Ambisonic Coding with Spatial Image Correction

Pierre Mahé, Stéphane Ragot, Sylvain Marchand, Jérôme Daniel

Ambisonic Coding with Spatial Image Correction

Pierre MAHÉ[1,2] Stéphane RAGOT[1] Sylvain MARCHAND[2] Jérôme DANIEL[1]

[1]*Orange Labs, Lannion, France* [2]*L3i, Université de La Rochelle, France*

pierre.mahe@orange.com, stephane.ragot@orange.com,
sylvain.marchand@univ-lr.fr, jerome.daniel@orange.com

***Abstract*—We present a new method to enhance multi-mono coding of ambisonic audio signals. In multi-mono coding, each component is represented independently by a mono core codec, this may introduce strong spatial artifacts. The proposed method is based on the correction of spatial images derived from the sound-field power map of original and degraded ambisonic signals. The correction metadata is transmitted as side information to restore the spatial image by post-processing. The performance of the proposed method is compared against naive multi-mono coding (with no side information) at the same overall bitrate. Experimental results are provided for the case of First-Order Ambisonic (FOA) signals and two mono core codecs: EVS and Opus. The proposed method is shown to provide on average some audio quality improvement for both core codecs. ANOVA results are provided as a complementary analysis.**

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

I. Introduction

With the emergence of new spatial audio applications (e.g. Virtual or Extended Reality or smartphone-based immersive capture) there is a need for efficient coding of immersive audio signals, especially for mobile communications. There are different ways to represent a spatial audio scene and ambisonics [1]–[3] has become an attractive format.

For that purpose, the most straightforward approach to code ambisonic audio signals is to extend existing mono or stereo codecs by "discrete coding", i.e. compress ambisonic components using separate core codec instances. When the core codec is mono, this approach is hereafter referred to as multi-mono coding. The objective of this paper is to enhance the quality of multi-mono coding in a backward compatible way with a new method based on spatial post-processing with side information.

At low bitrates, multi-mono/-stereo coding causes several quality artifacts [4] because correlation between components is not preserved. Spatial artifacts may be reduced by applying a fixed channel matrixing before/after multi-mono/-stereo coding as implemented in ambisonic extensions of Opus [5], [6] or e-AAC+ [7, clause 6.1.6]. Subjective test results reported in [8] suggest that matrixing methods may have some quality limitations when the ambisonic order increases and that matrixing does not remove all artifacts caused by the core codec.

An alternative to multi-mono/-stereo coding is to analyze the audio scene to extract sources and spatial information. The DirAC method [9] estimates the dominant source direction and diffuseness parameters from first-order ambisonic signals and re-creates the spatial scene based on a mono downmix and these spatial parameters. This method has been extended to High-Order Ambisonics (HOA) in HO-DirAC [9] where the sound field is divided into angular sectors; for each angular sector, one source is extracted. The recent COMPASS method [10] estimates the number of sources and derives a beamforming matrix by Principal Component Analysis (PCA) to extract sources. These approaches are efficient at low bitrates, but they have the disadvantage to reproduce only spatial cues, and not the original spatial image. Moreover, they rely on assumptions on the ambisonic scene (e.g. number of sources, positions). If these assumptions are not valid for some signals, coding performance may be strongly impacted.

In this work, we present a new coding method that may be interpreted as a spatial audio post-processing. The objective of the method is to restore the spatial image of an ambisonic signal altered by multi-mono coding. The image spatial correction is derived by comparing the sound-field power map of original and coded ambisonic signals.

In the field of ambisonic processing, numerous works studied spatial image visualization of an audio scene. A simple way to visualize spatial images is to compute the sound-field power map using steered response power (SRP) beamforming [11], [12]. Basic SRP may be replaced by a different type of beamforming such as MVDR [13]. For spatial audio effects, several solutions were proposed to manipulate and modify a spatial image. For instance, methods for ambisonic-domain spatial processing (e.g. amplification of one direction) are described in [14].

