Audio Engineering Society

Conference Paper

Presented at the 2022 International Conference on
Audio for Virtual and Augmented Reality
2022 August 15–17, Redmond, WA, USA

Parametric Ambisonic Encoding using a Microphone Array with a One-plus-Three Configuration

Leo McCormack[1], Raimundo Gonzalez[1,2], Janani Fernandez[1], Christoph Hold[1], and Archontis Politis[3]

[1] *Department of Signal Processing and Acoustics, Aalto University, Espoo, Finland*
[2] *Department of Neuroscience and Biomedical Engineering, Aalto University, Espoo, Finland*
[3] *Faculty of Information Technology and Communication Sciences, Tampere University, Finland*

Correspondence should be addressed to Leo McCormack (`leo.mccormack@aalto.fi`)

ABSTRACT

A parametric signal-dependent method is proposed for the task of encoding a studio omnidirectional microphone signal into the Ambisonics format. This is realised by affixing three additional sensors to the surface of the cylindrical microphone casing; representing a practical solution for imparting spatial audio recording capabilities onto an otherwise non-spatial audio compliant microphone. The one-plus-three configuration and parametric encoding method were evaluated through formal listening tests using simulated sound scenes and array recordings, given a binaural decoding workflow. The results indicate that, when compared to employing first-order signals obtained linearly using an open tetrahedral array, or third-order signals derived from a 19-sensor spherical array, the proposed system is able to produce perceptually closer renderings to those obtained using ideal third-order signals.

1 Introduction

In recent years, the Ambisonics format [1, 2] has gained renewed interest and is being increasingly adopted in the fields of immersive audio recording and reproduction. This is largely due to the rising popularity of emerging augmented and virtual reality (AR/VR) devices, where the format is well positioned due to: its ability to decouple the recording setup from the playback setup, and for its trivial sound-field rotation capabilities. Ambisonic processing pipelines are divided into three main stages: 1) the *encoding* stage, where microphone array signals capturing real sound scenes, and/or monophonic material accompanied by metadata (describing synthetic sound scenes), are converted into Ambisonic/spherical harmonic (SH) signals; 2) an optional *modification* stage, where linear [3] or parametric [4] spatial transformations may be applied to the Ambisonic sound scene, such as sound-field rotations, warping, translations, or directional loudness modifications; and finally 3) a *decoding* stage, where the Ambisonic sound scene is reproduced over the target playback setup (i.e. over headphones or over a multi-channel loudspeaker setup).

The focus of this paper concerns Ambisonic encoding using real microphone array recordings as input. Typically, spherical microphone arrays (SMAs) are employed for this task, due to their consistent spatial resolution for all directions on the sphere and their linear encoding convenience [5]. The most popular SMA configuration is an open tetrahedral arrangement of directional (typically cardioid) sensors, which is suitable for capturing up to first-order SH signals. However, subsequent decoding of lower-order recordings has been shown to lead to localisation ambiguity, and strong colourations due to high coherence between the loudspeaker channels [6, 7]. Therefore, arrays capable of higher-order Ambisonics capture, which can alleviate such problems, may be more preferable. However, due to the high costs associated with producing multiple phase-matched cardioid capsules, larger/higher-order SMAs typically employ cheaper omnidirectional sensors mounted onto a rigid spherical baffle. This baffle serves to force cardioid-like directivities and to also neatly enclose the electrical components and wiring into one compact package. There are a few commercially available rigid SMAs targeting up to fourth-order SH recording. However, the band-width of usable SH components can be narrow at higher-orders, due to spatial aliasing at high frequencies, and the need to regularise filter inversions to mitigate amplifying sensor noise at low frequencies. Such arrays also employ dozens of sensors and thus their associated costs can still be high.

With the Ambisonics format continuing to gain traction in AR/VR applications, there may arise a need for cheaper and more accessible Ambisonics recording solutions to be introduced to this growing audience. Therefore, in this paper, a signal-dependent ambisonic encoding approach is presented, which is intended to operate based upon: an existing studio omnidirectional microphone, which has additional microphones affixed to the microphone casing. A sound-field model comprising a single source mixed with diffuse sounds is then employed, as used, for example, by the Directional Audio Coding (DirAC) method [8]. However, contrary to DirAC, which is a well-known parametric ambisonic decoding method, the focus of the present study is in regard to ambisonic encoding of sound scenes that are recorded using cost effective hardware. Therefore, the spatial parameters are estimated, and the necessary spatial filtering techniques are formulated, directly in the space-domain; in a similar manner to [9, 10]. With the assumption that the additional sensors may be of lower-quality, sounds that are analysed as having clear directionality (i.e. corresponding to distinct sound sources) are encoded into the Ambisonics format using only the studio omnidirectional microphone signal. The remaining sound components are then assumed to correspond to diffuse sounds. It is recognised that decorrelating the omnidirectional microphone signal multiple times to reproduce such components may, in any case, incur signal quality degradation; therefore, all sensors in the array are used for encoding the diffuse components. Constraints to further minimise the amount of decorrelation required to conform to the assumed sound-field model are also applied [11]. The proposed microphone array configuration and parametric encoding method is then evaluated based on simulated array steering vectors and image-source based room simulations. The proposed system targeting third-order is compared alongside traditional linear ambisonic encodings of an open tetrahedral SMA with cardioid sensors (first-order) and a 19-channel rigid SMA (third-order), which are all decoded for playback over headphones[1] and rated with respect to binaural decodings of perfect/synthetic Ambisonic recordings (third-order).

