial Transformations for the Enhancement of Ambisonic Recordings," in *Proceedings of the 2nd International Conference on Spatial Audio* (Erlangen, Germany) (2014 Feb.).

[8] J. G. Tylka and E. Y. Choueiri, "Performance of Linear Extrapolation Methods for Virtual Sound Field Navigation," *J. Audio Eng. Soc.*, vol. 68, no. 3, pp. 138–156 (2020 Mar.). http://dx.doi.org/10.17743/jaes.2019.0054.

[9] B. Rafaely, *Fundamentals of Spherical Array Processing*, Springer Topics in Signal Processing, vol. 16 (Springer, Cham, Switzerland, 2019). https://doi.org/10.1007/978-3-319-99561-8.

[10] 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," presented at the *120th Convention of the Audio Engineering Society* (2006 May), paper 6857.

[11] C. T. Jin, N. Epain, and A. Parthy, "Design, Optimization and Evaluation of a Dual-Radius Spherical Microphone Array," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 22, no. 1, pp. 193–204 (2013 Oct.). https://doi.org/10.1109/TASLP.2013.2286920.

[12] M. A. Gerzon, "Periphony: With-Height Sound Reproduction," *J. Audio Eng. Soc.*, vol. 21, no. 1, pp. 2–10 (1973 Feb.).

[13] 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.

[14] J. Ivanic and K. Ruedenberg, "Rotation Matrices for Real Spherical Harmonics. Direct Determination by Recursion," *J. Phys. Chem.*, vol. 100, no. 15, pp. 6342–6347 (1996 Apr.). https://doi.org/10.1021/jp953350u.

[15] B. Bernschütz, A. V. Giner, C. Pörschmann, and J. Arend, "Binaural Reproduction of Plane Waves With Reduced Modal Order," *Acta Acust. united Acust.*, vol. 100, no. 5, pp. 972–983 (2014 Sep./Oct.). https://doi.org/10.3813/AAA.918777.

[16] Z. Ben-Hur, F. Brinkmann, J. Sheaffer, S. Weinzierl, and B. Rafaely, "Spectral Equalization in Binaural Signals Represented by Order-Truncated Spherical Harmonics," *J. Acoust. Soc. Am.*, vol. 141, no. 6, pp. 4087–4096 (2017 Jun.). https://doi.org/10.1121/1.4983652.

[17] M. Zaunschirm, C. Schörkhuber, and R. Höldrich, "Binaural Rendering of Ambisonic Signals by Head-Related Impulse Response Time Alignment and a Diffuseness Constraint," *J. Acoust. Soc. Am.*, vol. 143, no. 6, pp. 3616–3627 (2018 Jun.). https://doi.org/10.1121/1.5040489.

[18] C. Schörkhuber, M. Zaunschirm, and R. Höldrich, "Binaural Rendering of Ambisonic Signals via Magnitude Least Squares," in *Proceedings of the Annual German Conference on Acoustics (DAGA)*, vol. 44, pp. 339–342 (Munchen, Germany) (2018 Mar.).

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

[20] F. Schultz and S. Spors, "Data-Based Binaural Synthesis Including Rotational and Translatory Head-Movements," in *Proceedings of the AES 52nd International Conference: Sound Field Control-Engineering and Perception* (2013 Sep.), paper P-7.

[21] F. Winter, F. Schultz, and S. Spors, "Localization Properties of Data-Based Binaural Synthesis Including Translatory Head-Movements," in *Proceedings of the Forum Acusticum*, vol. 31 (Krakow, Poland) (2014 Sep.).

[22] L. Birnie, T. Abhayapala, V. Tourbabin, and P. Samarasinghe, "Mixed Source Sound Field Translation for Virtual Binaural Application With Perceptual Validation," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 29, pp. 1188–1203 (2021 Feb.). https://doi.org/10.1109/TASLP.2021.3061939.

[23] S. Braun and M. Frank, "Localization of 3D Ambisonic Recordings and Ambisonic Virtual Sources," in *Proceedings of the 1st International Conference on Spatial Audio* (Detmold, Germany) (2011 Nov.).

[24] A. Avni, J. Ahrens, M. Geier, et al., "Spatial Perception of Sound Fields Recorded by Spherical Microphone Arrays With Varying Spatial Resolution," *J. Acoust. Soc. Am.*, vol. 133, no. 5, pp. 2711–2721 (2013 May). https://doi.org/10.1121/1.4795780.

