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

Optimizing Higher-Order Directional Audio Coding with Adaptive Mixing and Energy Matching for Ambisonic Compression and Upmixing

Published in:
Proceedings of the 2023 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, WASPAA 2023

DOI:
10.1109/WASPAA58266.2023.10248179

Published: 01/01/2023

Document Version
Peer-reviewed accepted author manuscript, also known as Final accepted manuscript or Post-print

Please cite the original version:
Hold, C., McCormack, L., Politis, A., & Pulkki, V. (2023). Optimizing Higher-Order Directional Audio Coding with Adaptive Mixing and Energy Matching for Ambisonic Compression and Upmixing. In *Proceedings of the 2023 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, WASPAA 2023* (IEEE Workshop on Applications of Signal Processing to Audio and Acoustics; Vol. 2023-October). IEEE. https://doi.org/10.1109/WASPAA58266.2023.10248179

OPTIMIZING HIGHER-ORDER DIRECTIONAL AUDIO CODING WITH ADAPTIVE MIXING AND ENERGY MATCHING FOR AMBISONIC COMPRESSION AND UPMIXING

Christoph Hold,[1] *Leo McCormack,*[1] *Archontis Politis,*[2] *Ville Pulkki*[1]

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

ABSTRACT

In order to transmit sound-scenes encoded into the higher-order Ambisonics (HOA) format to low-bandwidth devices, transmission codecs are needed to reduce data requirements. Recently, the model-based higher-order directional audio coding (HO-DirAC) method was formulated for HOA input to HOA output. Compression is achieved by reducing the number of audio transport channels through spatial discretization. These transport channels are then used to reconstruct the scene on the receiving end based on accompanying spatial metadata. This reconstructed scene may also be optionally upmixed to a higher-order; leading to an enhancement in spatial-resolution. In this paper, the authors analyze certain sound-scenes that were especially challenging for the previously proposed HO-DirAC framework, which the authors postulate could be attributed to the lower-order reconstruction of diffuse sound-field components. Three optimizations for HO-DirAC are proposed, which all employ optimal adaptive mixing and/or energy matching of Ambisonic components based on spatial covariance matrices. The methods are formulated such that they are applied directly in the reconstruction of HOA from the spatially discrete transport audio signals. Notably, a dedicated low-complexity solution without additional side-information is derived. Instrumental evaluations confirm a reduced reconstruction error when using either of the proposed optimizations. These improvements were also demonstrated via a perceptual evaluation, whereby four, six, and twelve transport channels were used to reconstruct (and upmix to) fifth-order reference sound-scenes. The evaluation highlighted the high perceptual performance of the proposed optimizations, including the low-complexity version, thereby improving parametric spatial audio coding and reproduction.

Index Terms— spatial audio coding, higher order ambisonics, spherical harmonic domain

1. INTRODUCTION

Scene-based audio can encapsulate an arbitrary number of sound-sources into a defined number of audio channels. Often, these channels correspond to higher-order Ambisonic (HOA) audio streams, where limiting the number of channels constricts the spatial and timbral fidelity of the spatial audio scene [1]. Channel counts required for high resolution reproductions typically far exceed what is feasible for most practical applications. While coding the HOA signals directly can lead to significant compression [2, 3], the result typically scales with the number of HOA channels (quadratically), and (psycho-acoustic) intrinsic properties of the format can not be fully leveraged by such an approach. Hence, several dedicated spatial audio compression architectures for HOA have emerged over the years [4, 5, 6, 7, 8, 9, 10, 11].