This paper is organized as follows. Section II gives an overview of ambisonics, beamforming and soundfield powermap concepts. Section III explains the key principles of the proposed spatial image correction. Section IV provides a full description of the proposed coding approach with post-processing for any ambisonic order. Section V presents subjective test results for the FOA case and compares the proposed method with naive multi-mono coding.

II. Ambisonics

Ambisonics is based on a decomposition of the sound field on a basis of spherical harmonics. Initially limited to first order [1], the formalism has been extended to higher orders [2]. We refer to [3] for fundamentals of ambisonics. To perfectly reconstruct the sound field, an infinite order is required. In practice, the sound field representation is truncated to a finite

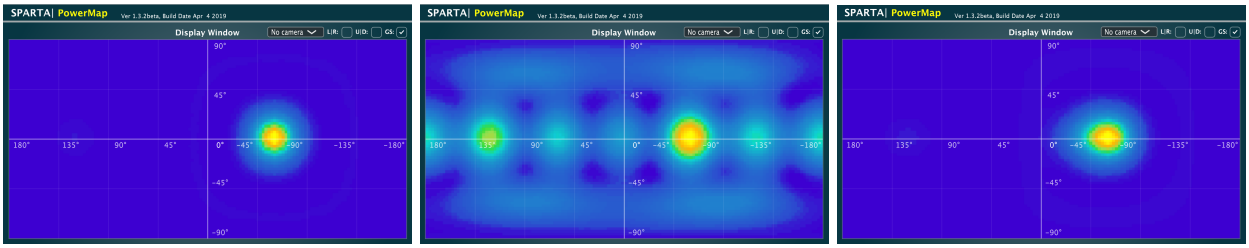

Fig. 1. *Power map of original (left), coded (middle) and post-processed (right) ambisonic signals for pink noise localized at* $(-75°, 0°)$.

order N. For a given order N the number of ambisonic components is $n = (N+1)^2$.

As an example, First-Order Ambisonic (FOA) ($N = 1$) has $n = 4$ components: *W*, *X*, *Y*, *Z*. A plane wave with pressure $s(t)$ at azimuth θ and elevation ϕ (with the mathematical convention) is encoded to the following B-format representation:

$$\mathbf{b}(t) = \begin{bmatrix} w(t) \\ x(t) \\ y(t) \\ z(t) \end{bmatrix}^T = \begin{bmatrix} 1 \\ \cos\,\theta\,\cos\,\phi \\ \sin\,\theta\,\cos\,\phi \\ \sin\,\phi \end{bmatrix}^T s(t) \tag{1}$$

More generally, a mono source $s(t)$ in direction (θ, ϕ) is "projected" (encoded) into the spherical harmonic domain at order N, using a weight vector $\mathbf{d}(\theta, \phi) = [d_1(\theta, \phi), \cdots, d_n(\theta, \phi)]$ corresponding to the contribution of each ambisonic component:

$$\mathbf{b}(t) = \mathbf{d}(\theta, \phi)\, \mathbf{s}(t) \tag{2}$$

where $\mathbf{b}(t)$ is an ambisonic signal and $\mathbf{d}(\theta, \phi)$ is the encoding vector, also called weight vector. Alternatively, the part from signal $\mathbf{s}(t)$ originating from a direction (θ, ϕ) may be extracted from the ambisonic signal $\mathbf{b}(t)$ with the inverse vector transformation, known as beamforming or spatial filtering [13].

