Audio Engineering Society

Conference Paper

Presented at the International Conference on
Spatial and Immersive Audio
2023 August 23–25, Huddersfield, UK

Parametric architecture for the transmission and binaural reproduction of microphone array recordings

Leo McCormack[1], Christoph Hold[1], and Archontis Politis[2]

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

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

ABSTRACT

This paper proposes a multi-directional parametric architecture for transmitting and reproducing microphone array recordings using a reduced number of transport audio channels. The approach enables the maximum number of directional source signals to be adjusted and either configured to be restrictive, in order to reduce the number of transmission channels, or alternatively set higher, in order to improve the accuracy of the model. Ambient sounds, which remain once the directional sounds are subtracted from the input, are represented by a dedicated monophonic residual signal. After transmission, the source signals are spatialised over the playback system based upon the accompanying spatial metadata; whereas the monophonic residual signal is reproduced using a spatially incoherent source spreading algorithm. A binaural perceptual evaluation then followed. The results suggest that high spatial audio quality may be attained when reproducing four-channel array recordings using two audio transmission channels, and 25-channel and 32-channel recordings when transmitting four audio channels.

1 Introduction

The increasing appetite amongst consumers to experience immersive spatial audio content – driven, in large part, by the more widespread availability of augmented and virtual reality (AR/VR) devices – has subsequently led to a strong demand for the development of perceptually accurate microphone array reproduction methods. Technologies supporting AR/VR systems often involve binaural reproduction and therefore need to take into consideration head-tracking data, which describes listener head-rotations and/or translations away from the recording point [1, 2]; while also potentially seeking to enable additional sound-field modification freedoms [3, 4]. Data-streams enabling an accurate and interactive reproduction on the playback side, however, typically demand a large number of transport audio channels. Since some applications may have strict data streaming limitations imposed onto them, further rendering strategies may need to be explored to provide the same immersive audio experience while using reduced data resources. Therefore, the focus of this paper relates to the development of a new rendering architecture, which may permit the transmission and binaural rendering of microphone array recordings using fewer transport channels; while maintaining high accuracy and the freedom to interactively control the rendering.

Regarding the transmission and subsequent reproduction of microphone array recordings, potential options

for this task may be loosely placed into three subcategories: 1) those employing statistical based approaches [5], which have more traditionally been applied to so-called *channel-based* material (i.e., audio material which is already represented in the target playback format); 2) those adopting a perceptually-motivated parametric sound-field model [6, 7], which are usually applied to so-called *scene-based* material (i.e., audio that describes the scene as a whole); and 3) hybrid combinations of these two paradigms [8, 9, 10]. For the first category of options, the multi-channel signal statistics are typically obtained through some form of principal component analysis (subspace decomposition), which is followed by the identification and removal of information regarded as redundant or irrelevant. The requirements of such methods are usually strongly geared towards maintaining perceived transparency between the original and reconstructed playback audio. Model-based reproduction approaches, on the other hand, typically operate by decomposing the input scene into a subset of signals accompanied by spatial metadata, which are then used to directly synthesise signals for the target playback setup(s). Evaluations of such parametric methods are then usually based upon a comparison between the synthesised playback signals and reference playback signals which are obtained, for example, through direct binaural recordings/simulations of the same sound scene [6, 11].

Given the present focus of this paper, it may be argued that reconstructing the microphone array recordings on the playback end, and then mapping them to the binaural channels, would likely be sub-optimal when compared to the direct binaural reproduction of the transmitted audio. Therefore, a sound-field model-based paradigm may be considered to be a stronger candidate for addressing the presently considered requirements. Perhaps the most well-known parametric method is Directional Audio Coding (DirAC) [6], which operates based upon a first-order Ambisonics [12] representation of the input microphone array recording. Here, direction-of-arrival (DoA) and diffuseness parameters are estimated in a time-frequency transform (TFT) domain. In the original DirAC formulation, the omnidirectional signal from the Ambisonics representation, along with these associated spatial parameters, are transmitted to the playback side. The omnidirectional signal may then be spatialised over the target reproduction system (loudspeakers in the original case) based upon the DoA estimates, and also spread to all channels in the playback setup and decorrelated. The time-frequency-dependent balance between these two audio streams is then dictated by the diffuseness parameters. The original DirAC formulation, therefore, potentially represents the most extreme data resource limited solution for transmitting microphone array signals; since a single channel of audio is transmitted. An extension to the DirAC-based transmission model was then explored recently in [13]. This approach involves steering fixed beamformers in a number of directions, and estimating the DoA and diffuseness parameters within directionally-constrained regions (sectors) on the sphere. This updated design was shown to improve the perceived spatial accuracy on the receiving end by transmitting four or more audio signals. However, the design of these fixed beamformers for fewer than four transmission channels is not a trivial exercise. An additional limitation is that DirAC-based solutions require that the input microphone array recordings are encoded into the Ambisonics domain, which can result in some loss in spatial resolution/performance.

