Hold, Christoph; Pulkki, Ville; Politis, Archontis; McCormack, Leo

Compression of Higher-Order Ambisonic Signals using Directional Audio Coding

Published in:
IEEE/ACM Transactions on Audio Speech and Language Processing

DOI:
10.1109/TASLP.2023.3328284

Published: 01/01/2024

Document Version
Publisher's PDF, also known as Version of record

Compression of Higher-Order Ambisonic Signals Using Directional Audio Coding

Christoph Hold, *Graduate Student Member, IEEE*, Ville Pulkki, Archontis Politis, and Leo McCormack

***Abstract*—Delivering high-quality spatial audio in the Ambisonics format requires extensive data bandwidth, which may render it inaccessible for many low-bandwidth applications. Existing widely-available multi-channel audio compression codecs are not designed to consider the characteristic inter-channel relations inherent to the Ambisonics format, and thus may not leverage this knowledge to optimise the compression. Therefore, this article proposes a spatial audio compression algorithm, based on a novel reformulation of the Higher-Order Directional Audio Coding (HO-DirAC) method, which is specifically intended for compressing higher-order Ambisonic audio streams. The methodology builds upon the concept of a spherical filter bank acting in the spherical harmonic domain. This results in directionally constrained sound-field estimates and parameterization, which may be utilized to reconstruct the input Ambisonic signals with minimal perceived loss of quality. The results of a listening experiment indicate high perceptual quality when using six or more audio transport channels to deliver fifth-order (36 channels) Ambisonic sound scenes. The proposed formulation is also designed with low computational complexity in mind and may therefore be well suited for compressing Ambisonic sound scenes for a wide range of applications.**

***Index Terms*—Ambisonics, spatial audio, audio coding.**

I. Introduction

SPATIAL audio has become well-established within many applications related to immersive media production and consumption. Transmitting high-resolution spatial audio, however, still remains a key challenge in cases where the available data bandwidth is limited. Moreover, the scene-based higher-order Ambisonics format has recently seen greater adoption as an alternative to traditional channel-based formats, such as stereophony and related multi-channel extensions; owing to its ability to store spatial sound scenes in a directionally continuous spherical harmonic (SH) representation, which decouples the recording setup from the playback system. Transmitting spatial audio in the Ambisonics format has therefore become a popular option, since the playback setup does not need to be known to the (compression) codec, and the format permits greater playback flexibility on the receiving end. However, as is demanded by the SH basis, the number of required audio signals scales quadratically with the SH representation order, which dictates the available spatial resolution of the Ambisonics format. Therefore, the bandwidth required for transmitting higher-order sound scenes at the desired spatial resolution can become considerable and potentially prohibitive within applications where data bandwidth is limited.

The practical challenges associated with such high-resolution sound scenes become more apparent upon reviewing relevant listening test studies. These studies indicate that sufficient Ambisonic orders required for a perceptually transparent representation of sound-fields demand channel counts which may be far beyond the typical capabilities of consumer devices; where, in summary, increasing the SH order is generally linked to an improved perceptual quality up to a perceptual threshold. Moreover, processing and transmitting data consumes energy, which may further limit the maximum possible order for low-power devices. While simply truncating the Ambisonic order will lead to a reduction in the number of necessary audio channels, this may be undesirable as it has also been shown to degrade the perceived audio quality [1], [2].

Existing audio compression codecs are generally unsuitable for the task of Higher-Order Ambisonics (HOA) compression, as they are not specifically designed with such signals in mind. While there are multi-channel audio compression codecs available [3], [4], the majority do not leverage the known inter-channel relationships that are specific and inherent to the Ambisonics format, such as the high spatial redundancy of encoded plane-wave components [5], which may be exploited to improve the compression coding gain and/or the perceived robustness of the compression algorithm. Recent investigations regarding localization of compressed Ambisonic sound scenes using multi-channel codecs [4] (OPUS) indicate that while the perceptual quality improves with higher Ambisonic order, the required bandwidth was also shown to increase substantially.

There are two main existing approaches which have been considered for the task of multi-channel audio compression in the context of HOA. The first class of compression algorithms operate on the statistical signal structure and typically transform the input signals into statistically orthogonal components, e.g., by means of a singular value decomposition (SVD), in order

Manuscript received 26 July 2022; revised 22 June 2023 and 30 August 2023; accepted 5 October 2023. Date of publication 17 November 2023; date of current version 8 December 2023. This work was supported by the Fraunhofer IIS, and Academy of Finland. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Jens Ahrens. *(Corresponding author: Christoph Hold.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by Aalto University Research Ethics Committee under Application No. D11217103.0412021, and performed in line with the Declaration of Helsinki.

Christoph Hold, Ville Pulkki, and Leo McCormack are with the Acoustics Lab, Department of Information and Communications Engineering, Aalto University, 02150 Espoo, Finland (e-mail: christoph.hold@aalto.fi; ville.pulkki@aalto.fi; leo.mccormack@aalto.fi).

Archontis Politis is with the Faculty of Information Technology and Communication Sciences, Tampere University, 33100 Tampere, Finland (e-mail: archontis.politis@tuni.fi).

Digital Object Identifier 10.1109/TASLP.2023.3328284

to determine the components of high statistical relevance to the signal. Compression may then be achieved by discarding components of less statistical relevance. As this approach is purely signal driven, HOA compression based on SVD may suffer from the ambiguity of SVD components, especially during update steps, which can introduce unwanted artifacts [6]. In order to smooth transitions and minimize an unstable sound-field image, one may apply a reordering of the SVD components, along with interpolation [6], [7]. This methodology has led to a number of proposed improvements for spatial audio, e.g., by decomposing the mel-frequency space [6], [8], [9], or by tailoring the decomposition method using independent component analysis [10]. For example, implementations of the MPEG-H codec associate direct sound components with sound sources and separate the scene into predominant sounds, coded as parameterized plane-waves, and a residual stream after extraction [11], [12], [13], [14]. The residual is then transformed into virtual loudspeaker signals, in order to feed perceptual coders, which are then converted back into the Ambisonics format [11], [13].

The second class of compression algorithms are model based, which introduce domain-specific knowledge and assumptions. Given that the human auditory perception is limited in certain known respects [15], compression may be achieved by determining and isolating only those components that are deemed to be perceptually relevant by the employed sound-field model. For example, such perceptual coding model could leverage minimum audible angles and spatial masking [16]. One established perceptually-motivated sound-field model, which assumes a single dominant source and an isotropic diffuse component for each time-frequency band, forms the basis for the Directional Audio Coding (DirAC) [17] method. Despite its simplicity, listening test studies have demonstrated high perceptual quality of the model, given relatively simple sound scenes as input [18]. However, complex sound scenes, which may easily violate the single-source assumption, can produce undesirable audible artifacts [19], [20]. Furthermore, the original DirAC formulation [17], [18], [21] was developed based on first-order Ambisonic (B-Format) signals only and thus cannot exploit the additional information available in higher-order Ambisonic encodings. Therefore, multiple extensions and alternatives to the first-order DirAC formulation were proposed. Most notably, Higher-Order Directional Audio Coding (HO-DirAC) segregates the sound-field into multiple directionally constrained components, wherein each of them apply the DirAC sound-field model [19]. However, the output of this formulation is optimized to enhance loudspeaker [19] or binaural [20] reproduction and is therefore not directly applicable to Ambisonic audio compression.

In addition, different sound-field models lead to different codecs, of which the Coding and Multidirectional Parameterization of Ambisonic Sound Scenes (COMPASS) [22] method is notable here. COMPASS first estimates the number of active sound sources, and then localizes and extracts them in the Ambisonics domain, leaving a residual signal. While shown to provide very high performance for spatial upmixing, i.e., extending the spherical harmonics representation order, the compression capabilities currently remain unexplored. Furthermore, the methodology introduces more assumptions and is therefore potentially more sensitive to audible estimation errors. The algorithm is also quite complex and potentially more computationally demanding than HO-DirAC, especially for higher input orders. All of the aforementioned alternative parametric sound-field models, however, have been employed within the context of reproduction enhancement and have yet to be investigated for the task of compression and transmission.

Given the widespread adoption of the Ambisonics format, there is a clear need for suitable and robust spatial audio codecs specifically intended for the format. Although statistical compression approaches have been explored extensively in recent years, perceptually-motivated model-based approaches have seen far fewer developments. Here, the strategy is to first seek to obtain an intermediate model-based representation of the Ambisonic sound scene requiring fewer audio channels. When accompanied with suitable metadata, these transported audio channels may be used to reconstruct the Ambisonic signals in a perceptually-motivated manner. Naturally, one could then explore the application of existing multi-channel coders to these intermediate transport signals, which may further improve compression compared to applying them to the Ambisonic signals directly.

We emphasise that only the original DirAC formulation [17] was specifically intended for the task of coding and compression; requiring a single audio transport channel accompanied by parameter metadata, as utilized in [23]. However, all subsequent formulations of DirAC, including existing higher-order extensions [19], [20], instead focused on the task of spatial enhancement and reproduction, and thus, all input channels were employed to reproduce the parameterized sound scene, with no compression achieved. The primary novelty of the presented work is, therefore, in the proposal of a new perceptually-motivated spatial audio compression codec, based on the HO-DirAC architecture, for the task of audio coding and compressing Ambisonic sound scenes. The proposal adopts the higher-order analysis of [19] and reformulates the higher-order synthesis conducted in [24] for HOA input and output by integrating the spherical filter bank allowing reconstruction described in [25], [26]. The proposed HO-DirAC formulation is accompanied by a perceptual evaluation, which demonstrates large potential compression gains through the reduction of transmission channels, with minimal degradation in perceived audio quality.

The remainder of this article is structured as follows. A synopsis of the work carried out in [25] and [26] is presented in Section II. Section III describes the proposed HO-DirAC methodology, including the theoretical background for the presented spherical filter bank approach. Thereafter, the listening test used to evaluate the proposed re-formulation is described in Section IV. After discussing the results in Section V, we conclude in Section VI.

II. Theoretical Background

Higher-order Ambisonic signals may be expressed using real-valued spherical harmonics $Y_n^m(\Omega)$ as given in [27, (2.53)]. Consider a signal incident from $\Omega = [\phi, \theta] \in \mathbb{S}^2$, with azimuth

angle ϕ and zenith/colatitude angle θ, which is spatially band-limited to SH order N. The SH basis functions may be used to represent the spherical function as a linear combination of SH spectrum coefficients obtained from the spherical harmonic transform (SHT). As the latter requires evaluating a continuous spherical integral, the SHT is discretized in practice by a quadrature rule. The chosen quadrature grid needs to be sufficient to discretize spherical polynomials of degree $2\,N$. When sampling a signal that is spatially band-limited to order N, the discrete SH transform up to order N in matrix notation is [28, (3.35)]

$$\boldsymbol{\chi} = \boldsymbol{Y}^{\mathsf{H}}\,\mathrm{diag}\,[\boldsymbol{\alpha}]\,\boldsymbol{x}\,, \tag{1}$$

where the $Q \times (N+1)^2$ matrix $\boldsymbol{Y}$ contains the SHs evaluated for grid directions Ω_q, $\boldsymbol{\chi}$ is the SH representation of $\boldsymbol{x}$, $[\cdot]^{\mathsf{H}}$ denotes the Hermitian transpose, $\mathrm{diag}[\cdot]$ constructs a diagonal matrix using the values of the enclosed vector, and the corresponding Q quadrature grid weights are denoted as $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_Q]$. In the special case of a uniform spherical grid, such as a spherical t-design as in [29] or [30], all weights are equal by definition, i.e., $\alpha_q = \frac{4\pi}{Q}$. The inverse spherical harmonic transform (iSHT) in matrix notation is then

$$\boldsymbol{x} = \boldsymbol{Y}\boldsymbol{\chi}\,, \tag{2}$$

obtaining spherically discrete signals $\boldsymbol{x}$ from the spherically continuous SH representation $\boldsymbol{\chi}$.