2 Signal model

It is first assumed that the input Q-channel microphone array signals $\mathbf{x}(t,f) = [x_1(t,f),...,x_Q(t,f)] \in \mathbb{C}^{Q\times 1}$ have been transformed into the time-frequency domain, where t and f denote the down-sampled time index and frequency, respectively. It is also assumed that the first channel x_1 will represent a signal corresponding to an omnidirectional sensor, whereas the remaining $Q-1$ sensor signals may exhibit arbitrary directivities and the sensors themselves may, in practice, be constructed using lower quality components.

The employed signal model assumes that the captured microphone array signals may comprise either a single source, $s(t,f)$, an ambient component encapsulating diffuse noise and reverberation, $\mathbf{d}(t,f) \in \mathbb{C}^{Q\times 1}$, or a combination of the two:

$$\mathbf{x}(t,f) = \mathbf{a}(\boldsymbol{\gamma},f)s(t,f) + \mathbf{d}(t,f) + \mathbf{n}(t,f), \quad (1)$$

where $\mathbf{a}(\boldsymbol{\gamma},f) \in \mathbb{C}^{Q\times 1}$ is the array steering vector for direction $\boldsymbol{\gamma}$, and $\mathbf{n}(t,f) \in \mathbb{C}^{Q\times 1}$ is the array sensor noise, which is assumed to be uncorrelated across sensors. It is henceforth assumed that the array steering vectors $\mathbf{A} = [\mathbf{a}(\boldsymbol{\gamma}_1,f),...,\mathbf{a}(\boldsymbol{\gamma}_V,f)] \in \mathbb{C}^{Q\times V}$ are available for a dense spherical grid of V directions, $\boldsymbol{\Gamma}_V = [\boldsymbol{\gamma}_1,...,\boldsymbol{\gamma}_V]$, uniformly distributed over the sphere. These array steering vectors may be obtained through free-field measurements or simulations of the microphone array in question.

[1] Binaural audio examples, using the proposed one-plus-three array configuration and encoding method, may be found here: `https://doi.org/10.5281/zenodo.6501944`

3 Parametric spatial analysis

The first stage of the proposed encoding method involves the estimation of the direction-of-arrival (DoA) of a dominant source, $\boldsymbol{\gamma}_{\text{DoA}}$, and a diffuseness parameter $\psi \in [0,1]$, over time and frequency. This spatial parameter estimation is conducted based on the subspace principles applied to the array spatial covariance matrix (SCM), which is first obtained as

$$\mathbf{C}_{\text{x}}(f) = \mathscr{E}[\mathbf{x}(t,f)\mathbf{x}^{\text{H}}(t,f)], \tag{2}$$

where $\mathscr{E}[.]$ denotes the expectation operator.

3.1 Frequency-averaging of the array SCMs

Parametric spatial audio algorithms are often implemented using a short-time Fourier transform (STFT) or a filter-bank, such as the Quadrature-Mirror filter-bank (QMF), which provide a uniform frequency resolution. However, given the target applications, it may be more desirable to perform the parametric analysis with a perceptually motivated frequency resolution; such as the equivalent-rectangular bandwidth (ERB) scale [12]. In this case, the SCMs for frequency bands corresponding to each ERB grouping may be averaged as

$$\mathbf{C}_{\text{x}}^{(\text{ERB})}(f_0) = \sum_{f \in \mathbf{f}_{\text{ERB}}} \mathbf{T}(f,f_0)\mathbf{C}_{\text{x}}(f)\mathbf{T}^{\text{H}}(f,f_0), \tag{3}$$

where f_0 is the nearest uniformly spaced frequency index for each ERB frequency, $\mathbf{f}_{\text{ERB}}$ are all of the frequency indices corresponding to the ERB band grouping, and [13, 14]

$$\mathbf{T}(f,f_0) = [\mathbf{A}(f_0)\mathbf{Y}^{\text{T}}][\mathbf{A}(f)\mathbf{Y}^{\text{T}}]^{\dagger}, \tag{4}$$

is a matrix that transforms the array basis functions for frequency f to be aligned with those for frequency f_0, thus allowing the summation of array SCMs over the frequency grouping. Note that $\mathbf{Y} \in \mathbb{R}^{(N+1)^2 \times V}$ denotes SH weights of order N corresponding to the directions of the same dense grid used to measure or simulate the array steering vectors, and † denotes the Moore-Penrose pseudoinverse.

3.2 Spatial whitening of the array SCMs

Given only sensor noise as input, the frequency averaged array SCMs described by Equation (3) should exhibit an identity-like structure. It is based upon this principle that many traditional subspace-based spatial parameter estimation approaches operate, since they typically assume that sources are to be detected and localised in the presence of sensor noise. However, if one is to assume for the present applications that the energy of the diffuse array signals may be higher than the array sensor noise (i.e. $\text{Tr}\big[\mathscr{E}[\mathbf{d}(t,f)\mathbf{d}^{\text{H}}(t,f)]\big] >> \text{Tr}\big[\mathscr{E}[\mathbf{n}(t,f)\mathbf{n}^{\text{H}}(t,f)]\big]$), it may be more beneficial to have this identity-like structure in those cases where the array is presented with a diffuse-field. Therefore, in this work, the SCMs are also spatially whitened as