A recent compression architecture is the spherical harmonic domain (SHD) input to output reformulation of higher-order directional audio coding (HO-DirAC) [12], which subdivides the spherical sound-scene into multiple (often uniformly) distributed sectors [13], and then applies a perceptually motivated parameterization on each of them. The formulation permits scaling the number of sectors, which in turn scales the number of audio transport channels (TCs) and required side-information used to restore the HOA signals. This architecture may provide high perceptual performance at a small fraction of the original data, which was previously compared against fifth-order reference HOA scenes [12]. The coder in [12] is formulated exploiting reconstruction properties of the sector design. While this allows coding higher-order directional sounds using the plane-wave (PW) assumption, the diffuse components in the scene may be reconstructed at lower orders. Therefore, presumably, some complex scenes could give rise to small timbral coloration and narrowing of the spatial image in the codec output. Covariance rendering by optimal mixing proposed in [14] has been formulated to ensure the rendered output matches target multichannel statistics dictated by the employed model, which can mitigate certain audible artifacts [15, 16]. The approach has been utilised for enhancing DirAC and for upmixing tasks in [15], for enhanced HO-DirAC loudspeaker reproduction in [13], and also for binaural reproduction in [17, 18, 16] utilising differing models. Previous work typically used all channels of the input to construct, e. g. , loudspeaker playback signals. However, we aim to avoid utilizing virtual loudspeakers, therefore, the current work may be not only more optimal in terms of channel mixing, but also avoiding potential sampling and aliasing problems.

We propose an optimization step, which applies to parametric spatial audio coding and upmixing tasks in the SHD. The proposed method can improve perceptual quality by matching the spatial covariance of the reconstructed audio to a target. The present paper develops a SHD optimal mixing solution to match the spatial covariance of a given input, reconstructed from a limited number of TCs. The end results are shown to more closely match the input RMS, now that a combination of reconstruction (low orders) and parametric covariance rendering (higher orders) is considered; thereby leveraging and combining the idiosyncratic benefits of both topologies.

2. METHOD

The HOA input signals $\boldsymbol{\chi}(t, f)$ are first converted into the time-frequency domain via an appropriate transform (such as [19]); where t and f are the time and frequency indices, respectively. Compression may be reached by reducing the number of transport channels, which may (in practice) be then further subjected to other (perceptual) audio codecs. This reduction of transport channels is achieved by spatially dividing the HOA signals using a spherical

979-8-3503-2372-6/23/$31.00

filterbank (SFB) [20], which comprises energy and/or amplitude preserving properties, as

$$\mathbf{x} = \mathbf{A}\boldsymbol{\chi}, \tag{1}$$

where $\mathbf{A}$ is a beamforming matrix. After transmission, the HOA signals may be reconstructed as

$$\tilde{\boldsymbol{\chi}} = \mathbf{B}\mathbf{x}, \tag{2}$$

where $\mathbf{B}$ is a reconstruction matrix. The latter can be derived as a perfect reconstruction, or a subset reconstruction (e. g. , only zeroth order), depending on the number of steering directions in $\mathbf{A}$ [20]. It is highlighted that perfect reconstruction is typically only achieved for lower orders than the input, due to estimator limitations and, more importantly, restrictions on the number of TCs. Typically, the beamformers in $\mathbf{A}$ are perceptually-motivated spatial filters reducing side- and back-lobes, such as the max-RE pattern [21] described by modal weights $\mathbf{c_n}$, which lead to increased estimator performance. The SFB analysis is expanded using $\mathrm{diag}_N[\cdot]$ repeating each entry for each order $2n+1$ times as

$$\mathbf{A} = \mathbf{Y}\,\mathrm{diag}_N\left[\mathbf{c_n}\right], \tag{3}$$

and the corresponding reconstruction is then given by

$$\mathbf{B} = \beta\,\mathrm{diag}_N\left[1/\mathbf{c_n}\right]\mathbf{Y}^{\mathsf{H}}, \tag{4}$$

where a factor β is derived in [22] to ensure preservation properties, with $\mathbf{Y} \in \mathbb{R}^{J\times L}$ the spherical harmonics matrix of order N with $L=(N+1)^2$ entries, and J the number of sector steerings. The latter also extracts the audio TCs, hence, compression gain can be achieved for $J<L$. Note that we may set $c_n = 1$ in (4) above the reconstruction order [12].