It is possible to transform the ambisonic representation to another sound field representation by selecting a combination of m beamforming vectors. The vectors are merged into a transformation matrix $\mathbf{D}$ of size $n \times m$.

$$\mathbf{a}(t) = \mathbf{D}\mathbf{b}(t) \tag{3}$$

where $\mathbf{D} = [\mathbf{d}(\theta_1, \phi_1), \cdots, \mathbf{d}(\theta_m, \phi_m)]^T$ is the transformation matrix, $\mathbf{b}(t) = [\mathbf{b}_1(t), \cdots, \mathbf{b}_n(t)]$ is the input matrix of n ambisonic components, $\mathbf{a}(t) = [\mathbf{a}_1(t), \cdots, \mathbf{a}_m(t)]$ is the output matrix of the sound field representation. The two representations are equivalent provided that the matrix $\mathbf{D}$ is unitary and $m \geq n$:

$$\mathbf{D}^\mathbf{T}\mathbf{D} = \mathbf{I} \tag{4}$$

where $\mathbf{I}$ is the identity matrix. If the matrix $\mathbf{D}$ does not satisfy this condition, spatial modification will occur. In this work, we will modify coefficients of the transformation matrix to adjust the power level in specific directions.

A power map can be used to measure sound field activity in an ambisonic signal. Many computation methods exist, such as SRP, MVDR, or CroPaC [15]. We focus here on SRP which scans a discrete set of directions over the sphere:

$$s_i(t) = \sum_{k=1}^{n} \mathbf{d}_i(k)\, b_k(t) \tag{5}$$

where $s_i(t)$ is the extracted signal of the i^{th} beam, $\mathbf{d}_i$ the weight vector for the beam direction and $b_k(t)$ the k^{th} ambisonic components. The power in the i-th direction is:

$$\mathbf{P}_i = \frac{1}{L} \sum_{t=1}^{L} s_i(t)^2 \tag{6}$$

where t is used as a discrete index in a L-sample frame. A more efficient way to compute $\mathbf{P}_i$ is based on the covariance matrix $\mathbf{C} = \mathbf{B}\,\mathbf{B}^\mathbf{T}$ of the ambisonic samples, the later being gathered in matrix $\mathbf{B} = [\mathbf{b}_k(t=1), \cdots, \mathbf{b}_k(t=L)]$:

$$\mathbf{P}_i = \mathbf{d}_i \mathbf{C} \mathbf{d}_i \tag{7}$$

The sound-field power map may be visualized by plotting $\mathbf{P}_i$ as a function of azimuth and elevation, e.g. with an equirectangular projection. An example of such a $20ms$ power map of the audio scene is shown in Fig. 1. A pink noise was generated and spatialized to $(\theta = -75°, \phi = 0°)$ at order $N = 3$. To illustrate this, we used here the publicly available VST plugin SPARTA [16] to generate the power map of the original, coded and post-processed signals. The ambisonic signal was encoded by multi-mono EVS [17] at $n\times$16.4 kbit/s. It can be seen that multi-mono coding results in phantom sources and spatial alterations. After the proposed post-processing, sound spatialization is restored.

III. Spatial Post-Processing in Ambisonic Domain

The principle of the proposed coding method is to correct the spatial degradation due to multi-mono coding by post-processing. The method computes a power map of the original ambisonic signal and a power map of the coded ambisonic signal after core codec processing. A spatial correction matrix is derived based on both power maps, with the objective to restore as well as possible the original signal spatialization. The transformation matrix is quantized and transmitted as side information on top of the multi-mono bitstream, in a backward compatible way. The proposed approach is formulated with the objective to find a transformation matrix $\mathbf{T}$ of size $n \times n$ such that the power map of $\mathbf{T}\hat{\mathbf{B}}$ is identical to that of $\mathbf{B}$, where $\mathbf{B}$ and $\hat{\mathbf{B}}$ are, respectively, the original and coded ambisonic signals. As shown in Section II, the sound-field power map can be obtained from the covariance matrix. Therefore, the problem be reformulated as follows:

$$\mathbf{d}_i \mathbf{C} \mathbf{d}_i = \mathbf{d}_i \widetilde{\mathbf{C}} \mathbf{d}_i \tag{8}$$

where i is the index of a spherical grid (θ_i, ϕ_i) on the unit sphere, $\mathbf{C}$ are the covariance matrix of the original and $\widetilde{\mathbf{C}} =$