$$\hat{\mathbf{C}}_{\text{x}}^{(\text{ERB})}(f_0) = \mathbf{W}(f_0)\mathbf{C}_{\text{x}}^{(\text{ERB})}(f_0)\mathbf{W}^{\text{H}}(f_0). \tag{5}$$

where $\mathbf{W} \in \mathbb{C}^{Q \times Q}$ is the matrix performing the spatial whitening operation, which is computed as [9]

$$\mathbf{W}(f_0) = \mathbf{\Sigma}^{-1/2}(f_0)\mathbf{V}^{\text{H}}(f_0), \tag{6}$$

using the eigenvalue decomposition (EVD) of

$$\sum_{f \in \mathbf{f}_{\text{ERB}}} \mathbf{T}(f,f_0)\mathbf{A}(f)\mathbf{A}^{\text{H}}(f)\mathbf{T}^{\text{H}}(f,f_0) = \mathbf{V}(f_0)\mathbf{\Sigma}(f_0)\mathbf{V}^{\text{H}}(f_0). \tag{7}$$

The frequency-averaged and spatially whitened array SCMs are then decomposed as

$$\hat{\mathbf{C}}_{\text{x}}^{(ERB)}(f_0) = \sum_{q=1}^{Q} \sigma_q(f_0)\mathbf{v}_q(f_0)\mathbf{v}_q^{\text{H}}(f_0), \tag{8}$$

where $\sigma_1 > ... > \sigma_Q$ are the eigenvalues sorted in descending order and $\mathbf{v}_1, ..., \mathbf{v}_Q \in \mathbb{C}^{Q \times 1}$ are their respective eigenvectors.

3.3 Diffuseness estimation

For estimating the diffuseness parameter, $\psi \in [0,1]$, the COMEDIE algorithm was employed [15], which operates based on the array SCM eigenvalues. It is noted, however, that the algorithm is (by default) intended for SH domain input, where $\mathbf{Y}\mathbf{Y}^{\text{T}} = \mathbf{I}$; due to the orthogonality of the SH weights used as steering vectors in this domain. Therefore, without the application of the spatial whitening operation described in the previous subsection, this estimator would not be suitable for the present case; since $\mathbf{A}\mathbf{A}^{\text{H}}$ is not guaranteed to be diagonal (especially at low-frequencies). However, given the application of the whitening operation, the assumptions made by this diffuseness estimation algorithm are also applicable directly in the sensor-domain; since $\mathbf{W}\mathbf{A}\mathbf{A}^{\text{H}}\mathbf{W}^{\text{H}} = \mathbf{I}$.

The COMEDIE algorithm is based on determining the diffuseness value through observing the variance of the eigenvalues as

$$\psi(f_0) = 1 - \frac{\beta(f_0)}{\beta_0} \tag{9}$$

with the normalisation $\beta_0 = 2(Q-1)$, the deviation $\beta = \frac{1}{\langle\sigma\rangle}\sum_{q=1}^{Q}|\sigma_q - \langle\sigma\rangle|$, and the mean $\langle\sigma\rangle = \frac{1}{Q}\sum_{q=1}^{Q}\sigma_q$.

3.4 Direction-of-arrival estimation

For estimating the DoA of the most dominant sound source, the multiple signal classification (MUSIC) [16] approach is applied per ERB band as

$$\mathbf{p}_{\text{MUSIC}} = \text{diag}[\mathbf{A}^{\text{H}}\mathbf{W}^{\text{H}}\mathbf{V}_{\text{n}}\mathbf{V}_{\text{n}}^{\text{H}}\mathbf{W}\mathbf{A}], \tag{10}$$

where $\mathbf{V}_{\text{n}} = [\mathbf{v}_2, ..., \mathbf{v}_Q] \in \mathbb{C}^{Q\times Q-1}$ is the noise subspace, which is made up of the eigenvectors corresponding to the smallest $Q-1$ eigenvalues; and diag[.] denotes the construction of a vector based on the diagonal entries of the enclosed matrix. The DoA, $\boldsymbol{\gamma}_{\text{DoA}}$, is then determined as the direction where the pseudo spectrum $\mathbf{p}_{\text{MUSIC}} \in \mathbb{C}^{V\times 1}$ is minimised.

4 Proposed parametric encoding

The proposed parametric encoding approach operates based upon two dedicated rendering streams, using the estimated spatial parameters described in Section 3. The first stream is specifically intended to encode sound components that have clear directionality, utilising only the signal corresponding to the omnidirectional sensor; whereas, the second stream aims to render the remaining ambient sound components in a diffuse manner using the signals of all sensors in the array. Note that the time and frequency indices are henceforth omitted for brevity of notation.

4.1 Source stream

The signal corresponding to the omnidirectional sensor, which is assumed to have high signal quality, is then encoded into the Ambisonics format at an arbitrary target order and modulated using the diffuseness parameter (i.e. a single-channel Wiener filter) as

$$\mathbf{a}_{\text{par,s}} = (1-\psi)\,\mathbf{y}(\boldsymbol{\gamma}_{\text{DoA}})\,x_1, \tag{11}$$

where $\mathbf{y}(\boldsymbol{\gamma}_{\text{DoA}}) \in \mathbb{R}^{(N+1)^2\times 1}$ are the spherical harmonic weights for the estimated DoA.