In the present work the aim is to match the spatial covariance matrix (SCM) of the reconstructed codec output $\tilde{\boldsymbol{\chi}}$, such that it is optimally close to the input SCM

$$\mathcal{E}\{\tilde{\boldsymbol{\chi}}\tilde{\boldsymbol{\chi}}^{\mathsf{H}}\} \approx \mathcal{E}\{\boldsymbol{\chi}\boldsymbol{\chi}^{\mathsf{H}}\}. \tag{5}$$

This is due to the assumption that a mismatch in input to reconstructed SCM would lead to a decrease in the perceptual performance of the spatial audio reproduction. However, in a coding scenario, we may not observe the input covariance at the decoder, and may also not want to transmit it. Note that the codec output might not match the input SCM for several reasons, including simplifications in the sound-field model or restrictions in the number of transport channels, and thus an insufficient reconstruction order.

2.1. Parameter Estimation

The input sound-field may be divided into J sectors, which are uniformly distributed on the sphere. Each sector $s \in [1,\ldots,J]$ consists of a sector pressure p_s and sector velocity $\mathbf{v}_s$, estimated from sector beamformers steered according to the SFB in $\mathbf{A}$. This subsequently leads to the estimation of the sector parameters [13], per time-frequency slot, as

$$\Omega_s = \angle\Re\{p_s\mathbf{v}_s^{\mathsf{H}}\}, \tag{6}$$

$$E_s = \frac{1}{2}(||p_s||^2 + \mathbf{v}_s^{\mathsf{H}}\mathbf{v}_s), \tag{7}$$

$$\psi_s = 1 - \frac{\Re\{p_s\mathbf{v}_s^{\mathsf{H}}\}}{E_s}. \tag{8}$$

It should be noted that each sector DoA Ω_s points towards an energy weighted average. The sector diffuseness ψ_s ranges from $[0,1]$, where a single impinging PW corresponds to 0. These parameter estimates may also be further post-processed. In the presented work, the diffuseness was filtered with a short median filter, and the DoA with a recursive spherical linear interpolation (SLERP), informed by the diffuseness estimate, similar to a Kalman filter. Of particular note is that the estimator incorporates sector velocity $\mathbf{v}_s$ in the sector energy estimate E_s, which is higher SHD order information only present at the encoder. HO-DirAC only transmits the transport audio signals $\mathbf{x}$ as in (1), which correspond to the sector pressures, alongside the DoA and diffuseness values for each sector. Note that for data efficient codecs, these are usually grouped and averaged according to auditory bands.

2.2. Optimal Mixing

In order to match the SCM corresponding to the reconstructed signals at the decoder to the full resolution input SCM, as in objective (5), several techniques are proposed. The following develops a solution for the presented SHD HO-DirAC problem. The SCM $\mathbf{C}_x$ is measured at the decoder

$$\mathbf{C}_x = \mathcal{E}\{\mathbf{x}\mathbf{x}^{\mathsf{H}}\}. \tag{9}$$

The encoder/decoder pair is derived to follow reconstruction properties [20]. If an energy-preserving design was chosen, we may reconstruct perfectly up to an order N_{dif}, i. e. , signals can be fully restored at the decoder using (2) and (4). Therefore, the low order entries in $\mathbf{C}_\chi$ can be matched, even without observing the input, as

$$\tilde{\mathbf{C}}_\chi = \mathcal{E}\{\tilde{\boldsymbol{\chi}}\tilde{\boldsymbol{\chi}}^{\mathsf{H}}\} = \mathbf{B}\mathbf{C}_x\mathbf{B}^{\mathsf{H}}. \tag{10}$$

If using fewer TCs, and following the preservation of amplitude criteria, then less components can be restored, e. g. , the zeroth order from $\tilde{\mathbf{C}}_{\chi 0} = \beta_A^2/4\pi\mathcal{E}\{\mathbf{x}\mathbf{x}^{\mathsf{H}}\}$.