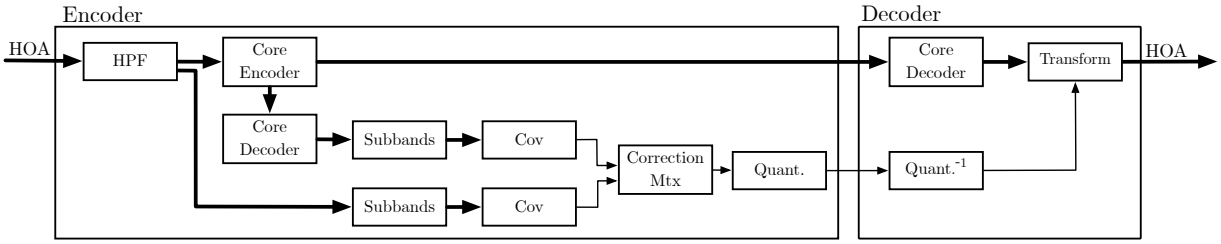

Fig. 2. *Overview of proposed coding method.*

$\mathbf{T}\hat{\mathbf{B}}\hat{\mathbf{B}}^T\mathbf{T}^T$, the restored ambisonic signal. From Eq. (8), one can verify that transformation matrix must satisfy:

$$\mathbf{T}\hat{\mathbf{C}}\mathbf{T}^T = \mathbf{C} \tag{9}$$

A possible solution to Eq. 9 can be found by Cholesky factorization of $\mathbf{C}$ and $\hat{\mathbf{C}}$ in terms of triangular matrices $\mathbf{L}$ and $\hat{\mathbf{L}}$ (with real and positive diagonal coefficients), where $\mathbf{C} = \mathbf{L}\mathbf{L}^T$ and $\hat{\mathbf{C}} = \hat{\mathbf{L}}\hat{\mathbf{L}}^T$. Thus, Eq. 9 becomes:

$$(\mathbf{T}\hat{\mathbf{L}})(\mathbf{T}\hat{\mathbf{L}})^T = \mathbf{L}\mathbf{L}^T \tag{10}$$

which gives the following solution:

$$\mathbf{T} = \mathbf{L}\hat{\mathbf{L}}^{-1} \tag{11}$$

Note that the Cholesky factorization requires $\mathbf{C}$ and $\hat{\mathbf{C}}$ to be positive definite matrices and these covariance matrices are only guaranteed to be positive semi-definite. A conditioning bias ϵ (e.g. 10^{-9}) is added to the diagonal of both matrices.

IV. General Codec Description

The proposed coding method operates on successive frames of 20 ms, which is the typical frame length used in mobile telephony. The overall codec architecture is shown in Fig. 2. We describe below a general implementation for any ambisonic order, while experiments are later reported for the FOA case ($N = 1$).

A. Encoder

The encoder input is an ambisonic signal in B-format. A 20Hz IIR high-pass filter (detailed in [18]) is applied on each ambisonic component to avoid any bias in the subsequent covariance matrix estimation. The input is multi-mono coded and decoded, by coding separately each ambisonic component by an independent core codec instance. The original signal and the coded signal are divided into subbands. To limit the number of subbands, Bark bands are merged to obtain 7 bands.

The encoder computes a spatial correction matrix $\mathbf{T}_b$ ($b = 1, \cdots, 7$) in each band. Note that we do not use a frame index to simplify notations in the following description. To guarantee smooth inter-frame transitions after spatial image correction, the determination of $\mathbf{T}_b$ is based on an analysis window covering 40 ms (current and previous frames) with a 50% overlap. In each subband the covariance matrices $\mathbf{C}_b$ and $\hat{\mathbf{C}}_b$ are estimated and the correction matrix $\mathbf{T}_b$ is computed as described previously in this section. The correction matrix $\mathbf{T}_b$ is normalized to preserve the energy level of the W ambisonic component in each subband; this ensures that the proposed correction method does not compensate for the core codec frequency response, especially in high frequency bands.