Therefore, in this paper, a new sound-field model-based transmission architecture is proposed, which directly analyses and decomposes the input microphone array recording into a reduced number (≤ 4) of transport audio channels. These signals are then synthesised directly to the binaural channels on the playback side, based upon the accompanying spatial metadata. The proposed architecture resembles a space domain reformulation of the Coding and Multi-Directional Decomposition of Ambisonic Sound Scenes (COMPASS) [11] method, while additionally adopting a hard threshold on the maximum number of beamformer signals (which are steered adaptively, rather than statically as in [13]) and an alternative ambient rendering pipeline requiring only a single transport audio channel.

2 Sound-field model

It is assumed that the input sound-field $\mathbf{x}(t,f) \in \mathbb{C}^{Q\times 1}$ is captured using an array of Q microphones and a TFT is applied [14], in order to represent the array signals over both (down-sampled) time t and frequency f.

The input sound-field may then be modelled as

$$\mathbf{x}(t,f) = \mathbf{A}_\mathrm{s}(f)\mathbf{s}(t,f) + \mathbf{d}(t,f), \tag{1}$$

where $\mathbf{d} \in \mathbb{C}^{Q\times 1}$ represents the microphone array signals containing only ambient sounds; $\mathbf{s} \in \mathbb{C}^{K\times 1}$ are the

signals of directional sound sources in the scene; and $\mathbf{A}_\mathrm{s} \in \mathbb{C}^{Q\times K}$ are the array transfer functions (ATFs) for the respective source directions, $\mathbf{\Gamma} = [\boldsymbol{\gamma}_1, ..., \boldsymbol{\gamma}_K]$; where $\boldsymbol{\gamma}_k \in S^2$ is the direction for the kth source. It is henceforth assumed that ATFs $\mathbf{A} = [\mathbf{a}(\boldsymbol{\gamma}_1), ..., \mathbf{a}(\boldsymbol{\gamma}_V)] \in \mathbb{C}^{Q\times V}$ are available for a dense grid of V directions, which may be obtained through measurements or simulations.

The array spatial covariance matrix (SCM) is given as

$$\begin{aligned}\mathbf{C}_\mathrm{x}(f) &= \mathscr{E}[\mathbf{x}(t,f)\mathbf{x}^\mathrm{H}(t,f)], \\ &= \mathbf{A}_\mathrm{s}(f)\mathbf{C}_\mathrm{s}(f)\mathbf{A}_\mathrm{s}^\mathrm{H}(f) + \mathbf{C}_\mathrm{d}(f),\end{aligned} \tag{2}$$

where $\mathscr{E}[.]$ denotes the expectation operator, which is typically achieved via temporal averaging over tens of milliseconds; $\mathbf{C}_\mathrm{s} = \mathscr{E}[\mathbf{s}\mathbf{s}^\mathrm{H}]$ is the SCM for the source signals; and $\mathbf{C}_\mathrm{d} \in \mathbb{C}^{Q\times Q}$ is the ambient array SCM.

In this study, it is assumed that these ambient array SCMs represent the capture of diffuse sounds (i.e., uncorrelated sound waves arriving from different directions), which also follow an isotopic energy distribution over the sphere, and thus they may be modelled as

$$\mathbf{C}_\mathrm{d}(f) = P_\mathrm{d}(f)\, \mathbf{D}_\mathrm{array}(f), \tag{3}$$

where $\mathbf{D}_\mathrm{array} = \mathbf{A}\mathbf{W}\mathbf{A}^\mathrm{H} \in \mathbb{C}^{Q\times Q}$ is the diffuse coherence matrix (DCM) of the array; with $\mathbf{W} = \mathrm{diag}[w_1, ..., w_V] \in \mathbb{R}^{V\times V}$ representing a diagonal matrix of integration weights to account for cases where the ATF measurement grid is not uniform; and P_d denotes the energy of the ambience in the scene. Note that the time and frequency indices are henceforth omitted for the brevity of notation.

2.1 Spherical harmonic domain

The above sound-field model is also directly applicable in the spherical harmonic (SH) domain. In this case, the ATFs may correspond instead to broad-band SH weights [15] and the input signals are the array signals after a suitable encoding operation has been applied [16, 1]. The number of input channels therefore becomes $Q = (N+1)^2$; where N is the SH order of expansion. Furthermore, in this domain, it is noted that the array DCM would become an identity matrix ($\mathbf{D}_\mathrm{array} = \mathbf{I}$).

3 Encoder

The task of the proposed encoder is to detect the number of sources independently across time and frequency, and to subsequently estimate their DoAs up to a user-specified maximum number. This information is then used to steer beamformers towards these sound sources in the scene, in order to isolate their signals. A residual component, encapsulating diffuse and/or weaker directional sounds, is represented by its own dedicated (monophonic) signal.

3.1 Spatial analysis

The number of sources is estimated using the SORTE algorithm [17], and the DoAs are estimated using multiple signal classification (MUSIC) [18]. The operation of these spatial analysis algorithms is based upon the eigenvalues and truncated eigenvectors of the array SCM, and thus the array SCMs are first decomposed as

$$\mathbf{C}_\mathrm{x} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\mathrm{H} = \sum_{k=1}^{K} \lambda_k \mathbf{v}_k \mathbf{v}_k^\mathrm{H} + \sum_{k=K+1}^{Q} \lambda_k \mathbf{v}_k \mathbf{v}_k^\mathrm{H}, \tag{4}$$

where $\lambda_1 > \lambda_2 > ... > \lambda_Q$ are the eigenvalues in descending order, and $\mathbf{v}_k$ are the respective eigenvectors.

