AN EFFICIENT PARAMETRIC CODEC FOR LOW-BITRATE FIRST-ORDER AMBISONICS

Wei-Ting Lai, Amy Bastine, Lachlan Birnie, Thushara D. Abhayapala, Prasanga N. Samarasinghe

Audio & Acoustic Signal Processing Group, The Australian National University, Canberra, Australia

ABSTRACT

Driven by the rapid growth of immersive teleconferencing and generative spatial audio, efficient low-bitrate coding of first-order Ambisonics (FOA) has become increasingly important. In this work, we develop a lightweight parametric FOA codec that retains the standard Directional Audio Coding analysis and synthesis while redesigning the spatial metadata quantization scheme. Rather than quantizing direction-of-arrival (DOA) and diffuseness independently, we combine them into a 3-D directivity vector and jointly quantize these vectors across frequency bands via residual vector quantization (RVQ). The RVQ codebooks are optimized within minutes via stage-wise k-means without backpropagation, yielding a constant-bitrate (CBR) representation that decouples metadata rate from the number of frequency bands. Evaluations show that our approach outperforms low-bitrate perceptual codecs in FOA reconstruction, remains competitive with neural codecs on the downstream Sound Event Localization and Detection task, and maintains robust performance when paired with different external monaural codecs. Given its lightweight training, strong performance, and CBR design, we consider the proposed method to be a favorable and reproducible baseline for future FOA codec research.

***Index Terms*—** first-order Ambisonics (FOA), Directional Audio Coding (DirAC), spatial audio codec

1. INTRODUCTION

Recently, low-bitrate audio codecs/tokenizers have become increasingly important [1–3], driven by growing demands in bandwidth-constrained communication, such as satellite communication [4], and generative audio modeling [5, 6]. Audio codecs aim to compress audio signals into compact discrete representations. Existing lossy codecs can be broadly categorized into perceptual and neural codecs. Perceptual codecs rely on psychoacoustic models to discard perceptually less important information [7, 8], whereas neural codecs learn compact representations directly from large-scale audio data [1–3]. In recent years, neural codecs have increasingly replaced conventional perceptual codecs, offering better perceptual quality at low bitrates through more efficient learned representations.

Meanwhile, immersive teleconferencing [9] and spatial audio generation [10–13] have increased the need for multichannel and spatial audio codecs [14–17]. First-order Ambisonics (FOA) provides an efficient spatial audio representation through a four-channel format that supports rendering to different loudspeaker and headphone configurations [18–20]. While neural codecs for monaural and stereo audio are well-established, such as EnCodec [2] and DAC [3], neural FOA coding remains relatively underexplored. However, independently applying monaural neural codecs to each FOA channel can degrade spatial fidelity [16]. FOA-VQGAN [16] is one of the few recent methods that directly learns discrete tokens from FOA waveforms, whereas existing FOA codecs, such as Opus [8, 9], MPEG-H [20], and Directional Audio Coding (DirAC) [21–26] use perceptual approaches [27].

For FOA coding at total bitrates below 10 kbps, the field remains relatively underexplored, with limited state-of-the-art methods to serve as baselines. Opus is a possible baseline [8,9]; however, its fixed spatial preprocessing followed by mono or stereo coding can limit and substantially degrade spatial fidelity at low bitrates [16,26], making it a non-ideal choice. Similarly, MPEG-H provides strong spatial reconstruction quality, but typically operates at bitrates above the low-bitrate range considered here, making it more suitable as a high-quality reference than a directly comparable baseline.

DirAC [21–26] provides a parametric alternative that decomposes FOA into audio signals and spatial parameters. The audio signals are coded by a signal codec, for example by coding the omnidirectional channel with an external monaural codec. The spatial parameters, including direction-of-arrival (DOA) and diffuseness, are separately transmitted by a metadata codec. Existing DirAC metadata codecs rely on perceptual rules to determine frequency grouping, parameter quantization resolution, and temporal update rates [23–26, 28]. To improve coding efficiency, recent implementations use variable-bitrate (VBR) metadata coding, where the transmitted bitrate varies according to the spatial characteristics of the audio content [24, 26]. However, such heuristic designs complicate bit allocation and reduce reproducibility as a baseline. This motivates a more efficient and reproducible FOA codec with constant-bitrate (CBR) DirAC spatial metadata.