The orthonormality of SHs follows from (1) and (2)

$$\boldsymbol{Y}^{\mathsf{H}}\,\mathrm{diag}\,[\boldsymbol{\alpha}]\,\boldsymbol{Y} = \boldsymbol{I}, \tag{3}$$

which implies that the integration of SH

$$\int_{\mathbb{S}^2} Y_n^m(\Omega)\,d\Omega = \sqrt{4\pi}\delta_{n0}\,, \tag{4}$$

vanishes for all components, except $n = m = 0$, written as δ_{n0} [28, (1.22)].

Quadrature grids can discretize the latter continuous integral, e.g., with J uniform nodes at Ω_j on a sufficient spherical t-design [29], [30] as

$$\int_{\mathbb{S}^2} f(\Omega)\,d\Omega = \frac{4\pi}{J}\sum_{j=1}^{J} f(\Omega_j)\,. \tag{5}$$

Parseval's relation establishes a connection between the discrete and spectral (SH) domain, e.g., for functions f and g as [28, (1.44)]

$$\int_{\mathbb{S}^2} f(\Omega)g^*(\Omega)\,d\Omega = \sum_{n=0}^{N}\sum_{m=-n}^{n} f_{nm}g_{nm}^* = \boldsymbol{g}^{\mathsf{H}}\boldsymbol{f}\,, \tag{6}$$

where the sum can be written compactly as the inner product of the two spectral coefficient vectors $\mathbf{g}$ and $\mathbf{f}$ with stacked coefficients, using the Ambisonic Channel Number (ACN) $n^2 + n + m$ convention [31].

The SH addition theorem is [28, (1.26)]

$$\sum_{m=-n}^{n} [Y_n^m(\Omega_1)]^*\,Y_n^m(\Omega_2) = \frac{2n+1}{4\pi}P_n(\cos(\Theta))\,, \tag{7}$$

with the angle $\Theta = \angle(\Omega_1, \Omega_2)$ and where P_n is the Legendre polynomial of degree n. With the latter, any directional weighting of axisymmetric spherical array beam patterns can be described by their $N+1$ beamforming coefficients or modal weighting coefficients c_n as [28, (5.24)]

$$w(\Theta) = \sum_{n=0}^{N}\frac{2n+1}{4\pi}c_n P_n(\cos(\Theta))\,. \tag{8}$$

III. Proposed Algorithm

This paper proposes a coding algorithm that transmits Ambisonic input ($\boldsymbol{\chi}$) to Ambisonic output ($\tilde{\boldsymbol{\chi}}$), as depicted in Fig. 1. The input signal is decomposed in terms of time-frequency, as well as direction. The codec leverages a time-frequency filter bank to obtain a time-frequency decomposition, and a spherical filter bank to obtain a directional decomposition (Section III-A) of the input. For the latter, a distributed set of beamformers directionally decomposes the amplitude density of a spherical sound-field. Spherically localized sector processing then enables extracting directionally constrained metadata describing the sound-field, previously proposed in HO-DirAC [19]. The decomposition, coding, and re-encoding structure allows controlling the number of channels describing an Ambisonic audio scene. Compression is achieved by transmitting fewer channels than provided at the input by utilizing the following signal, perceptual, and sound-field assumptions:

- The sound-field can be represented by a combination of directional and non-directional components, obtained at sampled points in time, frequency, and direction.
- Directional signal components exhibit a higher SH order than less-directional components.
- The perceptual spatial resolution of simultaneous sound events per time-frequency tile is limited.

This section will first detail the directional decomposition and re-encoding strategy, and then derive an encoder and decoder from it.

A. Spherical Filter Bank

A complete set of directional filters may decompose the sound-field into directionally constrained regions, forming an intermediate representation of the Ambisonic sound scene. Since the method should not introduce any skewing in the sound-field representation, the filtering needs to uniformly decompose the entire sphere. Furthermore, the objective is to formulate this filtering operation to be *invertible*, i.e., allowing restoring a spherical harmonic domain (SHD) signal. This uniform filtering over the entire sphere with its inversion properties eventually leads to the spherical filter bank interpretation explored in [25], [26] and builds the foundation upon which the codec algorithm is based.

1) Spherical Filter Bank Framework: Directional weighting and extraction of signals from the input sound-field, i.e., beamforming, is most conveniently carried out through operations applied directly in the SHD. Extracting a portion of the input SH signal is obtained as the inner product between the corresponding SH spectra (see (6)) of the beamformer weighting pattern and the SH signal. The weighting $w(\Omega)$ in the spherically discrete domain corresponds to the vector $\boldsymbol{w}$ in the SHD. Dropping the

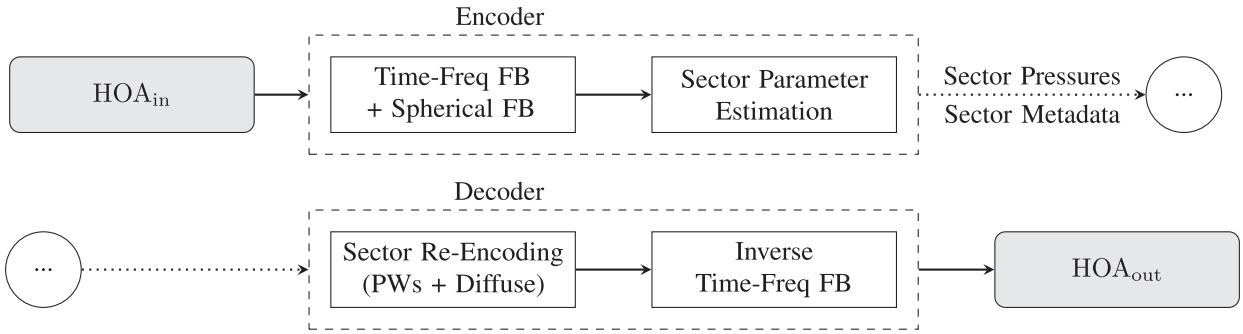

Fig. 1. Flow diagram of the proposed Ambisonic signal compression codec, where the solid lines denote audio signals and dotted lines denote transmission. Ambisonic input signal $\boldsymbol{\chi}$ passes through the encoder, which extracts (per time-frame) a set of intermediate signals via static beamforming, referred to as a spherical filter bank (Spherical FB), along with parameters for transmission. The decoder re-encodes the latter accordingly by synthesizing plane-waves (PWs) and a diffuse stream to form the output Ambisonic signal $\tilde{\boldsymbol{\chi}}$.

time and frequency dependency for convenience, this means that beamforming results in a scalar x with

$$x = \boldsymbol{w}^{\mathsf{H}} \boldsymbol{\chi} \,. \tag{9}$$

Furthermore, the SH spectrum coefficients of any axisymmetric beamformer is determined by a set of $N+1$ order weighting coefficients c_n (see (8)). Therefore, its SHD coefficients can be obtained directly for the pattern in any steering direction Ω', with the SH steering vector $\boldsymbol{y}$ stacking $L = (N+1)^2$ coefficients $y_n^m(\Omega')$, from

$$\boldsymbol{w} = \mathrm{diag}_N \left[\boldsymbol{c}_n\right] \boldsymbol{y} \,, \tag{10}$$

where the operator $\mathrm{diag}_N[\cdot]$ expands each vector entry $2n+1$ times to a diagonal matrix, effectively replicating the $N+1$ order/modal weightings $\boldsymbol{c}_n$ for each order.

In the filter bank framework, we define a $J \times L$ analysis matrix $\boldsymbol{A}$ and a $L \times J$ synthesis matrix $\boldsymbol{B}$. These are labelled in the codec depiction in Fig. 1 as a spherical filter bank (SFB) at the encoder, with the inverse operation (i.e., the sector re-encoding) at the decoder, respectively. The latter form a modal spectrum analysis and synthesis pair often occurring in signal processing. The analysis acts on signals in the SHD, providing J signals, one per sector $\xi = 1, \ldots, J$, as

$$\boldsymbol{x} = \boldsymbol{A} \boldsymbol{\chi} \,. \tag{11}$$

The now directionally discrete (beamformer) signals $\boldsymbol{x}$ may be analyzed or further manipulated, and offer an intuitive intermediate format. Ultimately, the signals $\boldsymbol{x}$ are re-encoded to the SHD as $\tilde{\boldsymbol{\chi}}$, so that both input and output are available in the same domain. Therefore, we define a synthesis matrix $\boldsymbol{B}$ to re-encode the signals $\boldsymbol{x}$ into the SH domain as

$$\tilde{\boldsymbol{\chi}} = \boldsymbol{B} \boldsymbol{x}. \tag{12}$$

The key challenge is to develop suitable pairs of the analysis matrix $\boldsymbol{A}$ and synthesis matrix $\boldsymbol{B}$, which meet individually defined criteria. For example, the analysis may be formulated to decompose the input signal utilizing specifically designed patterns, while the synthesis still perfectly reconstructs the SHD input signal. This article will also utilize other criteria relevant for coding. Assuming the same pattern for each beamformer (10), steered towards directions $\boldsymbol{\Omega}_\xi$, we may define an analysis matrix $\boldsymbol{A}$ as

$$\boldsymbol{A} = \boldsymbol{Y} \, \mathrm{diag}_N \left[\boldsymbol{c}_n^{\text{an}}\right] \,, \tag{13}$$

and the synthesis matrix $\boldsymbol{B}$ accordingly as

$$\boldsymbol{B} = \mathrm{diag}_N \left[\boldsymbol{c}_n^{\text{syn}}\right] \boldsymbol{Y}^{\mathsf{H}} \,. \tag{14}$$

Typically, the analysis pattern is a design choice, giving coefficients $\boldsymbol{c}_n^{\text{an}}$. Hence, the coefficients $\boldsymbol{c}_n^{\text{syn}}$ are then derived based upon coefficients $\boldsymbol{c}_n^{\text{an}}$, in order to meet certain input-to-output signal relations, as detailed later.

The beamformers at the encoder optimize different criteria, such as maximum directivity or optimal back-lobe suppression of each individual beam. The maximum directivity ($\max \mathrm{DI}$) beamformer can be considered standard practice and the coefficients are constant for all orders [28, (6.10)] with

$$c_n^{\max \mathrm{DI}} = \frac{4\pi}{(N+1)^2}. \tag{15}$$

Another name and interpretation of the same weighting is the higher-order hyper-cardioid pattern, or normalized plane-wave decomposition (PWD). Despite its desirable property of maximum directivity, the resulting considerable back- and side-lobes of this pattern may cause problems in the present encoder design. Therefore, other well-established axisymmetric patterns may optimize more suitable criteria. With similar properties as super-cardioids, the perceptually-motivated $\max \boldsymbol{r}_{\mathrm{E}}$ pattern became a popular choice in spatial audio. Its modal weighting coefficients are directly given as [32, (10)]

$$c_n^{\max \boldsymbol{r}_{\mathrm{E}}} = P_n \left(\cos \left(\frac{2.4068}{N + 1.51} \right) \right). \tag{16}$$

Crucially, this pattern exhibits a considerably less prominent back-lobe, hence less critical spatial interference, and is therefore preferred in this application and used henceforth. Most importantly, the beamformer shaping can improve the parameter estimation. Note that there are many more suitable pattern choices, for which some further details can be found, e.g., in [25]. Typically, the patterns are normalized to unit amplitude in the steering direction, which simplifies their output interpretation.