4.2 Ambient stream

For the ambient stream, all Q channels are employed, including the $Q-1$ sensors that may have lower signal quality than the omnidirectional sensor. However, since it is assumed that the ambient stream will represent the noisy and diffuse sound components of the scene, it is postulated that this may be less problematic in practice. Furthermore, if one wished to appropriately render diffuse-field conditions using only the high-quality omnidirectional sensor, the decorrelation operations required to attain a suitable rendering may, in any case, result in signal quality degradation.

The ambient stream is based on first performing a uniform L-direction plane-wave decomposition of the array signals, which is subjected to an energy-preserving constraint; in a similar manner to the method described in [17] and as conducted in [10] . A singular value decomposition of the L array steering vectors, $\mathbf{A}_{\text{d}} \in \mathbb{C}^{Q\times L}$, is first conducted as

$$\mathbf{A}_{\text{d}}^{\text{H}} = \mathbf{U}_{\text{d}}\mathbf{S}_{\text{d}}\mathbf{V}_{\text{d}}^{\text{H}}. \tag{12}$$

The singular values (stored in the diagonal entries of matrix $\mathbf{S}_{\text{d}} \in \mathbb{C}^{L\times Q}$) are then replaced by an identically sized matrix that has the value of 1 along the main diagonal and 0 elsewhere, $\hat{\mathbf{S}}_{\text{d}}$, in order to obtain the following unitary matrix

$$\mathbf{G}_{\text{d}} = \frac{1}{\sqrt{L}}\mathbf{U}_{\text{d}}\hat{\mathbf{S}}_{\text{d}}\mathbf{V}_{\text{d}}^{\text{H}}. \tag{13}$$

The plane-wave decomposed ambient signals, $\mathbf{z} = \mathbf{G}_{\text{d}}\mathbf{x} \in \mathbb{C}^{L\times 1}$, are then encoded into the Ambisonics format and modulated by the estimated diffuseness parameter as

$$\mathbf{a}_{\text{par,d}} = \psi E_{\text{d}}\mathbf{Y}_{\text{d}}\mathbf{G}_{\text{d}}\mathbf{x} = \psi E_{\text{d}}\mathbf{Y}_{\text{d}}\mathbf{z}, \tag{14}$$

where $\mathbf{Y}_{\text{d}} \in \mathbb{R}^{(N+1)^2\times L}$ are SH encoding weights for the same plane-wave decomposition directions, and $E_{\text{d}} = \mathbf{tr}[\mathbf{A}\mathbf{A}^{\text{H}}]^{-1/2}$ is a diffuse-field equalisation term, which expresses the inverse of the array diffuse field response (averaged over all directions). The term is used to mitigate colourations incurred by the physical microphone array configuration when capturing a diffuse-field.

4.3 Optimised decorrelation

To enforce diffuse properties, the intermediate plane-wave decomposed signals may be optionally decorrelated as

$$\hat{\mathbf{z}} = \mathscr{D}[\mathbf{G}_\mathrm{d}\mathbf{x}] = \mathscr{D}[\mathbf{z}], \tag{15}$$

where $\mathscr{D}[.]$ denotes a decorrelation operation on the enclosed signals.

However, since decorrelation operations may incur signal degradation, it may be beneficial to impose constraints that limit the amount of decorrelation applied to only what it required to obtain a diffuse rendering. Such a constraint may be realised through the optimised SCM matching framework presented in [11]

$$\hat{\mathbf{z}}^{(\mathrm{opt})} = \mathbf{M}\mathbf{z} + \mathbf{M}_\mathrm{r}\hat{\mathbf{z}}, \tag{16}$$

where $\mathbf{M} \in \mathbb{C}^{L\times L}$ and $\mathbf{M}_\mathrm{r} \in \mathbb{C}^{L\times L}$ are mixing matrices, which are obtained by solving [11]

$$\begin{aligned} &\underset{\mathbf{M},\mathbf{M}_\mathrm{r}}{\arg\min}\, \mathbb{E}[||\hat{\mathbf{z}}^{(\mathrm{opt})} - \mathbf{z}||^2], \quad \text{subject to} \\ &\mathbf{M}\mathbf{C}_\mathrm{z}\mathbf{M}^\mathrm{H} + \mathbf{M}_\mathrm{r}(\mathrm{Diag}[\mathbf{C}_\mathrm{z}])\mathbf{M}_\mathrm{r}^\mathrm{H} = \mathrm{Diag}[\mathbf{C}_\mathrm{z}], \end{aligned} \tag{17}$$

where $\mathbf{C}_\mathrm{z} = \mathbb{E}[\mathbf{z}\mathbf{z}^\mathrm{H}] \in \mathbb{C}^{L\times L}$ and Diag[.] denotes constructing a diagonal matrix based on the diagonal entries of the enclosed matrix.

The framework operates by first attempting, as much as possible, to attain the target inter-channel relationships for a diffuse-field, (Diag[$\mathbf{C}_\mathrm{z}$]), without applying any decorrelation; i.e. through only a linear mapping of the non-decorrelated $\mathbf{z}$ signals using $\mathbf{M}$. The decorrelated $\hat{\mathbf{z}}$ plane-wave decomposed signals are then introduced only to the degree necessary, in order to fulfil the remaining inter-channel dependency targets using $\mathbf{M}_\mathrm{r}$. The resultant optimised decorrelated signals $\hat{\mathbf{z}}^{(\mathrm{opt})}$, therefore conform to the target diffuse conditions, while using only the minimum amount of decorrelation required to attain them. Note that the derivation, and a `MATLAB` implementation of a solution to this problem, may be found in [11].