In this work, we propose a lightweight parametric FOA codec that introduces an efficient CBR spatial metadata quantizer while retaining the rest of the standard DirAC pipeline. We combine DOA and diffuseness parameters into a 3-D directivity vector and jointly quantize them across bands using a residual vector quantizer (RVQ) [1, 29], where the codebooks are fitted offline via k-means [30] within minutes. Results show that the proposed method maintains robust performance for both FOA reconstruction and downstream Sound Event Localization and Detection (SELD) [31, 32] when paired with different monaural neural signal codecs. We consider the proposed method to be a favorable baseline for future FOA codec comparisons, as it supports efficient CBR configurations across different target rates.

2. PROBLEM FORMULATION

Let the FOA signal in ACN/SN3D format at time frame $t \in \{1, \dots, T\}$ and frequency bin $k \in \{1, \dots, K\}$ be $\mathbf{a}_{t,k} = [W_{t,k}, Y_{t,k}, Z_{t,k}, X_{t,k}]^{\mathsf{T}} \in \mathbb{C}^4$.

DirAC provides a parametric representation of FOA through an analysis stage and a synthesis stage [21, 26]. The analysis stage disentangles the FOA soundfield $\mathbf{a}_{t,k}$ into an omnidirectional component $W_{t,k}$ and bandwise spatial metadata over B equivalent rectangular bandwidth (ERB) frequency bands [33]. For each band $b \in \{1, \dots, B\}$ at frame t, the bandwise intensity vector $\mathbf{I}_{t,b}$ and sound-

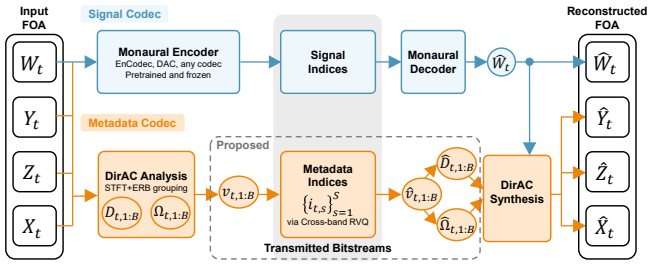


Fig. 1. Overview of the proposed parametric FOA codec.

field energy $E_{t,b}$ are computed as

$$\mathbf{I}_{t,b} = \sum_{k\in\mathcal{K}_b} \mathrm{Re}\{W_{t,k}\mathbf{U}_{t,k}^*\}, \tag{1}$$

$$E_{t,b} = \tfrac{1}{2}\sum_{k\in\mathcal{K}_b} \left(|W_{t,k}|^2 + \|\mathbf{U}_{t,k}\|_2^2\right), \tag{2}$$

where $\mathcal{K}_b$ denotes the STFT bins in band b and $\mathbf{U}_{t,k} = [Y_{t,k}, Z_{t,k}, X_{t,k}]^\mathsf{T}$ denotes the first-order spherical harmonic (SH) components.

From $\mathbf{I}_{t,b}$ and $E_{t,b}$, the DOA $\mathbf{\Omega}_{t,b} \in \mathbb{S}^2$ and diffuseness $D_{t,b} \in [0,1]$ are derived as

$$\mathbf{\Omega}_{t,b} = -\frac{\mathbf{I}_{t,b}}{\|\mathbf{I}_{t,b}\|_2}, \qquad D_{t,b} = 1 - \frac{\|\mathbf{I}_{t,b}\|_2}{E_{t,b}}. \tag{3}$$

Collecting these continuous-valued spatial parameters across all B bands yields the spatial metadata for frame t, denoted by $\mathcal{M}_t = \{(\mathbf{\Omega}_{t,b}, D_{t,b})\}_{b=1}^B$.

The omnidirectional channel $W_{t,k}$ and spatial metadata $\mathcal{M}_t$ are encoded independently. An external monaural signal codec encodes and decodes $W_{t,k}$ to yield the reconstructed omnidirectional signal $\hat{W}_{t,k}$, whereas a spatial metadata quantizer $Q(\cdot)$ maps the continuous metadata $\mathcal{M}_t$ to a set of discrete quantization indices $\mathbf{z}_t$, from which the reconstructed metadata $\hat{\mathcal{M}}_t$ is obtained via the dequantization mapping $Q^{-1}(\cdot)$ as

$$\mathbf{z}_t = Q(\mathcal{M}_t), \qquad \hat{\mathcal{M}}_t = Q^{-1}(\mathbf{z}_t) = \{(\hat{\mathbf{\Omega}}_{t,b}, \hat{D}_{t,b})\}_{b=1}^B. \tag{4}$$

