Parametric Ambisonic Encoding of Arbitrary Microphone Arrays

Leo McCormack, Archontis Politis, Raimundo Gonzalez, Tapio Lokki and Ville Pulkki

***Abstract*—This article proposes a parametric signal-dependent method for the task of encoding microphone array signals into Ambisonic signals. The proposed method is presented and evaluated in the context of encoding a simulated seven-sensor microphone array, which is mounted on an augmented reality headset device. Given the inherent flexibility of the Ambisonics format, and its popularity within the context of such devices, this array configuration represents a potential future use case for Ambisonic recording. However, due to its irregular geometry and non-uniform sensor placement, conventional signal-independent Ambisonic encoding is particularly limited. The primary aims of the proposed method are to obtain Ambisonic signals over a wider frequency band-width, and at a higher spatial resolution, than would otherwise be possible through conventional signal-independent encoding. The proposed method is based on a multi-source sound-field model and employs spatial filtering to divide the captured sound-field into its individual source and directional ambient components, which are subsequently encoded into the Ambisonics format at an arbitrary order. It is demonstrated through both objective and perceptual evaluations that the proposed parametric method outperforms conventional signal-independent encoding in the majority of cases.**

***Index Terms*—microphone array processing, ambisonic encoding, parametric spatial audio**

I. Introduction

THE capture and reproduction of spatial sound scenes has broad applicability in the fields of immersive audio, telepresence, and virtual and augmented reality. Traditional approaches to this task rely on recording the sound scene using an array of microphones, followed by mapping their signals directly to the respective channels of the intended playback setup. The orientation and directivities of the microphones are selected such that their interchannel differences, when delivered over the playback setup, dictate the listener's perception of the spatial sound scene in the desired manner. Examples of this channel-based workflow include employing binaural microphones for headphone playback, and multi-microphone arrangements for stereo [1] and surround loudspeaker formats [2]–[4]. However, such approaches may be considered inflexible, as there is often no clear solution for reproducing a recording intended for one specific playback setup over a different playback setup, or account for a different listener head orientation in the case of binaural microphone array recordings.

Leo McCormack, Raimundo Gonzalez, Tapio Lokki, and Ville Pulkki are with the Department of Signal Processing and Acoustics, Aalto University, Espoo, Finland.

Archontis Politis is with the Department of Information Technology and Communication Sciences, Tampere University, Finland.

Scene-based alternatives, on the other hand, aim to circumvent these limitations by describing the captured sound scene using a format that is independent of the array and playback setups. Perhaps the most wide-spread scene-based framework is the one popularised under the name of Ambisonics [5]. This refers to the two-step processing paradigm of: 1) employing a linear signal-independent mapping of the input microphone signals to intermediate spherical harmonic (SH) signals [6], often referred to as Ambisonic *encoding*; and 2) a linear mapping of these SH signals to the target binaural [7] or loudspeaker [8] setup, which is commonly referred to as Ambisonic *decoding*. Other linear signal-independent alternatives include beamforming designs that resemble head-related transfer functions (HRTFs) for headphone rendering [9], [10], or loudspeaker panning functions [11], [12]. However, contrary to Ambisonics, the decoding filters then need to be designed specifically for the particular recording array or device; or, alternatively, the array specifications may also be transmitted to the reproduction side. Since the Ambisonics framework has the benefit of decoupling the recording and the playback setups, it can afford greater practical flexibility and portability. Furthermore, spatial transformations, such as sound-field rotations [13], which are important for head-tracked virtual or augmented reality applications, are well defined and easily realised compared to other spatial audio formats.

The maximum spatial resolution afforded by a linear signal-independent Ambisonic workflow is, however, inherently limited by the number of microphones that comprise the array, since this dictates the maximum SH encoding order [6]. The Ambisonics format is also only truly portable in cases where the channel directivities (i.e. the SHs) are broad-band. However, when linearly encoding real microphone arrays, there are certain frequency-dependent limitations that affect this portability. These limitations are dictated by the array geometry and the placement of the microphones. For instance, there is a maximum frequency beyond which the SH directivities can no longer be obtained. This limit is often referred to as the spatial aliasing frequency [14], which is, in turn, also dependent on the SH order and degree. Furthermore, due to microphone sensor noise, regularisation of the encoding gains is required in practice, especially at lower frequencies and higher SH orders, which further limits the usable band-width of operation. Non-uniform arrangements of sensors and/or irregular array geometries also lead to direction-dependent differences in spatial resolution. This latter issue is the main motivation for why spherical microphone arrays (SMAs) with near-uniform sensor arrangements are more widely employed in practice. However, while there do exist commercial SMA offerings

capable of capturing up to fourth-order SHs, such arrays are uncommon and are often expensive and/or offer higher-order components for only narrow frequency bandwidths. Therefore, the majority of commercially available SMAs often comprise four sensors arranged in an open tetrahedral fashion, and are thus limited to first-order SH acquisition. Perceptual studies investigating the coupling of lower-order linear encoding with linear Ambisonic decoding have reported: the introduction of strong colourations, localisation inaccuracies, and a loss of perceived envelopment and spaciousness [15]–[19].

To overcome the perceptual limitations of a signal-independent low-order Ambisonics workflow, several signal-dependent alternatives for the decoding stage have been proposed. These alternatives operate by employing an assumed sound-field model and applying time-frequency domain processing techniques. Their intention is to map the input SH signals to the target playback format in an adaptive, signal-dependent, and often perceptually informed manner, in order to improve the perceived spatial accuracy of the reproduction. Directional Audio Coding (DirAC) [20] was the first proposed parametric decoding method, which operated on first-order SH signals as input. Its sound-field model assumes that the input scene may comprise a single plane-wave and/or an isotropic diffuse component per time-frequency tile. In practice, the method employs intensity-based analysis [21] to determine the plane-wave direction-of-arrival (DoA) and a diffuseness measure. Components that are analysed to be diffuse are routed to all channels of the target setup and subjected to decorrelation operations, whereas non-diffuse components are spatialised directly over the target setup through application of vector-base amplitude panning [22]. The DirAC model was then later extended to higher-orders in [23], [24], to resolve multiple simultaneous plane-waves by partitioning the sound-field into directionally constrained sectors [25], [26].

Other parametric Ambisonic decoding methods include the High Angular Resolution Planewave Expansion (HARPEX) [27] approach; which operates on first-order SH signals and assumes a sound-field model comprising two plane-waves for each narrow band frequency. By comparison, the Sparse-Recovery method [28] aims to resolve as few plane-waves as possible through an optimisation process, while ensuring that the sound scene is sufficiently described despite its sparse representation. The COding and Multi-Parameterisation of Ambisonic Sound Scenes (COMPASS) method [29] aims to resolve a time-variable number of plane-waves per frequency (based on source detection algorithms [30]). Along with extracting and spatialising the source components, the method also employs an additional directional ambient stream based on what remains after the source components are subtracted from the input sound-field. A similar model was also explored in [31], but with the addition of spatial post-filtering to improve the segregation of the source and directional ambient components. A linearly and quadratically constrained least-squares decoding solution was also proposed in [32], [33], which operated in a similar fashion to [24] but without the need for explicitly estimating a diffuseness parameter or requiring signal decorrelation.

It should be highlighted, however, that all of the parametric solutions mentioned thus far, are intended to enhance only the decoding stage of the Ambisonics pipeline. Signal-dependent Ambisonic encoding, on the other hand, has seen far fewer developments, with existing proposals primarily focusing on extending SH acquisition beyond the spatial aliasing frequency of SMAs; for example, using a tetrahedral array in [34], and higher-order SMAs in [35]. A general solution was also proposed in [36], which employed a signal model and subsequent spatial filtering to divide the sound-field into its individual source and ambient components. The model is similar to the parametric decoding methods described in [29], [31], except, the intention was to instead enhance the SH signals directly on the capturing side, rather than later relying on a parametric decoding method to render linearly encoded SH signals to the playback setup. The method used the decomposed spatial components encoded into SH signals, in order to replace the linearly encoded SH signals for frequency ranges where the linear signal-independent encoding was sub-optimal; as dictated by the objective evaluation metrics described in [37]. These existing signal-dependent encoding methods, however, all still impose the same maximum encoding order that would otherwise be dictated by the number of sensors associated with conventional linear encoding, and also considered only SMAs in their evaluations.

In general, Ambisonic encoding has primarily focused upon the use of SMAs, due to the practicality of mounting microphones on a sphere and its linear signal-independent encoding convenience [6]. However, with the Ambisonics format continuing to gain popularity, owing to its portability and flexibility, there may soon arise a need for ambisonic recording to be integrated into devices where spatial sound capture is not their primary purpose; for example: in 360 degree video cameras, mobile phones, head-mounted displays (HMDs) and other wearables related to augmented reality applications [38]–[43]. While linear ambisonic encoding for arbitrary microphone placements and mounting bodies is possible [44], it may be sub-optimal and limited in terms of its maximum order and usable bandwidth of operation, which would subsequently compromise the reproduction performance on the decoding side. Therefore, in this article, a general parametric encoding method is proposed, which draws influence from the COMPASS method described in [29], and the work of [36]. The primary novelty of the proposed method is in its general formulation, which allows it to cater to arbitrary array geometries and sensor placements; in order to obtain ambisonic signals of higher-order and over a wider frequency bandwidth than would otherwise be possible through a linear solution. The proposed method is also described and evaluated in the context of a case study, through the encoding of an array of seven microphones non-uniformly arranged over the irregular geometry of a HMD worn by a manikin. This particular sensor arrangement and array geometry represents a potential future scenario for ambisonics recording, which would otherwise be especially limited by conventional linear signal-independent encoding.

This article is arranged as follows: Section II describes how arbitrary microphone arrays may be linearly encoded into SH signals, and how such an encoding may be objectively

evaluated. The microphone array employed for this study is then described in Section III. The parametric signal model employed is detailed in Section IV. The spatial analysis and synthesis stages of the proposed method are then described in Section V and Section VI, respectively. Objective metrics and perceptual evaluations are detailed in Section VII, with the results and discussions provided in Section VIII. The article is then concluded in Section IX.

II. Conventional linear Ambisonic encoding

It is assumed that the input Q microphone array signals, $\mathbf{x}(t,f) \in \mathbb{C}^{Q\times 1}$ have been first transformed into the time-frequency domain, where t denotes the down-sampled time index and f denotes frequency. The conventional approach of encoding microphone array signals into Nth order ambisonic signals $\mathbf{a}_{\text{lin}} \in \mathbb{C}^{(N+1)^2\times 1}$ may be described with the following linear signal-independent mapping

$$\mathbf{a}_{\text{lin}}(t,f) = \mathbf{E}(f)\mathbf{x}(t,f), \tag{1}$$

where $\mathbf{E} \in \mathbb{C}^{(N+1)^2\times Q}$ is a frequency-dependent matrix of encoding weights. For SMAs, analytical descriptions of the geometry and sensor directivities may be used to derive $\mathbf{E}$, and more information can be found in e.g. [6], [45]–[47]. However, for irregular geometries, such as the array employed for this present study, a general approach is required. Here, the directional characteristics of the array are described through a dense grid of V array steering vectors, $\mathbf{A} = [\mathbf{a}(\boldsymbol{\gamma}_1), ..., \mathbf{a}(\boldsymbol{\gamma}_V)] \in \mathbb{C}^{Q\times V}$, which may be derived from numerical simulations or array measurements; where $\mathbf{a}(\boldsymbol{\gamma}) \in \mathbb{C}^{Q\times 1}$ is the steering vector of the array for direction $\boldsymbol{\gamma}$. The encoding matrix may be computed through a least-squares closed-form solution as [37], [44]

$$\mathbf{E}(f) = \mathbf{Y}\mathbf{W}\mathbf{A}^{\text{H}}(f)\big[V\mathbf{D}(f) + \beta\mathbf{I}_Q\big]^{-1}, \tag{2}$$

where $\mathbf{D}(f) = (1/V)\mathbf{A}(f)\mathbf{W}\mathbf{A}^{\text{H}}(f) \in \mathbb{C}^{Q\times Q}$ is the diffuse coherence matrix (DCM) of the array, $\mathbf{W} \in \mathbb{R}^{V\times V}$ is an optional diagonal weighting matrix to account for a non-uniform measurement grid, β is a regularisation parameter, $\mathbf{I}_Q \in \mathbb{R}^{Q\times Q}$ denotes an identity matrix, and $\mathbf{Y} \in \mathbb{R}^{(N+1)^2\times V}$ are the SH weights for all measurement directions.