Any axisymmetric pattern may be normalized in its steering direction by a weighted band-limited spherical Dirac $a_{\text{norm}} = \sum_{n=0}^{N} c_n \frac{2n+1}{4\pi}$. Note that the presented framework does not require unit amplitude in the steering direction and can therefore accommodate other normalizations.

2) Preservation Objectives – Encoder: The proposed encoder design splits the SH input into multiple directionally constrained sector signals, where the latter shall preserve the total sound-field properties. The objective translates to observe the sum over all ξ encoder weightings $w_\xi(\Omega)$ for all points on the unit sphere Ω, which may preserve amplitude as

$$\sum_{\xi=1}^{J} \beta_{\text{A}} w_\xi(\Omega) = 1,\ \forall \Omega \in \mathbb{S}^2\,, \tag{17}$$

or preserve energy (i.e., domain of squares) as

$$\sum_{\xi=1}^{J} \beta_{\text{E}} w_\xi^2(\Omega) = \sum_{\xi=1}^{J} \left[\sqrt{\beta_{\text{E}}} w_\xi(\Omega)\right]^2 = 1,\ \forall \Omega \in \mathbb{S}^2\,, \tag{18}$$

where β_{A} and β_{E} are the factors for amplitude and energy preservation, respectively. Since the directional weighting is carried out as beamforming directly in the SHD, the beamformer can not achieve complete separation between beams in the limited expansion orders that are present at the encoder. This interaction causes a scaling error and, therefore, introducing preservation factors β_{A} and β_{E} enables restoring the sum of the outputs to scale to unity. The factors are derived such that it allows both arbitrary axisymmetric filter patterns, as well as arbitrary scaling thereof, such as unity gain in the steering direction. To this end, defining preservation objectives (17) and (18), demands that the sum over all patterns scales to unity, as this property is not guaranteed from the beam shape design. Furthermore, choosing a uniform steering layout collapses each factor to a single, direction independent scalar value. For axisymmetric patterns, arranged on a uniform grid, the sum over all encoder sectors is restored to unity by (derived in [25])

$$\beta_{\text{A}} = \left[\sum_{\xi=1}^{J} w_\xi(\Omega)\right]^{-1} = \frac{4\pi}{c_0\, J} = \frac{\sqrt{4\pi}}{w_{00}\, J}\,, \tag{19}$$

and

$$\beta_{\text{E}} = \left[\sum_{\xi=1}^{J} w_\xi^2(\Omega)\right]^{-1} = \frac{4\pi}{\boldsymbol{w}^{\text{H}}\boldsymbol{w}\, J}\,. \tag{20}$$

3) Reconstruction Criteria – Decoder: The proposed codec is based on a spherical filter bank, which acts on SHD signals followed by a re-encoding back to the SHD. Perfect reconstruction is achieved if the signal restored from the filter bank $\tilde{\boldsymbol{\chi}}(t)$ exactly matches the SH input signal $\boldsymbol{\chi}(t)$, i.e.,

$$\tilde{\boldsymbol{\chi}}(t) \stackrel{!}{=} \boldsymbol{\chi}(t),\ \forall t\,, \tag{21}$$

where t denotes time. Alternatively, a filter-bank input-to-output signal relation may reconstruct the signal energy expectation $\mathbb{E}[\cdot]$, such that

$$\mathbb{E}\left[|\tilde{\boldsymbol{\chi}}(t)|^2\right] \stackrel{!}{=} \mathbb{E}\left[|\boldsymbol{\chi}(t)|^2\right]\,. \tag{22}$$

It follows that perfect reconstruction also reconstructs energy. However, if the reconstructed output signal cannot match the input signal perfectly, e.g., because of intermediate (signal-dependent) processing, the input-to-output signal relation may still match in terms of energy. The compression framework will subsequently introduce reconstruction of a certain subset of channels. Due to the nature of SHs, we may gradually reconstruct a subset of SHD signals. Furthermore, we deem the reconstruction of the lower SHD components to be perceptually more important. Later, the directional parameterization will aid in an informed approximation of the higher orders.

4) Number of Sectors: From the latter introduced concepts we can directly ascertain the number of required sectors (and hence the number of transport channels). The encoder uses uniform steering over the sphere, as explained previously, where the minimal number of sectors J depends on the encoder preservation objectives (17) and (18). Amplitude preservation as in (17) requires a quadrature of at least degree $N+1$, whereas energy preservation as in (18) requires a quadrature of at least degree $2\,N$ [25]. However, quadratures of higher degree integrate lower orders correctly, and a quadrature allowing energy preservation also allows amplitude preservation, since $2\,N \geq N+1$ for all $N \geq 1$.

The minimum number of required channels at the decoder depends on the reconstruction criteria. Perfect reconstruction is essentially a modified SH transform pair, implying that perfect reconstruction can only be achieved on a sufficiently dense grid for order $2\,N$. More relaxed criteria can still be achieved with less dense grids. In the coding context, particularly a re-encoding obeying amplitude preservation at the encoder is interesting (as in (17)), because it requires fewer grid points for $N > 1$ and, consequently, fewer signals to be transmitted. In practice, we choose a transport channel count (based on a supporting grid) leading to a scaling of the transmitted data; with the grid dictating the analysis order used to construct $\boldsymbol{A}$.

5) Amplitude Reconstruction: The decoder needs to reconstruct the Ambisonic signals from the transmitted sector signals by re-encoding them into the SHD. Perfect reconstruction is considered when the input and output appears to be identical (see (21)).

Given the considered compression context, we aim to design for fewer sectors than input SHs, since these sector pressure signals make up the transport channels for the presented framework. This means that typically $J < L_{\text{in}}$. Therefore, $\boldsymbol{A}$ is usually designed to decompose an order which is lower than the input order, i.e., $L < L_{\text{in}}$, which in turn implies that only the lower orders may be fully reconstructed. Generally, we can face two scenarios. First, the number of sectors is greater or equal than SHD coefficients of the decomposing order of $\boldsymbol{A}$. With $J \geq L$, given a sensible beamformer layout, the matrix $\boldsymbol{A}$ has full column-rank.

Perfect reconstruction in relation to the chosen analysis matrix $\boldsymbol{A}$ is generally given by the left inverse or pseudoinverse denoted as $[\cdot]^\dagger$, hence

$$\boldsymbol{B} = [\boldsymbol{A}]^\dagger\,. \tag{23}$$

Second, the number of sectors may be smaller than the number of SHs in $\boldsymbol{A}$, i.e., $J < L$, which may occur in the amplitude preserving sector steering explored above. The pseudoinverse then gives a least-squares solution recovering a subset, thus, not reconstructing all SH components, where the specifics depends on the geometry of the sector steering. It follows that $\mathrm{trace}[\boldsymbol{BA}] = J$, rather than being equal to L, indicating a loss of energy compared to the perfect reconstruction case. Due to the structure of SHs, however, the lower order components are typically matched, and deviations start to occur towards higher orders. The pseudoinverse also takes the energy of aliasing into account, which may cause the lower diagonal elements of $\boldsymbol{BA}$ to deviate from identity.

However, there are a number of reasons why an analytical solution is useful in practice. Perfect reconstruction may not always be desired, e.g., in the case of energy reconstruction, as is evident from (22). Fortunately, the analysis matrix $\boldsymbol{A}$ is defined in a way that allows formulating $\boldsymbol{B}$ directly and without explicit matrix inversion. Therefore, we next develop a solution that is mostly equivalent to the aforementioned pseudoinverse, but offers further insight and control over the reconstruction properties. Furthermore, we may explicitly target the reconstruction of a subset of (lower order) signal components.

In order to simplify the problem, we assume the same sector directions are used for the synthesis and analysis. We will further first assume full order reconstruction, i.e., full rank in $\boldsymbol{A}$. Since the design choice is restricted to axisymmetric beamformers on a regular grid, we may also restrict the synthesis matrix as introduced in (14), noting again that we may instead also employ numerical solutions. The unknown is then the coefficients $\boldsymbol{c}_n^{\text{syn}}$, which will enable perfect reconstruction in the inverse spherical filter bank. Under the perfect reconstruction of (21), from analysis (11), and synthesis (12), it follows that

$$\boldsymbol{BA} \overset{!}{=} \boldsymbol{I}\,. \tag{24}$$

First, expanded with (13) and (14), this shows

$$\mathrm{diag}_N\left[\boldsymbol{c}_n^{\text{syn}}\right] \boldsymbol{Y}^{\mathsf{H}} \boldsymbol{Y}\, \mathrm{diag}_N\left[\boldsymbol{c}_n^{\text{an}}\right] \overset{!}{=} \boldsymbol{I}\,. \tag{25}$$

The orthonormality property defined in (3), on a uniform grid, reveals that

$$\mathrm{diag}_N\left[\boldsymbol{c}_n^{\text{syn}}\right] \mathrm{diag}_N\left[\boldsymbol{c}_n^{\text{an}}\right] \overset{!}{=} \frac{J}{4\pi}\boldsymbol{I}\,, \tag{26}$$

and therefore shows for real-valued modal weighting that

$$c_n^{\text{syn}} \propto \frac{1}{c_n^{\text{an}}}\,, \text{ if } c_n^{\text{an}} \neq 0\,. \tag{27}$$

Intuitively, this means that the analysis pattern effect can be reverted, provided that the SH spectrum components are not set to zero.

From the previous observations, we can present a solution for perfect reconstruction in the current framework. Meeting the grid requirements for perfect reconstruction, and connecting with the previously shown β_{A} from the encoding, $c_n^{\text{syn}} = \beta_{\text{A}} c_0^{\text{an}}/c_n^{\text{an}}$ solves for perfect reconstruction. It follows that

$$\tilde{\boldsymbol{\chi}} = \mathrm{diag}_N\left[\boldsymbol{c}_n^{\text{syn}}\right] \boldsymbol{Y}^{\mathsf{H}} \boldsymbol{x}(t) = \beta_{\text{A}} \boldsymbol{B} \boldsymbol{x}(t)\,, \tag{28}$$

where $\boldsymbol{B}$ uses (14) with $c_n^{\text{syn}} = c_0^{\text{an}}/c_n^{\text{an}}$. Essentially, we see two parts in (28): β_{A} restoring the sum of signals of the encoder patterns to unity and $\boldsymbol{B}$ defining the re-encoding. The matrix $\boldsymbol{B}$ can also be designed with coefficients $\boldsymbol{c}_n^{\text{syn}}$ not aiming for perfect reconstruction of all components, which we will demonstrate in the following.