The remaining (higher-order) components of target $\mathbf{C}_\chi$ are generally not observable at the decoder and are therefore based on a model, comprised of a directional $\mathbf{C}_{\mathrm{dir}}$ and diffuse $\mathbf{C}_{\mathrm{dif}}$

$$\mathbf{C}_\chi = \mathbf{C}_{\mathrm{dir}} + \mathbf{C}_{\mathrm{dif}}, \tag{11}$$

which are assumed to be uncorrelated. Over K estimates per block, the directional component is expanded under PW assumption to any order N_{dir} forming SH vector $\mathbf{y}$, as

$$\mathbf{C}_{\mathrm{dir}} = \frac{a4\pi}{KL}\sum_{s=1}^{J}\sum_{k=1}^{K}\mathbf{y}(\Omega_{s,k})(1-\psi_{s,k})E_{s,k}\mathbf{y}^{\mathsf{H}}(\Omega_{s,k}), \tag{12}$$

while the diffuse component may be modeled using a diffuse field SCM $\mathbf{F}_s$ (or only $\mathbf{I}$) with

$$\mathbf{C}_{\mathrm{dif}} = \frac{a}{KL}\sum_{s=1}^{J}\sum_{k=1}^{K}\psi_{s,k}E_{s,k}\mathbf{F}_{s,k}. \tag{13}$$

Both include a factor correcting the analysis normality error and order mismatch, determined as

$$a = \frac{(N_{\mathrm{out}}+1)^2}{\mathrm{trace}(\mathbf{A}^{\mathsf{H}}\mathbf{A})}. \tag{14}$$

Note that in the previously proposed formulation [12], only the lower-order diffuse stream components could be restored, whereas the covariance model in this work spans all SHD components.

The question then remains regarding the sector energy E_s. Here, depending on the application, we may either transmit the energy per sector (or a downsampled version of it), or re-estimate it at the decoder side. The former option should result in the highest quality, since it incorporates full order resolution from the encoder, but it is at the expense of additional metadata requirements. Whereas, given the preservation of the input, re-estimating at the decoder may still be valid in most cases which means no additional transport data is necessary. Re-estimation may be achieved as $\tilde{E}_s = ||x||^2$, although more informed model options could be imagined.

In [15], the authors proposed a solution to the time-frequency dependent mixing problem, solving for a matrix $\mathbf{M}$ which mixes the channels, connecting input to output SCMs $\mathbf{C}_x, \mathbf{C}_\chi$. The adaptive mixing is evaluated per time-frequency tile and takes the form

$$\tilde{\boldsymbol{\chi}} = \mathbf{M}\mathbf{x} + \mathbf{r}, \tag{15}$$

where $\mathbf{M}$ is a mixing matrix and $\mathbf{r}$ is the residual. The residual injects energy, if $\mathbf{M}$ can not meet the target sufficiently, and usually stems from decorrelated signals. However, decorrelation can not be easily applied directly to the SHD, and when applied in the discrete domain it is challenging to avoid spatially aliasing the SHD. We will hence assume that decorrelation is unnecessary [15] and instead match the target energy using a diagonal gain matrix as $\mathbf{M}' = \mathbf{G}\mathbf{M}$, noting that the boost is typically small.

The matrix $\mathbf{M}$ depends on the three matrices $\mathbf{C}_x$, $\mathbf{C}_\chi$, and $\mathbf{Q}$. The first two have been developed and the prototype matrix $\mathbf{Q}$ linearly maps the signals comprising $\mathbf{C}_x$ to those of the output format with $\mathbf{C}_\chi$. This linear mapping may not meet the target multichannel statistics of $\mathbf{C}_\chi$, however, the mapped signals serve as a constraint, since there are otherwise infinite solutions of matrices $\mathbf{M}$ that can match the input to output signal statistics [14]. Note that $\mathbf{M}$ reduces to $\mathbf{Q}$ if the output signal statistics are met, in which case $\mathbf{C}_\chi = \mathbf{Q}\mathbf{C}_x\mathbf{Q}^\mathsf{H}$. We may consequently formulate the whole HO-DirAC reconstruction as a time- and frequency dependent mixing from the TCs $\mathbf{x}$ to the HOA signals $\hat{\boldsymbol{\chi}}$

$$\hat{\boldsymbol{\chi}} = \mathbf{B}\,\mathrm{diag}(\boldsymbol{\psi}_s)\mathbf{x} + \beta_A \mathbf{Y}(\boldsymbol{\Omega}_s)\,\mathrm{diag}(1-\boldsymbol{\psi}_s)\mathbf{x}. \tag{16}$$