The SORTE detection algorithm [17] first calculates the differences between eigenvalues as

$$\nabla\lambda_i = \lambda_i - \lambda_{i+1}, \quad \text{for } i = 1, ..., Q-1. \tag{5}$$

The number of sources is then determined as

$$K_\mathrm{SORTE} = \underset{k}{\mathrm{argmin}}\ f(k) \quad \text{for } k = 1, ..., Q-3, \tag{6}$$

given

$$f(k) = \begin{cases} \frac{\sigma_{k+1}^2}{\sigma_k^2}, & \sigma_k^2 > 0 \\ +\infty, & \sigma_k^2 = 0 \end{cases}, \quad \text{for } k = 1, ..., Q-2, \tag{7}$$

$$\sigma_k^2 = \frac{1}{Q-k} \sum_{i=k}^{Q-1} \left(\nabla\lambda_i - \frac{1}{Q-k} \sum_{i=k}^{Q-1} \nabla\lambda_i \right)^2. \tag{8}$$

Note that spatial whitening operations may optionally be included at this stage, in order to transform the array SCMs to be in a state which is more inline with the assumptions made by the employed detection algorithm; for more information, the reader is referred to [19].

For the DoA estimation, a MUSIC pseudo spectrum is generated as

$$P_\mathrm{MUSIC}(\boldsymbol{\gamma}) = \frac{1}{||\mathbf{V}_\mathrm{n}^\mathrm{H}\mathbf{a}(\boldsymbol{\gamma}, f_0)||^2}, \quad \text{for } \boldsymbol{\gamma} \in \mathbf{\Gamma}, \tag{9}$$

where $\mathbf{V}_\mathrm{n} \in \mathbb{C}^{Q\times(Q-K_\mathrm{SORTE})}$ is the noise subspace, which is constructed using the eigenvectors corresponding to the lowest $Q-K_\mathrm{SORTE}$ eigenvalues. Given a user-defined maximum number of source channels, which needs to obey $K_\mathrm{max} \leq \lfloor Q/2 \rfloor$, the $K = \min[K_\mathrm{SORTE}, K_\mathrm{max}]$ DoA estimates are obtained by ascertaining which directions correspond to the highest K peaks in the pseudo spectrum. Note that a higher maximum number of sources may lead to a better modelling of the scene, whereas a lower maximum source number reduces the number of transmission signals required.

3.2 Source beamformers

Once the number of sources and their respective DoAs have been estimated, the source signals may be obtained by applying an appropriate matrix of beamforming weights $\mathbf{B}_\mathrm{s} \in \mathbb{C}^{K\times Q}$ as

$$\mathbf{s} = \mathbf{B}_\mathrm{s}\mathbf{x}. \tag{10}$$

In this study, the following beamformers were selected

$$\mathbf{B}_\mathrm{s} = (\mathbf{A}_\mathrm{s}^\mathrm{H}\mathbf{A}_\mathrm{s} + \beta\mathbf{I})^{-1}\mathbf{A}_\mathrm{s}^\mathrm{H}, \tag{11}$$

where $\beta \geq 0$ is a regularisation term. Note that when $K = 1$ this design reverts to a matched-filter beamformer, which also corresponds to a hyper-cardioid beamformer when formed in the SH domain. Whereas, when $K > 1$, each row of beamforming weights is computed through the imposition of unity gain constraints towards their respective DoA, while placing null-constraints towards the other DoAs. Therefore, the regularisation term not only addresses cases where $\mathbf{A}_\mathrm{s}^\mathrm{H}\mathbf{A}_\mathrm{s}$ may become singular (e.g., at low frequencies), but also controls the extent of this source separation.

3.3 Ambient beamformer

An estimate of the ambient array signals is then obtained by spatially subtracting the source signal estimates from the input array signals as [11]

$$\mathbf{d} = (\mathbf{I} - \mathbf{A}_\mathrm{s}\mathbf{B}_\mathrm{s})\mathbf{x}. \tag{12}$$

A monophonic representation of these ambient sounds is then obtained by applying a zeroth-order Ambisonics encoder $\mathbf{w}_0 \in \mathbf{C}^{1\times Q}$ (i.e., omnidirectional beamformer)

$$d_0 = \mathbf{w}_0(\mathbf{I} - \mathbf{A}_\mathrm{s}\mathbf{B}_\mathrm{s})\mathbf{x} = \mathbf{w}_0\mathbf{d}. \tag{13}$$

This process is akin to a null-former which takes the patterns of the source beamformers into account. In this study, a standard least-squares Ambisonics encoder was selected [16]

$$\mathbf{w}_0 = \mathbf{1}\mathbf{W}\mathbf{A}^\mathrm{H}\left[\mathbf{D}_\mathrm{array} + \zeta\mathbf{I}\right]^{-1}, \tag{14}$$

where $\mathbf{1} \in \mathbb{R}^{1\times V}$ is a vector of ones (i.e., the zeroth order SH weights without the $1/\sqrt{4\pi}$ term); and $\zeta \geq 0$ is a regularisation term.