The DirAC synthesis then uses the dequantized spatial parameters $\hat{\mathcal{M}}_t$ together with the decoded omnidirectional signal $\hat{W}_{t,k}$ to synthesize the first-order SH channels. The complete reconstructed FOA signal is formed as

$$\hat{\mathbf{a}}_{t,k} = \begin{bmatrix} \hat{W}_{t,k} \\ \hat{\mathbf{U}}_{t,k} \end{bmatrix}, \tag{5}$$

where the omnidirectional component $\hat{W}_{t,k}$ is directly preserved, and the spatial channels $\hat{\mathbf{U}}_{t,k} = [\hat{Y}_{t,k}, \hat{Z}_{t,k}, \hat{X}_{t,k}]^\mathsf{T}$ are synthesized by superimposing directional component $\hat{\mathbf{U}}_{t,k}^{\mathrm{dir}}$ and diffuse component $\hat{\mathbf{U}}_{t,k}^{\mathrm{diff}}$ as

$$\hat{\mathbf{U}}_{t,k} = \hat{\mathbf{U}}_{t,k}^{\mathrm{dir}} + \hat{\mathbf{U}}_{t,k}^{\mathrm{diff}}. \tag{6}$$

The directional branch synthesizes spatial components corresponding to a plane wave arriving from $\hat{\mathbf{\Omega}}_{t,b}$ as

$$\hat{\mathbf{U}}_{t,k}^{\mathrm{dir}} = \sqrt{1-\hat{D}_{t,b}}\,\hat{W}_{t,k}\,\mathbf{y}_1(\hat{\mathbf{\Omega}}_{t,b}), \tag{7}$$

where $\mathbf{y}_1(\hat{\mathbf{\Omega}}_{t,b}) = [Y_1^{-1}(\hat{\mathbf{\Omega}}_{t,b}), Y_1^0(\hat{\mathbf{\Omega}}_{t,b}), Y_1^1(\hat{\mathbf{\Omega}}_{t,b})]^\mathsf{T} \in \mathbb{R}^3$ represents the first-order real SH basis vector.

The diffuse branch synthesizes an isotropic diffuse field from mutually decorrelated versions of $\hat{W}_{t,k}$ as

$$\hat{\mathbf{U}}_{t,k}^{\mathrm{diff}} = \sqrt{\hat{D}_{t,b}}\,\mathbf{d}_{t,k}(\hat{W}_{t,k}), \tag{8}$$

where $\mathbf{d}_{t,k}(\cdot) \in \mathbb{C}^3$ is the DirAC diffuse renderer that generates mutually decorrelated signals to simulate ambient reverberation [21].

Conventional DirAC metadata codecs quantize DOA and diffuseness separately using perceptually designed parameter resolutions [23, 26]. Specifically, $D_{t,b}$ is mapped to a non-uniform codebook $\mathcal{C}_D$, while $\mathbf{\Omega}_{t,b}$ is quantized onto a spherical grid $\mathcal{G}_\Omega$ whose angular resolution is adapted according to diffuseness. The number of parameter bands B is also reduced by merging adjacent frequency bands. In low-bitrate transmission scenarios, however, these heuristic psychoacoustic designs of $\mathcal{C}_D$, $\mathcal{G}_\Omega$, and B lead to VBR transmission and inefficient parameter allocation. The goal of this paper is therefore to develop an efficient CBR DirAC metadata quantizer $Q(\cdot)$ that improves the reconstruction fidelity of $\hat{\mathbf{a}}_{t,k}$ under strict low-bitrate constraints.

3. PROPOSED METHOD

We propose a lightweight FOA codec that retains the standard DirAC analysis and synthesis [26], redesigning only the spatial metadata quantizer $Q(\cdot)$. Specifically, spatial parameters are represented as 3-D directivity vectors and jointly quantized across bands via RVQ [1, 29], as illustrated in Fig. 1.

3.1. 3-D Directivity Vector

Unlike conventional DirAC metadata coding, which quantizes DOA and diffuseness separately, we combine them into a 3-D directivity vector based on the directional scaling in (7). Specifically, we define

$$\mathbf{v}_{t,b} = \sqrt{1-D_{t,b}}\,\mathbf{\Omega}_{t,b}. \tag{9}$$

The direction of $\mathbf{v}_{t,b}$ represents the DOA, while its norm corresponds directly to the directional gain used in DirAC synthesis. Thus, as diffuseness increases, $\|\mathbf{v}_{t,b}\|_2$ decreases and directional errors are naturally downweighted during vector quantization, without requiring an explicit diffuseness-dependent DOA resolution.