Since this encoding approach may lead to the attenuation of frequencies above the spatial aliasing limit f_{al}, the aliased frequencies may be optionally diffuse-field equalised to retain a flat magnitude response on average, as described in [48] and also recommended in the original sound-field microphone report by Gerzon [49], as

$$\mathbf{E}^{(eq)}(f) = \text{Diag}[V\mathbf{E}(f)\mathbf{D}(f)\mathbf{E}^{\text{H}}(f)]^{-1/2}\ \mathbf{E}(f), \quad \text{for } f > f_{al}, \tag{3}$$

where $\text{Diag}[\cdot]$ denotes constructing a diagonal matrix based on the diagonal elements of the enclosed square matrix. The spatial aliasing frequency limit of the array may be specified based on analytical formulae in the case of SMAs, or, in the general case, through observation of the encoding performance metrics described in the following subsection.

A. Objective evaluation of conventional Ambisonic encoders

In order to gain insight into the performance of a linear signal-independent Ambisonic encoder, two well established objective metrics may be employed, namely: the spatial correlation and diffuse level differences [37], [44]. These metrics are computed through comparison between the microphone array encoded patterns and ideal SH patterns over a dense grid of directions. The spatial correlation is effectively a measure of spatial similarity, with the metric ranging between 0 and 1, and may be computed as

$$\mathbf{c}(f) = \text{diag}[\mathbf{E}(f)\mathbf{A}(f)\mathbf{W}\mathbf{Y}^{\text{T}}]\ \odot\ \text{diag}[V\mathbf{E}(f)\mathbf{D}(f)\mathbf{E}^{\text{H}}(f)]^{-1/2}, \tag{4}$$

where $\text{diag}[\cdot]$ denotes constructing a vector from the diagonal elements of the enclosed square matrix, $\odot$ denotes the Hadamard product, and $\mathbf{c}(f) \in \mathbb{R}^{(N+1)^2\times 1}$ are the resultant spatial correlation values for each SH component. Low spatial correlation values indicate that the encoded patterns have deviated from the ideal patterns, which is typically the case above the spatial aliasing frequency of the array. The upper usable frequency limit for each SH component may therefore be determined as the frequency where this metric begins to trend towards 0.

Since higher-order components generally require significant gain amplification at low frequencies, regularisation is often employed in practice. This allows a compromise to be made between minimising sensor noise amplification and the provision of a sufficiently wide operating frequency range of usable SH components. The diffuse level difference metric is therefore useful in the determination of the lower usable frequency bound for each SH component, which may be determined as the frequency where the metric begins to deviate from 0 dB. The level difference metric may be computed as

$$\boldsymbol{\delta}(f) = 10\ \log_{10}\big(\text{diag}[V\mathbf{E}(f)\mathbf{D}(f)\mathbf{E}^{\text{H}}(f)]\big), \tag{5}$$

where $\boldsymbol{\delta}(f) \in \mathbb{R}^{(N+1)^2\times 1}$ are the level differences for each SH component.

III. The array in question

While the parametric encoding method proposed in this article is general, and thus applicable to a wide-range of microphone arrays of arbitrary geometry, including SMAs, the focus of this work is primarily in regard to encoding arrays of irregular geometry and with non-uniformly distributed sensor placements. Therefore, an array of seven sensors arranged on the surface of an HMD worn by a manikin, was first designed and 3D modelled; as depicted in Fig. 1 (left). Five sensors were arranged on the left, right, front, back and top orientations of the HMD, and two more sensors were placed in the forward facing directions in order to obtain a higher degree of frontal spatial resolution. The far-field pressure response of the array was then simulated[1] for 841 directions, following a 28th order Fliege design [50], using the Boundary Element Method (BEM) module of COMSOL Multiphysics. The array

[1]Note that the simulated array responses and other associated files may be downloaded from here: https://doi.org/10.5281/zenodo.6382345.

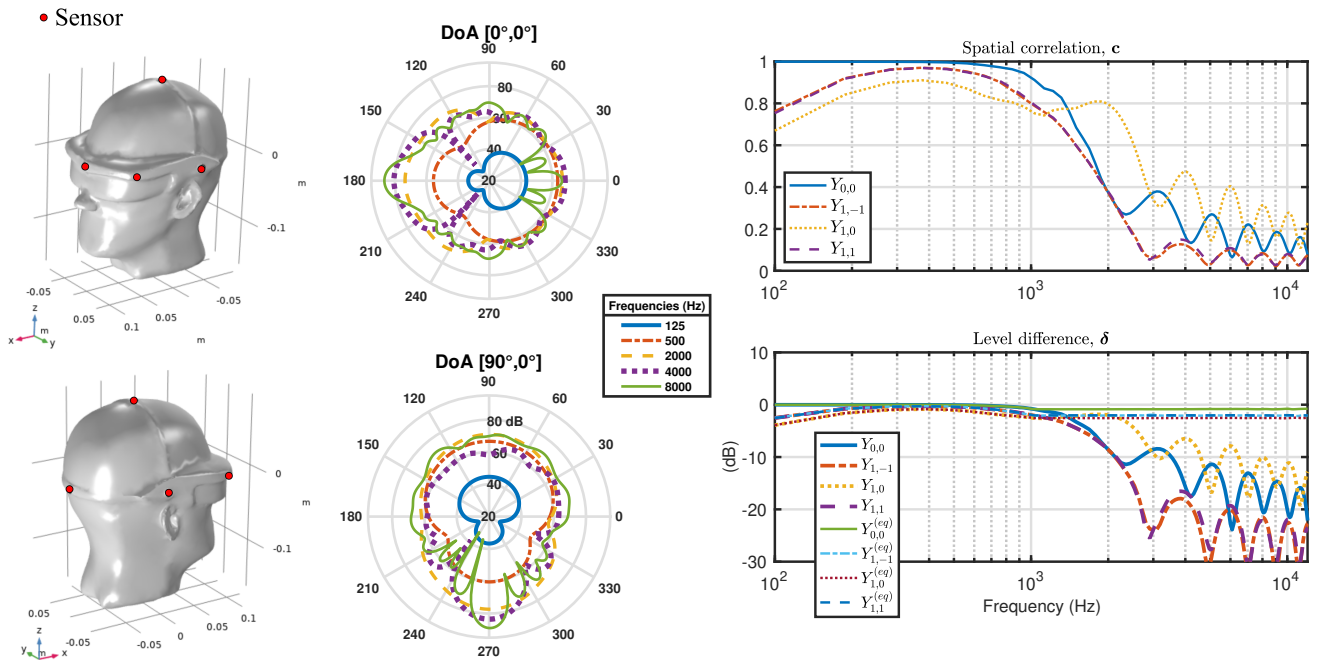

Fig. 1: Left: A picture of the microphone array in question, with the sensor positions depicted as red dots. Middle: Directivity of the scattered pressure from the surface of the array for two incident plane-wave directions on the horizontal plane aligned with the frame of the HMD. Right: A depiction of the objective metrics for the least-squares Ambisonic encoder, $\mathbf{E}$, as given by Equation (2), derived using the steering vectors for the array in question. Note that the results with the $^{(eq)}$ superscript are of the diffuse-field equalised encoder, $\mathbf{E}^{(eq)}$, as per Equation (3), with the spatial aliasing frequency of 1 kHz.

was simulated for 128 frequencies (uniformly spaced between 93.75 Hz-12 kHz) in total, with a meshing resolution of $\frac{1}{6}$ of the wavelength of each simulated frequency. The scattered pressure measured along the horizontal plane aligned with the HMD is presented in Fig. 1 (middle) as a directivity pattern for two different incident plane-wave directions, which indicates that the directivity of the scattered field of the array can change according to the DoA of the incident wave. This direction-dependent scattering, which is a product of the asymmetrical design employed, differs from the widely utilised rigid SMA configuration where the baffles produce similar scattered directivities for all incident directions.

Note that this particular array design was chosen as it represents a likely future use case in the context of augmented reality applications. It is also an array that is particularly problematic for the conventional linear Ambisonic encoding approach. The challenges associated with linear signal-independent encoding may be demonstrated by computing the performance metrics[2] described in Section II-A; the results for which are provided in Fig. 1 (right). It can be observed that not all components of the same order are encoded in the same manner, which is something that is distinctly different from SMAs, and thus subsequently translates into a non-uniform spatial resolution for different directions. Furthermore, with SMAs, the components of a lower-order typically have a wider operational bandwidth than their higher-order components. However, this is not the case for this irregular array; as the z-axis dipole $Y_{1,0}$ component appears to exhibit adequate encoding performance up to higher frequencies than the omni-directional $Y_{0,0}$ component. Such properties are due to the irregular microphone placement and directionally diverse scattering arising due to the geometry of the HMD and the head of the manikin. The metrics also indicate that SH domain beamformers of first-order directivity cannot be reliably generated above approximately 1 kHz. This is also confirmed when the directivity patterns of beamformers derived from linearly encoded Ambisonics are plotted, as depicted in Fig. 2 (left). In contrast, when the microphone sensors are used directly, beamformers with higher directivity may be employed, which may also be generated beyond the spatial aliasing frequency of a linear encoding; as shown in Fig. 2 (right). This is therefore an early indication that a parametric encoding method based on space-domain beamforming, could potentially yield improved spatial resolution, and over a wider frequency band-width, when compared to conventional linear encoding.

[2]Note that the linear signal-independent encoder and objective evaluation metrics were computed using the MATLAB library found here: https://github.com/polarch/Spherical-Array-Processing

IV. Signal Model

The narrow-band spatial covariance matrices (SCMs) of the signal vectors are given by $\mathbf{C}_\mathbf{x}(t,f) = \mathbb{E}[\mathbf{x}(t,f)\mathbf{x}^\mathrm{H}(t,f)] \in \mathbb{C}^{Q\times Q}$, which in practice are computed over a number of temporal frames. Note that the time-frequency indices are omitted henceforth for brevity of notation.