4.4 Final rendering

The final parametrically encoded Ambisonic sound scene is then obtained as the superposition of both previously described streams

$$\begin{aligned} \mathbf{a}_\mathrm{par} &= \mathbf{a}_\mathrm{par,s} + \mathbf{a}_\mathrm{par,d}^{(\mathrm{opt})}, \\ &= (1-\psi)\,\mathbf{y}(\boldsymbol{\gamma}_\mathrm{DoA})\,x_1 + \psi E_\mathrm{d}\mathbf{Y}_\mathrm{d}\hat{\mathbf{z}}^{(\mathrm{opt})}. \end{aligned} \tag{18}$$

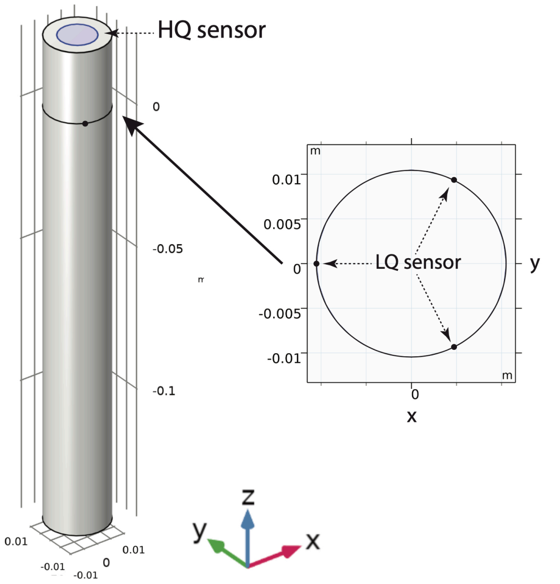

Fig. 1: One-plus-three microphone array design simulation. The high-quality (HQ) sensor is on the top base of the cylinder (highlighted in blue) and the (potentially) lower-quality (LQ) sensors are mounted onto the cylindrical casing.

5 Evaluation

The proposed one-plus-three microphone array configuration and parametric encoding method were evaluated based on formal listening tests using simulated steering vectors of the one-plus-three microphone array.

5.1 The one-plus-three microphone array

The geometry for the one-plus-three array, as depicted in Figure 1, comprised a cylinder 17 cm in height and 2.1 cm in diameter. One sensor was placed on the top surface and three additional sensors were distributed equally around the curved surface of the cylinder, 2.5 cm below the top surface. The top circular sensor covers the circular top surface with a diameter of 1.25 cm, which is approximately 1/2 an inch. It is assumed that the three (potentially) lower-quality sensors would be smaller, and were therefore simulated as smaller areas on the baffle of the array. The geometry of this array was selected as it emulates the design of 1/2 inch single-capsule microphones commonly used in studio recordings. The response of the array was simulated using the Pressure Acoustics, Boundary-Element

Method (BEM) module of COMSOL Multiphysics [18] for 625 directions, following a 25th order regularly distributed Fliege-Maier design [19]. For each direction, the response of the array was simulated for 128 frequencies between 93.5 Hz and 12 kHz. The responses of the array were then collected and processed in MATLAB, where they were re-sampled from 24 kHz to 48 kHz using the resample function.

5.2 Implementation of the proposed method

The implementation of the encoding algorithm was realised using the alias-free STFT design described in [20] configured with a hop-size of 128 (48 kHz sample rate) with 90% overlapping windows. The array SCMs were computed and averaged over time in blocks of 4096 samples (i.e. 32 down-sampled time indices, approx. 85 ms), and also over frequency using the ERB scale [12]. The spatial parameters were estimated for every 2048 samples block and for each ERB band as described in Section 3. The subsequent parametric encoding was conducted as described in Section 4, using a $L = 12$ uniform spherical grid for the plane-wave decomposition and performing the decorrelation using cascaded lattice all-pass filters and fixed sub-band delays as described in [21].

5.3 Listening test design

An image-source based room simulator[2] was configured for two different acoustics. The first was an anechoic (*dry*) environment with a broad-band reverberation time (RT60) of 0 s, while the second was a $[10 \times 6 \times 3.5]$ m sized, moderately reverberant room (*rev*) with RT60 times of $[0.52, 0.59, 0.39, 0.20, 0.16, 0.13]$ s in octave bands from 125 Hz to 4 kHz. The receiver position was placed $[-0.55, -0.57, 0.23]$ m from the centre of the room, with one source placed directly 1 m in-front of the receiver position and a second source placed 1 m to the left of the receiver position; i.e. two sources separated by 90° on the horizontal plane. Three sets of contrasting audio signals were employed as the two sound source signals: 1) a shaker accompanied by strings (*music*), 2) a male and a female speaker (*speech*), and 3) a waterfall and a piano (*mix*). Since the one-plus-three array was simulated up to 12 kHz, all stimuli were low-pass filtered at 12 kHz. These listening test scenes are summarised in Table 1.

Table 1: Listening test scenes.