At the decoder, the DirAC parameters are recovered from the quantized vector as

$$\hat{D}_{t,b} = 1 - \|\hat{\mathbf{v}}_{t,b}\|_2^2, \qquad \hat{\mathbf{\Omega}}_{t,b} = \frac{\hat{\mathbf{v}}_{t,b}}{\|\hat{\mathbf{v}}_{t,b}\|_2}. \tag{10}$$

When $\|\hat{\mathbf{v}}_{t,b}\|_2 = 0$, the DOA is irrelevant to synthesis and can be assigned an arbitrary unit direction, e.g., $\hat{\mathbf{\Omega}}_{t,b} = [1,0,0]^\mathsf{T}$.

3.2. Cross-Band Residual Vector Quantization

Encoding each band vector $\mathbf{v}_{t,b}$ independently requires separate bit allocation across bands. Instead, we stack all B band vectors into a multiband directivity matrix $\mathbf{V}_t = [\mathbf{v}_{t,1}, \ldots, \mathbf{v}_{t,B}]^\mathsf{T} \in \mathbb{R}^{B\times 3}$ and quantize $\mathbf{V}_t$ jointly using cross-band RVQ [1, 29].

The quantizer consists of S stages. For stage $s \in \{1, \ldots, S\}$, let $\mathcal{C}_s = \{\mathbf{C}_{s,m}\}_{m=0}^{M-1} \subset \mathbb{R}^{B\times 3}$ denote a pre-fitted codebook of M cross-band correction patterns, with $\mathbf{C}_{s,0} = \mathbf{0}$ reserved as an idle codeword.

Starting from $\hat{\mathbf{V}}_t^{(0)} = \mathbf{0}$, each RVQ stage tests every codeword by adding it to the reconstruction from the previous stage. We denote the resulting matrix at stage s as $\tilde{\mathbf{V}}_{t,m}^{(s)} = \hat{\mathbf{V}}_t^{(s-1)} + \mathbf{C}_{s,m}$.

To ensure that each reconstructed band remains within the unit sphere, we define the row-wise projection as

$$\left[\Pi(\tilde{\mathbf{V}}_{t,m}^{(s)})\right]_{b,:} = \frac{\left[\tilde{\mathbf{V}}_{t,m}^{(s)})\right]_{b,:}}{\max\left(1, \|\left[\tilde{\mathbf{V}}_{t,m}^{(s)})\right]_{b,:}\|_2\right)}. \tag{11}$$

Let $e_t = \sum_{b=1}^B E_{t,b}$ denote the total energy of frame t. For

Table 1. FOA reconstruction performance on SpatialVCTK, STARSS23, and MEIR. Lower is better for all metrics.

Method	W-ch	$\mathbf{R}_{\mathrm{sp}}$	S	$\mathbf{R}_{\mathrm{tot}}$	SpatialVCTK					STARSS23				MEIR			
					STFT	Mel	Ang	Diff	WER	STFT	Mel	Ang	Diff	STFT	Mel	Ang	Diff
MPEG-H [20]	–	–	–	≈48	1.31	0.82	**2.11°**	0.02	0.20	1.45	**0.81**	**10.37°**	**0.05**	2.39	1.47	27.53°	**0.16**
FOA-VQGAN [16]	–	–	–	0.9	1.62	1.39	3.96°	–	0.67	–	–	–	–	1.82	1.43	25.83°	–
Opus [8]	–	–	–	≈24	2.42	1.70	10.04°	0.11	0.23	2.50	1.73	28.46°	0.12	4.16	2.48	56.53°	0.27
	–	–	–	≈32	2.22	1.44	5.31°	0.05	0.19	2.35	1.52	19.18°	0.08	4.02	2.27	38.67°	0.24
DirAC [26]	EnCodec [2]	0.75	–	6.75	2.06	2.21	29.21°	**0.00**	0.15	2.04	1.67	47.42°	0.30	2.24	1.45	31.22°	0.24
	EnCodec	1.5	–	7.5	1.18	0.94	3.56°	**0.00**	0.15	1.92	1.60	39.19°	0.32	1.97	1.31	18.24°	0.25
	DAC [3]	0.75	–	6.75	1.91	2.02	29.21°	**0.00**	**0.14**	1.66	1.22	47.48°	0.30	2.17	1.42	31.16°	0.24
	DAC	1.5	–	7.50	**0.91**	**0.60**	3.56°	**0.00**	**0.14**	1.53	1.13	39.19°	0.32	1.87	1.28	**18.15°**	0.25
Proposed	EnCodec	0.75	5	6.75	1.23	0.97	3.51°	0.02	0.15	1.65	1.44	37.98°	0.20	1.78	1.24	20.84°	0.22
	EnCodec	1.5	10	7.5	1.22	0.95	2.72°	0.02	0.15	1.65	1.43	34.89°	0.20	1.78	1.24	19.61°	0.22
	EnCodec	3.0	20	9	1.21	0.94	2.18°	0.02	0.15	1.64	1.43	33.06°	0.20	1.78	1.24	18.84°	0.23
	DAC	0.75	5	6.75	0.99	0.67	3.51°	0.02	**0.14**	1.26	0.96	37.80°	0.20	1.69	1.21	20.76°	0.22
	DAC	1.5	10	7.5	0.97	0.65	2.72°	0.02	**0.14**	1.26	0.96	34.61°	0.20	1.68	1.21	19.51°	0.22
	DAC	3.0	20	9.0	0.96	0.63	2.17°	0.02	**0.14**	**1.25**	0.95	32.73°	0.20	**1.68**	**1.20**	18.75°	0.23