It is assumed that a number $K < Q$ of active signals from sound sources $\mathbf{s} = [s_1, ..., s_K] \in \mathbb{C}^{K\times 1}$ at each time-frequency

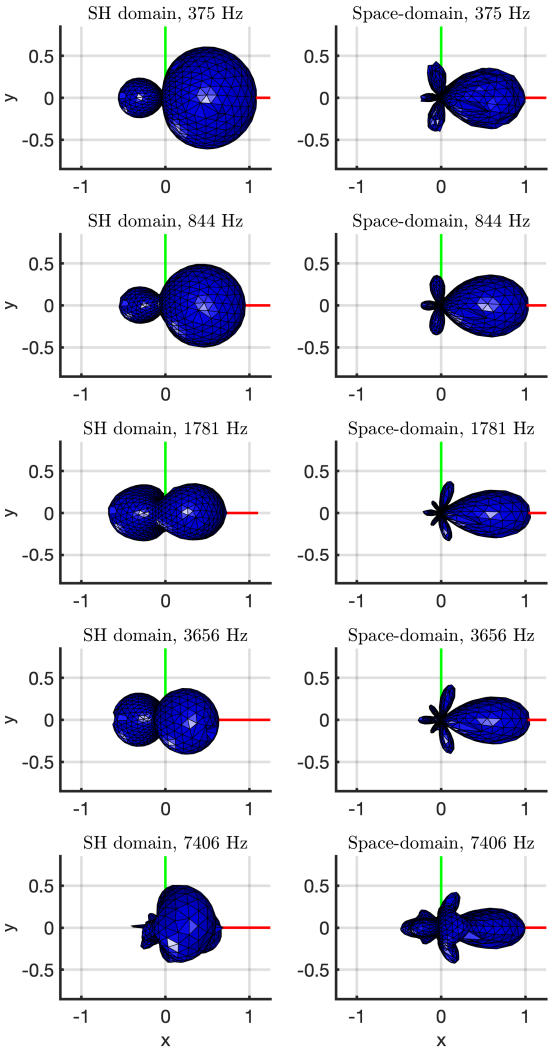

Fig. 2: Example directivity patterns of beamformers when using linearly encoded SH domain signals (left) or the microphone signals directly (right), for five different frequencies and using the array in question. Note that the SH domain beamformers are hyper-cardioid (maximum directivity) beamformers with diffuse-field equalisation enabled above the spatial aliasing frequency (1 kHz), while the space-domain beamformers are as described in [51].

tile, are incident from directions $\mathbf{\Gamma}_\mathrm{s} = [\gamma_1, ..., \gamma_K]$. The array signal vector is therefore described as

$$\mathbf{x} = \mathbf{A}_\mathrm{s}\mathbf{s} + \mathbf{d} + \mathbf{n}, \tag{6}$$

where $\mathbf{A}_\mathrm{s} = [\mathbf{a}(\gamma_1), ..., \mathbf{a}(\gamma_K)] \in \mathbb{C}^{Q\times K}$ contains the array steering vectors for the source directions; $\mathbf{d} \in \mathbb{C}^{Q\times 1}$ is the diffuse signal vector, which comprises reverberation and spatially diffuse sounds with no clear directionality; and $\mathbf{n} \in \mathbb{C}^{Q\times 1}$ is the sensor noise signal vector, which is assumed to be uncorrelated between sensors.

Assuming uncorrelated source signals, their second-order statistics are given by the diagonal SCM $\mathbf{C_s} = \mathbb{E}[\mathbf{s}\mathbf{s}^\mathrm{H}] \in \mathbb{C}^{K\times K}$, which has a total source signal power $P_\mathrm{s} = \mathbf{tr}[\mathbf{C_s}]$. The array SCM solely arising from these source components is given as

$$\mathbf{C_{x,s}} = \mathbb{E}[\mathbf{A}_\mathrm{s}\mathbf{s}\mathbf{s}^\mathrm{H}\mathbf{A}_\mathrm{s}^\mathrm{H}] = \mathbf{A}_\mathrm{s}\mathbf{C_s}\mathbf{A}_\mathrm{s}^\mathrm{H}. \tag{7}$$

The diffuse array signal vector is then modelled as

$$\mathbf{d} = \mathbf{A}\mathbf{W}^{1/2}\mathbf{z}, \tag{8}$$

where $\mathbf{z} \in \mathbb{C}^{V\times 1}$ are the diffuse signal components incident from all directions in the measurement grid. Assuming uncorrelated diffuse signal components, their SCM is given as $\mathbf{C_z} = \mathbb{E}[\mathbf{z}\mathbf{z}^\mathrm{H}] \in \mathbb{C}^{V\times V}$, and the total diffuse signal power is therefore $P_\mathrm{d} = \mathbf{tr}[\mathbf{C_z}]$. Note that in the case of an isotropic diffuse signal vector, the SCM becomes $\mathbf{C_z} = (P_\mathrm{d}/V)\mathbf{I}_V$. The SCM for the diffuse signals, as captured by the array, is then given as

$$\mathbf{C_d} = \mathbb{E}[\mathbf{d}\mathbf{d}^\mathrm{H}] = \mathbf{A}\mathbf{W}^{1/2}\mathbf{C_z}\mathbf{W}^{1/2}\mathbf{A}^\mathrm{H} = P_\mathrm{d}\mathbf{D}. \tag{9}$$

The array noise SCM is then

$$\mathbf{C_n} = \mathbb{E}[\mathbf{n}\mathbf{n}^\mathrm{H}] = P_\mathrm{n}\mathbf{I}_Q, \tag{10}$$

with equal noise power P_n across all sensors.

The overall array signal SCM, based on this assumed model, is therefore

$$\mathbf{C_x} = \mathbf{A}_\mathrm{s}\mathbf{C_s}\mathbf{A}_\mathrm{s}^\mathrm{H} + \mathbf{C_d} + \mathbf{C_n}. \tag{11}$$

V. Parametric Spatial Analysis

A. Spatial whitening of the array SCM

The proposed parametric analysis is based on the subspace principles of array signal processing, from which the number of active sound sources and their direction-of-arrivals (DoAs) are estimated. It is noted, however, that the employed subspace techniques assume that the array SCM will exhibit an identity-like structure, with its eigenvalues all being P_n, when the sound sources in the scene are inactive. These algorithms are therefore well-suited to the task of estimating the number of sources and their directions in the presence of sensor noise. However, in the present scenario, it is assumed that directional components are instead mixed with both sensor noise and diffuse sounds; with the latter not necessarily conforming to this identity-like structure, as demonstrated by Equation (9). If one is to further assume that sensor noise may be negligible (i.e. $P_\mathrm{d} >> P_\mathrm{n}$) for the intended applications of the proposed method, then it may be more beneficial to instead have the array SCMs exhibit an identity-like structure when the array is placed under isotropic diffuse-field conditions. Therefore, prior to estimating the required spatial parameters, a spatial whitening operation is applied. This operation is to ensure that the array SCMs, given an isotropic diffuse-field input, would instead conform to the following

$$\mathbf{C_x}^{(w)} = \mathbf{T}\mathbf{C_x}\mathbf{T}^\mathrm{H} = P_\mathrm{d}\mathbf{T}\mathbf{D}\mathbf{T}^\mathrm{H} = P_\mathrm{d}\mathbf{I}_Q. \tag{12}$$

where $\mathbf{T} \in \mathbb{C}^{Q\times Q}$ is the signal-independent ideal diffuse-field spatial whitening matrix, which is computed as

$$\mathbf{T} = \mathbf{\Lambda}^{-1/2}\mathbf{R}^\mathrm{H}, \tag{13}$$

given the eigenvalue decomposition $\mathbf{D} = \mathbf{R}\mathbf{\Lambda}\mathbf{R}^\mathrm{H}$.

The subspace decomposition is then applied to the array SCMs after the ideal diffuse-field whitening as

$$\mathbf{C_x}^{(w)} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{\mathrm{H}} = \sum_{k=1}^{K} \lambda_k \mathbf{v}_k \mathbf{v}_k^{\mathrm{H}} + \sum_{k=K+1}^{Q} \lambda_k \mathbf{v}_k \mathbf{v}_k^{\mathrm{H}}, \quad (14)$$

where K refers to the number of sources, λ are the eigenvalues sorted in descending order, and $\mathbf{v}$ are the respective eigenvectors. With the current assumptions, the largest K eigenvalues should be $\mathrm{diag}[\mathbf{C_s}]$, while the smallest $Q-K$ eigenvalues should all be equal to P_{d}. Examples of eigenvalues for both the whitened and un-whitened array SCMs, for up to three white noise sources in a diffuse field, are presented in Fig. 3 using the array in question. It is noted that for a diffuse-field input, the eigenvalues are not necessary all equal in practice. However, the whitened array SCM do more closely conform to the subspace assumptions for these diffuse-field conditions. This also extends to the source(s) mixed with diffuse sound cases, where the $Q-K$ smallest eigenvalues (highlighted with a grey background) are notably flatter when the whitening operation is applied in the 1 kHz and 2 kHz examples. However, at higher frequencies, where $\mathbf{D}$ in any case begins to trend towards an identity matrix, the whitening operation may not provide any benefit; as is shown in the 4 kHz example.

B. Source signal detection

The estimation of the number of sound sources, often referred to as *detection* in sensor array processing literature, may be based on analysis of the SCM eigenvalues and thresholding [52], eigenvalue statistics [30], or operations performed directly on the eigenvectors [53]. Alternative approaches are based upon information theoretic criteria [54]. For this work, the SORTE algorithm is employed, as it has been demonstrated to be a robust detector in [30], and does not require any parameter tuning. The first step relies on determining the differences between the eigenvalues as

$$\nabla\lambda_i = \lambda_i - \lambda_{i+1}, \quad \text{for } i = 1, ..., Q-1. \quad (15)$$

The number of sources is then given by

$$K = \underset{k}{\mathrm{argmin}}\ f(k) \quad \text{for } k = 1, ..., Q-3, \quad (16)$$

with

$$f(k) = \begin{cases} \frac{\sigma^2_{k+1}}{\sigma^2_k}, & \sigma^2_k > 0 \\ +\infty, & \sigma^2_k = 0 \end{cases}, \quad \text{for } k = 1, ..., Q-2, \quad (17)$$

$$\sigma^2_k = \frac{1}{Q-k} \sum_{i=k}^{Q-1} \left(\nabla\lambda_i - \frac{1}{Q-k} \sum_{i=k}^{Q-1} \nabla\lambda_i \right)^2. \quad (18)$$

C. Source direction estimation

Once the number of sound sources has been determined, establishing their DoAs can be based on first generating activity-maps based on, for example, scanning the same dense grid of directions $\mathbf{\Gamma} = [\gamma_1, ..., \gamma_V]$ as used to simulate (or measure) the array. Such activity-maps may be based on computing the energy of conventional beamformers, such as the filter-and-sum [55], or minimum-variance distortionless response (MVDR) [56] beamformers. However, since the subspace principles are employed for the source detection task, a spatial pseudospectrum [57]–[59] represents a practical alternative and often leads to sharper activity-maps than those generated by steered-response power approaches. In this work, the MUltiple-Signal Classification (MUSIC) approach [58] is employed as

$$P_{\mathrm{MUSIC}}(\boldsymbol{\gamma}) = \frac{1}{||\mathbf{V}_{\mathrm{n}}^{\mathrm{H}} \mathbf{Ta}(\boldsymbol{\gamma})||^2} \quad \text{for } \boldsymbol{\gamma} \in \mathbf{\Gamma}, \quad (19)$$

where $\mathbf{V}_{\mathrm{n}}$ refers to the noise subspace, defined as the eigenvectors corresponding to the smallest $Q-K$ eigenvalues. Peak-finding may then be employed to numerically extract the K source DoA estimates from the pseudospectrum.