The previously shown perfect reconstruction, demanding recovering all SHD spectrum components, imposes the strictest form of reconstruction criteria. We indicated earlier, however, that analyzing the amplitude of a sound-field pressure density with a set of beamformers of order N only requires a uniform discretization quadrature of degree $N+1$. Even though this relaxed grid condition does not allow re-synthesizing the exact input on the same grid locations, (since perfect reconstruction demands a minimum quadrature of degree $2\,N$), the signal is still contained in the beamformer outputs in the sense of (17). It follows that whereas the relaxed grid condition is not capable of perfectly reconstructing all SH signals, one may still reconstruct lower order components. While the aforementioned pseudoinverse gives the least-squares matching J of the L components, we found that we can also target the reconstruction as a function of order explicitly (similar to truncating the columns of $\boldsymbol{A}$ before finding the pseudoinverse). While acknowledging that $\boldsymbol{A}$ is not of full rank in this case, the developed solution still yields sufficient results in practice, while retaining control and the intuition of reverting the modal weighting in $\boldsymbol{A}$. The solution can therefore be seen as a partial inversion of analysis matrix $\boldsymbol{A}$, which is formulated within the strict bounds of only a simple modal weighting onto the orthonormal SH basis, evaluated on a uniform grid. For low orders $\tilde{N} \leq \lfloor \sqrt{J} - 1 \rfloor$, we may hence recover with $c_n^{\text{syn}} = c_0^{\text{an}}/c_n^{\text{an}}$, and for the remaining components we may set $c_n^{\text{syn}} = 1$. The latter could be manipulated to adjust the $\mathrm{trace}[\boldsymbol{BA}]$. An important special case is the solution for recovering the $n = 0$ component during re-encoding to the SH domain, which is proportional to the (omni-directional) sound-field pressure p_0. The scaling acting on p_0 is found from [26]

$$c_0^{\text{syn}} \frac{J}{4\pi} c_0^{\text{an}} p_0 = \beta_{\text{A}}^{-1} c_0^{\text{syn}} p_0\,. \tag{29}$$

From the above, we can spot again that when incorporating the preservation factor β_{A} during reconstruction as in (28), the sound-field pressure is reconstructed for any synthesis pattern normalized as

$$c_0^{\text{syn}} = 1\,, \tag{30}$$

even if the sector grid is not sufficiently dense for perfect reconstruction. This essential observation allows designing underdetermined systems with very few channels that still reconstruct the perceptually relevant pressure component. We postulate that reconstructing at least the omni-directional signal is tied to preserving timbral properties and mitigating coloration in the coded result.

6) Energy Reconstruction: The previous subsection described reconstructing a signal, where the sum of all parts matches the input at each time t; i.e., the reconstruction of amplitude. This criterion, however, is often not applicable, as the

intermediate sector signals might be altered, or even replaced, potentially in a signal-dependent manner. In these cases, one may instead seek to match the input-to-output energy as in (22). In the presented algorithm, this case appears when the decoder introduces decorrelation between the sector signals. Decorrelation is a common operation used to increase the perceived number of channels to more than were transmitted, or when the decoder makes the assumption of decorrelated signal components which are not met in practice. Thus, decorrelators are a common component of perceptual audio codecs. Note that the sector signals s_ξ are obtained from spatially band-limited signals, i.e., from a finite-order SH expansion. Any modification of s_ξ must not exceed the spatial band-limitation of the re-encoding, otherwise the output will suffer from truncation artifacts. Any decorrelation must therefore obey or re-establish this requirement, where the resulting decorrelated (but spatially band-limited) signals are written as $\tilde{\boldsymbol{x}}(t)$.

Analogous to (28), the energy preserving re-synthesis is carried out as

$$\tilde{\boldsymbol{\chi}}(t) = \sqrt{\beta_\mathrm{E}}\boldsymbol{B}\tilde{\boldsymbol{x}}(t)\,, \tag{31}$$

with $\boldsymbol{B}$ again composed of $c_n^{\mathrm{syn}} = c_0^{\mathrm{an}}/c_n^{\mathrm{an}}$. In contrast to the previously derived reconstruction (28), the factor β_E now ensures that the input-to-output signal energy matches (see (22)). Comparing (28) and (31) now also shows an interesting property of the presented framework, since the difference between both reconstruction methods only manifests itself as switching between β_A and β_E, making the decoder workflow comprehensive and intuitive to adapt.

B. Encoder

The encoder extracts the transmitted audio streams from the higher-order Ambisonic signals, along with a set of parameters, as illustrated in Fig. 1. The encoder shares the sound-field sector-processing methodology detailed in [19] and [24], which extracts a local direction of arrival (DoA) and diffuseness estimate per sector and each time-frequency tile. The input audio passes through the alias-free short-time Fourier transform (afSTFT), with its implementation detailed in [33, Ch.1], configured with a hopsize of 128 and blocksize of 2048 samples at 48 kHz sampling frequency, resulting in 16 down-sampled values per band and block. To increase frequency resolution at lower frequencies beyond uniform discretization, the afSTFT implementation further subdivides the lower bands (as is typical in perceptual audio coding filter banks), resulting in 133 bands in total. The encoder output data thus scales with the number of time-frequency filters as well as the number of beamformers/spherical filters. Leaving the number of time-frequency sub-bands fixed, the spherical filter bank resolution dictates directly the number of audio transport channels and thus the maximum compression factor of the encoder, since the directionally filtered sound-field pressure along with their local sound-field parameters comprise the encoder output.

a) Parameter Estimation: The pressure signal of a sound-field may be extracted proportionally from an omni-directional receiver, i.e., $w_w(\Omega) = 1$, whereas the velocity vector is proportional to signals extracted by three dipole patterns along the $[x, y, z]$-direction, i.e., $[w_x(\Omega), w_y(\Omega), w_z(\Omega)]$. This set of patterns is proportional to the first-order SH and the corresponding output signals are proportional to the so called *B-format* signals, typically described as the four channels with the ordering *w, x, y, z*.

While first-order DirAC operates on a single global time-frequency-dependent DoA and diffuseness parameterization, both extracted from the sound-field pressure and velocity, HO-DirAC utilizes directionally constrained estimates, thereby leveraging higher-order sound-field information. The input HOA signal is converted into a set of directionally constrained first-order SHD sets, narrowed by the analysis filters of the spherical filter bank $w_\xi(\Omega)$. This results in a set of sector signals describing the directionally constrained acoustic pressure and velocity components [34]. We denote a single sector as ξ and the latter two components as p_ξ and $\boldsymbol{v}_\xi$, respectively. By directionally constraining the sector pressure and velocity, the DoA and diffuseness estimates likewise represent directionally constrained, or local, estimates. This technique allows for multiple simultaneous estimates distributed over the sphere, as presented in [34] and then implemented in DirAC in [19]. This computationally efficient, and perceptually-motivated strategy is therefore chosen for the encoder.

The directionally constrained pseudo/active intensity vector $\boldsymbol{i}_\xi$ of sector ξ is proportional to the measured real part $\Re$ of

$$\boldsymbol{i}_\xi \propto -\Re\{p_\xi^{\mathsf{H}}\boldsymbol{v}_\xi\}\,. \tag{32}$$

The opposite vector direction directly estimates the sector DoA $\Omega_\xi^{\mathrm{DoA}}$ of its predominant signal component

$$\Omega_\xi^{\mathrm{DoA}} = \angle\left[-\boldsymbol{i}_\xi\right]\,. \tag{33}$$

In the case of simultaneous sound sources competing in the same time-frequency tile, the vector, points towards the energetic average. The sector signals p_ξ and $\boldsymbol{v}_\xi$ also provide an estimate of the directionally constrained sound-field sector energy E_ξ and diffuseness parameter ψ_ξ

$$\psi_\xi = 1 - \frac{||\boldsymbol{i}_\xi||}{E_\xi} = 1 - \frac{2||\boldsymbol{i}_\xi||}{|p_\xi|^2 + \boldsymbol{v}_\xi{}^{\mathsf{H}}\boldsymbol{v}_\xi}, \tag{34}$$

which is an indicator of the relation between a single plane-wave and diffuse sound-field components. Hence, ψ_ξ can be interpreted as the directionality of the (local) sound-field energy flow, which measures $\psi = 0$ for a single impinging plane-wave, and $\psi = 1$ for zero observed net flow.

The sector patterns producing sector signals p_ξ and $\boldsymbol{v}_\xi$ are generated by spatially multiplying directional weightings of each $w_\xi(\Omega)$ onto the omni-directional $w_w(\Omega)$ and dipole $[w_x(\Omega), w_y(\Omega), w_z(\Omega)]$ patterns. Recall that we defined the SHD coefficients of $w_\xi(\Omega)$ in a single steering direction as $\boldsymbol{w}_\xi$ (10), with the latter stacked into $\boldsymbol{A}$. The sector velocity patterns $\boldsymbol{w}_\xi^{x,y,z}$ may be obtained by transforming the product of the spherical multiplication into the SHD, e.g., for the dipole

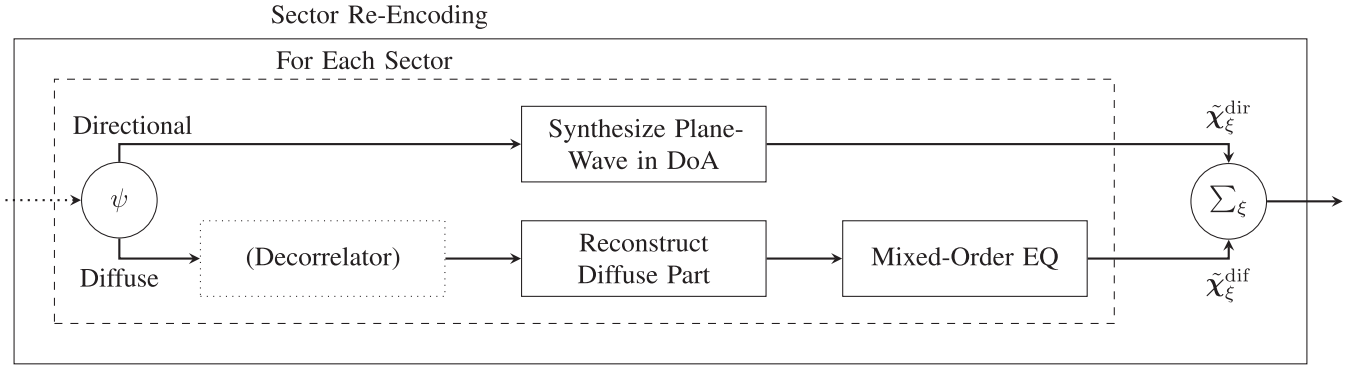

Fig. 2. Sector re-encoding of the proposed decoder, detailing the flow diagram shown in Fig. 1. The transmitted input is marked by the dotted line, solid lines visualize audio signals. Each block is explained in Section III-C. The transmitted parameterization and audio is used to synthesize plane-waves for the directional stream, and to reconstruct the diffuse SHD parts, before summing both over sectors.

along x, as

$$\boldsymbol{w}_\xi^x = \mathcal{SHT}\{w_\xi(\Omega) w_x(\Omega)\} \,. \tag{35}$$

The latter is deterministic for the chosen encoder design and can be pre-computed. Note that this spherical multiplication with a first-order pattern produces SHD coefficients one order higher than $\boldsymbol{w}_\xi$. In practice, this leads to the stacked form with zero padded $\boldsymbol{w}_\xi^w$ coefficients, where the sector signals are obtained from the HOA input $\boldsymbol{\chi}$ as

$$\begin{bmatrix} p_\xi \\ v_\xi^x \\ v_\xi^y \\ v_\xi^z \end{bmatrix} = \begin{bmatrix} \left[\boldsymbol{w}_\xi^w \, 0 \dots 0\right] \\ \boldsymbol{w}_\xi^x \\ \boldsymbol{w}_\xi^y \\ \boldsymbol{w}_\xi^z \end{bmatrix} \boldsymbol{\chi} \,. \tag{36}$$

b) Post-Filtering: The sound-field parameters used as metadata may be subject to additional smoothing and filtering. However, this topic is only marginally explored in the presented study. A fourth-order median filter on the diffuseness estimate corrects for major outliers that most likely originate from numerical problems. Furthermore, a diffuse input sound-field was shown in [34] to produce a lower directionally constrained diffuseness ψ_ξ than in the unconstrained case, and hence the encoder clips towards this theoretical value.