Pulling out $\mathbf{x}$ expresses the model compactly as a single matrix

$$\hat{\boldsymbol{\chi}} = \mathbf{Q}\mathbf{x}, \tag{17}$$

which can be utilized directly as the prototype mixing matrix $\mathbf{Q}$, upon which the covariance rendering refines. In contrast to previous work, the prototype $\mathbf{Q}$ is formulated as time- and frequency-variant.

We interpret the mixing matrix also as a time-frequency adaptive transform matrix from the TCs to HOA $\mathbf{x} \rightarrow \tilde{\boldsymbol{\chi}}$. This makes it obvious that the optimal solution is given directly as the SFB reconstruction $\mathbf{B}$, wherever possible, meaning the $L = N_{\text{in}}^2$ channels for an energy preserving sector design, and up to $L = J$ for an amplitude preserving design. For the implementation, regularization was increased while limiting amplification in $\mathbf{G}$ to 6 dB, the SCMs were estimated utilizing 8 blocks of 128 samples, and $\mathbf{M}$ was finally smoothed with $2/3$ current, $1/3$ previous solution (not recursive).

2.3. SHD Energy Matching

Despite being optimal in a least-squares sense, the previously shown full mixing solution might not be necessary in practice, or even not desired, due to high computational requirements and potential issues with noise. This paper, therefore, develops a simplified version based on adaptive scaling of the energy in the model based mixing $\mathbf{Q}$.

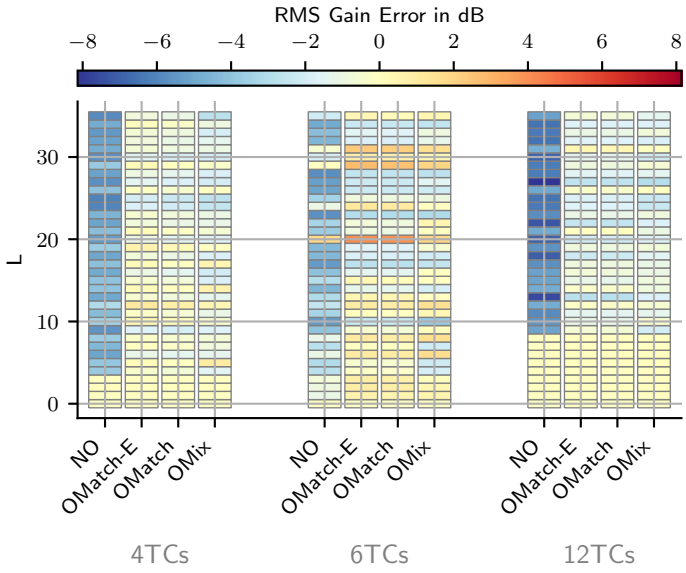

Figure 1: Gain Error from RMS over SHD ($L = n^2 + 2n + m$), compared to the input scene *Orchestra*, with and without covariance rendering. Noting the perfect match up to the reconstruction order, beyond that OM helps to match the energy per order.

TCs	4	6	12
NO	$\mathbf{-2.3}$ [0.5; 4.7]	$\mathbf{-1.7}$ [0.3; 3.0]	$\mathbf{-2.0}$ [0.3; 4.5]
OMix	$\mathbf{-0.5}$ [0.2; 2.0]	$\mathbf{-0.3}$ [0.1; 0.8]	$\mathbf{-0.3}$ [0.0; 1.1]
OMatch	$\mathbf{-0.4}$ [0.1; 2.0]	$\mathbf{-0.2}$ [0.0; 0.8]	$\mathbf{-0.4}$ [0.0; 1.1]
OMatch-E	$\mathbf{-0.4}$ [0.0; 2.1]	$\mathbf{-0.3}$ [0.1; 1.1]	$\mathbf{-0.4}$ [0.0; 1.2]

Table 1: Average dB reconstruction RMS deviations to the reference input (5th order) over multiple items. The mean is shown in bold, next to the minimum and maximum absolute error.