Name	Acoustics	Stimuli
dry_music	Anechoic	shaker & strings
dry_speech	Anechoic	male & female speaker
dry_mix	Anechoic	waterfall & piano
rev_music	Reverberant	shaker & strings
rev_speech	Reverberant	male & female speaker
rev_mix	Reverberant	waterfall & piano

With the simulator now configured, array steering vectors for the one-plus-three microphone array were then used to simulate recordings of the sound scenes from the perspective of the receiver position. This was conducted by quantising and convolving the incoming image-sources with the nearest array steering vector in the $V = 625$ grid. Similarly, reference recordings for an ideal SH receiver were obtained by directly encoding them using third-order SH weights (*ideal_ref_o3*). For additional insights, an open tetrahedral array of radius 2 cm with cardioid sensors, and a rigid-baffle 19-sensor SMA of radius 4.9 cm with sensors placed on the vertices of a dodecahedron, except for the lowest vertex, were also used to obtain recordings of the simulated scenes. These two additional arrays represent array configurations that are currently popular for Ambisonic recording.

The simulated one-plus-three microphone and two SMAs recordings were then encoded into the Ambisonics format. For the one-plus-three microphone, the proposed parametric encoding approach was employed targeting third-order (*1+3_par_o3*). Whereas, for the two SMAs, the open-source SPARTA Array2SH audio plug-in described in [22] was used; this plug-in applies a linear encoding matrix based on analytical descriptors of the array sensor placement, radius, and its baffle type (open or rigid). The audio plugin also applies diffuse-field equalisation above the spatial aliasing frequency of the SMA, as described in [23] and recommended by Gerzon in [2]. The open tetrahedral array was linearly encoded into first-order (*open_lin_o1*), while the rigid 19-sensor SMA was linearly encoded into third-order (*rigid_lin_o3*). All encoded scenes were then decoded for static headphones playback using the Magnitude-Least Squares (MagLS) decoder, as implemented in the open-source SPARTA AmbiBIN audio plug-in de-

[2] The image-source based room simulator may be found here: https://github.com/polarch/shoebox-roomsim

Table 2: The rendering pipelines under test.

Name	Array	Encoding method	Decoding method
ideal_ref_o3	Ideal SH receiver	Direct encoding to third-order	Magnitude least-squares
1+3_par_o3	Proposed one-plus-three array	Proposed encoding to third-order	Magnitude least-squares
rigid_lin_o3	Rigid SMA with 19-sensors	Linear encoding to third-order	Magnitude least-squares
open_lin_o1	Open tetrahedral SMA	Linear encoding to first-order	Magnitude least-squares
omni	Omnidirectional microphone	None	Routed to left and right

scribed in [24]. As an anchor-like condition, and also highlighting the absence of spatial recording capabilities of omnidirectional studio microphones, the signal corresponding to the top-most (high-quality) sensor of the one-plus-three array was simply routed to the left and right binaural channels. These listening test cases are summarised in Table 2.

The multiple-stimulus listening test was then divided into three parts: **spatial**, **timbre**, and **overall**. In the spatial part of the test, all of the binaural signals from the rendering pipelines under test were equalised to the reference binaural signals based on their averaged magnitude responses; thus, achieving timbral equivalence across the test cases, with the listening test subjects asked to rate them based on their spatial similarity with the reference. In the timbral part, the reference test cases were instead duplicated and equalised based on the mean magnitudes of the rendering pipelines under test; therefore, attaining spatial equivalence, with the test subjects requested to rate them based only on their timbral similarity with the reference. In the final part of the test, termed overall, the test cases were simply normalised to the reference based on their broad-band root-mean-square values averaged across both channels, with the test subjects asked to rate them based on personal preference. Note that the employed three-part listening test procedure is also described in further detail in [9].

6 Results and discussion

The listening test results, based on ten participants, are presented in Fig. 2. It can be seen that the ratings of renderings for the proposed one-plus-three microphone array have medians that are closer in value to the ideal reference renderings in the spatial results, with the exceptions being the test scenes involving speech stimuli. The median values of the timbre results are almost identical with the ideal reference rendering ratings for all the test scenes and microphone arrays, with the exception of the rigid SMA.

Subsequent Friedman tests conducted on the spatial results showed there were statistically significant differences ($p < 0.05$) between the ratings for the three simulated microphone arrays and the reference rendering for all the test scenes. It should be noted that the ratings for the omnidirectional anchor case were omitted from these (and subsequent) analyses. Post ad-hoc analysis with Dunnett's test revealed that the differences between ratings for the ideal reference rendering and the one-plus-three microphone array rendering were statistically significant only for the speech test scenes, while the ratings for the tetrahedral SMA were significantly different from the ratings for the one-plus-three microphone array in the cases of the music and mix test scenes, but did not reach significance for the speech test scenes. The differences in ratings for the rigid SMA renderings and the one-plus-three microphone array renderings were statistically significant for the music and reverberant mix tests scenes but not for the other three test scenes.