This zeroth order Ambisonic encoder was also diffuse-field equalised above the spatial aliasing frequency, f_a, in order to flatten its magnitude response (on average) at high frequencies, with [20, 21]

$$\mathbf{w}_0^{(eq)}(f) = \sqrt{\frac{|\mathbf{w}_0(f_\mathrm{a})\mathbf{D}_\mathrm{array}(f_\mathrm{a})\mathbf{w}_0^\mathrm{H}(f_\mathrm{a})|}{|\mathbf{w}_0(f)\mathbf{D}_\mathrm{array}(f)\mathbf{w}_0^\mathrm{H}(f)|}}\,\mathbf{w}_0(f), \quad \text{for } f > f_a. \tag{15}$$

Note that if the ATFs are broad-band SHs, then the Ambisonic encoder reverts to a vector comprising zeros for all entries except the first element (i.e., reverting to just taking the omnidirectional component of the SH signals). The above equalisation would also be intrinsically bypassed, since $\mathbf{D}_\mathrm{array}$ does not vary across frequency when using a broad-band SH basis.

4 Transmission - Coding

The proposed rendering architecture requires that K_max source signals, $\mathbf{s}$, and the single-channel ambient signal, d_0, are transmitted to the decoder side; i.e. $K_\mathrm{max}+1$ channels in total. In this study, it is suggested that four-channel microphone array recordings (and first-order Ambisonic recordings) should set $K_\mathrm{max} = 1$, which would therefore result in a 50% channel reduction. For higher sensor count arrays (and higher-order Ambisonic recordings) it is suggested that $K_\mathrm{max} \leq 3$, thus resulting in a maximum of four audio transport channels. Conventional lossy/lossless multi-channel codecs could then, conceivably, be applied to the transport signals to reduce signal data bandwidth requirements further; although, investigating this was beyond the scope of the present (architecture focused) study.

Besides the number of extracted sources, metadata comprising the time and frequency dependent DoAs must also be transmitted. The latter can be quantised and expressed as indices into the V directional grid, which should be common to both the spatial analysis (i.e.,

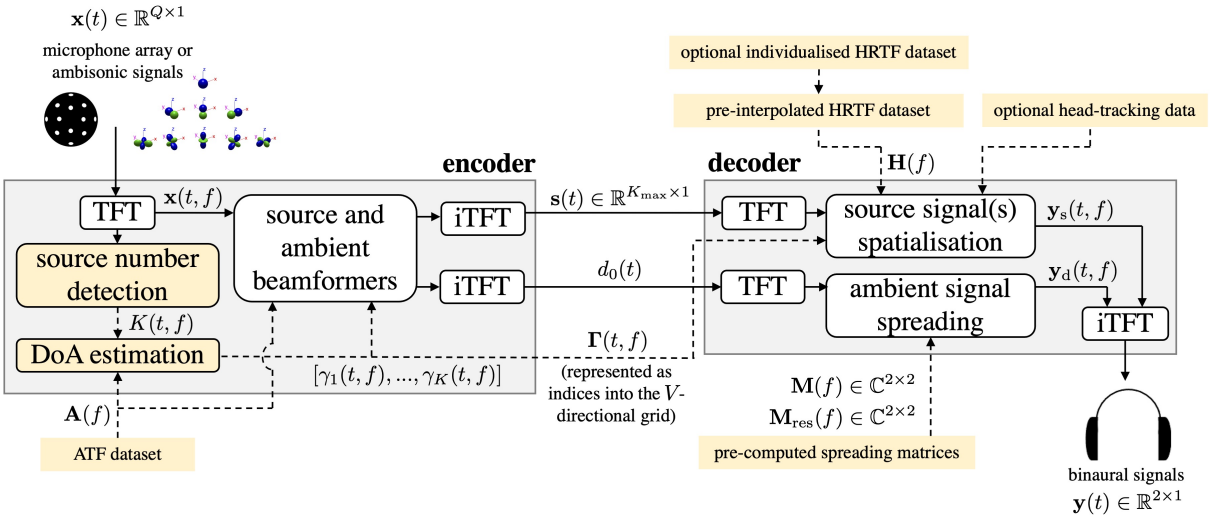

Fig. 1: Block diagram of the proposed rendering architecture.

MUSIC scanning) and to the dataset of spatialisation gains (discussed in the following section). Opportunities for reducing/optimising bandwidth requirements for this data stream are numerous. For instance, these spatial parameters may be estimated once per block of time frames, which could span large regions of time; for example, 42.6̇ ms (2048 samples at a sampling rate of 48 kHz), as adopted in the present study. Additionally, capping $K_{\max} \leq 3$ and $V \leq 2048$ would mean that only a maximum of 2 bits and 11 bits, respectively, would be required for coding the number of sources and corresponding DoA indices. Huffman code books precomputed for typical scenarios could then be used to optimise the bit-allocation. Furthermore, by taking into account the limitations of human spatial perception, and/or taking into consideration limitations in beamformer spatial selectivity, closely located DoAs could be merged into one average DoA. The spatial parameter estimation may also be conducted over frequency band groupings, such as perceptually-motivated equivalent rectangular bandwidths (ERB) [11], rather than over the full frequency resolution of the time-frequency transform (which is often uniformly spaced).