W-ch denotes the external monaural signal codec used to encode and decode the omnidirectional W channel. R_{sp} and R_{tot} denote the nominal spatial-metadata and total bitrates in kbps, respectively. R_{tot} values for Opus and MPEG-H are approximate due to their VBR operation. Ang and Diff denote angular and diffuseness errors, respectively, and WER is reported in percent. The best result per metric is shown in **bold**, and the best result per metric at $R_{\mathrm{sp}} = 0.75$ kbps is underlined.

$e_t > 0$, the distortion of each estimated matrix is measured as

$$d_t\Big(\mathbf{V}_t, \Pi(\tilde{\mathbf{V}}_{t,m}^{(s)})\Big) = \frac{\sum_{b=1}^{B} E_{t,b} \left\| [\mathbf{V}_t]_{b,:} - [\Pi(\tilde{\mathbf{V}}_{t,m}^{(s)})]_{b,:} \right\|_2^2}{e_t}, \tag{12}$$

where frames with $e_t = 0$ are ignored.

Each RVQ stage greedily selects the codeword with the minimum distortion as

$$i_{t,s} = \arg\min_m d_t\left(\mathbf{V}_t, \Pi(\tilde{\mathbf{V}}_{t,m}^{(s)})\right), \tag{13}$$

and the reconstruction is updated as

$$\hat{\mathbf{V}}_t^{(s)} = \Pi(\tilde{\mathbf{V}}_{t,i_{t,s}}^{(s)}). \tag{14}$$

After S stages, the final quantized output is $\hat{\mathbf{V}}_t = \hat{\mathbf{V}}_t^{(S)}$, whose b^{th} row is $\hat{\mathbf{v}}_{t,b}$.

3.3. Codebook Fitting and CBR Operation

The RVQ codebooks are fitted offline using stage-wise k-means [30]. For each stage s, the codebook $\mathcal{C}_s$ is fitted based on the reconstruction from the previous stage $\hat{\mathbf{V}}_t^{(s-1)}$ to minimize the total energy-weighted distortion $\sum_t e_t\, d_t\Big(\mathbf{V}_t, \hat{\mathbf{V}}_t^{(s)}\Big)$.

Once fitted, the codebooks remain fixed and directly compress the spatial metadata of arbitrary input FOA signals without further optimization. With a frame update period of T_{u}, the quantizer outputs S discrete indices $\{i_{t,s}\}_{s=1}^{S}$ per frame t, yielding a constant spatial metadata bitrate $R_{\mathrm{sp}} = (S \log_2 M)/T_{\mathrm{u}}$. Because R_{sp} is completely decoupled from the number of parameter bands B, the system can employ a fine frequency resolution without increasing the metadata bitrate budget.

4. EXPERIMENTAL VALIDATION

4.1. Experimental Setup

This section describes the experimental setup for evaluating the proposed method on FOA reconstruction and downstream SELD, including codebook fitting, datasets, baselines, and evaluation metrics.

For offline codebook fitting of the proposed quantizer, we generate a simulated FOA dataset using the image-source method in Pyroomacoustics [34, 35]. Dry source signals are drawn from LibriSpeech (speech) [36], ESC-50 (general sounds) [37], and ASE1K (transients) [38]. A total of 120,000 6-second ACN/SN3D FOA scenes are synthesized at 24 kHz (≈ 200 h), with reverberation times $T_{60} \in [0, 1.2]$ s, random source gains in $[-6, 6]$ dB, and additive white Gaussian noise at an SNR of 30 dB. Most scenes contain 1–3 static sources, while 12,000 scenes contain 10 concurrent sources to simulate dense cocktail-party scenarios.