Attempts to smooth the DoA vectors, e.g., with a Kalman filter or vector spherical linear interpolation (SLERP), are not applied in this version, since the smoothing was found to introduce artifacts related to the timing constants or lagging parameter changes. Smoothing may, however, improve performance in more informed use-cases, such as target-speaker tracking, where a smooth parameter evolution is more appropriate.

C. Decoder

The decoder conceptually comprises two parts, the *directional* and *diffuse* stream. Both stem from the sound-field sector estimates provided by the previously described encoder, making them local estimates of the directional and diffuse sound-field components. As detailed before, the encoder transmits J sector pressure signals p_ξ, along with a set of DoA Ω_ξ^{DoA} and diffuseness ψ_ξ estimates per time-frequency tile.

The transmitted sector pressure signal p_ξ is hence split, into a directional component

$$p_\xi^{\text{dir}} = (1 - \psi_\xi)^g \; p_\xi \,, \tag{37}$$

and diffuse component

$$p_\xi^{\text{dif}} = (\psi_\xi)^g \; p_\xi \,. \tag{38}$$

In the typical DirAC framework, these two streams are assumed to be uncorrelated, and the mixing factor is therefore chosen as $g = 1/2$. However, the presented algorithm does not make this assumption, and will consequently set $g = 1$.

Both streams need individual decoding, which corresponds directly to re-encoding the sector pressure signals to their corresponding Ambisonic streams, i.e., modal spectrum synthesis, as detailed in the following and depicted in Fig. 2.

1) Directional Components: The directional stream is interpreted as local plane-wave components. The encoder estimated their DoA as Ω_ξ^{DoA}. The re-encoding to order N_{dir} is therefore given as the plane-wave re-encoding, similar to [24]:

$$\tilde{\boldsymbol{\chi}}_\xi^{\text{dir}} = \text{diag}_N\left[\boldsymbol{c}_n^{\text{syn,dir}}\right] \mathbf{Y}^{\mathsf{H}}\left(\Omega_\xi^{\text{DoA}}, N_{\text{dir}}\right) \, p_\xi^{\text{dir}} \,. \tag{39}$$

The coefficients $\boldsymbol{c}_n^{\text{syn,dir}}$ are hence set to unity, which also satisfies the condition $c_0^{\text{syn,dir}} = 1$ of (30). Any re-encoding other than plane-waves did not impart noticeable improvements, however, since the SH truncation artifacts increase drastically at these low SH orders, potentially outweighing benefits of re-encoding SH order shaping.

Under the assumption of synthesizing plane-wave components, their re-encoding order may be chosen freely. If $N_{\text{dir}} = N_{\text{in}}$, the codec operates as a transmission algorithm. Whereas, $N_{\text{dir}} > N_{\text{in}}$ achieves spatial enhancement or upmixing.

2) (Locally) Diffuse Components: The diffuse sector pressure components are re-encoded to order N_{dif} and in steering directions Ω_ξ based on their sector of origin as

$$\tilde{\boldsymbol{\chi}}_\xi^{\text{dif}} = \text{diag}_N\left[\boldsymbol{c}_n^{\text{syn,dif}}\right] \mathbf{Y}^{\mathsf{H}}(\Omega_\xi, N_{\text{dif}}) \, p_\xi^{\text{dif}} = \boldsymbol{B} p_\xi^{\text{dif}} \,. \tag{40}$$

To act as an inverse to the encoder spherical filter bank, the order should be chosen based on the sector design order $N_{\text{dif}} = N_{\text{in}} - 1$.

Choosing $\boldsymbol{c}_n^{\text{syn,dif}}$ needs special attention and reference to the presented theory, since there are multiple options that dictate the performance. One option is aiming for perfect reconstruction of the diffuse stream. Canceling out the influence of the encoder beamformers, (28) achieves perfect reconstruction of the diffuse components within $\boldsymbol{\chi}$ up to order N_{dif} from the spherical filter bank outputs p_ξ. However, this requires a grid dense enough to support order $2N_{\text{dif}}$ and the corresponding transport channel count. Note that because $N_{\text{dif}} < N_{\text{in}}$ compression is still possible, however, relaxing the reconstruction criteria from perfect reconstruction enables higher compression gain. Therefore, the second option is re-encoding aiming for the relaxed amplitude reconstruction condition of only preserving the (omnidirectional) sound-field pressure component. As detailed before, this allows for even fewer transport channels, since the sector grid requirements are less demanding, i.e., requiring only a grid supporting order $N_{\text{dif}} + 1$. Analogous to the solution presented for the directional components, we may re-encode the sector signals as plane-waves, or as any alternative that is normalized, both of which satisfy $c_0^{\text{syn,dif}} = 1$ (see (30)). In the presented evaluation, we chose to use the perfectly reconstructing $\boldsymbol{c}_n^{\text{syn}}$, i.e., $c_n^{\text{syn}} \propto 1/c_n^{\text{an}}$, for the energy preserving grid ($2\,N$ quadrature) and plane-wave re-encoding for the relaxed grid condition ($N + 1$ quadrature).

Even though well motivated in the literature, and in contrast to previous HO-DirAC implementations, the present method implies no need for decorrelators on the diffuse streams. We found no significant audible benefits using decorrelators and hence favored omitting them for three main reasons. Firstly, all decorrelators we tested introduced some coloration and/or other audible artifacts. Secondly, independent decorrelation of the sector signals may not preserve the spherical order they originated from, hence a subsequent SH re-encoding could suffer from order limitation artifacts, which are more difficult to handle without further assumptions (see Section III-A6). Lastly, decorrelators are an additional computational expense, which was a consideration when developing the proposed algorithm.

3) Mixed-Order Equalization: As derived before, the synthesis order of the diffuse stream is always lower than the synthesis order of the directional stream. Although this is in line with the coding assumptions, HOA signals suffer from order truncation artifacts that are most prominently perceived as a high-frequency roll-off. Because of the order mismatch, signal components sound slightly different between both re-encoded streams, most notably slightly duller in the diffuse stream. We therefore propose to include a mixed-order equalization that aims to mitigate the high-frequency roll-off compared to the directional stream re-encoding. We settled for a simple time-frequency equalization based on a basic binaural model that simplifies the expected diffuse-field magnitude response in relation to the SH order. The order-truncation mitigation equalization is thereby directly obtained from the weighted modal relation of $\boldsymbol{c}_n^{\text{syn,dif}}$ to $\boldsymbol{c}_n^{\text{syn,dir}}$. The frequency response is derived from the coloration equalization described for tapered SH spectrum rendering in [35]. In the present equalization, however, the target order is the order of the directional stream, as opposed to a very high order used typically in binaural rendering equalization. The diffuse-field assumption of the model matches the diffuseness assumption of the components it is equalizing and under typical listening situations, and we found that this strategy provided satisfactory results.

4) Summation: The output of the sector re-encodings $\tilde{\boldsymbol{\chi}}_\xi^{\text{dir}}$ and $\tilde{\boldsymbol{\chi}}_\xi^{\text{dif}}$ is finally summed together over the sectors, where the preservation factors β_{A} and β_{E} ensure correct scaling. The directional stream sums coherently and therefore requires β_{A} as in (28). The diffuse stream in the proposed method also sums coherently and hence again we chose β_{A} in the summation. Note, however, that when decorrelating the diffuse stream signals, the factor β_{E} is appropriate under the assumption of incoherent signal summation, as in (31). The two assumptions, and hence factors, can also be combined in order to reflect non-ideal decorrelation from practical implementations. After passing the inverse time-frequency filter bank, the decoder delivers the Ambisonic output signal $\tilde{\boldsymbol{\chi}}$.

5) Summary: The codec controls and decreases the number of transport audio channels, while additionally transmitting intuitive to interpret metadata. With appropriate coding and quantization of the transport signals and associated metadata, improved compression may be achieved compared to coding and streaming all HOA signals.

IV. Evaluation

A. Perceptual Evaluation Test Design and Setup

A formal listening test assesses the perceived sound quality difference between the (unaltered) reference and the output of the codec proposed herein. We varied the number of audio channels and channel metadata during transmission while observing the participants' ratings. The listening test aimed to find a point of saturation, beyond which using more channels during transmission did not increase the ratings, which would indicate that no considerable change in audio quality was perceivable. As of now, with the available resources, we cannot test directly for a transparent codec. This would require more participants, more diverse listening test items, and an overall higher evaluation effort, beyond the scope of this study. Therefore, we chose to measure perceived sound quality (degradation) using a multiple stimuli with hidden reference and an anchor-like stimulus setup. The latter presents all items in a comparison and we may directly assess the quality impact as we restrict the number of transport channels.

All items were decoded to a 37-channel spherical loudspeaker setup in $2\,\text{m}$ radius around the listeners. This setup adequately supports Ambisonic decoding up to order five with uniform coverage, when fifth-order Ambisonic streams are decoded with the All-Round Ambisonic Decoder (AllRAD) [32] method. The setup consists of *Genelec 8331 A* loudspeakers in the anechoic room *WILSKA* at the Acoustics Lab, Aalto University, Finland.

TABLE I
TRANSPORT CHANNEL COUNTS (REFERENCE $N = 5$ IS 36)

Input SH Order	AP	EP/PR
$N = 2$	4	4
$N = 3$	6	12
$N = 5$	12	36

B. Tested Number of Audio Channels

The number of transmission channels tested depends on the preserved input order and the chosen encoder design. The two designs are abbreviated as *AP* and *EP*, for amplitude preservation and energy preservation, respectively, at the encoder. Reflecting the available resolution of the loudspeaker layout, the order of the decoding was set to five, since we expected to uncover the most differences for the highest available order playback. Therefore, we limited the evaluation to the fifth-order and explored the effect of varying the transport channel count, i.e., compression factor, as well as the preservation/reconstruction variant.

As discussed before, an energy preserving encoder design enables perfect reconstruction (of the diffuse stream), but requires more transport channels than the relaxed amplitude preserving condition. The latter can only reconstruct a subset of the input, and the presented evaluation opts for reconstruction of only the zeroth order sound-field pressure. It is not clear how these trade-offs compare perceptually, and therefore, we chose to test both design criteria, leading to the transmitted channel counts of Table I. It is noted that these channel counts were chosen purely for convenience due to available t-designs, and these designs are not the minimal design choices provided by critical sampling grids. Next to the fifth-order, i.e., 36-channel, input signal reference, a single transport channel marks the lower anchor-like condition. This low-quality condition only transmits a single signal from an omni-directional beamformer at the encoder and is in fact similar to the first-order DirAC formulation [17]. The metadata scales linearly with the transport channel count and consists of the sector DoA (azimuth and elevation angle) and the diffuseness value for each time-frequency tile. Note that in practice these could be quantized, down-sampled, and/or grouped into perceptual bands.

C. Listening Test Items

The presented evaluation used a total of four items[1], which were specifically designed to be challenging input sound scenes; in order to make limitations of the proposed method audible.