The transformation matrices $\mathbf{T}_b$ are coded using by μ-law companded scalar quantization with 8 bits per coefficients. Note that the matrices $\mathbf{L}_b$ and $\hat{\mathbf{L}}_b$, defined previously, are triangular, therefore $\mathbf{T}_b$ is also triangular. Only part of $\mathbf{T}_b$ needs to be transmitted.

B. Decoder

After multi-mono decoding, the decoded signal is divided into subbands. In parallel, the quantized transform matrix $\hat{\mathbf{T}}_b$ is received by the decoder. Then, transform matrix is applied to correct the spatial image.

C. Details on metadata coding for FOA

For the FOA case ($n = 4$), 10 coefficients of $\mathbf{T}_b$ need to be transmitted for each band. Moreover, the first coefficient (upper left corner) of $\mathbf{T}_b$ is normalized by W energy, thus it does not need to be transmitted. Only 9 correction coefficients are coded (with 8-bit scalar quantization) in 7 subbands, which results in a bitrate of 25.2 kbit/s for the correction metadata, on top of the multi-mono bitstream. Note that the diagonal entries of $\mathbf{T}_b$ are always positive and it would be possible to save 1 sign bit per subband, however this optimization is discarded because it does not have a significant effect.

V. Experimental Results for FOA

A. Test setup

The proposed post-processing is designed to operate with any core codec. In this test we used two core codecs for multi-mono coding: EVS [17] and Opus [5]. With such low bitrate codecs, one cannot measure performance with objective metrics as SNR or MSE [19]. A subjective test is required. Several subjective test methodologies have been developed for mono or stereo, e.g. ITU-T P.800 (ACR, DCR, CCR) [20], ITU-R BS.1534 (MUSHRA) [21]. For immersive audio coding, test methodologies are less mature and often adapted from existing ones. To evaluate the proposed coding method, we conducted a Comparison Category Rating (CCR)-like subjective test [20]. For each test item, subjects were asked to compare the audio quality of two conditions (A and B) operated at the same overall bitrate, where one is naive multi-mono coding, and the other is multi-mono coding with post-processing. The test interface is shown in Fig.3. The grading scale ranged of -3 to 3 ("*B is much better than A*" to "*A is much better than B*"), and only integer scores were allowed.

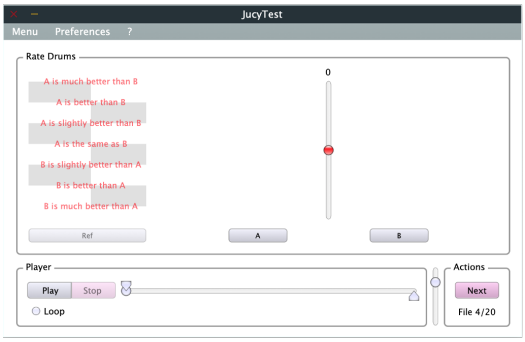

Fig. 3. *Test interface.*

In both conditions, spatial and timbral artifacts may occur. To evaluate which condition is the most fidelity to the original signal, we added an explicit reference with the original item (without coding). This ground-truth condition was explicitly labeled as the reference ("Ref") for subjects.