[25] V. Pulkki, A. Politis, M.-V. Laitinen, J. Vilkamo, and J. Ahonen, "First-Order Directional Audio Coding (DirAC)," in V. Pulkki, S. Delikaris-Manias, and A. Politis (Eds.), *Parametric Time-Frequency Domain Spatial Audio*, pp. 89–140 (Wiley, Hoboken, NJ, 2017), 1st ed. https://doi.org/10.1002/9781119252634.ch5.

[26] S. Berge and N. Barrett, "High Angular Resolution Planewave Expansion," in *Proceedings of the 2nd International Symposium on Ambisonics and Spherical Acoustics*, (Paris, France) (2010 May).

[27] 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.).

[28] C. Schörkhuber and R. Höldrich, "Linearly and Quadratically Constrained Least-Squares Decoder for Signal-Dependent Binaural Rendering of Ambisonic Signals," in *Proceedings of the AES International Conference on Immersive and Interactive Audio* (2019 May), paper 22.

[29] L. McCormack and A. Politis, "Estimating and Reproducing Ambience in Ambisonic Recordings," in *Proceedings of the 30th European Signal Processing Conference (EUSIPCO)*, pp. 314–318 (Belgrade, Serbia) (2022 Aug.).

[30] T. Pihlajamäki and V. Pulkki, "Projecting Simulated or Recorded Spatial Sound Onto 3D-Surfaces," in *Proceedings of the AES 45th International Conference: Applications of Time-Frequency Processing in Audio* (2012 Mar.), paper 4-5.

[31] T. Pihlajamaki and V. Pulkki, "Synthesis of Complex Sound Scenes With Transformation of Recorded Spatial Sound in Virtual Reality," *J. Audio Eng. Soc.*, vol. 63, no. 7/8, pp. 542–551 (2015 Aug.). https://doi.org/10.17743/jaes.2015.0059.

[32] M. Kentgens, A. Behler, and P. Jax, "Translation of a Higher Order Ambisonics Sound Scene Based on Parametric Decomposition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 151–155 (Barcelona, Spain) (2020 May).

[33] L. McCormack, A. Politis, and V. Pulkki, "Parametric Spatial Audio Effects Based on the Multi-Directional Decomposition of Ambisonic Sound Scenes," in *Proceedings of the 24th International Conference on Digital Audio Effects (DAFx20in21)*, pp. 214–221 (Vienna, Austria) (2021 Sep.).

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

[35] J. Ahrens, H. Helmholz, D. L. Alon, and S. V. A. Garí, "Spherical Harmonic Decomposition of a Sound Field Using Microphones on a Circumferential Contour Around a Non-Spherical Baffle," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 30, pp. 3110–3119 (2022 Sep.). https://doi.org/10.1109/TASLP.2022.3209940.

[36] A. Bastine, L. Birnie, T. D. Abhayapala, P. Samarasinghe, and V. Tourbabin, "Ambisonics Capture Using Microphones on Head-Worn Device of Arbitrary Geometry," in *Proceedings of the 30th European Signal Processing Conference (EUSIPCO)*, pp. 309–313 (Belgrade, Serbia) (2022 Aug.). https://doi.org/10.23919/EUSIPCO55093.2022.9909803.

[37] 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 on Audio for Virtual and Augmented Reality* (2020 Aug.), paper 3-1.

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

[39] J. Fernandez, L. McCormack, P. Hyvärinen, A. Politis, and V. Pulkki, "Enhancing Binaural Rendering of Head-Worn Microphone Arrays Through the Use of Adaptive Spatial Covariance Matching," *J. Acoust. Soc. Am.*, vol. 151, no. 4, pp. 2624–2635 (2022 Apr.). https://doi.org/10.1121/10.0010109.

[40] L. Madmoni, J. Donley, V. Tourbabin, and B. Rafaely, "Binaural Reproduction From Microphone Array Signals Incorporating Head-Tracking," in *Proceedings of the Immersive and 3D Audio: from Architecture to Automotive (I3DA)*, pp. 1–5 (Bologna, Italy) (2021 Nov.). https://doi.org/10.1109/I3DA48870.2021.9610940.