For the timbre results, Friedman tests conducted on the ratings revealed statistically significant differences for the music and mix test scene renderings. No significant differences were found between the ratings for the speech test scenes. Post ad-hoc analysis with Dunnett's test revealed the differences between ratings for the ideal reference rendering and the one-plus-three microphone array renderings did not reach significance for any of the test scenes. The tetrahedral SMA's ratings were significantly different from the one-plus-three ratings only in the case of the reverberant music scene. The ratings for the rigid SMA renderings were found to be significantly different from the ratings for the

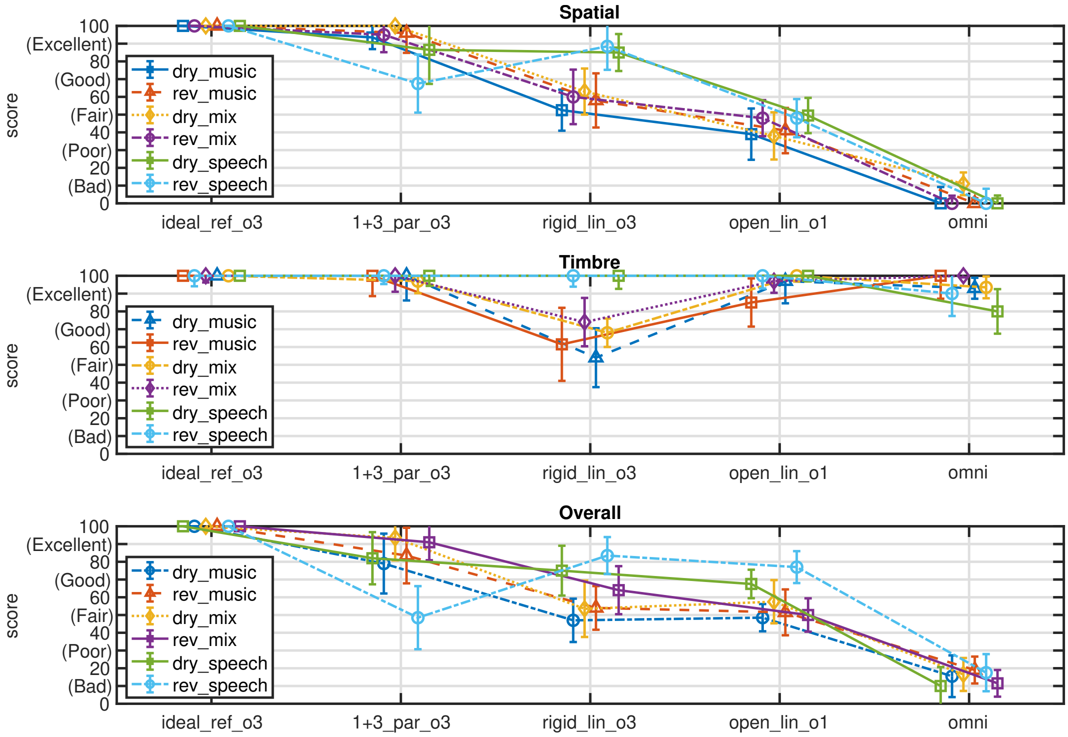

Fig. 2: Listening test results displaying medians and 95% confidence intervals, based on 10 test subjects.

one-plus-three microphone array renderings for all four music and mix sound scenes.

The statistical analyses imply that the simulated renderings of the proposed system are perceptually similar to the ideal references for most sound scenes in terms of spatial characteristics, with the exception being sound scenes composed of only speech sound sources. Furthermore, the proposed system performed in a similar manner to the ideal reference in terms of timbre for all the test scenes. Compared to the two SMAs using linear encoding strategies, the proposed system achieved ratings which were higher for the majority of the sound scenes in terms of spatial characteristics, with the exception being the speech sound scenes, where the proposed system performed in a manner similar to existing commercially available solutions. In terms of timbre characteristics, the renderings of the proposed system did not have any perceivable colouration or artefacts.

It may be inferred, therefore, that the proposed system, consisting of a parametric encoding method in conjunction with a one-plus-three microphone array configuration, may yield better or similar performance to popular SMA configurations coupled with conventional linear encoding procedures.

7 Prototype

In order to demonstrate a practical implementation of the proposed microphone array configuration, a prototype was constructed. This prototype involved affixing a 3D printed ring, housing three MEMS microphones, onto the cylindrical casing of an AKG CK 62-ULS omni-directional microphone capsule and AKG C480 B pre-amp; as shown in Fig. 3. Especially if such an omni-directional microphone is already at hand, the assembled prototype may represent a cost effective solution for obtaining a spatial audio compliant recording device. It is also noted that the prototype requires the same number of analogue-to-digital converters as used by existing first-order tetrahedral arrays; however, when coupled with the proposed rendering techniques, it may also attain higher spatial resolution compared to traditional signal-independent Ambisonic recording solutions. A study investigating the application of the proposed rendering techniques using this prototype,

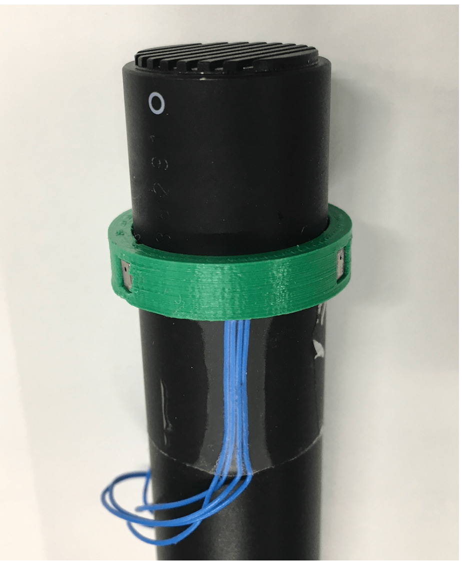

Fig. 3: Prototype of the proposed one-plus-three microphone array configuration, where the high-quality omni-directional capsule is accompanied by three MEMS-microphones (on the 3D printed green ring).

and how closely the perceived performance matches the simulated case used for the present evaluations, is a topic of future work.