By applying $\mathbf{Q}$ onto $\mathbf{C}_x$

$$\mathcal{E}\{\mathbf{Q}\mathbf{x}(\mathbf{Q}\mathbf{x})^\mathsf{H}\} = \mathbf{Q}\mathbf{C}_x\mathbf{Q}^\mathsf{H}, \tag{18}$$

the matching becomes

$$\mathbf{M} = \sqrt{\frac{\mathcal{E}_n\{\mathrm{diag}(\mathbf{C}_\chi)\}}{\mathcal{E}_n\{\mathrm{diag}(\mathbf{Q}\mathbf{C}_x\mathbf{Q}^\mathsf{H})\}}}\,\mathbf{Q} = \mathbf{D}\mathbf{Q}, \tag{19}$$

where $\mathcal{E}_n$ is the expectation over order, e. g. , the mean of all degrees per order, and the element-wise square-root. Then $\mathbf{D}$ is a diagonal matrix matching the target energy by refining the prototype mixing $\mathbf{Q}$. It should be noted that $\mathcal{E}_n\{\mathrm{diag}(\mathbf{C}_\chi)\}$ may be significantly simplified if energy matching is performed only over SHD order, since $C_\chi(n) = \sum_{s=1}^{J} E_s[a_1(1-\psi_s)\frac{(2n+1)}{4\pi} + a_2\psi_s\mathbf{F}_s(n)]$ with the corresponding factors from (12) - (13). Despite the latter not matching the energy of each SHD degree, because of the angle dependency and hence dependency in degree of $\mathbf{C}_\chi$, we found similar performance in practice, while further simplifying the computation.

3. EVALUATION

An instrumental, as well as perceptual validation, was conducted to compare the previously developed full adaptive mixing solution

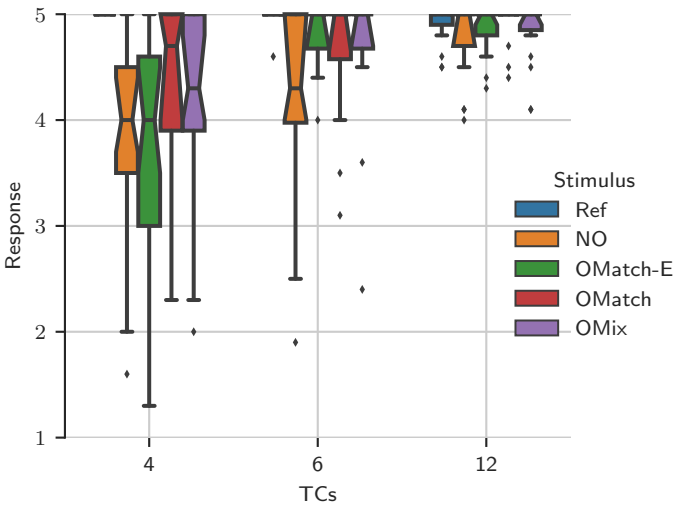

Figure 2: Listening experiment results, boxplot showing the median and quartiles, the notches indicate 95% CI from bootstrapping.