To extract DirAC metadata, a 1024-point STFT with a Hann window and a 120-sample hop size is applied, grouping frequency bins into $B = 36$ parameter bands across 0–12 kHz based on ERB [33]. The extracted metadata are temporally averaged to a fixed update period of $T_{\mathrm{u}} = 40$ ms. To reduce computational cost, we sample 8 out of the 150 frames from each clip. After discarding zero-energy silent frames, a total of $T = 932{,}102$ frames are retained for codebook fitting. Unless otherwise stated, we use $S = 5$ RVQ stages with $M = 64$ codewords per stage as the default configuration, resulting in $R_{\mathrm{sp}} = 0.75$ kbps. At each RVQ stage, k-means is run for 18 iterations from three deterministic initializations, and the solution with the lowest distortion is retained. The complete codebook fitting process takes approximately 7 min on an NVIDIA RTX 3070 GPU.

Reconstruction performance is evaluated on SpatialVCTK, created by spatializing VCTK speech recordings under free-field conditions [39], STARSS23 [32], which contains real-world FOA recordings, and FOA-MEIR [40], constructed using real measured room impulse responses and spatial background noise.

We compare against four baseline codecs. MPEG-H 3D Audio [20] is included as a high-bitrate reference for FOA coding. For conventional DirAC, we reproduce the coding scheme in [26]. Since the original work does not fully specify the bitrate-control details, we implement a CBR variant following its key principle that fewer DOA bits are allocated as diffuseness increases. FOA-VQGAN is included as a representative end-to-end neural FOA codec. As its implementation is not publicly available, we report the published results [16] under matched evaluation conditions where applicable. We additionally evaluate Opus using four-channel mapping family 1 [8], constrained VBR, and a 10-ms frame duration.

To examine the framework's robustness when paired with different monaural signal codecs, the W channel is encoded using pretrained EnCodec [2] or DAC [3] at 6 kbps for both DirAC and the proposed method. The monaural codecs are kept fixed, and their decoded $\hat{W}$ outputs are directly used for DirAC synthesis. We additionally evaluate the impact of varying the number of RVQ stages S and the number of parameter bands B, with smaller B obtained by pooling adjacent ERB bands, to analyze the rate–distortion trade-off and verify whether preserving finer frequency resolution under a

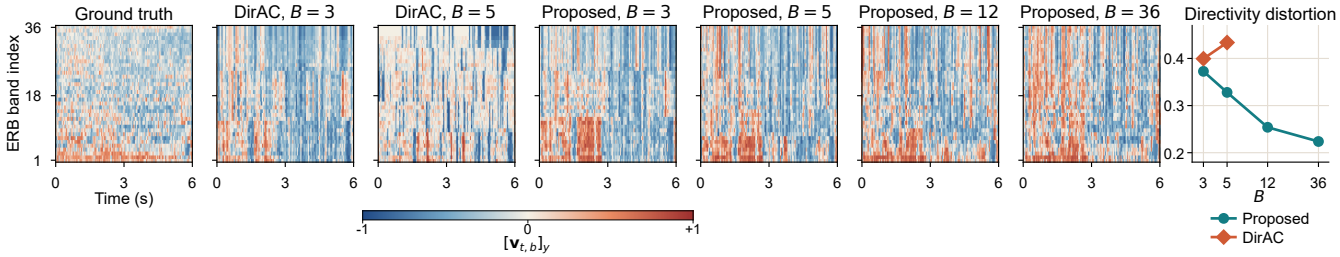


Fig. 2. Reconstruction of the 3-D directivity vector for a STARSS23 clip ($R_{\text{sp}} = 750$ bps for both methods). For visualization, we show the y-axis component $[\mathbf{v}_{t,b}]_y$. The rightmost curve reports the energy-weighted distortion $\mathcal{D} = \sum_t e_t d_t / \sum_t e_t$.

decoupled constant bitrate benefits spatial reconstruction.

FOA reconstruction quality is evaluated using STFT and Mel-spectral losses implemented with Auraloss [41]. For SpatialVCTK, we additionally evaluate Word Error Rate (WER) by transcribing the reconstructed $\hat{W}$ channel using the pretrained multilingual Whisper tiny model [42]. Spatial fidelity is measured using angular and diffuseness errors computed between the reference and reconstructed signals. The angular error compares the directions of the active-intensity vectors using a metric adapted from [16], while the diffuseness error is defined as the mean absolute error between the corresponding $D_{t,b}$ values. Audio samples are available online[1].