[41] J. Vilkamo and T. Backstrom, "Time-Frequency Processing: Methods and Tools," in V. Pulkki, S. Delikaris-Manias, and A. Politis (Eds.), *Parametric Time-Frequency Domain Spatial Audio*, pp. 1–24 (Wiley, Hoboken, NJ, 2017). https://doi.org/10.1002/9781119252634.ch1.

[42] H. Teutsch, *Modal Array Signal Processing: Principles and Applications of Acoustic Wavefield Decomposition*, Lecture Notes in Control and Information Sciences, vol. 348 (Springer, Berlin, Germany, 2007). https://doi.org/10.1007/978-3-540-40896-3.

[43] E. G. Williams, *Fourier Acoustics: Sound Radiation and Nearfield Acoustical Holography* (Academic Press, London, UK, 1999). https://doi.org/10.1016/B978-0-12-753960-7.X5000-1.

[44] L. S. Davis, R. Duraiswami, E. Grassi, et al., "High Order Spatial Audio Capture and Its Binaural Head-Tracked Playback Over Headphones With HRTF Cues," presented at the *119th Convention of the Audio Engineering Society* (2005 Oct.), paper 6540.

[45] C. D. Salvador, S. Sakamoto, J. Trevino, and Y. Suzuki, "Design Theory for Binaural Synthesis: Combining Microphone Array Recordings and Head-Related Transfer Function Datasets," *Acoust. Sci. Technol.*, vol. 38, no. 2, pp. 51–62 (2017 Jun.). https://doi.org/10.1250/ast.38.51.

[46] I. Ifergan and B. Rafaely, *Theoretical Framework for Beamformer Distribution in Beamforming Based Binaural Reproduction*, Ph.D. thesis, Ben-Gurion University of the Negev, Beer-Sheva, Israel (2020, Sep.).

[47] F. Zotter, H. Pomberger, and M. Noisternig, "Energy-Preserving Ambisonic Decoding," *Acta Acust. united Acust.*, vol. 98, no. 1, pp. 37–47 (2012 Jan./Feb.). https://doi.org/10.3813/AAA.918490.

[48] J. W. Strutt, "On Our Perception of Sound Direction," *Philos. Mag.*, vol. 13, no. 74, pp. 214–32 (1907).

[49] 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.

[50] R. O. Duda and W. L. Martens, "Range Dependence of the Response of a Spherical Head Model," *J. Acoust. Soc. Am.*, vol. 104, no. 5, pp. 3048–3058 (1998 Nov.). https://doi.org/10.1121/1.423886.

[51] J. Daniel, "Spatial Sound Encoding Including Near Field Effect: Introducing Distance Coding Filters and a Viable, New Ambisonic Format," in *Proceedings of the AES 23rd International Conference: Signal Processing in Audio Recording and Reproduction* (2003 May), paper 16.

[52] O. L. Frost, "An Algorithm for Linearly Constrained Adaptive Array Processing," *Proc. IEEE*, vol. 60, no. 8, pp. 926–935 (1972 Aug.). https://doi.org/10.1109/PROC.1972.8817.

[53] A. Politis, L. McCormack, and V. Pulkki, "Enhancement of Ambisonic Binaural Reproduction Using Directional Audio Coding With Optimal Adaptive Mixing," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 379–383 (New Paltz, NY) (2017 Dec.). https://doi.org/10.1109/WASPAA.2017.8170059.

[54] P. Lladó, T. McKenzie, N. Meyer-Kahlen, and S. J. Schlecht, "Predicting Perceptual Transparency of Head-Worn Devices," *J. Audio Eng. Soc.*, vol. 70, no. 7/8, pp. 585–600 (2022 Jul.). http://dx.doi.org/10.17743/jaes.2022.0024.

[55] M. H. Ashcraft and G. A. Radvansky, *Cognition* (Pearson Education, Boston, MA, 2014), 6th ed.

[56] M. A. Nees, "Have We Forgotten Auditory Sensory Memory? Retention Intervals in Studies of Nonverbal Auditory Working Memory," *Front. Psychol.*, vol. 7, paper 1892 (2016 Dec.). https://doi.org/10.3389/fpsyg.2016.01892.

[57] M. Sams, R. Hari, J. Rif, and J. Knuutila, "The Human Auditory Sensory Memory Trace Persists About 10 Sec: Neuromagnetic Evidence," *J. Cogn. Neurosci.*, vol. 5, no. 3, pp. 363–370 (1993 Jul.). https://doi.org/10.1162/jocn.1993.5.3.363.