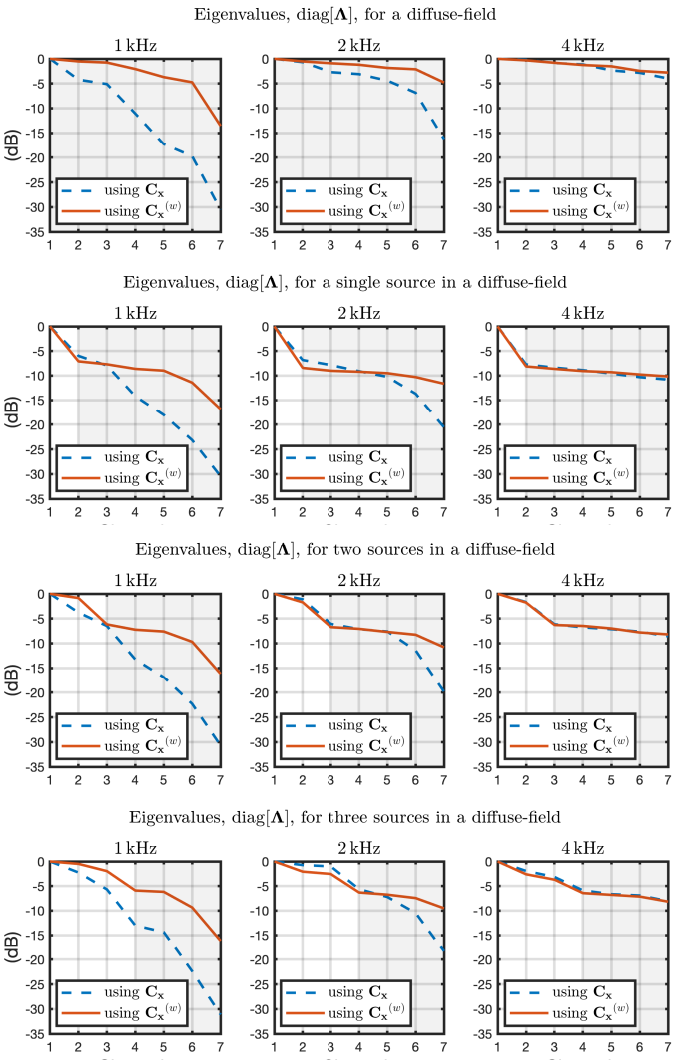

Fig. 3: An example of the effect of spatial whitening on the eigenvalues of the array SCM for three frequencies, given up to three (top-bottom) equal-power white noise source signals in a diffuse-field with $\mathbf{tr}[\mathbf{C_s}] = \mathbf{tr}[\mathbf{C_z}]$, and using the array in question. The first, second, and third sources were incrementally introduced in the following directions: $[0, 90, -90]$ degrees azimuth.

VI. Parametric spatial synthesis

A. Source rendering

Once the number of sources has been detected and their respective DoAs have been determined, spatial filters may be constructed to obtain estimates of the source signals. The extracted source signals may then be encoded into SH signals as incident plane-waves from the same respective DoAs. Various beamforming designs are possible with their own advantages and disadvantages. In the simplest case, beamformers may be steered towards the K DoAs using a matched filter (MF) approach, and thus the source beamforming matrix $\mathbf{W}_\mathrm{s} \in \mathbb{C}^{K\times Q}$ is simply

$$\mathbf{W}_\mathrm{s}^{(MF)} = \mathrm{Diag}(\mathbf{A}_\mathrm{s}^\mathrm{H}\mathbf{A}_\mathrm{s})^{-1}\mathbf{A}_\mathrm{s}^\mathrm{H}, \tag{20}$$

where the matrix of the source steering vectors $\mathbf{A_s} \in \mathbb{C}^{Q\times K}$ is constructed by taking a subset of the dense array response measurements corresponding to the estimated DoAs. The diagonal normalisation matrix ensures that unit gain in achieved in the focusing direction for each beamformer. However, while such a design is numerically robust, it does not offer the highest suppression of the ambient sound and of sources in the other estimated directions when $K > 1$. To improve this aspect, a linearly-constrained minimum power (LCMP) solution [56] may be employed with the constraint $\mathbf{W}_\mathrm{s}\mathbf{A}_\mathrm{s} = \mathbf{I}_K$, resulting in

$$\mathbf{W}_\mathrm{s}^{(LCMP)} = [\mathbf{A}_\mathrm{s}^\mathrm{H}(\mathbf{C_x} + \beta\mathbf{I}_Q)^{-1}\mathbf{A}_\mathrm{s}]^{-1}\mathbf{A}_\mathrm{s}^\mathrm{H}(\mathbf{C_x} + \beta\mathbf{I}_Q)^{-1}, \tag{21}$$

where β denotes a regularisation term to avoid any ill-conditioned inversions. Equivalently, and as more commonly formulated in the literature, the beamforming matrix may be expressed as $\mathbf{W}_\mathrm{s}^{(LCMP)} = [\mathbf{w}_1, ..., \mathbf{w}_K]$, with the weight vectors required to extract the kth source signal obtained based on minimising the array output power $\mathbf{w}_k = \arg\min[\mathbf{w}^\mathrm{H}\mathbf{C_x}\mathbf{w}]$ under the linear constraint $\mathbf{A}_\mathrm{s}^\mathrm{H}\mathbf{w} = \mathbf{c}$, where the $\mathbf{c}$ vector has 1 at the kth entry and zeros elsewhere. It is further noted that it is possible for the LCMP solution to become unstable if two or more DoA estimates fall too close together. In such cases, heuristic approaches may be devised to cull or merge the DoA vectors to improve the robustness of the beamforming solution. Alternatively, if such instabilities are identified, then a single-column minimum power distortionless response (MPDR) solution may instead be employed for each source; although, this approach may then overestimate the energy of sources in the scene. Note that examples of extracted source signal energies for up to three simultaneous white noise sources in a free-field, when using the array in question and the LCMP beamformer design, are depicted in Fig. 4. It can be observed that at lower-frequencies, the beamformers are unable to fully separate the source signals; resulting in them containing also up to 3 dB of the signal energy from other source(s). However, given that practical scenes typically comprise source signals that are sparser across frequency and more intermittent over time, these examples may be considered to represent a worst-case scenario for free-field conditions.

Once the source signals have been extracted, they are then encoded into the Ambisonics format as

$$\mathbf{a}_\mathrm{s} = \mathbf{Y}_\mathrm{s}\mathbf{W}_\mathrm{s}\mathbf{x}, \tag{22}$$

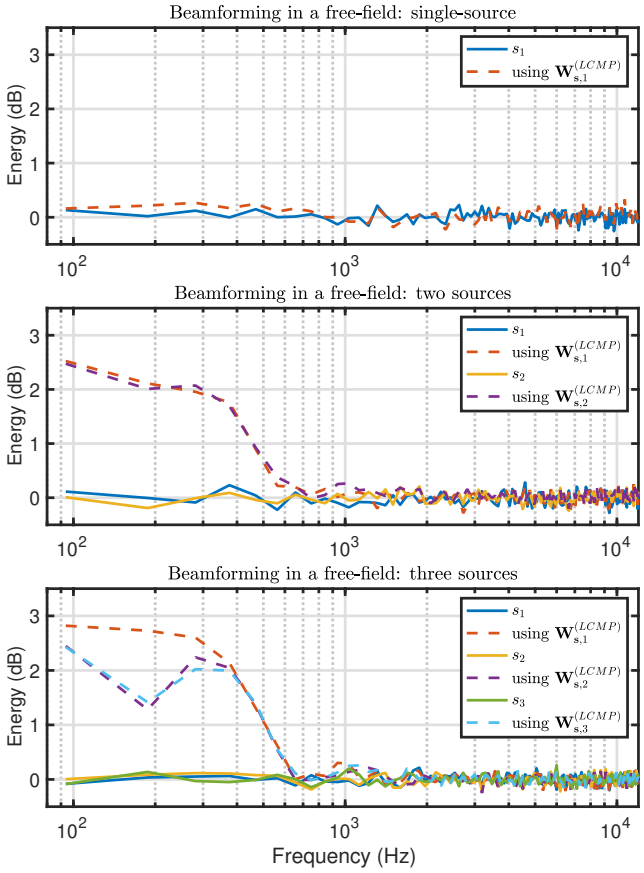

Fig. 4: Examples of beamformer energy plotted over frequency for one (top), two (middle), and three (bottom) uncorrelated white noise source signals in a free-field, using the beamforming solution described by Equation (21) ($\beta = 0.01\ \mathbf{tr}[\mathbf{C_x}]$) and the array in question. The first, second, and third sources were incrementally introduced in the following directions: $[0, 90, -90]$ degrees azimuth.

where $\mathbf{Y}_\mathrm{s} = [\mathbf{y}(\boldsymbol{\gamma}_1), ..., \mathbf{y}(\boldsymbol{\gamma}_K)] \in \mathbb{R}^{(N+1)^2\times K}$ are the encoding SH weights for the respective source directions. Note that, unlike conventional linear signal-independent encoding, there is no maximum order dictated by the number of sensors in the array, and thus the encoding order may be arbitrarily selected by the user.

B. Ambient rendering

To encode the residual sound scene component, which encapsulates ambient sound and weakly directional sources, a two-stage strategy is followed. Firstly, the residual array signals are obtained after the source components have first been subtracted from the input sound-field. This source subtraction is conducted via a spatial filtering matrix $\mathbf{W}_\mathrm{d} \in \mathbb{C}^{Q\times Q}$, which is derived as

$$\mathbf{W}_\mathrm{d} = \mathbf{I}_Q - \mathbf{A}_\mathrm{s}\mathbf{W}_\mathrm{s}, \tag{23}$$

with an estimate of the residual array signals then given by

$$\mathbf{d} = \mathbf{W}_\mathrm{d}\mathbf{x}. \tag{24}$$

Secondly, a plane-wave decomposition of these residual signals is conducted over a uniformly distributed set of

$L \geq (N+1)^2$ directions, which are subsequently re-encoded into ambisonic signals of the target order. The plane-wave decomposition may be performed using unity gain beamformers following Equation (20), based on the respective steering response matrix $\mathbf{A}_\mathrm{d} \in \mathbb{C}^{Q\times L}$, which yields the signals $\mathbf{z}_\mathrm{d} = \mathbf{A}_\mathrm{d}^\mathrm{H}\mathbf{d} \in \mathbb{C}^{L\times 1}$. It is noted, however, that the beamformer directivity patterns achieved through Equation (20) are inherently frequency-dependent. Therefore, due to the fixed number of plane-wave decomposition directions, it is possible that some frequencies may be over-represented due to greater overlapping of the beamformer patterns. Conversely, at other frequencies, the beamformers patterns may instead become too narrow to capture the residual sound-field energy without losses. Additionally, if the employed microphone array features an irregular geometry and/or non-uniform sensor placement, then the directivity patterns and the energy captured by the beamformers will also be direction-dependent. Therefore, since it is assumed that the residual signals are mostly made up of diffuse ambient components, energy-preservation prior to re-encoding may be deemed to be more important than the unity response constraint imposed by Equation (20). To ensure this energy-preserving property of the beamforming matrix, the following singular value decomposition is first conducted

$$\mathbf{A}_\mathrm{d}^\mathrm{H} = \mathbf{U}_\mathrm{d}\mathbf{\Sigma}_\mathrm{d}\mathbf{V}_\mathrm{d}^\mathrm{H}. \tag{25}$$

This is followed by discarding the matrix containing the singular values $\mathbf{\Sigma}_\mathrm{d}$ and truncating the $\mathbf{U}_\mathrm{d}$ matrix, in order to force the array steering vector matrix to be unitary with

$$\hat{\mathbf{A}}_\mathrm{d} = \frac{1}{\sqrt{L}}\mathbf{U}_\mathrm{d}^{(tr)}\mathbf{V}_\mathrm{d}^\mathrm{H}, \tag{26}$$

where $\mathbf{U}_\mathrm{d}^{(tr)} \in \mathbb{C}^{L\times Q}$ is the truncated version of $\mathbf{U}_\mathrm{d}$, whereby only the first Q columns are retained. Note that this energy-preservation constraint is similar to the method proposed in [60], which instead employed broad-band SH vectors. An example of this energy-preserving plane-wave decomposition, when the array in question is under diffuse-field conditions, is depicted in Fig. 5. The figure demonstrates that the energy-preservation constraint leads to a more consistent capture of diffuse energy across both frequency and direction, when compared to using the unity response constraint.