- *Orchestra:* An orchestral piece by Anton Bruckner, having a high source count and prominent (image-source model) reverberation.
- *Applause:* Clapping from 1000 sources, arranged on the horizontal plane around the listener, without reverberation.
- *Music:* An electronic pop music piece, cinematic, created from 24 stem tracks using plane-wave encoding.
- *Scene:* Simultaneous speech, noise, piano, and clapping in slightly reverberant conditions; obtained utilising a hybrid reverberation model.

The first two items are designed to be maximally critical and expose any shortcomings of the codec, such as revealing artifacts arising due to model violations and/or erroneous parameter estimates, whereas the latter two items represent challenging, yet more realistic scenarios. The presented items are the outcome of an extensive pre-screening process of a plethora of diverse items. It should be mentioned that informal testing indicated comparable performance also on fourth-order spherical microphone array recordings. However, widely available spherical microphone array recordings can not support the fifth-order reference case (dictated by the reproduction loudspeaker array) and, furthermore, may impart other shortcomings originating from the array to SHD conversion unrelated to the presented method. Further examples have been made available online.

The *Orchestra* piece was created by simulating a cuboid room (dimensions: $[35.7 \times 19.8 \times 17.4]$ m) with an image-source model. The sources were arranged in a typical European orchestra setting, in an arc of radius 3 m and 6 m around the receiver position. The receiver marks a possible conductor's perspective, where the orchestra spans around the frontal plane. The dimensions of the room resemble the Vienna Konzerthaus, and the reverberation time was simulated accordingly in the range $RT_{60} = 2.93$ s to $RT_{60} = 1.14$ s (in octave bands). The resulting impulse responses were then convolved with dry individual source recordings obtained from [36].

Applause was generated by randomized dry clapping samples at 1000 simultaneous positions on the horizontal plane around the listener. Due to no further reverberation, the resulting test item is both dry and dense, with broadband impulsive sounds.

In order to include a more realistic scenario, the test includes an electronic popular music piece, labeled as *Music*. The piece was available as 24 stem tracks and intended for spatial audio research [37]. A mix created for the tested loudspeaker setup aimed to fully utilize the spatial capabilities of the loudspeaker layout. While the vocal and bass drum track were positioned in the center, the other instruments and effects were positioned surrounding the listener. No additional spatial audio effects were applied. The participants did, however, describe the piece as having a cinematic feel, owing to its use of sound effects on the audio stems and inclusion of panned sound images, which makes this item interesting for evaluating the codec within a context more in-line with a cinema experience.

Item *Scene* featured four sound sources comprising a sound scene. The scene consisted of a single speaker, clapping, piano, and noise on the horizontal plane at angles $[-30, 30, 0, 120]$ degrees azimuth, respectively. The reverberation consisted of early reflections from an image-source model and a late reverberation tail from a stochastic reverberation model. The models simulated a room with a reverberation time of $RT_{60} = 1.2$ s, which dropped towards the high frequencies to $RT_{60} = 0.32$ s. The image-source reverberation was faded linearly (50 ms) into the exponentially decaying stochastic reverb after 76 ms. The latter cross-over time marks the conservative mixing time estimate t_{mp95} from [38]. This item was included because

[1]Audio items and additional material is publicly available on a companion website: http://research.spa.aalto.fi/publications/papers/hoac/

the simultaneous sound sources originate from competing directions, where the steady sources (such as the white noise) interact with impulse-like sounds (such as the clapping).

D. Listening Test Procedure

Before the test, every participant received written instructions and gave their consent to the ethical guidelines. Participants were informed about the hidden reference scenario and instructed to rate accordingly (at least) one item as 100. The listeners received a computer tablet allowing them to switch seamlessly between stimuli during playback, and to move the rating sliders. The tablet was placed on a slightly elevated stand, so that the test participants would avoid looking downwards when operating the interface; thus potentially facilitating a more comfortable and critical listening experience. The experiment showed four pages, one for each listening test item, with eight sliders for each stimulus, both in double-blind and randomized order. The labels showed *Bad* $(0-20)$, *Poor* $(20-40)$, *Fair* $(40-60)$, *Good* $(60-80)$, and *Excellent* $(80-100)$. The stimuli were presented in a 10 s loop with loudness adjusted to approximately 80 dB(A). After the experiment, the conductor asked the participants to give comments about the experienced differences. The whole procedure took approximately 30 minutes to complete.

E. Perceptual Evaluation Results and Analysis

An informal pre-test indicated small differences between the reference and conditions with six or more transport channels. Only very dense arrangements appeared to reveal a difference between 6 and 12 channels or more. Four transport channels still sounded reasonable in many cases, whereas some (sparse) example scenes indicated that even a single channel may be still acceptable (i.e., as originally used by first-order DirAC).

Sixteen subjects (14 m, 1f, 1n) rated eight stimuli for four items, totaling 512 responses. All statistical tests in this study are reported at $\alpha = 0.05$. Responses of conditions where a subject failed to identify the Reference (≤ 90) were excluded, which happened four out of 64 times. In general, participants reported that the differences were small, except in one or two conditions.

Fig. 3 shows all collected responses as opaque dots, along with a box plot and violin plot. The box plot indicates the median, as well as the quartiles, where outliers are marked past 1.5 the inter-quartile range. The violin plot on top additionally visualizes the density of an estimated underlying distribution. Fig. 4 splits the ratings amongst stimuli, revealing more details about their consonance. The box plots in Fig. 4 also feature notches that indicate the 95% confidence intervals around the median, estimated from bootstrapping ($n = 10000$).

A Shapiro-Wilk test on the responses implies that the data should not be described as sampled from a standard normal distribution ($p < 0.001$ for each stimulus). Hence, further statistical analysis should not employ parametric statistical models based on the normality assumption, but rather non-parametric methods. A Friedman test indicates differences between groups, and may be seen as a non-parametric repeated measures analysis of variance (ANOVA). It indicates that the stimuli produced statistically different results ($p < 0.001$ for H_0: stimulus had

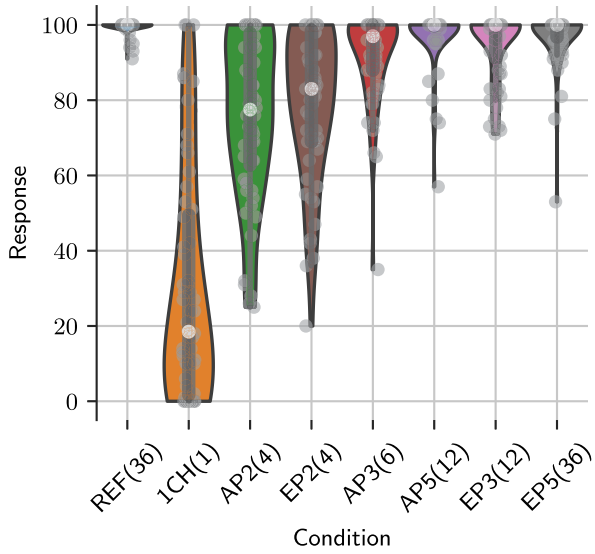

Fig. 3. Results of the perceptual evaluation. Inner parts (black) show boxplots with quartiles, where the white circle marks the median. A violin plot approximates the underlying distribution of the individual responses marked as opaque dots. The conditions utilized REF = 36, 1CH = 1, AP2 = 4, EP2 = 4, AP3 = 6, AP5 = 12, EP3 = 12, EP5 = 36 audio channels during transmission.

no effect on rating). In order to detect which groups differ, the non-parametric Wilcoxon signed-rank test implies statistically different responses (H_0: related paired samples come from the same distribution). The results regarding overall performance in Fig. 3 report significance according to the two-sided test, Benjamini-Hochberg corrected in order to control the false-discovery rate (FDR) in multiple comparisons. In this work, since we are also interested in ranking the results, the Wilcoxon signed-rank may also test for a one-sided *greater* difference. Fig. 5 visualizes the one-sided test results divided according to the test items, where the scale is clipped in order to visualize differences in the statistically relevant range. Note that in Fig. 5 no further correction for multiple testing is applied, since we may only conclude trends at this level of granularity, and because of the one-sided comparison.

Overall, Fig. 3 illustrates the general tendency of more channels leading to higher ratings, which saturates towards the reference condition *REF*. Condition *1CH* was identified as an anchor-like condition with a median rating below 20, even though rated across the whole scale. The median for six or more channels is close to the median of the reference (> 90), hence showing considerable saturation towards the uncompressed reference.

As already indicated in the informal pre-test, test items which were not specifically designed to challenge the codec hardly evoked any clear differences. Most participants also reported that the differences were particularly small for the item *Music*. This content dependency is further highlighted in Fig. 4. Grouping the responses according to the test item reveals that *Music* seems the least critical item and *Scene* saturates already for four transport channels, as the confidence interval around the median starts to overlap with the reference rating. The item *Applause* and *Orchestra* provoked the largest range in ratings, and thus can be regarded as the most critical items.

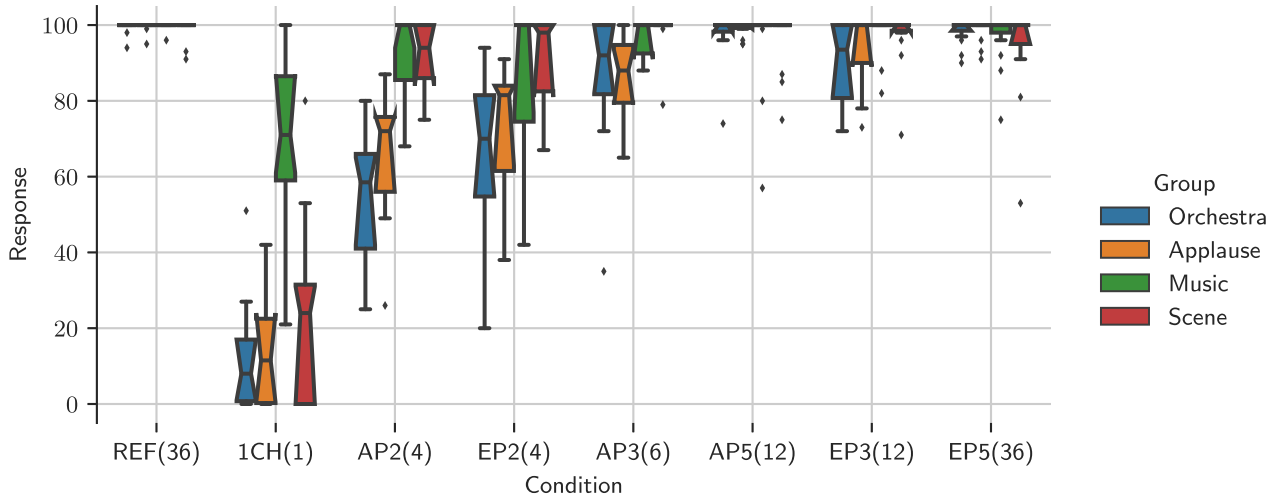

Fig. 4. Listening-experiment responses, differentiated into the test items indicated by color. The boxplots show the median, where the notches indicate confidence intervals around the median estimated from bootstrapping. The conditions utilized REF = 36, 1CH = 1, AP2 = 4, EP2 = 4, AP3 = 6, AP5 = 12, EP3 = 12, EP5 = 36 audio channels during transmission.