5 Decoder

The decoder is responsible for reproducing the transmitted audio signals over the target playback system. Note that a block diagram of the proposed rendering architecture is depicted in Fig. 1.

5.1 Source stream

The source signals may be simply spatialised over the target playback setup with

$$\mathbf{y}_\mathrm{s} = \mathbf{H}_\mathrm{s}\mathbf{s}, \tag{16}$$

where $\mathbf{H}_\mathrm{s} = [\mathbf{h}(\gamma_1),...,\mathbf{h}(\gamma_K)] \in \mathbb{C}^{2\times K}$ is a matrix of spatialisation gains for the target reproduction system. In this study, the spatialisation gains are head-related transfer functions (HRTFs), $\mathbf{h}(\gamma_k)$, which correspond to each kth DoA. Note that, given the present context, it is beneficial to pre-interpolate the HRTFs $\mathbf{H} = [\mathbf{h}(\gamma_1),...,\mathbf{h}(\gamma_V)] \in \mathbb{C}^{2\times V}$ to the same V-directional grid used for the DoA estimation. This pre-interpolation step allows for reduced run-time computational complexity, and the transmission of DoA indices for a common grid (rather than general azimuth and elevation angles) is likely to require lower data bandwidth.

It is noted that sound-field rotations (in order to accommodate head-tracking), and other directional transformations (such as listener translations, source tracking informed spatial editing, and side-chain morphing [2, 4]), may also be applied to the DoAs at this point. Furthermore, in principle, the above spatialisation gains could be replaced with amplitude-panning gains corresponding to an arbitrary loudspeaker array. Alternatively, SH vectors of arbitrary order may be used to target the Ambisonics format, or ATFs may be used to represent these direct sound components in the original microphone array format. In this study, and for the

perceptual evaluations, however, the focus was placed on the binaural rendering task.

5.2 Ambient stream

The monophonic signal encapsulating ambient sounds, d_0, is reproduced with the intention to spread the signal over the sphere in a spatially incoherent (i.e., diffuse) and isotropic manner. For the presented architecture, the signal spreading method described in [22] was selected. The method minimises the amount of decorrelated signal energy introduced into the output; thus, retaining the desired spatially incoherent spreading trait, while also preserving high signal quality.

A spatially coherent spreading of the ambient signal was selected as the basis for the prototype signals

$$\mathbf{y}_{\text{proto}} = d_0 \, \frac{1}{V} \sum_{v=1}^{V} w_v \, \mathbf{h}(\gamma_v) = d_0 \, \mathbf{h}_{\text{coh}}. \qquad (17)$$

The prototype SCM, resulting from this spatially coherent spreading, is therefore

$$\mathbf{C}_{\text{proto}} = P_{\text{d}} \, \mathbf{h}_{\text{coh}} \mathbf{h}_{\text{coh}}^{\text{H}} = \mathscr{E}[\mathbf{y}_{\text{proto}} \mathbf{y}_{\text{proto}}^{\text{H}}], \qquad (18)$$

where $P_{\text{d}} = \mathscr{E}[|d_0|^2]$ is an estimate of the ambient signal energy.

The task is to then mix the prototype signals, such that the SCM of the resulting signals, $\mathbf{y}_{\text{d}} \in \mathbb{C}^{2\times 1}$, instead adheres to the employed sound-field model; i.e.,

$$\mathbf{C}_{\text{target}} = P_{\text{d}} \, \mathbf{D}_{\text{bin}} = \mathscr{E}[\mathbf{y}_{\text{d}} \mathbf{y}_{\text{d}}^{\text{H}}], \qquad (19)$$

where $\mathbf{D}_{\text{bin}} = \mathbf{H}\mathbf{W}\mathbf{H}^{\text{H}} \in \mathbb{C}^{2\times 2}$ is the signal-independent binaural DCM.

The problem is therefore outlined as

$$\mathbf{y}_{\text{d}} = \mathbf{M}\mathbf{y}_{\text{proto}} + \mathbf{M}_{\text{res}} \mathscr{D}[\mathbf{y}_{\text{proto}}], \qquad (20)$$

where $\mathscr{D}[.]$ denotes decorrelation operations on the enclosed signals; and $\mathbf{M} \in \mathbb{C}^{2\times 2}$ and $\mathbf{M}_{\text{res}} \in \mathbb{C}^{2\times 2}$ are the primary and residual spreading matrices, respectively. The solution to this outlined problem may be found in [23, 22]. Essentially, the approach first attempts to manipulate the prototype signals to make them conform to the target SCM using a linear combination of the prototype signals. After this first stage, decorrelated versions of the prototype signals are mixed into the output, but only to the degree necessary to fulfil the remaining target inter-channel dependencies.

It is worth highlighting that, since the only signal-dependent term is P_{d}, which is common to both the prototype and target SCMs, the matrices $\mathbf{M}$ and $\mathbf{M}_{\text{res}}$ may be pre-computed and stored per frequency based on any nonzero value of $P_{\text{d}} \neq 0$. Therefore, the main computational complexity on the decoder side arises from the time-frequency (and inverse) transforms, the matrix multiplications with the signal spreading matrices, and the decorrelation operations applied to the coherently spread prototype signals. Note that an alternative approach, requiring only one channel of decorrelation, was proposed recently in [24], which may also be suitable for this signal spreading task.