The plane-wave signal vector is then encoded into ambisonic signals as

$$\mathbf{a}_\mathrm{d} = E_\mathrm{d}\mathbf{Y}_\mathrm{d}\mathbf{z}_\mathrm{d} = E_\mathrm{d}\mathbf{Y}_\mathrm{d}\hat{\mathbf{A}}_\mathrm{d}\mathbf{W}_\mathrm{d}\mathbf{x}, \tag{27}$$

where $\mathbf{Y}_\mathrm{d} \in \mathbb{R}^{(N+1)^2\times L}$ is a matrix of SH weights for the respective plane-wave directions, and $E_\mathrm{d} = \mathbf{tr}[\mathbf{D}]^{-1/2}$ is a diffuse-field equalisation term. Note that, optionally, the decomposed residual array signals may also be subjected to a channel-wise decorrelation operation $\hat{\mathbf{z}}_\mathrm{d} = \mathbb{D}[\mathbf{z}_\mathrm{d}]$, in order to enforce the diffuse properties assumption, before they are encoded into the Ambisonics format.

C. Overall rendering

The final parametrically encoded Ambisonic signals are then obtained as

$$\mathbf{a}_\mathrm{par}(t,f) = \mathbf{a}_\mathrm{s}(t,f) + \mathbf{a}_\mathrm{d}(t,f). \tag{28}$$

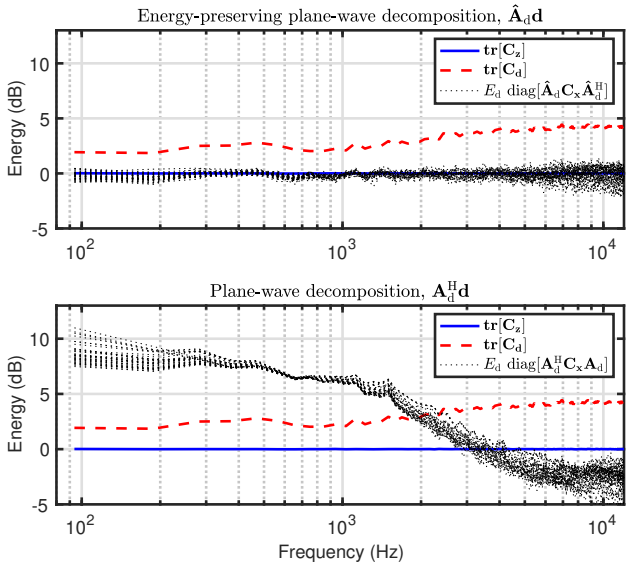

Fig. 5: A depiction of the energy of $\mathbf{z}_\mathrm{d}$ plotted over frequency for $L = 60$ directions when the array in question is placed in an isotropic diffuse-field. The top plot employed the energy-preserving steering vectors $\hat{\mathbf{A}}_\mathrm{d}$, while the bottom plot used $\mathbf{A}_\mathrm{d}^\mathrm{H}$. For visual reference, the total energy of the input diffuse-field $\mathbf{tr}[\mathbf{C_z}]$, and the total energy of the diffuse-field as captured by the microphone array $\mathbf{tr}[\mathbf{C_d}]$, are also plotted.

Naturally, this decoupling of the two streams also allows for the possibility of re-balancing them, for example, to apply more gain to the source stream, which would be akin to de-reverberation, or to emphasise the ambient stream to exaggerate the reverberance of the scene. Other parametric based spatial audio effects and/or sound-field modifications are also possible based upon the manipulation of the estimated spatial parameters prior to synthesis [61]. The parametrically encoded signals may also be substituted by linearly signal-independent encoded signals for the frequency bandwidths at which conventional encoding is optimal; as explored in [36], based on the objective metrics depicted in Fig. 1.

VII. Evaluation

The evaluation of the proposed encoding method was approached through: the calculation of objective metrics, and by conducting formal listening tests. Both evaluations utilised the microphone array described in Section III.

A. Objective metrics evaluation

To evaluate the objective performance of the proposed method, synthetic microphone array recordings of different scenarios were created. These were based on uncorrelated white noise source signals of varying number and directions, which were mixed with an isotropic diffuse field. The diffuse field was modelled based on uncorrelated white noise sources in all $V = 841$ measurement directions, accompanied by the appropriate integration weights for the employed spherical grid [50]. The gains for the source signal(s) the diffuse-field signals

were then adjusted to attain specific direct-to-diffuse (DDR) ratios, which were computed as

$$\mathrm{DDR} = 10\,\log_{10}\left(\frac{\mathbf{tr}[\mathbf{C_s}]}{\mathbf{tr}[\mathbf{C_z}]}\right) = 10\,\log_{10}\left(\frac{\sum_{k=1}^{K}\mathbb{E}[|s_k|^2]}{\sum_{\nu=1}^{V}\mathbb{E}[|z_\nu|^2]}\right). \tag{29}$$

For this study, the following DDRs were targeted: $[0, 6, 12, \mathrm{Inf}]$ dB. Note that all objective metrics were based on computing $\mathbf{C_x}$ over one second of input microphone array audio (sampling rate of 48 kHz), given a short-time Fourier transform (STFT) with a window size of 512 samples with no overlap; i.e. averaged over $\lfloor 48000/512 \rfloor = 93$ down-sampled time frames per frequency. The plane-wave decomposition of the ambient signals was based on selecting the $L = 60$ nearest measurements for the directions corresponding to a minimum t-design [62] of degree 10. The decorrelation of $\hat{\mathbf{z}}$, prior to re-encoding them in Equation (27), was conducted based on directly randomising their phase uniformly in the range $[-\pi, \pi)$. In cases where two DoA estimates fell within the same $\pi/(2\sqrt{Q})$ angle, one of the DoA estimates was randomly omitted in order to improve the stability of the employed beamforming solution. The beamformers also used $\beta = 0.01\ \mathbf{tr}[\mathbf{C_x}]$ as the regularisation term. Note that all $V = 841$ measurement directions were also used when computing $\mathbf{D}$, and for the grid-scanning conducted by the DoA estimator described in Section V-C.

The first objective metrics of interest relate to the parameter analysis performance, which refers to the method's ability to correctly detect the true number of sources and estimate their true DoAs. This was conducted based on computing the root-mean-square-error (RMSE) values as

$$\mathrm{RMSE}_K = \sqrt{\frac{1}{N_f}\sum_{f=1}^{N_f}|K(f) - \hat{K}(f)|^2}, \tag{30}$$

$$\mathrm{RMSE}_{DoA} = \sqrt{\frac{1}{N_f}\sum_{f=1}^{N_f}\left|\cos^{-1}\mathbf{u}^{\mathrm{T}}(f)\hat{\mathbf{u}}(f)\right|^2}, \tag{31}$$

where N_f refers to the employed number of frequency bins (up to the 12 kHz simulation limit), K is the true number of sources, $\mathbf{u}$ is the true source direction in Cartesian coordinates of unit length, and $\hat{K}$ and $\hat{\mathbf{u}}$ are the estimated source number and source direction vector, respectively. Note that in cases where more than one DoA estimate was made, the error metric was computed for all combinations between the estimates and ground truths and the lowest $\min(\hat{K}, K)$ error values were selected, followed by taking the mean to obtain a combined average. In total, 1000 iterations of randomised source directions were simulated, in order to obtain one averaged error value for each source number (up to $K = 3$) and DDR combination.

Perceptually motivated objective metrics were also computed, in order to evaluate how accurately the proposed method synthesises the target SH signals; given a binaural rendering workflow. The metrics were based on first linearly decoding the SH signals to the binaural channels $\mathbf{z}_{\mathrm{bin}} \in \mathbb{C}^{2\times 1}$ as

$$\mathbf{z}_{\mathrm{bin}}(t, f) = \mathbf{D}_{\mathrm{bin}}(f)\mathbf{a}(t, f), \tag{32}$$

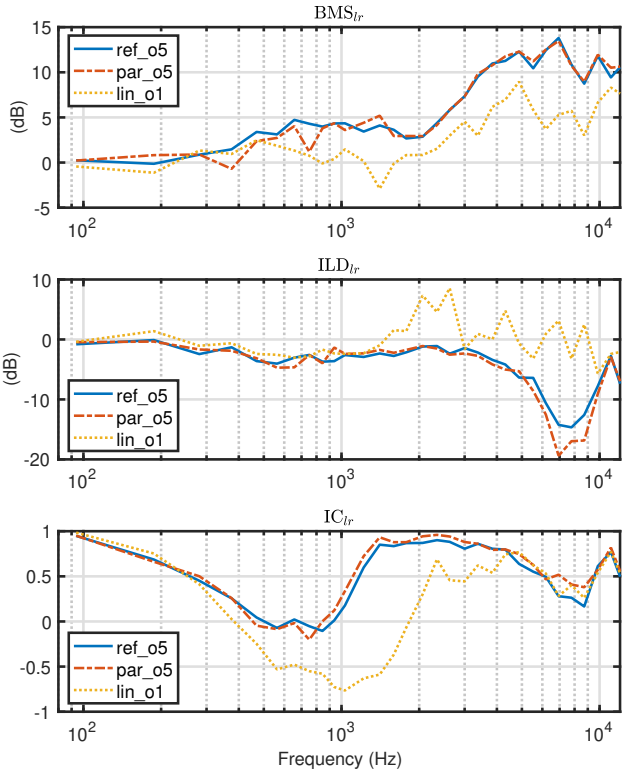

Fig. 6: Binaural metrics for a scene comprising two sources, one directly in-front and one directly to the left of the array in question, with a DDR of 6 dB, when using: the proposed method targeting fifth-order (*par_o5*), linear first-order encoding (*lin_o1*), and reference fifth-order encoding (*ref_o5*).

where $\mathbf{D}_{\mathrm{bin}} \in \mathbb{C}^{2\times(N+1)^2}$ denotes a frequency-dependent binaural decoding matrix. Note that the magnitude least-squares design proposed in [7] was employed for this task. The binaural SCM is then given by

$$\mathbf{C}_{\mathrm{bin}}(f) = \begin{pmatrix} c_{z_{ll}}(f) & c_{z_{lr}}(f) \\ c_{z_{rl}}(f) & c_{z_{rr}}(f) \end{pmatrix} = \mathbb{E}[\mathbf{z}_{\mathrm{bin}}(t, f)\mathbf{z}_{\mathrm{bin}}^{\mathrm{H}}(t, f)], \tag{33}$$

from which the following binaural metrics can be computed:

$$\mathrm{BMS}_{lr}(f) = 10\,\log_{10}[c_{z_{ll}}(f) + c_{z_{rr}}(f)], \tag{34}$$

$$\mathrm{ILD}_{lr}(f) = 10\,\log_{10}[c_{z_{ll}}(f)/c_{z_{rr}}(f)], \tag{35}$$

$$\mathrm{IC}_{lr}(f) = \frac{\mathrm{real}[c_{z_{lr}}(f)]}{\sqrt{c_{z_{ll}}(f)c_{z_{rr}}(f)}}, \tag{36}$$

where BMS_{lr} is the binaural mean spectrum (BMS), which corresponds to the timbral colouration of the encoding and decoding processing; ILD_{lr} is the inter-aural level difference (ILD) between the left and right ears, which relates directly to the inter-channel level differences between the two binaural channels; and IC_{lr} is the inter-aural coherence (IC), which relates directly to the inter-channel coherence. Note that an example of these binaural metrics for one scenario is depicted in Fig. 6.