Both core codecs (EVS and Opus) were operated in super-wideband (SWB) mode at two different bitrates (R_1 and R_2). For EVS, R_1 = 97.6 (4×24.4) kbit/s and R_2 = 128.0 (4×32.0) kbit/s. For Opus, R_1 = 128.0 (4×32.0) kbit/s and R_2 = 160.0 (4×40.0) kbit/s. These bitrates are overall bitrates; the proposed coding method used the same overall bitrate budget which was split between multi-mono coding and metadata coding. All test conditions are summarized in Table I. At a given overall bitrate, the proposed coding method with post-processing method allocates a lower bitrate to multi-mono coding to represent ambisonic components because some budget is reserved for metadata. The EVS-SWB bitrate granularity is limited to a discrete set $(9.6, 13.2, 16.4, 24.4, 32, 48, 96, 128)$, therefore some bitrate was left unused for the proposed method at R_1 and R_2. We preferred to keep the same metadata coding (at 25.2 kbit/s) for EVS and Opus.

The input FOA signals were sampled at 32 kHz. All (original and coded) FOA test items were binauralized for headphone presentation. We selected the SPARTA binaural renderer with default settings [16]. The test items consisted of 10 challenging ambisonic items, already used and decribed in the work reported in [18]. To add more spatial variety, we added two items: "Castanets", a signal from the EBU SQAM database [22] spatialized at several static positions around the listener ; "Noise", a pink noise source rotating smoothly around the listener in the horizontal plane.

The presentation of items and A/B conditions were randomized for each subject. All subjects conducted the listening test with the same professional audio hardware in a dedicated listening room. In total 24 listeners participated (11 expert and 13 naive).

B. Subjective test results

Overall test results are shown in Fig. 4. In most cases the proposed coding method provides improvement over naive multi-mono for both codecs (EVS, Opus) and bitrates (R_1, R_2). The post-processing only corrects the spatial image, therefore this quality improvement can be attributed to spatial

TABLE I
Pairs of condition for CCR-like test and related bitrates (in kbit/s).

Core codec	Multi-mono bitrate	Proposed coding method
EVS	4×24.4 (97.6)	$4 \times 16.4 + 25.2$ (90.8)
EVS	4×32.0 (128.0)	$4 \times 24.4 + 25.2$ (122.8)
Opus	4×32.0 (128.0)	$4 \times 25.7 + 25.2$ (128.0)
Opus	4×40.0 (160.0)	$4 \times 33.7 + 25.2$ (160.0)

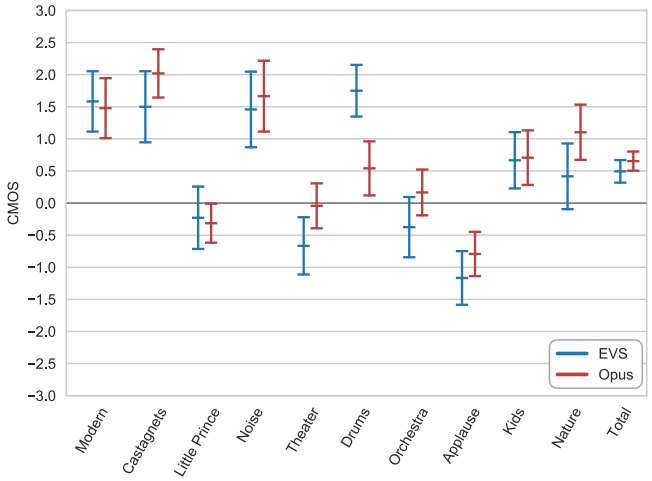

Fig. 4. *Overall results mean scores with 95% confidence intervals: a positive (resp. negative) score indicates that the proposed coding method improves (resp. degrades) with respect to naive multi-mono.*

image restoration. With the proposed coding method, localization artifacts for naive multi-mono, reported in [4], were almost removed. Three items were, on average, significantly degraded: *Applause*, *Theater* and *Little Prince*. The spatial correction estimation covers 40 ms; for items with multiple transients in different directions over this period of time, as in *Applause*, the estimated correction is an average of multiple directions and it may amplify a wrong part of space. Moreover, the correction may also result in some kind of spatial pre/post-echo if other sounds mixed with strong percussive events are amplified. The degradation may also be explained by the reduced bit allocation to the core codec when post-processing is used, especially for the *Applause* item which is complex to code.