The final output time-domain signals may then be obtained by summing the two audio streams together $\mathbf{y} = \mathbf{y}_{\text{s}} + \mathbf{y}_{\text{d}}$, followed by the application of an appropriate inverse time-frequency transform (iTFT).

6 Evaluation

A binaural multiple-stimulus listening test was conducted to assess the perceived differences between: a previous formulation of COMPASS [11, 25], which uses all Q channels for synthesising the playback signals, and the new architecture proposed in this paper, which uses $K_{\text{max}} + 1$ channels for synthesising the playback signals. The present authors also sought to assess the perceived differences of the proposed architecture when using different microphone arrays/receivers[1].

6.1 Implementation of the proposed method

The chosen TFT was the alias-free STFT (short-time Fourier transform) described in [14], which was configured with a hop size of $2.\dot{6}$ ms (128 samples at a 48 kHz samplerate) and window size of $5.\dot{3}$ ms. The array SCMs and spatial parameters were estimated for every $42.\dot{6}$ ms (2048 samples at 48 kHz) block of time frames per frequency band. An additional one-pole filter, with a coefficient value of 0.5, was applied to average the array SCMs further. The employed HRTFs were of a KU100 dummy-head (the data for which may be found via [26]). Cascaded lattice all-pass filters were employed for the decorrelation operations, as suggested in [5]. The beamformers used $\beta = 0.1$, whereas the Ambisonic encoder used $\zeta = 0.3$. A uniform directional grid corresponding to a t-design of degree 60 ($V = 1860$) was chosen for both the encoder and decoder processing.

[1]The listening test audio files may be downloaded from here: `https://zenodo.org/record/7956701`

Table 1: The receivers and rendering configurations under test. Note that # refers to the number of audio channels required to be transmitted to enable a flexible and (optionally) interactive reproduction.

Name	Receiver	Rendering method	Q	#
hidden_ref	Binaural microphone	Direct binauralisation of the simulated sound scene	2	n/a
compass_o4	Fourth order Ambisonics	COMPASS with covariance matched ambience [25]	25	25
proposed_o4	Fourth order Ambisonics	Proposed rendering architecture using $K_{\text{max}} = 3$	25	4
proposed_em32	32-sensor rigid SMA	Proposed rendering architecture using $K_{\text{max}} = 3$	32	4
proposed_o1	First order Ambisonics	Proposed rendering architecture using $K_{\text{max}} = 1$	4	2
proposed_tetra	4-sensor open SMA	Proposed rendering architecture using $K_{\text{max}} = 1$	4	2
magls_o1	First order Ambisonics	Magnitude-least squares decoding	4	4

6.2 Test conditions

Two different rooms were simulated using the image-source method. The first was a *small* room, with dimensions $[6 \times 5 \times 3.1]$ m, which had RT60 times of $[0.33, 0.39, 0.26, 0.20, 0.07, 0.04]$ s in octave bands from 125 Hz to 4 kHz. The second was a *medium* sized room ($[10 \times 6 \times 3.15]$ m), which had RT60 times of $[0.52, 0.59, 0.39, 0.20, 0.16, 0.13]$ s. The listener/receiver position was set to $[-0.42, -0.44, 0.18]$ m translated from the centre of the room, and three source positions were placed 1 m away, directly to the left, in-front, and right of the listener. Three different sets of stimuli were selected. These were (left to right): 1) a *band* comprising a broad-band shaker, bass guitar, and synthesised strings; 2) a *mix* comprising a water fountain, female speech, and an acoustic piano; and 3) a *speech* case involving an English male, English female, and Danish male speaking simultaneously. All of the combinations of rooms and sets of stimuli were used for the perceptual study, resulting in 6 test scenes in total.

Five different receivers were then placed at the listener/receiver position in the simulated rooms; these were: 1) a binaural receiver using the same HRTF dataset as used by the rendering methods under test (reference and *hidden_ref*); 2) an ideal SH (Ambisonic) receiver of fourth-order (*o4*); 3) an ideal SH (Ambisonic) receiver of first-order (*o1*); 4) a spherical microphone array (SMA) comprising 32-sensors mounted onto a rigid baffle of 42 mm radius, corresponding to an Eigenmike32 (*em32*); and 5) a SMA comprising four cardioid sensors in an open tetrahedral arrangement with a 20 mm radius (*tetra*).

These array recordings were then processed using the *proposed* rendering architecture, setting $K_{\text{max}} = 1$ for the first-order and tetrahedral receivers, and $K_{\text{max}} = 3$ for the other higher channel count receivers. As an additional control, the COMPASS formulation (*compass*) described in [25], which uses covariance matching for reproducing ambience, and requires all Q channels for the rendering, was used to render the fourth-order Ambisonic recordings. The magnitude least-squares (*magls*) method [27], which was applied to the first-order Ambisonic recordings, was included as a baseline. These rendering methods/configurations under test, along with their required number of transport audio channels, are summarised in Table 1.