These binaural metrics were computed based on: the array signals parametrically encoded into fifth-order SH signals using the proposed method, the array signals linearly encoded

to first-order SH signals following Equation (2) (with diffuse-field equalisation above the spatial aliasing limit as described by Equation (3)), and a fifth-order SH reference based on directly encoding the source and diffuse signals used to simulate the array recording. Note that all $V = 841$ measurement directions were used to compute $\mathbf{E}$ (with $\beta = 0.3$). The error values for the three binaural metrics, RMSE_{BMS}, RMSE_{ILD}, RMSE_{IC}, were then calculated in a similar manner to Equation (30), using the metric values derived from the binaural decoding of the reference fifth-order SH encoding as the true values. The metrics were also computed and averaged over 1000 iterations of random source directions. However, contrary to the parameter analysis evaluation, the metrics were averaged over frequency using the perceptually-motivated equivalent rectangular bandwidths (ERB) scale.

B. Perceptual evaluation

A multiple-stimulus binaural listening test was also conducted in order to evaluate the perceptual encoding performance of the proposed method, given a binaural rendering workflow. Note that, contrary to parts of the objective evaluations, these perceptual evaluations were conducted based solely on estimated spatial parameters. For the implementation of the proposed method[3] used for the listening tests: the sampling rate, the $L = 60$ directions for the residual rendering, the employed culling scheme for the DoA estimates, and the beamformer regularisation term, were all configured to be the same as in Section VII-A. Whereas: the time-frequency transform, temporal averaging of $\mathbf{C}_\mathbf{x}$, the updating of the spatial parameters and mixing matrices, and the decorrelation approach, were instead altered to better suit the dynamic sound scenes used for the listening test. The employed time-frequency transform was the $90\,\%$ overlap alias-free STFT design[4] described in [63], which was configured to use a hop size of 128 samples, with the hybrid filtering feature enabled; thus providing 133 frequency bands in total. The temporal averaging of the array SCM was conducted in blocks, based on combining the current block of 2048 time-domain samples with the previous block of 2048 samples; thereby averaging $\mathbf{C}_\mathbf{x}$ over $4096/128 = 32$ down-sampled time frames per frequency band. The proposed spatial analysis and synthesis were then updated and applied for each block of 2048 time-domain samples. Signal decorrelation was conducted based on assigning random delays per channel and per frequency band, with longer delays employed at lower frequencies and shorter delays at high frequencies; as used previously for similar studies conducted by the present authors [23], [29], [31].

To create the listening test scenes, three different contrasting sets of four source stimuli were first selected: 1) a four-piece funk band, 2) four simultaneous speakers, and 3) a mixed source scenario comprising a piano, speech, a water fountain, and clapping. Since the array in question was simulated up to $12\,\mathrm{kHz}$, all stimuli were low-pass filtered at $12\,\mathrm{kHz}$. These filtered stimuli were then directly convolved with the array measurements corresponding to fixed directions $[0, 0; 90, 0; -90, 0; 45, 50;]$ degrees (azimuth, elevation) and summed, in order to obtain a simulated array recording of the anechoic sound scene. The stimuli were also directly encoded into fifth-order SH signals in these same directions, in order to serve as the anechoic reference case. To also include a more realistic acoustical environment, a shoe-box room simulator[5], based on the image-source method, was employed. The wall absorption coefficients were configured in octave bands, to obtain reverberation times (RT60) of $[0.5, 0.55, 0.5, 0.35, 0.2, 0.15]\,\mathrm{s}$ ($125\,\mathrm{Hz}$ to $4\,\mathrm{kHz}$) for a $[10 \times 7 \times 4]\,\mathrm{m}$ (Width $\times$ Depth $\times$ Height) sized room. The receiver position was set to the centre of the room, with the four source positions set in the same directions as with the anechoic case, $1\,\mathrm{m}$ away from the receiver. The direct paths and modelled room reflections were then quantised to the employed $V = 841$ measurement grid and directly convolved with the respective array measurements, in order to obtain a simulated array recording of the reverberant scene. The direct path and reflections were also directly encoded into fifth-order SH signals, which served as the reverberant reference test case.

The simulated array recordings of the aforementioned sound scenes were subsequently encoded into fifth-order SH signals using the proposed parametric (*IA_par_o5*) method, and also into first-order SH signals using the conventional linear (*IA_lin_o1*) approach, as described by Equation (2). As an additional control condition, a tetrahedral array of cardioid-pattern sensors with a radius of $2\,\mathrm{cm}$, as commonly employed for ambisonic recording in practice, was also used to obtain simulated recordings and encoded into fifth-order SH using the proposed method (*tetra_par_o5*). Note that this tetrahedral array was simulated based on analytical descriptors [45], [46] for the same $V = 841$ directions, in order to have parity with the grid used to simulate the array in question. This condition was intended to reveal any improvements of the proposed method when using an array type that is commercially and widely available, and often employed for capturing first-order linearly encoded recordings (*tetra_lin_o1*). Additionally, this SMA may demonstrate differences between the method applied to the irregular array under study, and a more regular

TABLE I: Listening test scenes.

Name	Room	Source stimuli
band_dry	Anechoic	bass guitar, drums, shaker, strings
band_rev	Reverberant	bass guitar, drums, shaker, strings
speech_dry	Anechoic	two male and two female speakers
speech_rev	Reverberant	two male and two female speakers
mix_dry	Anechoic	clapping, water fountain, piano, speech
mix_rev	Reverberant	clapping, water fountain, piano, speech

TABLE II: Listening test cases.

Name	Array	Encoding method
hidden_ref_o5	Ideal SH receiver	Direct fifth-order
IA_par_o5	Irregular array in question	Proposed fifth-order
IA_lin_o1	Irregular array in question	Conventional first-order
tetra_par_o5	Open tetrahedral array	Proposed fifth-order
tetra_lin_o1	Open tetrahedral array	Conventional first-order

[3]Much of the implementation of the proposed method was based on MATLAB code found here: https://github.com/polarch/COMPASS-ref

[4]The employed alias-free STFT design may be found here: https://github.com/jvilkamo/afSTFT

[5]The shoe-box room simulator utilised in this study may be found here: https://github.com/polarch/shoebox-roomsim

array that exhibits a uniform spatial resolution. All encoded SH signals and the reference SH signals were then decoded to the binaural channels using the magnitude least-squares method proposed in [7].

In total, there were six test scenes, as summarised by Table I, and five test cases, as summarised in Table II. The listening test was then conducted in three parts:

- **Spatial**: where the test cases were frequency-dependently equalised to the reference case. The listening subjects were then instructed to assess the test cases based on their spatial accuracy, and ignore any remaining timbral differences.
- **Timbre**: where the magnitude response of each test case was imposed onto the reference case, therefore ensuring that all the test cases presented were spatially equivalent. The listening subjects were then instructed to rate the cases based only on timbral differences.
- **Overall**: test cases were simply normalised to the reference based on their average broad-band root-mean-square signals powers. The listening subjects were then asked to rate the cases based on personal preference.

Fourteen subjects participated in the listening test, all of whom were naive as to the hypothesis of the study, reported having normal hearing, and had previous experience participating in perceptual studies. The scale of the listening test was set between 0 and 100, and had the verbal anchors: bad, poor, fair, good, and excellent between the respective 20 point intervals. The test subjects were instructed to rate each test case with respect to the reference, and relative to each other, while in consideration of the specific perceptual attribute under test (spatial, timbre, or overall). The average length for completing all three parts of the test was approximately 40 minutes. The tests were conducted in specially-built sound dampened listening booths (background noise level of $\mathrm{L_{A,eq,30\,s}} = 22.0\,\mathrm{dB}$ SPL(A)) located at Aalto University, using Sennheiser HD600 headphones.

VIII. Results and Discussion

The results for the objective parameter analysis evaluation are presented in Fig. 7. It can be observed that, with the exception of the 3 sources and 0 dB DDR case, the RMSE_{DoA} errors remain quite consistent; even as more sound sources are introduced into the simulation. The standard deviations are high, which is likely a product of the irregular array geometry and non-uniform sensor placements, but are otherwise consistent across the different numbers of sources and DDR values. The error and standard deviation for the 3 sources case at 0 dB DDR, however, are notably higher and wider; although, it is highlighted that this represented the most challenging case that was tested. The perceptual ramifications of these estimation errors, however, may be more suitably inferred from the results of listening tests described below. Regarding the evaluation of RMSE_K, given positive DDR values the errors were found to be low and the standard deviations are narrow; suggesting that the source number estimator is suitable for detecting sources within moderate to low energy diffuse-fields. Whereas, in the 0 dB DDR case, the errors indicate that the employed source

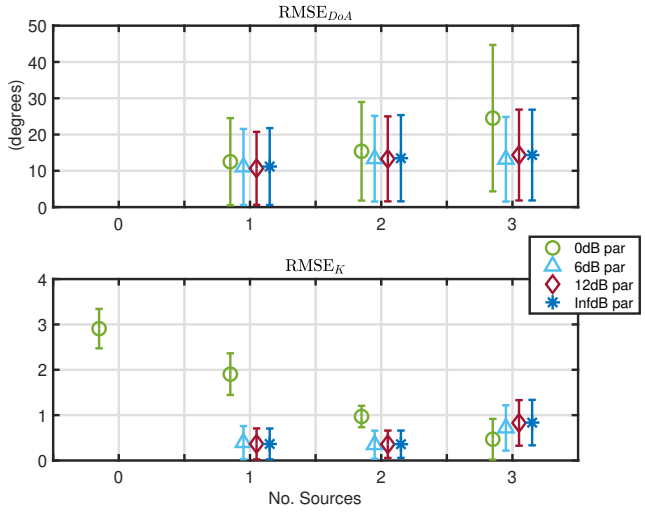

Fig. 7: RMSE and standard deviations results for the objective spatial analysis evaluation, which were averaged over frequency bins between $[0, 12]$ kHz and 1000 iterations of randomly selected source directions.

number estimator may over-estimate, or is otherwise unable to reliably detect the true number of sources. This 0 dB DDR issue may have also influenced the following objective binaural metrics results to some degree.