To evaluate whether the compressed representations retain task-relevant spatial information beyond reconstruction fidelity, we further assess the proposed method on downstream SELD, following the DCASE 2023 Challenge protocol on STARSS23 [32]. We freeze both the RVQ codebooks of the proposed spatial metadata codec and the pretrained monaural signal codec, concatenate their latent representations, and train a lightweight probe, comprising three temporal convolutional layers and two fully connected layers, to predict Multi-ACCDOA [43] at a 100-ms label resolution. We report the F-score and localization error (LE) with a 20° spatial threshold.

4.2. FOA Reconstruction Results

Table 1 shows the FOA reconstruction performance on the three evaluation datasets. With only 750 bps of spatial metadata, the proposed method remains competitive with the 48 kbps high-bitrate MPEG-H reference. In contrast, while conventional DirAC performs well at 1.5 kbps, its performance degrades substantially at 750 bps, highlighting the difficulty of designing effective bit allocation under low bitrate constraints. By separately coding the audio signal and spatial metadata, both the proposed method and DirAC remain robust when paired with different monaural signal codecs. Moreover, because the W channel is encoded independently by the selected signal codec, its content preservation, as reflected by WER, is determined by the signal codec only. This allows these two methods to directly benefit from pretrained monaural codecs, resulting in more consistent WER performance than the end-to-end neural FOA codec. Spatial errors are generally higher on STARSS23 and MEIR, where multi-source and reverberant conditions make spatial parameter estimation more challenging. This highlights the need for more reliable spatial metrics for evaluating reconstructed FOA signals in complex acoustic scenes. Although we align the evaluation protocol and metrics as closely as possible, differences in experimental settings may remain, and the comparison should therefore not be interpreted as an absolute ranking between FOA-VQGAN and the other methods.

4.3. Effect of Parameter-Band Resolution

We select a clip from STARSS23 to illustrate the reconstruction of the 3-D directivity vector at different parameter-band resolutions B, as shown in Fig. 2. Under the same metadata bitrate of 0.75 kbps, the two methods exhibit contrasting trends as B increases. For the proposed method, directivity distortion progressively decreases from $B = 3$ to 36, suggesting that finer parameter-band resolution better preserves frequency-dependent spatial variations. This benefit comes without a bitrate penalty because the proposed representation decouples the spatial metadata rate from B and jointly quantizes the cross-band 3-D directivity vectors. In contrast, increasing B from 3 to 5 in conventional DirAC increases distortion, since the fixed bitrate must be distributed across more bands, forcing a coarser DOA grid $\mathcal{G}_\Omega$. We did not further implement $B > 5$ at 0.75 kbps, as at $B = 5$, the average DOA allocation is already reduced to 3 bits, corresponding to only eight directions. At the matched $B = 3$ and 5, the proposed method still yields lower distortion than DirAC, suggesting more efficient use of the limited codewords through joint cross-band quantization.

Table 2. STARSS23 SELD results. Results from the uncompressed reference are underlined. Best results among codecs are in bold.

Method	W-ch	F-score ↑	LE ↓
DCASE baseline [44]	–	<u>29.9</u>	<u>22°</u>
FOA-VQGAN [16]	–	11.1	37°
Proposed	EnCodec	**14.2**	**30°**
	DAC	13.8	34°

4.4. Downstream SELD

Table 2 shows the downstream SELD performance. When paired with either EnCodec or DAC, the k-means-quantized spatial metadata achieves performance comparable to the end-to-end neural FOA codec, with EnCodec giving slightly better results. While a gap remains when compared to the DCASE baseline trained on uncompressed continuous features, these results provide initial evidence that the proposed FOA codec can be flexibly paired with different monaural signal codecs while preserving key spatial cues for downstream tasks.

5. CONCLUSION

In this paper, we proposed a constant-bitrate spatial metadata quantizer for a lightweight parametric FOA codec. DOA and diffuseness are combined into 3-D directivity vectors and jointly quantized across frequency bands using RVQ with codebooks fitted by stage-wise k-means. With only 750 bps of spatial metadata, the proposed method achieves competitive FOA reconstruction performance compared with conventional methods operating at higher bitrates, alongside downstream SELD performance comparable to FOA-VQGAN. Performance remains robust with either EnCodec or DAC as the monaural codec. Future work includes exploring spatial audio generation as a downstream task and extending the framework to higher-order Ambisonics.