6.3 Test procedure

The perceptual study was conducted in purpose-built acoustically dry booths. The test participants wore Sennheiser HD650 headphones, and were presented with an interface featuring a slider for each respective test case, which also depicted the verbal anchors: “Bad", “Poor", “Fair", “Good", and “Excellent" in steps of 20 points from 0 to 100. The subjects were able to switch between the test cases, and were instructed to rate them based on their similarity to the reference condition. It was noted that the subjects should pay attention to the overall sound quality, and to attributes such as sound source localisation, externalisation, timbral colouration, and reverberation characteristics. They were also asked to give due consideration to the verbal anchors provided. The participants took, on average, approximately 20 minutes to complete the tests.

7 Results and discussion

The results of the perceptual study, based on a total of 12 participants, are presented in Fig 2. Note that

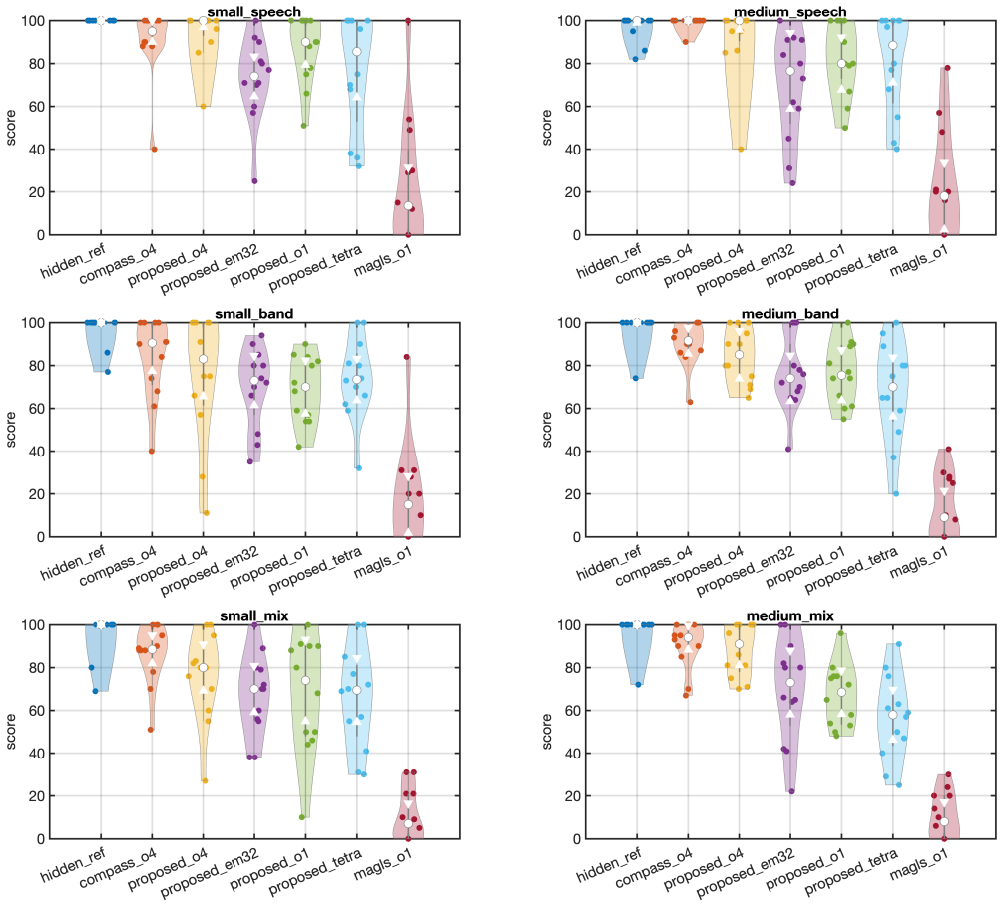

Fig. 2: Listening test results plotted separately for each test scene.

the medians are represented with white circles, the interquartile range is shown as a black vertical line, the 95% confidence intervals are depicted as white triangle markers, and the individual data points are shown as coloured dots[2].

When using the fourth-order Ambisonic receiver, it can be observed that the COMPASS method (using all channels) and the proposed rendering architecture (using limited transport channels) were rated similarly. Their median values were also in the upper range of the perceptual scale, which was denoted with the "Excellent" verbal anchor, for all six test scenes. This suggests that limiting the proposed rendering, by using only four channels to synthesise the binaural output signals, did not notably affect the perceived quality.

The remaining test cases were based on receivers which comprised fewer channels (i.e., the tetrahedral and first-order ambisonic receivers) and/or receivers featuring physical limitations associated with a practical array configuration (i.e., the tetrahedral and Eigenmike32 receivers). These three test cases were all rated lower than the fourth-order ambisonics receiver cases. However, they were all rated as being similar to each other, suggesting that the reduction in transmission channels may not be the factor influencing the performance rating. It is possible that the perceived reduction in quality was more related to lower robustness in estimating the model parameters, and to problems resulting from physical limitations, such as spatial aliasing at higher frequencies. However, it is highlighted that the median scores were still placed within the range denoted by either the “Good" or “Excellent" verbal anchors in the majority of cases. All test cases using the proposed rendering architecture were rated higher than the first-order MagLS condition, which requires either the same or a higher number of channels to construct the binaural playback signals.