The results for the binaural metrics evaluations are shown in Fig. 8, using both the analysed parameters (left) and the known/Oracle spatial parameters (right). For both the analysed parameters and Oracle cases, it can be observed that the proposed parametric encoding yields lower RMSE values for all DDR values that are above 0 dB, and for all three binaural metrics, when compared to the linearly encoded baseline. However, for the 0 dB DDR cases, the error is higher, especially for the purely diffuse ($K = 0$) case, when using the estimated spatial parameters. The error for this particular case is significantly lower when using the Oracle parameters, thus suggesting that the aforementioned issues regarding the employed source number estimator may be to the detriment of the overall encoding method for such conditions. Therefore, the proposed method could benefit from the addition of a diffuse-conditions detector, which would allow the source number detector to be bypassed (i.e. force $K = 0$) in cases where the sound-field is analysed to be highly diffuse. A topic of future work could therefore involve investigating the use of such detectors; for example, the estimator described in [64] may be suitable for this task, provided that spatial whitening of the SCM is conducted, as described by Equation (12), and with the selection of an appropriate threshold value.

The results for the multiple stimulus listening test are presented in Fig. 9. The parametric rendering was rated notably higher than the linear signal-independent encoding in terms of both the spatial and timbral attributes, and also based on the overall preference of the listeners. The hidden references were consistently assigned scores near to 100, whereas the linearly encoded irregular array was likely interpreted as a low quality anchor and rated near to 0. The linearly encoded tetrahedral

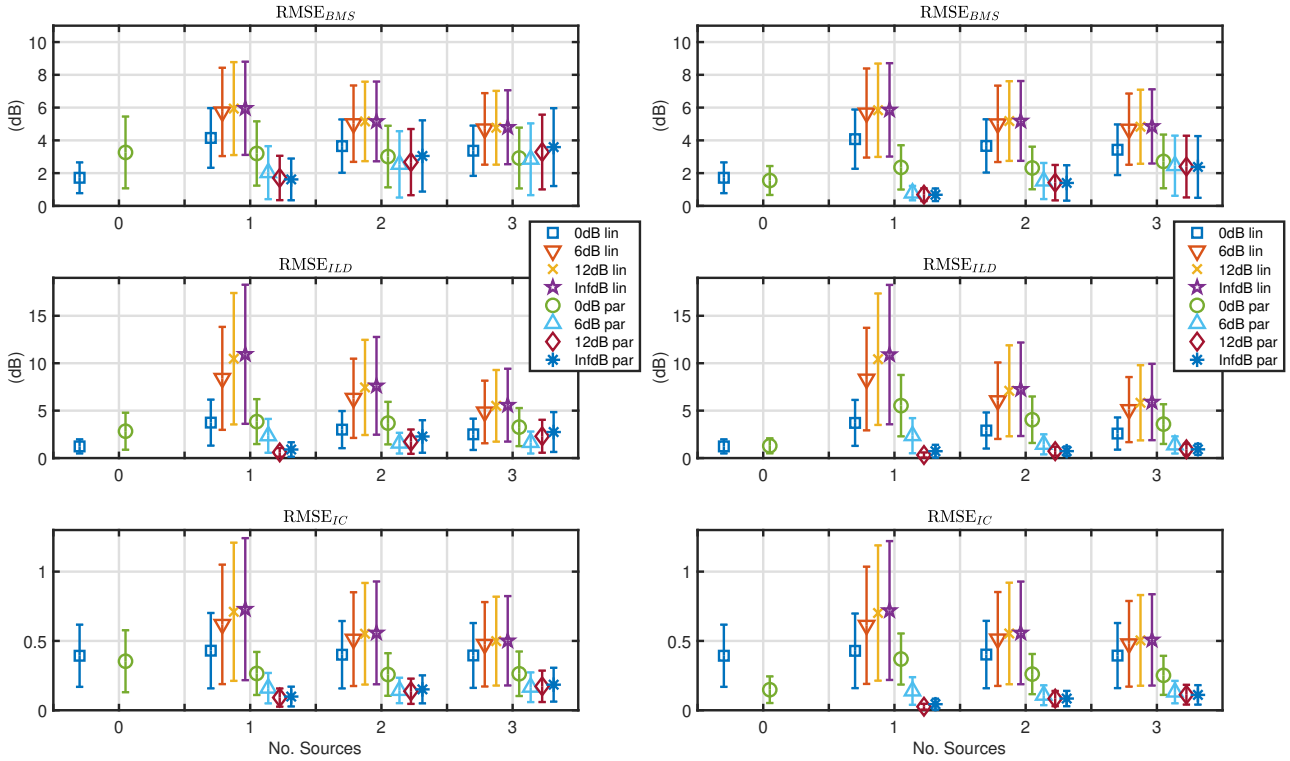

Fig. 8: RMSE and standard deviations results for the objective binaural metrics evaluation, computed based on the ideal fifth-order reference. Averaged in ERB frequency bands (up to 12 kHz), and 1000 iterations of random source directions. Left: using the parametric analysis, right: with known parameters (Oracle).

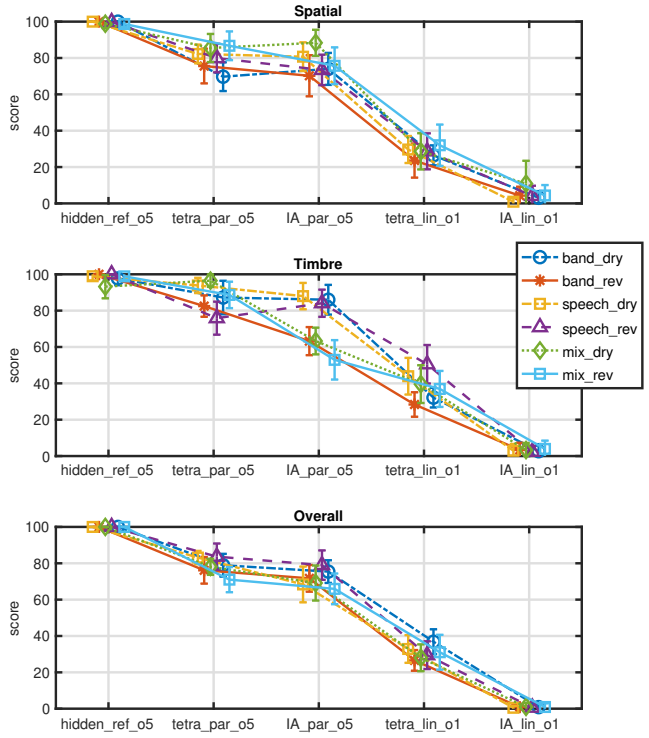

Fig. 9: Means and 95% confidence intervals for the listening test results, based on fourteen participants.

array fared better than the linearly encoded irregular array, which is likely a result of its uniform arrangement of sensors and smaller radius, which achieves a direction-independent spatial aliasing frequency of approximately 6 kHz; rather than the direction-dependent approximate 1 kHz spatial aliasing limit exhibited by the irregular array. For the spatial part of the listening test, the proposed parametrically encoded array signals for both arrays performed similarly, and were assigned scores within the good and excellent verbal anchors. The timbral part of the test indicated that the irregular array introduced noticeable timbral colourations for certain sound scenes, since they were rated lower than the parametrically encoded tetrahedral array; notably, both of the *mixture* scenes were rated lower. However, it should be highlighted that broad-band transient stimuli (such as clapping) typically require a responsive analysis for an adequate parameterisation and rendering, and such sounds tend to more readily reveal any artefacts arising due to signal decorrelation. Whereas the broad-band noise source (the waterfall) and musical source (piano) instead benefit from longer temporal averaging windows. Therefore, this particular sound scene may be considered especially challenging, since there are conflicting configuration requirements for the various contrasting source signals. However, the results for the overall part of the test suggest that the spatial attributes of the proposed encoding approach were more favoured by the test participants compared to the timbral attributes, since the overall scores were more inline with those of the spatial part of the listening test.

IX. Conclusion

This article proposes a parametric signal-dependent method for encoding the signals of an array of microphones into Ambisonic signals. The method is highly general by design, and is intended to yield improved performance over conventional linear signal-independent encoding, especially when employing irregular microphone array geometries and/or non-uniform microphone placements. The proposed method conducts a multi-directional parameterisation of the captured sound scene, and employs spatial filtering to divide the scene into its individual source and directional ambient components. The source components are then encoded into the Ambisonics domain at an arbitrary output order. The ambient components are first projected onto a uniform spherical arrangement of points, optionally decorrelated, and then encoded at the same target output order. The output ambisonic signals are then obtained by summing these two streams.

The proposed method was evaluated in the context of binaurally decoding ambisonic signals, which were obtained by encoding simulated recordings of a non-uniform arrangement of seven microphones affixed to a head-mounted display worn by a manikin. The evaluation was based on first analysing objective binaural metrics. Here, the objective binaural cues were computed based on first targeting fifth-order ambisonic output using the proposed parametric method and first-order using conventional linear signal-independent encoding, followed by decoding them to the binaural channels. The objective binaural cues were then compared against those derived from a fifth-order directly encoded reference case. It was found that the proposed encoding method outperformed conventional linear Ambisonic encoding for all of the scenarios tested, where the direct-to-diffuse ratio was above 0 dB. For the 0 dB case, the improvement in performance of the proposed method, compared to the linear encoding, was less apparent. However, when substituting the processing with known spatial parameters, the computed error values of the proposed method were either similar to, or lower than, the linearly encoded baseline. This therefore suggests that there is room for further improvements in the proposed spatial analysis for such conditions. The proposed method was then evaluated based on formal listening tests. It was found that the test subjects rated the parametrically encoded fifth-order cases to be perceptually closer to ideal/reference fifth-order cases, when decoded to the binaural channels and compared against first-order linearly encoded and decoded baseline cases. These improved results hold for both the perceived spatial and timbral attributes for a number of sound scenes, comprising a diverse range of different source stimuli for both anechoic and reverberant environments.

Acknowledgments

This research has received funding from the Aalto University Doctoral School of Electrical Engineering. The authors also extend their thanks to the reviewers for their insightful comments and suggestions during the review process.

References

[1] K. De Boer, "Stereophonic sound reproduction," *Philips Technical Review*, vol. 5, pp. 107–114, 1940.

[2] A. Fukada, "A challenge in multichannel music recording (seminar presentation)," in *AES 19th Int. Conference: Surround Sound: Techniques, Technology, Perception*, 2001.

[3] K. Hamasaki and K. Hiyama, "Reproducing spatial impression with multichannel audio," in *Audio Engineering Society Conference: 24th International Conference: Multichannel Audio, The New Reality*. Audio Engineering Society, 2003.

[4] H. Lee, "Multichannel 3D microphone arrays: A review," *Journal of the Audio Engineering Society*, vol. 69, no. 1/2, pp. 5–26, 2021.

[5] M. A. Gerzon, "Periphony: With-height sound reproduction," *Journal of the audio engineering society*, vol. 21, no. 1, pp. 2–10, 1973.

[6] B. Rafaely, *Fundamentals of spherical array processing*. Springer, 2015, vol. 8.

[7] C. Schörkhuber, M. Zaunschirm, and R. Höldrich, "Binaural rendering of Ambisonic signals via magnitude least squares," in *Proc. DAGA*, vol. 44, 2018, pp. 339–342.