[58] H. Lee, M. Frank, and F. Zotter, "Spatial and Timbral Fidelities of Binaural Ambisonics Decoders for Main Microphone Array Recordings," in *Proceedings of the AES International Conference on Immersive and Interactive Audio* (2019 Mar.), paper 75.

[59] M. Blochberger and F. Zotter, "Particle-Filter Tracking of Sounds for Frequency-Independent 3D Audio Rendering From Distributed B-Format Recordings," *Acta Acust.*, vol. 5, paper 20 (2021 Apr.). https://doi.org/10.1051/aacus/2021012.

[60] L. McCormack, A. Politis, T. McKenzie, C. Hold, and V. Pulkki, "Object-Based Six-Degrees-of-Freedom Rendering of Sound Scenes Captured With Multiple Ambisonic Receivers," *J. Audio Eng. Soc.*, vol. 70, no. 5, pp. 355–372 (2022 May). https://doi.org/10.17743/jaes.2022.0010.

THE AUTHORS

Leo McCormack

Nils Meyer-Kahlen

David Lou Alon

Zamir Ben-Hur

Sebastià V. Amengual Garí

Philip W. Robinson

Leo McCormack is a doctoral candidate for the Department of Signal Processing and Acoustics at Aalto University, Finland, researching parametric spatial audio technologies. He received his M.Sc. degree in Computer Communications and Information Sciences, majoring in Acoustics and Audio Technology at Aalto University, Finland, and his B.Sc. in Music Technology and Audio Systems at the University of Huddersfield, UK. He was also an intern at Fraunhofer IIS, Erlangen, Germany, in 2013–2014, and at Meta Reality Labs Research, Redmond, during the summer of 2022. His research interests include microphone array signal processing for sound-field reproduction and acoustic scene analysis.

Nils Meyer-Kahlen is a doctoral candidate for the Department of Signal Processing and Acoustics at Aalto University in Finland. Before joining the lab in 2019, he completed his B.Sc. and M.Sc. in Electrical Engineering and Audio Engineering at the Technical University and the University of Music and Performing Arts in Graz, Austria. In 2022, he was an intern at Meta Reality Labs Research, Redmond. His main research interest is virtual acoustics for augmented reality, from both a technological and perceptual point of view.

David Lou Alon is a Research Scientist at Meta Reality Labs Research, investigating spatial audio technologies. He received his Ph.D. in electrical engineering from Ben Gurion University (Israel, 2017) in the field of spherical microphone array processing. His research areas include head-related transfer functions, spatial audio capture, binaural reproduction, and headphone equalization for VR and AR application.

Zamir Ben-Hur is currently a research scientist at Meta Reality Labs Research, working on spatial audio technologies. He received a B.Sc. (summa cum laude), M.Sc., and Ph.D. degrees in electrical and computer engineering in 2015, 2017, and 2020, respectively, from Ben-Gurion University of the Negev, Beer-Sheva, Israel. His research interests include spatial audio signal processing for binaural reproduction with improved spatial perception.

Sebastià V. Amengual is currently a research scientist at Reality Labs Research working on room acoustics, spatial audio and auditory perception. He received a Diploma Degree in Telecommunications with a major in Sound and Image in 2014 from the Polytechnic University of Catalonia (UPC) in 2014, completing his Master's thesis at the Norwegian University of Science and Technology (NTNU). His doctoral work at the Detmold University of Music focused on investigating the interaction of room acoustics and live music performance using virtual acoustic environments. His research interests lie in the intersection of audio, perception, and music.

Philip W. Robinson is a research science manager in audio presence at Meta Reality Labs Research (RLR) in Redmond, WA. Prior to joining RLR, he incorporated virtual acoustics simulation and reproduction systems into building design processes at the architecture firm of Foster + Partners. He was a Fulbright Scholar and post-doctoral researcher at Aalto University in Finland, where he studied perception of concert hall acoustics, spatial auditory resolution, and echo thresholds. He has been a visiting researcher at École polytechnique fédérale de Lausanne (EPFL) in Switzerland and Hanyang University in South Korea. He received a Ph.D. from Rensselaer Polytechnic Institute in Troy, NY, in 2012. In a previous life, he was a registered architect in his home state of New Mexico. He remains passionate about architecture, the study of which gave him a great interest in perception of environments, real or virtual.