8 Conclusion

This paper proposes a solution for imparting ambisonic recording capabilities onto an omnidirectional studio microphone. This is realised by affixing three or more additional sensors to the studio microphone casing, coupled with the use of spatial analysis and adaptive signal processing to isolate directional and non-directional sound components in the captured scene; which can then be encoded into the target Ambisonics format at an arbitrary order. The (presumably) higher quality omnidirectional signal is used for encoding the directional components, while both the omnidirectional signal and the secondary array signals are used to deliver an enhanced ambience encoding.

The proposed microphone array configuration and the encoding method were evaluated by conducting formal listening tests, based on a binaural rendering workflow and using simulated array steering vectors. The proposed system was compared to binaural renderings of: ideal simulated Ambisonic encodings, and that of a linearly encoded open tetrahedral array and also a linearly encoded rigid 19-sensor spherical microphone array. The results, based on 10 subjects, indicate that the proposed microphone array configuration and parametric approach is able to ultimately lead to binaural renderings which are perceptually more similar to the reference case, compared to the other linearly encoded arrays, in the majority of the tested cases.

Additionally, a prototype of the proposed array configuration was constructed, which may serve as a practical example and form the basis of a future study.

References

[1] Gerzon, M. A., “Periphony: With-height sound reproduction,” *Journal of the Audio Engineering Society*, 21(1), pp. 2–10, 1973.

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

[3] Kronlachner, M. and Zotter, F., “Spatial transformations for the enhancement of Ambisonic recordings,” in *Proceedings of the 2nd International Conference on Spatial Audio, Erlangen*, 2014.

[4] McCormack, L., Politis, A., and Pulkki, V., “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, 2021.

[5] Rafaely, B., *Fundamentals of spherical array processing*, volume 8, Springer, 2015.

[6] Bertet, S., Daniel, J., Parizet, E., and Warusfel, O., “Investigation on localisation accuracy for first and higher order ambisonics reproduced sound sources,” *Acta Acustica united with Acustica*, 99(4), pp. 642–657, 2013.

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

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

[9] Fernandez, J., McCormack, L., Hyvärinen, P., Politis, A., and Pulkki, V., "Enhancing binaural rendering of head-worn microphone arrays through the use of adaptive spatial covariance matching," *The Journal of the Acoustical Society of America*, 151(4), pp. 2624–2635, 2022.

[10] McCormack, L., Politis, A., Gonzalez, R., Lokki, T., and Pulkki, V., "Parametric Ambisonic Encoding of Arbitrary Microphone Arrays," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30, pp. 2062–2075, 2022, doi:10.1109/TASLP.2022.3182857.

[11] Vilkamo, J., Bäckström, T., and Kuntz, A., "Optimized covariance domain framework for time–frequency processing of spatial audio," *Journal of the Audio Engineering Society*, 61(6), pp. 403–411, 2013.

[12] Moore, B. C. and Glasberg, B. R., "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns," *The journal of the acoustical society of America*, 74(3), pp. 750–753, 1983.

[13] Doron, M. A. and Nevet, A., "Robust wavefield interpolation for adaptive wideband beamforming," *Signal Processing*, 88(6), pp. 1579–1594, 2008.

[14] Beit-On, H. and Rafaely, B., "Focusing and frequency smoothing for arbitrary arrays with application to speaker localization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28, pp. 2184–2193, 2020.

[15] Epain, N. and Jin, C. T., "Spherical harmonic signal covariance and sound field diffuseness," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(10), pp. 1796–1807, 2016.

[16] Schmidt, R., "Multiple emitter location and signal parameter estimation," *IEEE transactions on Antennas and Propagation*, 34(3), pp. 276–280, 1986.

[17] Zotter, F., Pomberger, H., and Noisternig, M., "Energy-preserving ambisonic decoding," *Acta Acustica united with Acustica*, 98(1), pp. 37–47, 2012.

[18] Multiphysics, C., "Introduction to comsol multiphysics®," *COMSOL Multiphysics, Burlington, MA, accessed Feb*, 9(2018), p. 32, 1998.

[19] Fliege, J. and Maier, U., "A Two-Stage Approach for Computing Cubature Formulae for the Sphere," in *Mathematik 139T, Universitat Dortmund, Fachbereich Mathematik, Universitat Dortmund, 44221*, 1996.

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

[21] Herre, J., Purnhagen, H., Breebaart, J., Faller, C., Disch, S., Kjörling, K., Schuijers, E., Hilpert, J., and Myburg, F., "The reference model architecture for MPEG spatial audio coding," in *Audio Engineering Society Convention 118*, 2005.

[22] McCormack, L., Delikaris-Manias, S., Farina, A., Pinardi, D., and Pulkki, V., "Real-time conversion of sensor array signals into spherical harmonic signals with applications to spatially localized sub-band sound-field analysis," in *Audio Engineering Society Convention 144*, Audio Engineering Society, 2018.

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

[24] McCormack, L. and Politis, A., "SPARTA & COMPASS: Real-time implementations of linear and parametric spatial audio reproduction and processing methods," in *Audio Engineering Society Conference: 2019 AES International Conference on Immersive and Interactive Audio*, Audio Engineering Society, 2019.