[8] F. Zotter and M. Frank, "All-round ambisonic panning and decoding," *Journal of the Audio Engineering Society*, vol. 60, no. 10, pp. 807–820, 2012.

[9] J. Chen, B. D. Van Veen, and K. E. Hecox, "External ear transfer function modeling: A beamforming approach," *The Journal of the Acoustical Society of America*, vol. 92, no. 4, pp. 1933–1944, 1992.

[10] E. Rasumow, M. Blau, S. Doclo, M. Hansen, D. Püschel, V. Mellert *et al.*, "Perceptual evaluation of individualized binaural reproduction using a virtual artificial head," *Journal of the Audio Engineering Society*, vol. 65, no. 6, pp. 448–459, 2017.

[11] J. Backman, "Microphone array beam forming for multichannel recording," in *Audio Engineering Society Convention 114*. Audio Engineering Society, 2003.

[12] Y. Hur, J. S. Abel, Y.-c. Park, and D. H. Youn, "A bank of beamformers implementing a constant-amplitude panning law," *The Journal of the Acoustical Society of America*, vol. 136, no. 3, pp. EL212–EL217, 2014.

[13] J. Ivanic and K. Ruedenberg, "Rotation matrices for real spherical harmonics. direct determination by recursion," *The Journal of Physical Chemistry A*, vol. 102, no. 45, pp. 9099–9100, 1998.

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

[15] O. Santala, H. Vertanen, J. Pekonen, J. Oksanen, and V. Pulkki, "Effect of listening room on audio quality in ambisonics reproduction," in *Audio Engineering Society Convention 126*. Audio Engineering Society, 2009.

[16] S. Braun and M. Frank, "Localization of 3d ambisonic recordings and ambisonic virtual sources," in *1st International Conference on Spatial Audio,(Detmold)*, 2011.

[17] A. Avni, J. Ahrens, M. Geier, S. Spors, H. Wierstorf, and B. Rafaely, "Spatial perception of sound fields recorded by spherical microphone arrays with varying spatial resolution," *The Journal of the Acoustical Society of America*, vol. 133, no. 5, pp. 2711–2721, 2013.

[18] S. Bertet, J. Daniel, E. Parizet, and O. Warusfel, "Investigation on localisation accuracy for first and higher order ambisonics reproduced sound sources," *Acta Acustica united with Acustica*, vol. 99, no. 4, pp. 642–657, 2013.

[19] P. Stitt, S. Bertet, and M. van Walstijn, "Off-centre localisation performance of ambisonics and HOA for large and small loudspeaker array radii," *Acta Acustica united with Acustica*, vol. 100, no. 5, pp. 937–944, 2014.

[20] V. Pulkki, A. Politis, M.-V. Laitinen, J. Vilkamo, and J. Ahonen, "First-order directional audio coding (DirAC)," in *Parametric Time-Frequency Domain Spatial Audio*, V. Pulkki, S. Delikaris-Manias, and A. Politis, Eds. John Wiley & Sons, 2017, pp. 89–138.

[21] F. J. Fahy and V. Salmon, "Sound intensity," *The Journal of the Acoustical Society of America*, vol. 88, no. 4, pp. 2044–2045, 1990.

[22] V. Pulkki, "Virtual sound source positioning using vector base amplitude panning," *Journal of the audio engineering society*, vol. 45, no. 6, pp. 456–466, 1997.

[23] A. Politis, J. Vilkamo, and V. Pulkki, "Sector-based parametric sound field reproduction in the spherical harmonic domain," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 5, pp. 852–866, 2015.

[24] A. Politis, L. McCormack, and V. Pulkki, "Enhancement of ambisonic binaural reproduction using directional audio coding with optimal adaptive mixing," in *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2017, pp. 379–383.

[25] A. Politis and V. Pulkki, "Acoustic intensity, energy-density and diffuseness estimation in a directionally-constrained region," *arXiv preprint arXiv:1609.03409*, 2016.

[26] L. McCormack, S. Delikaris-Manias, A. Politis, D. Pavlidi, A. Farina, D. Pinardi, and V. Pulkki, "Applications of spatially localized active-intensity vectors for sound-field visualization," *J. Audio Engineering Society*, vol. 67, no. 11, pp. 840–854, 2019.

[27] S. Berge and N. Barrett, "High angular resolution planewave expansion," in *Proc. of the 2nd Int. Symp. on Ambisonics and Spherical Acoustics*, 2010, pp. 6–7.

[28] A. Wabnitz, N. Epain, and C. T. Jin, "A frequency-domain algorithm to upscale ambisonic sound scenes," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012, pp. 385–388.

[29] A. Politis, S. Tervo, and V. Pulkki, "COMPASS: Coding and multi-directional parameterization of ambisonic sound scenes," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 6802–6806.

[30] K. Han and A. Nehorai, "Improved source number detection and direction estimation with nested arrays and ULAs using jackknifing," *IEEE Trans. Signal Processing*, vol. 61, no. 23, pp. 6118–6128, 2013.

[31] L. McCormack and S. Delikaris-Manias, "Parametric first-order ambisonic decoding for headphones utilising the cross-pattern coherence algorithm," in *EAA Spatial Audio Signal Processing Symposium*, 2019, pp. 173–178.

[32] C. Schörkhuber and R. Höldrich, "Linearly and quadratically constrained least-squares decoder for signal-dependent binaural rendering of ambisonic signals," in *Audio Engineering Society Conference: 2019 AES International Conference on Immersive and Interactive Audio*. Audio Engineering Society, 2019.

[33] P. M. Giller and C. Schörkhuber, "A super-resolution ambisonics-to-binaural rendering plug-in," *Fortschritte der Akustik-DAGA, Rostock*, 2019.

[34] C. Schörkhuber and R. Höldrich, "Signal-dependent encoding for first-order ambisonic microphones," *Fortschritte der Akustik, DAGA, Kiel*, pp. 1037–1040, 2017.

[35] J. Lin, X. Wu, and T. Qu, "Anti spatial aliasing HOA encoding method based on aliasing projection matrix," in *2020 IEEE 3rd International Conference on Information Communication and Signal Processing (ICICSP)*. IEEE, 2020, pp. 321–325.

[36] A. Politis, S. Tervo, T. Lokki, and V. Pulkki, "Parametric multidirectional decomposition of microphone recordings for broadband high-order ambisonic encoding," in *Audio Engineering Society Convention 144*. Audio Engineering Society, 2018.

[37] S. Moreau, J. Daniel, and S. Bertet, "3D sound field recording with higher order ambisonics–objective measurements and validation of a 4th order spherical microphone," in *120th Convention of the AES*, 2006, pp. 20–23.

[38] P. Calamia, S. Davis, C. Smalt, and C. Weston, "A conformal, helmet-mounted microphone array for auditory situational awareness and hearing protection," in *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2017, pp. 96–100.

[39] R. M. Corey, N. Tsuda, and A. C. Singer, "Acoustic impulse responses for wearable audio devices," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 216–220.

[40] J. Donley, V. Tourbabin, J.-S. Lee, M. Broyles, H. Jiang, J. Shen, M. Pantic, V. K. Ithapu, and R. Mehra, "Easycom: An augmented reality dataset to support algorithms for easy communication in noisy environments," *arXiv preprint arXiv:2107.04174*, 2021.

[41] J. Ahrens, H. Helmholz, D. L. Alon, and S. V. A. Garí, "A head-mounted microphone array for binaural rendering," in *2021 Immersive and 3D Audio: from Architecture to Automotive (I3DA)*. IEEE, 2021, pp. 1–7.

[42] K. Nagatomo, M. Yasuda, K. Yatabe, S. Saito, and Y. Oikawa, "Wearable seld dataset: Dataset for sound event localization and detection using wearable devices around head," *arXiv preprint arXiv:2202.08458*, 2022.

[43] 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," *The Journal of the Acoustical Society of America*, vol. 151, no. 4, pp. 2624–2635, 2022.

[44] A. Politis and H. Gamper, "Comparing modeled and measurement-based spherical harmonic encoding filters for spherical microphone arrays," in *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2017, pp. 224–228.

[45] E. G. Williams, *Fourier acoustics: sound radiation and nearfield acoustical holography*. Academic press, 1999.

[46] H. Teutsch, *Modal array signal processing: principles and applications of acoustic wavefield decomposition*. Springer, 2007, vol. 348.

[47] C. T. Jin, N. Epain, and A. Parthy, "Design, optimization and evaluation of a dual-radius spherical microphone array," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 22, no. 1, pp. 193–204, 2013.

[48] C. Schörkhuber and R. Höldrich, "Ambisonic microphone encoding with covariance constraint," in *Proceedings of the International Conference on Spatial Audio*, 2017, pp. 7–10.

[49] M. A. Gerzon, "The design of precisely coincident microphone arrays for stereo and surround sound," in *Audio Engineering Society Convention 50*. Audio Engineering Society, 1975.

[50] J. Fliege and U. Maier, "The distribution of points on the sphere and corresponding cubature formulae," *IMA Journal of Numerical Analysis*, vol. 19, no. 2, pp. 317–334, 1999.

[51] H. Cox, R. Zeskind, and T. Kooij, "Practical supergain," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, no. 3, pp. 393–398, 1986.

[52] W. Chen, K. M. Wong, and J. P. Reilly, "Detection of the number of signals: A predicted eigen-threshold approach," *IEEE Trans. Signal Processing*, vol. 39, no. 5, pp. 1088–1098, 1991.

[53] J.-S. Jiang and M. A. Ingram, "Robust detection of number of sources using the transformed rotational matrix," in *2004 IEEE Wireless Communications and Networking Conference (IEEE Cat. No. 04TH8733)*, vol. 1. IEEE, 2004, pp. 501–506.

[54] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Automatic Control*, vol. 19, no. 6, pp. 716–723, 1974.

[55] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust localization in reverberant rooms," in *Microphone arrays*. Springer, 2001, pp. 157–180.

[56] O. L. Frost, "An algorithm for linearly constrained adaptive array processing," *Proceedings of the IEEE*, vol. 60, no. 8, pp. 926–935, 1972.

[57] A. Barabell, "Improving the resolution performance of eigenstructure-based direction-finding algorithms," in *ICASSP'83. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 8. IEEE, 1983, pp. 336–339.

[58] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE transactions on antennas and propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[59] M. Kaveh and A. Barabell, "The statistical performance of the music and the minimum-norm algorithms in resolving plane waves in noise," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, no. 2, pp. 331–341, 1986.

[60] F. Zotter, H. Pomberger, and M. Noisternig, "Energy-preserving ambisonic decoding," *Acta Acustica united with Acustica*, vol. 98, no. 1, pp. 37–47, 2012.

[61] 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)*, 2021, pp. 214–221.

[62] R. H. Hardin and N. J. Sloane, "Mclaren's improved snub cube and other new spherical designs in three dimensions," *Discrete & Computational Geometry*, vol. 15, no. 4, pp. 429–441, 1996.

[63] J. Vilkamo and T. Bäckstrom, "Time-frequency processing: Methods and tools," in *Parametric Time-Frequency Domain Spatial Audio*, V. Pulkki, S. Delikaris-Manias, and A. Politis, Eds. John Wiley & Sons, 2017, pp. 1–24.

[64] N. Epain and C. T. Jin, "Spherical harmonic signal covariance and sound field diffuseness," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 10, pp. 1796–1807, 2016.