The two other items (*Theater*, *Little Prince*) exhibit some strong reverberation that may affect post-processing.

C. ANOVA statistical analysis

A further analysis was done to determine if the performance of the proposed coding method is influenced by codecs and bitrates. A first variance analysis was conducted on CMOS scores with the within-group factors "Codec" (2 levels: EVS/Opus), "Bitrate" (2 levels: bitrate1/bitrate2) and "Item Content" (10 levels) and the Between-group factor "Expertise" (2 levels: Naive and Expert). This showed that there is no significant effect of the factor "Expertise" or significant interaction between this factor and the others.

Another variance analysis was conducted without the factor "Expertise", but still with the within-group factors "Codec",

TABLE II
ANOVA results for within-group factors "Codec", "Bitrate" and "Item".

Effect	Deg. of freedom	F-ratio	p-value
Codec	1	3.71	0.067
Bitrate	1	10.12	0.004
Item	9	27.64	0.000
Codec * Bitrate	1	28.39	0.000
Codec * Item	9	3.43	0.001
Bitrate * Item	9	2.63	0.007
Codec * Bitrate * Item	9	1.22	0.283

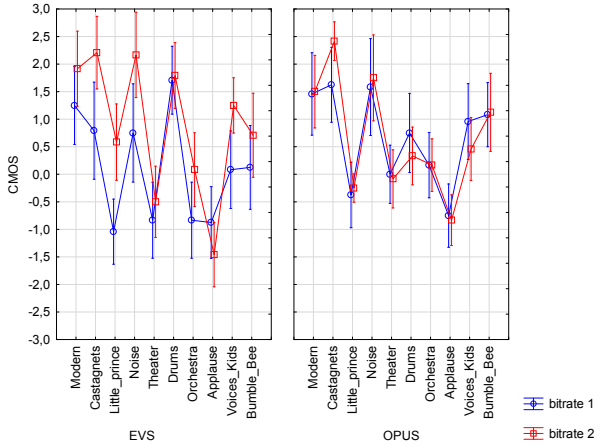

Fig. 5. *Test results as a function of codecs (EVS, Opus) bitrates (R_1, $R2$ labeled "bitrate1" and "bitrate2") and items.*

"Bitrate" and "Item". As shown in Table II, this showed a significant effect of factors "Bitrate" and "Item", as well as significant two-way interactions. Those effects and interactions are illustrated in Fig. 5. The results showed that there is no dependency on the codec, and this suggests that the proposed post-processing method may work for any codec used for multi-mono coding.

It is worth noting that the bitrate influence is strongly visible for the EVS codec. Indeed, multi-mono coding is allocated a lower bitrate when post-processing is used and this reduced allocation impacts more strongly quality than spatial correction improves it. This can be explained by the performance vs. bitrate behavior of EVS. In addition, the spatial correction may amplify coding noise or artifacts when applied on a too degraded ambisonic signal. This indicates that ambisonic components need to be coded at a sufficient quality for post-processing to be efficient.

VI. Conclusion

This article presented a coding method with post-processing to enhance multi-mono coding by means of spatial image correction. The sound-field power map of the original and coded signals are computed to correct the spatialization of coded signals. A CCR-like subjective test was conducted for two core codecs (EVS and Opus) at two overall bitrates. Results showed an improvement of the overall audio quality over naive multi-mono coding. A further ANOVA analysis demonstrated that performance improvement does not depend on the core codec. This suggests that the proposed method may be used with other core codecs. It also appeared that the proposed coding method requires a minimum bitrate to be allocated to multi-mono coding to enable significant spatial improvement by post-processing; at a given overall bitrate, the bitrate taken for spatial correction meta-data may penalize the performance improvement expected from post-processing.

Acknowledgments

The authors thank Lætitia Gros for her help and expertise on subjective methodologies and ANOVA analysis.