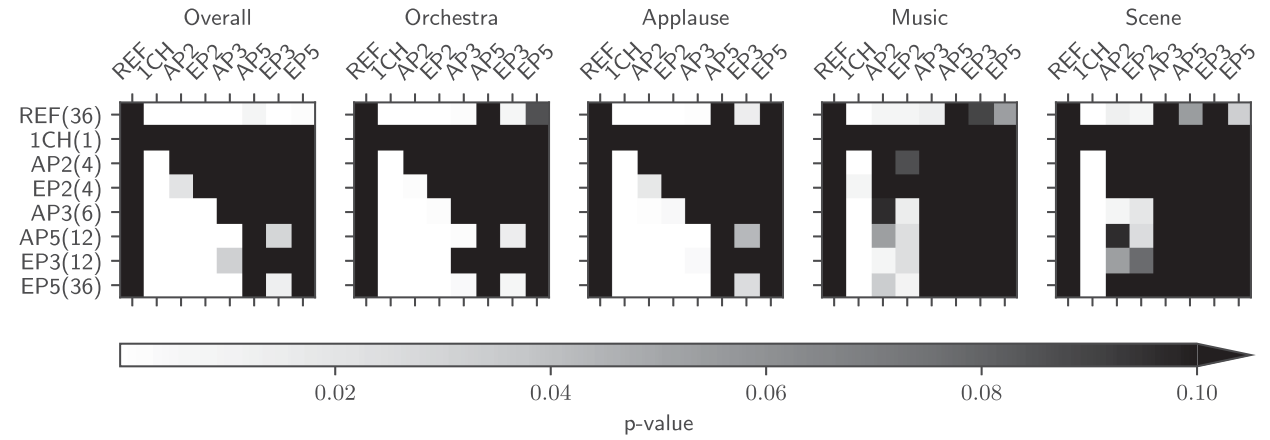

Fig. 5. One-sided *greater* Wilcoxon signed-rank test results between codec conditions under test, differentiated into the test items, indicating statistically significant higher ranking. The p-value scale is clipped at 0.1 for better visualization, and no correction for multiple testing is applied.

First, the generally high subjective quality is apparent again, with most of the items ranked in the range of a *good* or *excellent* rating. Even for the single audio channel transmission *1CH*, the median of the item *Music* was just above 70 points, and therefore rated as subjectively *good* audio quality. This particular high ranking might be related to the relatively sparse instrumentation with different sonic textures that may mask some of the processing artifacts. In the four-channel audio transmission cases *AP2* and *EP2*, the ratings seem to already saturate towards the reference. Here, again *Applause* and *Orchestra* produce greater differences. Most interestingly, for the latter two items, *EP2* seems to rank higher than *AP2*, which is supported by the results in Fig. 5, as well as by the statistically different (FDR corrected) results in Fig. 3. For the items *Music* and *Scene*, the median already saturates at the top of the scale. When using four transport channels, *EP2* perfectly reconstructs the first-order diffuse stream components, whereas *AP2* only preserves the sum of pressure amplitude, i.e., zeroth order. We may therefore infer a slight benefit of the energy preserving perfect reconstruction paradigm. Another interesting comparison is between *AP5* and *EP3*, since both conditions use 12 audio transport channels. Condition *EP3* preserves only second-order diffuse stream components but achieves their perfect reconstruction. On the other hand, condition *AP5* preserves higher-order components but cannot achieve their perfect reconstruction. Both conditions clearly saturate towards the end of the scale, but the most critical stimulus *Orchestra* seems to still produce certain differences. Although we may not conclude a different overall performance (FDR corrected), the statistical inference from Fig. 5 indicates that *AP5* performs better for item *Orchestra* ($p = 0.01$). A similar conclusion may be drawn for *Applause* ($p = 0.04$).

Participants rated codec conditions *AP5* and *EP5* the highest and no statistical difference can be reported between them (compare Fig. 5, and FDR corrected on all ratings). Fig. 5 further implies that for most test cases there is no statistically significant difference between both conditions *AP5* and *EP5*

and the reference *REF*. The only item where *REF* compared as statistically significantly better is for *Scene*, although some lower channel transmissions did not have the same inference, thus questioning the statistical conclusions. For the item *Scene*, one participant also described a slight change in the noise color. The latter might originate from the simplistic mixed-order timbre equalization. Furthermore, we noticed for this item a slight modulation noise, due to the omission of any vector smoothing during analysis. This modulation was only described as audible by the author in localized broadband static noise scenarios and could be mitigated through more advanced treatment during encoding.

V. Discussion

The presented codec targets compressing spatial audio streams, where both the input and output signals are delivered in the SHD, i.e., as higher-order Ambisonics. The methodology is based upon segregating the input sound-field into directionally constrained sectors, whereby a spherical filter bank establishes preservation and reconstruction properties between input and output signals. The input encoding extracts spatial audio parameters such as a DoA and a diffuseness estimate for multiple sound-field sectors.

Under the assumption of limited perceptual resolution, compression was achieved by limiting the number of transmitted sound-field sector signals. As the listening test in Section IV-E implies, the perceived difference in the compressed audio saturates towards the uncompressed reference audio. The test showed that the difference is dependent on the type of audio signal, but more importantly, it is dependent on the number of transmission channels. For the presented test, the data indicated that six channels transmitted with the proposed codec approached the reference case of 36 channels. This corresponds to a compression ratio of factor 6, or a rate saving of $83\,\%$, neglecting the (necessary) transmission of metadata. Note that the compression of this metadata is a topic of future work. The most critical items may benefit from increasing to 12 transport channels. We also have strong indications that the number of necessary transported channels stays in that range, even when increasing the Ambisonic channel count, i.e., increasing the Ambisonic order. This returns to the assumption that the human perceptual resolution is limited in terms of simultaneous sound events in space per time-frequency tile. Comparing the results with other studies, a different methodology mentions a similar baseline of channels for HOA transport [14].

The number of transport audio channels corresponds to the maximum number of re-encoded DoAs per time-frequency tile as the directional stream. For the (locally) diffuse stream, the number of channels dictates the achievable reconstruction properties. For example, 12 channels enables an energy preserving encoder design with second-order perfect reconstruction of the locally diffuse input signal portion, while allowing extraction and subsequent re-encoding of directional components to arbitrary (higher) orders. The listening test results also indicate that, for most audio items, the diffuse field may be represented at a lower Ambisonic order without severe perceptual impact, supporting the general sentiment within the research community. In other words, the codec can leverage the concept that higher orders usually decay quickly in typical sound-fields without strong directional components. We also explored another preservation constraint, which, although not fully preserving the input energy, can match the input in terms of the (sum of) pressure. The latter allows for considerably fewer channels to be transmitted, while still achieving good perceptual quality in typical listening scenarios. In fact, the latter could achieve higher perceptual ratings in the 12-channel transmission case compared to the 12-channel energy preserving design. This can be explained by the ability of the encoder to preserve higher-order signal components (fourth vs. second). The differences mostly manifest themselves in the reconstruction of the diffuse components, which are no longer perfectly reconstructing. However, the perceptual evaluation challenges the relevance of perfect reconstruction, since we only observed a benefit in direct comparison (*AP2* vs. *EP2:* first-order reconstruction from four channels). Generally, items with more energy in higher orders rendered in the diffuse stream provoked more differences in the perceptual quality ratings. Hence, extracting and re-encoding the usually perceptually more prominent directional components benefits the perceived audio quality, which is where the presented parametric spatial audio codec can operate rather efficiently. Note that even though we only recovered the $n = 0$ components for *AP*, more orders could have been restored, e.g., up to first-order for the 6 transport channels case.

This leads to the discussion of perceptual codecs and the choice of which signal components to discard. The SH basis is an orthogonal basis, and therefore no intrinsic or signal-independent correlation can be exploited within. A diffuse sound-field exhibits incoherent signals as its SH spectrum, and therefore the theoretical redundancy compression coding gain is minimal. On the other hand, a plane-wave is fully described by its scalar pressure and DoA, making it theoretically possible to transmit its HOA signal with only a single channel plus metadata. Of course, in reality, sound-fields are a mixture of these extreme cases. The presented HO-DirAC sound-field model could be considered as an adequate choice for the vast majority of tested items (even those beyond the scope of the reported items). Only particularly challenging items that were specifically designed to violate the codec and its assumptions evoked distinctly audible artifacts. This is highlighted by the results for item *Music*, the only item not specifically designed to break the codec and resembling a more typical real-world scenario. Here, even the single-channel transmission case reminiscent of early first-order DirAC performed well.

The presented algorithm is also connected to SHD band-width extension, since the directional stream is always synthesized to a higher order than the order from which the signals were extracted. This is due to the trade-off in the sound-field analysis, where the sector pressure is naturally defined as one order lower than the sector intensity necessary for the active-intensity based DoA extraction. However, it is the sector pressure that is subject to the spherical filter bank design ensuring reconstruction properties, and therefore transmitted and used for reconstruction. Another interpretation of the algorithm is that it reconstructs the

lower order input HOA signals, while simultaneously spatially enhancing components estimated as being highly directional (by synthesizing them at a higher order).

While we believe the presented framework shows potential, we also uncovered some directions for future work. The currently proposed algorithm does not leverage higher orders for encoding, even if available. For example, *AP3* and *EP3* utilize the third-order components during sector analysis, despite fifth-order components being available. This aspect needs to be investigated further, since the orthogonal basis may still support the available information utilized in the framework. Therefore, any potential shortcoming is proposed for future work, since the practical benefit of this currently remains unclear. Furthermore, the current spherical filter bank is based on uniform steering grids, specifically spherical t-designs. While these designs contribute very little aliasing, they are not minimal in the sense of points and therefore audio channels. Critical and other sampling schemes could theoretically enable a lower number of channels, however, at the expense of fewer re-encoded directional components. Since we identified this number as a crucial parameter influencing the perceived quality rating, we consider reducing it as a trade-off that needs to be investigated in the future. Besides improving parameter estimation, we also identified the mixed-order re-encoding scenario showing potential for further improvements of the method, since currently the energy per order might be re-balanced based on the estimated directional-to-diffuse ratio. This could be addressed, for example, using methods targeting to match the spatial covariance in the SHD.

Certainly, an immediate avenue for future work is to include more parts necessary for a complete codec. This would mean, first and foremost, compressing the audio and metadata streams. Since the transport channels are (conveniently) ordinary audio signals extracted from discrete directions, conventional multi-channel audio compression strategies are available. As opposed to the input HOA signals, these transport channels can be assumed to be partially independent and weakly correlated, and hence available perceptual codecs may perform optimally. We found that applying perceptual audio coding on beamformer outputs, as opposed to on the SHD signals, greatly reduces spatial artifacts; since coding artifacts do not directly affect the SHD signal relations and therefore remain more directionally localized. The proposed framework further allows to balance the coder between audio transport channels and additional parameterization data. While non-parametric audio coders need to solely operate on the audio data, in the current framework we may utilize the combination of a subset of audio transport channels and metadata and therefore have more bandwidth available for coding the (fewer) audio transport channels. The proposed metadata is perceptually-motivated and intuitive to process. Initial tests with frequency band grouping, averaging, quantizing and down-sampling the parameterization can overcome the otherwise substantially increasing audio data. However, applying additional codec layers impart further compression artifacts that might interact with the proposed parametric spatial audio processing, and were therefore not investigated in the present study. It is also worth mentioning in this context that profiling the algorithm revealed that the longest computation time was spent on evaluating the SHs in order to form the direct-stream synthesis matrix, which can be pre-computed and sought from a look-up table. An implementation comprising all codec components could then be part of a more extensive comparative codec evaluation.