alongside the model parameter informed matching simplification. Figure 1 provides insight regarding the RMS mismatch over order, observed between the reference input and restored output. It shows that the reconstruction properties of the SFB hold for lower-orders. In the presented item a lack of higher-order energy emanates for the previously formulated version *NO* (without optimization). Furthermore, it illustrates how the presented optimizations can restore higher-order components and thus decrease the reconstruction error. Evaluations for SHD HO-DirAC with *OMix* (sec.2.2), *OMatch* (sec.2.3), and *OMatch-E* (with re-estimated sector-energies at the decoder, i. e. , no additional side-information) are shown. Table 1 summarizes the reconstruction error, from $\mathrm{RMS}_{nm}(\mathrm{out})/\,\mathrm{RMS}_{nm}(\mathrm{in})$, where the mean is in bold, next to the minimum and maximum absolute error over all items. The error was evaluated based on five diverse items, typically used to test parametric spatial audio algorithms. It highlights the increased performance by reduced reconstruction error for any of the proposed optimizations, where all could improve on the baseline. It also reveals that the previous formulation can already achieve reconstruction very close to the reference (small min. error), for scenes that confirm to the HO-DirAC model and presumably without significant higher-order diffuse stream energy. It is highlighted here that an RMS reconstruction error is not necessarily an indicator of low perceptual quality, as the model is heavily perceptually inspired, e. g. , by rendering conflicting DoAs as an energetic mean. All items are made available online.

3.1. Perceptual Evaluation

A listening test established the perceptual distances between an uncompressed fifth-order reference case and the presented methods. A multiple stimulus with hidden reference perceptual test design was selected for this task, using the labels: Imperceptible; Perceptible, but not annoying; Slightly annoying; Annoying; Very annoying. The listeners were presented with 5 sliders randomly corresponding to a hidden reference, or HO-DirAC conditions as in Tab. 1, switching instantaneously. The fifth-order streams were decoded by AllRAD [21] to a spherical loudspeaker setup, supporting decoding to fifth-order, and were playing at just below $< 80\,\mathrm{dB(C)}$. Two items were tested, which were identified as being especially critical in previous evaluations [12, 23]; i. e. , they evoked the largest perceptual differences. These two items, a dense orchestral piece, and a band in a reverberant environment consisting of simultaneous drums, shaker, violins and bass, played in a 10 s loop. In total, 11 participants (reported age avg. 28 years) rated the conditions. All were experienced listeners. The test was, however, found to be especially challenging, which also formed part of the motivation for including only two scenes; as this allowed the test duration to be approximately 20 min. Participants could not, however, identify the hidden reference case a total of six times (< 4.5), which led to the removal of those particular trails.

4. PERCEPTUAL STUDY RESULTS AND DISCUSSION

The results of the perceptual experiment are shown in Fig. 2. It should be noted that the previous solution in [12] has already been shown to attain high perceptual quality in typical scenarios, while certain sound-scenes produced minor timbral colorations of non-directional sound-field components. While the framework is able to restore directional signals to full order (and can potentially even increase their order), it can only fully restore the diffuse parts up to the reconstruction order. As the reconstruction order is typically much lower, depending on the number of TCs, we attribute this mismatch for the perceptible difference. Therefore, the proposed optimization approaches largely seek to aim to improve matching the energy of these diffuse-field components at higher-orders. A Kruskal-Wallis test implies group differences amongst TCs and Stimuli (with and without *Ref*). The results confirm that an increase in TCs generally leads to improved perceptual quality. The four TC mode represented a challenging case, indicating that these particular scenes under test saturated the available TCs. For the 12 TC mode, the results suggest that all tested variants of HO-DirAC were largely transparent with the reference, which is also supported by participants rating the reference lower than other stimuli. The perceptual gains afforded by the three proposed post-processing approaches were, therefore, only perceptually confirmed in the four and six TC modes, with perceptual gains more clearly shown for the latter. While the *OMix* solution can potentially change the inter-order relations, this benefit is not evident in the performance. OMatch and OMatch-E showed similar improvements in the objective metrics, only confirmed perceptually for 12 TCs and, surprisingly, for 6 TCs (no full energy preservation of the input here). Using four TCs, the perceptual performance likely already saturates due to the limited available TCs and DoA estimates, which the proposed techniques can not overcome.