References

[1] M. A. Gerzon, "Periphony: With-height sound reproduction," *AES Journal*, 1973.
[2] J. Daniel, "Représentation de champs acoustiques, application à la transmission et à la reproduction de scènes sonores complexes dans un contexte multimédia," Ph.D. dissertation, Université Paris 6, 2000.
[3] B. Rafaely, *Fundamentals of spherical array processing*. Springer, 2019.
[4] P. Mahé, S. Ragot, and S. Marchand, "First-order ambisonic coding with quaternion-based interpolation of PCA rotation matrices," in *EAA Spatial Audio Signal Processing Symposium*, Sep. 2019.
[5] J.-M. Valin, G. Maxwell, T. B. Terriberry, and K. Vos, "High-Quality, Low-Delay Music Coding in the Opus Codec," *CoRR*, 2016.
[6] J. Skoglund, "Ambisonics in an Ogg Opus Container," IETF RFC 8486, 2018.
[7] 3GPP TS 26.918, "Virtual Reality (VR) media services over 3GPP," 2018.
[8] T. Rudzki, I. Gomez-Lanzaco, P. Hening, J. Skoglund, T. McKenzie, J. Stubbs, D. Murphy, and G. Kearney, "Perceptual Evaluation of Bitrate Compressed Ambisonic Scenes in Loudspeaker Based Reproduction," in *AES Conference: International Conference on Immersive and Interactive Audio*, 2019.
[9] V. Pulkki, A. Politis, M.-V. Laitinen, J. Vilkamo, and J. Ahonen, "First-order directional audio coding (DirAC)," in *Parametric Time-Frequency Domain Spatial Audio*. John Wiley & Sons, 2018, ch. 5.
[10] A. Politis, S. Tervo, and V. Pulkki, "COMPASS: Coding and Multidirectional Parameterization of Ambisonic Sound Scenes," in *Proc. ICASSP*, 2018.
[11] D. P. Jarrett, E. A. Habets, and P. A. Naylor, "3D source localization in the spherical harmonic domain using a pseudointensity vector," in *Proc. EUSIPCO*, 2010.
[12] L. McCormack, A. Politis, and V. Pulkki, "Sharpening of Angular Spectra Based on a Directional Re-assignment Approach for Ambisonic Sound-field Visualisation," in *Proc. ICASSP*, 2019.
[13] B. Rafaely, "Beamforming with Noise Minimization," in *Fundamentals of Spherical Array Processing*. Springer, 2019.
[14] M. Kronlachner, "Spatial transformations for the alteration of ambisonic recordings," *Master's thesis, Graz University of Technology*, 2014.
[15] S. Delikaris-Manias and V. Pulkki, "Cross pattern coherence algorithm for spatial filtering applications utilizing microphone arrays," *IEEE Transactions on Audio, Speech, and Language Processing*, 2013.
[16] L. McCormack and A. Politis, "SPARTA & COMPASS: Real-time implementations of linear and parametric spatial audio reproduction and processing methods," in *AES International Conference on Immersive and Interactive Audio*, 2019.
[17] S. Bruhn and al., "Standardization of the new 3GPP EVS codec," in *Proc. ICASSP*, 2015.
[18] P. Mahé, S. Ragot, and S. Marchand, "First-Order Ambisonic Coding with PCA Matrixing and Quaternion-Based Interpolation," in *Proc. DAFx*, Sep. 2019.
[19] ITU-R BS.1387–1, "Method for Objective Measurement of Perceived Audio Quality," Nov. 2001.
[20] ITU-T P.800 Rec., "Methods for subjective determination of transmission quality," Aug. 1996.
[21] ITU-R Rec. BS.1534–3, "Method for the subjective assessment of intermediate quality level of coding systems," Oct. 2015.
[22] EBU - TECH 3253, "Sound Quality Assessment Material recordings for subjective tests," 2008.