[2]https://github.com/bastibe/Violinplot-Matlab

8 Summary

This paper presents a parametric rendering architecture, which may have application in the low bit-rate transmission and reproduction of microphone array recordings. The proposed architecture is divided into dedicated encoder and decoder stages. The encoder estimates the direction from which prominent directional sounds emanate, and then isolates their signals using beamformers (up to a user defined maximum number). These directional signals, along with their respective estimated direction-of-arrival (DoAs) estimates, may then be transmitted to the decoder side. Ambient sounds are then described and transmitted via a dedicated single signal. This monophonic signal is an omnidirectional representation of the input array recording after the directional sounds have been subtracted. The decoder is then tasked with spatialising the directional signals over the target playback setup based on the DoA estimates, while spreading the ambient signal in a spatially incoherent and isotropic manner.

A binaural multiple stimulus listening test evaluated the perceived quality of the proposed rendering architecture. Several configurations of the proposed architecture were investigated, varying the number of transmission channels and using three different types of receivers; namely: a four-channel spherical array transmitted using two transport channels, a 32-channel rigid spherical array transmitted using four transport channels, and ideal first- and fourth-order Ambisonic receivers transmitted using two and four transport channels, respectively. The results of the perceptual study indicate that the proposed architecture, when using the fourth-order receiver (with four transmission channels), produces renderings which are similar to both direct binaural reference conditions, and to renders which are based on a comparable parametric method using 25 channels for the reproduction. Reduced perceptual quality was identified for the receivers comprising fewer channels, and for those which suffer from physical limitations (leading to e.g., spatial aliasing). However, all test cases using the proposed rendering architecture were still rated notably higher than a state-of-the-art signal-independent method, which uses the same or a higher number of channels for synthesising the binaural signals.

References

[1] Zotter, F. and Frank, M., *Ambisonics: A Practical 3D Audio Theory for Recording, Studio Production, Sound Reinforcement, and Virtual Reality*, Springer Nature, 2019, doi:10.1007/978-3-030-17207-7.

[2] Pihlajamäki, T. and Pulkki, V., "Projecting simulated or recorded spatial sound onto 3D-surfaces," in *Audio Engineering Society Conference: 45th International Conference: Applications of Time-Frequency Processing in Audio*, Audio Engineering Society, 2012.

[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] 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.

[6] 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.

[7] Sen, D., Peters, N., Kim, M., and Morrell, M., "Efficient compression and transportation of scene-based audio for television broadcast," in *Audio Engineering Society Conference: 2016 AES International Conference on Sound Field Control*, Audio Engineering Society, 2016.

[8] Herre, J., Kjörling, K., Breebaart, J., Faller, C., Disch, S., Purnhagen, H., Koppens, J., Hilpert, J., Rödén, J., Oomen, W., et al., "MPEG surround-the ISO/MPEG standard for efficient and compatible multichannel audio coding," *Journal of the Audio Engineering Society*, 56(11), pp. 932–955, 2008.

[9] Rudzki, T., Gomez-Lanzaco, I., Stubbs, J., Skoglund, J., Murphy, D. T., and Kearney, G., "Auditory localization in low-bitrate compressed ambisonic scenes," *Applied Sciences*, 9(13), p. 2618, 2019.

[10] Xu, J., Niu, Y., Wu, X., and Qu, T., "Higher order ambisonics compression method based on independent component analysis," in *Audio Engineering Society Convention 150*, Audio Engineering Society, 2021.

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

[12] Gerzon, M. A., "Periphony: With-height sound reproduction," *Journal of the audio engineering society*, 21(1), pp. 2–10, 1973.

[13] Hold, C., Pulkki, V., Politis, A., and McCormack, L., "Compression of Higher-Order Ambisonic Signals using Directional Audio Coding," *Submitted for review*, 2022.

[14] 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.

[15] Rafaely, B., *Fundamentals of spherical array processing*, volume 8, Springer, 2015, doi:10.1007/978-3-319-99561-8.

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

[17] Han, K. and Nehorai, A., "Improved source number detection and direction estimation with nested arrays and ULAs using jackknifing," *IEEE Trans. Signal Processing*, 61(23), pp. 6118–6128, 2013, doi:10.1109/TSP.2013.2283462.

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

[19] 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*, 2022.

[20] 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.

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

[22] McCormack, L., Politis, A., and Pulkki, V., "Rendering of source spread for arbitrary playback setups based on spatial covariance matching," in *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, IEEE, 2021.

[23] 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.

[24] Anemüller, C., Adami, A., and Herre, J., "Efficient Binaural Rendering of Spatially Extended Sound Sources," *Journal of the Audio Engineering Society*, 71(5), pp. 281–292, 2023.

[25] McCormack, L. and Politis, A., "Estimating and Reproducing Ambience in Ambisonic Recordings," in *30th European Signal Processing Conference (EUSIPCO)*, pp. 314–318, EURASIP, 2022.

[26] Armstrong, C., Thresh, L., Murphy, D., and Kearney, G., "A perceptual evaluation of individual and non-individual HRTFs: A case study of the SADIE II database," *Applied Sciences*, 8(11), p. 2029, 2018.

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