5. CONCLUSION

This paper investigated transmitting Ambisonic recordings using a reduced number of transport channels, their reconstruction, and subsequent spatial upmixing to higher-orders. A previous formulation of HO-DirAC served as the foundation for this study, upon which, three optimisation approaches were proposed. The explored enhancements are based on optimal adaptive mixing and/or energy matching of the SCMs and resulted in reduced reconstruction errors. High perceptual performance was demonstrated when coding second-order Ambisonics into four transport channels, and third-order into six or twelve transport channels, and subsequently reconstructing and upmixing the transmitted audio to fifth-order. It was also demonstrated that the proposed enhancement may operate without any additional side-information compared to the previous HO-DirAC formulation.

6. 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," *The Journal of the Acoustical Society of America*, vol. 133, no. 5, pp. 2711–2721, 2013.

[2] B. Lee, T. Rudzki, J. Skoglund, and G. Kearney, "Context-based evaluation of the opus audio codec for spatial audio content in virtual reality," *Journal of the Audio Engineering Society*, vol. 71, no. 4, pp. 145–154, 2023.

[3] T. Rudzki, I. Gomez-Lanzaco, J. Stubbs, J. Skoglund, D. T. Murphy, and G. Kearney, "Auditory localization in low-bitrate compressed ambisonic scenes," *Applied Sciences*, vol. 9, no. 13, p. 2618, 2019.

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

[5] D. Sen, N. Peters, M. Y. Kim, and M. Morell, "Efficient Compression and Transportation of Scene Based Audio for Television Broadcast," in *AES International Conference on Sound Field Control*, 2016.

[6] A. Daniel, R. Nicol, and S. McAdams, "Multichannel Audio Coding Based on Minimum Audible Angles," in *40th International Conference: Spatial Audio: Sense the Sound of Space*, 2010, pp. 1–10.

[7] S. Zamani, "Signal Coding Approaches for Spatial Audio and Unreliable Networks," Ph.D. dissertation, University of California Santa Barbara, 2019.

[8] S. Zamani and K. Rose, "Spatial Audio Coding without Recourse to Background Signal Compression," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019.

[9] J. Herre, J. Hilpert, A. Kuntz, and J. Plogsties, "MPEG-H 3D Audio - The New Standard for Coding of Immersive Spatial Audio," *IEEE Journal on Selected Topics in Signal Processing*, vol. 9, no. 5, 2015.

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

[11] E. Hellerud, A. Solvang, and U. P. Svensson, "Spatial redundancy in Higher Order Ambisonics and its use for lowdelay lossless compression," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2009.

[12] C. Hold, V. Pulkki, A. Politis, and L. McCormack, "Compression of higher-order ambisonic signals using directional audio coding," *Under review*, 2022.

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

[14] J. Vilkamo, T. Bäckström, and A. Kuntz, "Optimized covariance domain framework for time–frequency processing of spatial audio," *Journal of the Audio Engineering Society*, vol. 61, no. 6, pp. 403–411, 2013.

[15] J. Vilkamo and V. Pulkki, "Minimization of decorrelator artifacts in directional audio coding by covariance domain rendering," *Journal of the Audio Engineering Society*, vol. 61, no. 9, pp. 637–646, 2013.

[16] L. McCormack and A. Politis, "Estimating and reproducing ambience in ambisonic recordings," in *2022 30th European Signal Processing Conference (EUSIPCO)*. IEEE, 2022, pp. 314–318.

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

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

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

[20] C. Hold, S. J. Schlecht, A. Politis, and V. Pulkki, "Spatial filter bank in the spherical harmonic domain: Reconstruction and application," in *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2021, pp. 361–365.

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

[22] C. Hold, A. Politis, L. McCormack, and V. Pulkki, "Spatial Filter Bank Design in the Spherical Harmonic Domain," in *2021 29th European Signal Processing Conference (EUSIPCO)*, no. August. IEEE, 2021.

[23] L. McCormack, C. Hold, and A. Politis, "Parametric architecture for the transmission and binaural reproduction of microphone array recordings," in *Audio Engineering Society Conference: AES 2023 International Conference on Spatial and Immersive Audio*. Audio Engineering Society, 2023.