VI. Conclusion

In this paper, we described a parametric spatial audio compression algorithm which targets higher-order Ambisonic input to output. The codec allows varying the intermediate number of transport audio channels and parameters. The proposed novelty lies in the encoder and decoder parts, which operate as a spherical filter bank and its inverse, respectively. These are used to extract directionally constrained sound-field signals and parameters, which are then used to reconstruct the input. A formal listening test indicated that the presented codec achieves excellent results already for six transmitted audio channels plus their metadata, even for challenging test items. Beyond that, the test showed heavy perceived quality saturation towards the 36-channel reference condition, implying that the presented codec may represent a promising new compression strategy for Higher-Order Ambisonics (HOA) audio signals.

Acknowledgment

The authors thank the Fraunhofer IIS and all test participants. We are particularly thankful for the detailed suggestions and expertise shared by the anonymous reviewers.

References

[1] 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," *J. Acoust. Soc. Amer.*, vol. 133, no. 5, pp. 2711–2721, 2013.

[2] J. Ahrens and C. Andersson, "Perceptual evaluation of headphone auralization of rooms captured with spherical microphone arrays with respect to spaciousness and timbre," *J. Acoust. Soc. Amer.*, vol. 145, no. 4, pp. 2783–2794, 2019.

[3] J. Herre et al., "MPEG surround-the ISO/MPEG standard for efficient and compatible multichannel audio coding," *J. Audio Eng. Soc.*, vol. 56, no. 11, pp. 932–955, 2008.

[4] T. Rudzki, I. Gomez-Lanzaco, J. Stubbs, J. Skoglund, D. T. Murphy, and G. Kearney, "Auditory localization in low-bitrate compressed ambisonic scenes," *Appl. Sci. (Switzerland)*, vol. 9, no. 13, 2019, Art. no. 2618.

[5] E. Hellerud, A. Solvang, and U. P. Svensson, "Spatial redundancy in higher order ambisonics and its use for lowdelay lossless compression," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2009, pp. 269–272.

[6] S. Zamani, T. Nanjundaswamy, and K. Rose, "Frequency domain singular value decomposition for efficient spatial audio coding," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2017, pp. 126–130.

[7] N. Peters, D. Sen, M.-Y. Kim, O. Wuebbolt, and S. M. Weiss, "Scene-based audio implemented with higher order ambisonics (HOA)," in *Proc. IEEE SMPTE Annu. Tech. Conf. Exhib.*, 2015, pp. 1–13.

[8] S. Zamani, "Signal coding approaches for spatial audio and unreliable networks," Ph.D. dissertation, Univ. of California Santa Barbara, Santa Barbara, CA, USA, 2019.

[9] S. Zamani and K. Rose, "Spatial audio coding without recourse to background signal compression," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2019, pp. 720–724.

[10] J. Xu, Y. Niu, X. Wu, and T. Qu, "Higher order ambisonics compression method based on independent component analysis," in *Proc. 150th Audio Eng. Soc. Conv.*, 2021, pp. 1–7.

[11] J. Herre, J. Hilpert, A. Kuntz, and J. Plogsties, "MPEG-H 3D audio - the new standard for coding of immersive spatial audio," *IEEE J. Sel. Topics Signal Process.*, vol. 9, no. 5, pp. 770–779, Aug. 2015.

[12] D. Sen, N. Peters, M. Y. Kim, and M. Morell, "Efficient compression and transportation of scene based audio for television broadcast," in *Proc. AES Int. Conf. Sound Field Control*, 2016, pp. 1–8.
[13] S. R. Quackenbush and J. Herre, "MPEG standards for compressed representation of immersive audio," *Proc. IEEE*, vol. 109, no. 9, pp. 1578–1589, Sep. 2021.
[14] R. L. Bleidt et al., "Development of the MPEG-H TV audio system for ATSC 3.0," *IEEE Trans. Broadcast.*, vol. 63, no. 1, pp. 202–236, Mar. 2017.
[15] J. Blauert, *Spatial Hearing*. Cambridge, MA, USA: The MIT Press, 1985.
[16] A. Daniel, R. Nicol, and S. McAdams, "Multichannel audio coding based on minimum audible angles," in *Proc. 40th Int. Conf.: Spatial Audio: Sense Sound Space*, 2010, pp. 1–10.
[17] V. Pulkki, "Spatial sound reproduction with directional audio coding," *J. Audio Eng. Soc.*, vol. 55, no. 6, pp. 503–516, 2007.
[18] J. Vilkamo, T. Lokki, and V. Pulkki, "Directional audio coding: Virtual microphone-based synthesis and subjective evaluation," *AES: J. Audio Eng. Soc.*, vol. 57, no. 9, pp. 709–724, 2009.
[19] A. Politis, J. Vilkamo, and V. Pulkki, "Sector-based parametric sound field reproduction in the spherical harmonic domain," *IEEE J. Sel. Topics Signal Process.*, vol. 9, no. 5, pp. 852–866, Aug. 2015.
[20] A. Politis, L. McCormack, and V. Pulkki, "Enhancement of ambisonic binaural reproduction using directional audio coding with optimal adaptive mixing," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2017, pp. 379–383.
[21] M. V. Laitinen and V. Pulkki, "Binaural reproduction for directional audio coding," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2009, pp. 337–340.
[22] A. Politis, S. Tervo, and V. Pulkki, "COMPASS: Coding and multidirectional parameterization of ambisonic sound scenes," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 6802–6806.
[23] J. Ahonen, V. Pulkki, and T. Lokki, "Teleconference application and b-format microphone array for directional audio coding," in *Proc. AES Int. Conf.*, 2007, pp. 1–10.
[24] L. McCormack, V. Pulkki, A. Politis, O. Scheuregger, and M. Marschall, "Higher-order spatial impulse response rendering: Investigating the perceived effects of spherical order, dedicated diffuse rendering, and frequency resolution," *J. Audio Eng. Soc.*, vol. 68, no. 5, pp. 338–354, 2020.
[25] C. Hold, A. Politis, L. McCormack, and V. Pulkki, "Spatial filter bank design in the spherical harmonic domain," in *Proc. IEEE 29th Eur. Signal Process. Conf.*, 2021, pp. 106–110.
[26] C. Hold, S. J. Schlecht, A. Politis, and V. Pulkki, "Spatial filter bank in the spherical harmonic domain: Reconstruction and application," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2021, pp. 361–365.
[27] A. Politis, "Microphone array processing for parametric spatial audio techniques," Ph.D. dissertation, Aalto Univ., Espoo, Finland, 2016.
[28] B. Rafaely, *Fundamentals of Spherical Array Processing* (Springer Topics in Signal Processing Series), vol. 16. Cham, Switzerland:Springer, 2019.
[29] R. H. Hardin and N. J. Sloane, "McLaren's improved snub cube and other new spherical designs in three dimensions," *Discrete Comput. Geometry*, vol. 15, no. 4, pp. 429–441, 1996.
[30] M. Gräf and D. Potts, "On the computation of spherical designs by a new optimization approach based on fast spherical fourier transforms," *Numerische Mathematik*, vol. 119, no. 4, pp. 699–724, 2011.
[31] M. Chapman et al., "A standard for interchange of ambisonic signal sets: Including a file standard with metadata," in *Proc. Int. Symp. Ambisonics Spherical Acoust.*, 2009, pp. 1–6.
[32] F. Zotter and M. Frank, "All-round ambisonic panning and decoding," *J. Audio Eng. Soc.*, vol. 60, no. 10, pp. 807–820, 2012.
[33] V. Pulkki, S. Delikaris-Manias, and A. Politis, *Parametric Time–Frequency Domain Spatial Audio*. Hoboken, NJ, USA: Wiley, 2018.
[34] A. Politis and V. Pulkki, "Acoustic intensity, energy-density and diffuseness estimation in a directionally-constrained region," 2016, *arXiv:1609.03409*.
[35] C. Hold, H. Gamper, V. Pulkki, N. Raghuvanshi, and I. J. Tashev, "Improving binaural ambisonics decoding by spherical harmonics domain tapering and coloration compensation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 261–265.
[36] J. Pätynen, V. Pulkki, and T. Lokki, "Anechoic recording system for symphony orchestra," *Acta Acustica United Acustica*, vol. 94, no. 6, pp. 856–865, 2008.
[37] F. Melchior, D. Marston, C. Pike, D. Satongar, and Y. W. Lam, "A library of binaural room impulse responses and sound scenes for evaluation of spatial audio systems," in *Proc. German Annu. Conf. Acoust.*, 2014, pp. 555–556.
[38] A. Lindau, L. Kosanke, and S. Weinzierl, "Perceptual evaluation of physical predictors of the mixing time in binaural room impulse responses," in *Proc. 128th Conv. Audio Eng. Soc.*, 2010, pp. 1–17.

Christoph Hold (Graduate Student Member, IEEE) received the B.Sc. degree in electrical engineering and the M.Sc. degree in audio communication and technology in 2019 from the Technische Universität Berlin, Berlin, Germany. He is currently working towards the Doctoral degree with the Acoustics Lab, Aalto University, Espoo, Finland, focusing on parametric spatial audio processing. He specialized in signal processing and virtual acoustics with Technische Universität Berlin. From 2015 to 2017, he was a Research Assistant with TU Berlin, followed by two research internships (2017 and 2018) with Microsoft Research, Redmond, WA, USA, and Fraunhofer IIS Erlangen, Germany, in 2022. His research interests include high quality spatial audio, its coding and its perception. He was the Chair of the AES Berlin Student Section and part of the 142nd and 154th AES Convention committee.

Ville Pulkki is currently a Professor with the Department of Information and Communications Engineering, Aalto University, Helsinki, Finland. He has been working on the field of spatial audio for more than 25 years. He developed the vector-base amplitude panning method in his Ph.D. (2001) and directional audio coding after the Ph.D. with his research group. He also has contributions in perception of spatial sound, laser-based measurement of room responses, and binaural auditory models. He was the recipient of the Samuel L. Warner Memorial Medal Award from the Society of Motion Picture and Television Engineers and AES Silver Medal Award. He enjoys being with his family, building his summer house, playing various musical instruments, and acting, dancing and singing in musical ensembles.

Archontis Politis received the M.Sc. degree in sound and vibration studies from the Institute of Sound and Vibration Research, University of Southampton, Southampton, U.K., in 2008, and the Doctor of Science degree in spatial audio processing from Aalto University, Espoo, Finland, in 2016. He is currently an Assistant Professor with Tampere University, Tampere, Finland. From 2008 to 2009, he was a Researcher in a joint collaboration between the Glasgow School of Arts and Arup Acoustics, Glasgow, U.K., performing research on virtual acoustics. In 2015, he was a Visiting Researcher with the University of Maryland Institute for Advanced Computer Studies, MA, USA, and in the same year he completed a research internship with Microsoft Research, Redmond, WA, USA, on spatial audio technologies. He was the Editor of a book on Parametric Spatial Audio Processing, as Organizer in the DCASE scientific challenge, and has chaired various special sessions in international conferences. His research interests include spatial audio technologies, virtual acoustics, array signal processing, and acoustic scene analysis.

Leo McCormack received the B.Sc. degree in music technology and audio systems with the University of Huddersfield, Huddersfield, U.K., and the M.Sc. degree in computer communications and information sciences, majoring in acoustics and audio technology with Aalto University, Espoo, Finland. He is currently a Postdoctoral Researcher with the Acoustics Lab, Aalto University, researching parametric spatial audio technologies. His research interests include microphone array signal processing for sound-field capture and reproduction, and acoustic scene analysis.