[1] https://weiting-lai.github.io/MetaFOA

6. REFERENCES

[1] N. Zeghidour, A. Luebs, A. Omran, J. Skoglund, and M. Tagliasacchi, "SoundStream: An end-to-end neural audio codec," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 495–507, 2021.

[2] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, "High fidelity neural audio compression," *Trans. Mach. Learn. Res.*, 2023.

[3] R. Kumar, P. Seetharaman, A. Luebs, I. Kumar, and K. Kumar, "High-fidelity audio compression with improved RVQGAN," *Adv. Neural Inf. Process. Syst.*, vol. 36, pp. 27 980–27 993, 2023.

[4] A. V. McCree and T. P. Barnwell, "A mixed excitation LPC vocoder model for low bit rate speech coding," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 4, pp. 242–250, 1995.

[5] Z. Borsos, R. Marinier, D. Vincent, E. Kharitonov, O. Pietquin, M. Sharifi, D. Roblek, O. Teboul, D. Grangier, M. Tagliasacchi, and N. Zeghidour, "AudioLM: A language modeling approach to audio generation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 2523–2533, 2023.

[6] J. Copet, F. Kreuk, I. Gat, T. Remez, D. Kant, G. Synnaeve, Y. Adi, and A. Défossez, "Simple and controllable music generation," *Adv. Neural Inf. Process. Syst.*, vol. 36, pp. 47 704–47 720, 2023.

[7] K. Brandenburg, "MP3 and AAC explained," in *Proc. AES 17th Int. Conf. High-Quality Audio Coding*, 1999, pp. 1–12.

[8] J.-M. Valin, K. Vos, and T. Terriberry, "Definition of the Opus Audio Codec," IETF RFC 6716, Sep. 2012.

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

[10] X. Jiang, C. Han, Y. A. Li, and N. Mesgarani, "Exploring self-supervised contrastive learning of spatial sound event representation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2024, pp. 1281–1285.

[11] B. Devnani, S. Seto, Z. Aldeneh, A. Toso, E. Menyaylenko, B.-J. Theobald, J. Sheaffer, and M. Sarabia, "Learning spatially-aware language and audio embeddings," *Adv. Neural Inf. Process. Syst.*, vol. 37, pp. 33 505–33 537, 2024.

[12] A. Dementyev, W. Zulfikar, S. Hersek, P. Getreuer, A. Kumar, and V. Kumar, "PhaseCoder: Microphone geometry-agnostic spatial audio understanding for multimodal LLMs," *arXiv preprint arXiv:2601.21124*, 2026.

[13] Z. Zhu, Y. Chen, Y. Shao, W. Guo, C. Pan, Y. Zhang, Y. Wang, W. Liu, H. Zhang, C. Zeng, W. Cheng, Y. Liu, R. Yang, S. Yves, L. Bo, and Z. Zhao, "Spatial-Omni: Spatial audio understanding integration in multimodal LLMs via FOA encoding," *arXiv preprint arXiv:2606.10738*, 2026.

[14] Z. Xu, Y. Xu, V. Kothapally, H. Wang, M. Yang, and D. Yu, "Spatialcodec: Neural spatial speech coding," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2024, pp. 1131–1135.

[15] A. Ratnarajah, S.-X. Zhang, and D. Yu, "Banc: Towards efficient binaural audio neural codec for overlapping speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2025, pp. 1–5.

[16] P. Sudarsanam, S. Braun, and H. Gamper, "FOA tokenizer: Low-bitrate neural codec for first order ambisonics with spatial consistency loss," *arXiv preprint arXiv:2510.22241*, 2025.

[17] F. Grötschla, A. Sen, A. Lombardi, G. Cámbara, and A. Schwarz, "Vcnac: A variable-channel neural audio codec for mono, stereo, and surround sound," *arXiv preprint arXiv:2601.14960*, 2026.

[18] M. A. Gerzon, "Periphony: With-height sound reproduction," *J. Audio Eng. Soc.*, vol. 21, no. 1, pp. 2–10, 1973.

[19] D. B. Ward and T. D. Abhayapala, "Reproduction of a plane-wave sound field using an array of loudspeakers," *IEEE Trans. Speech, Audio Process.*, vol. 9, no. 6, pp. 697–707, 2001.

[20] J. Herre, J. Hilpert, A. Kuntz, and J. Plogsties, "MPEG-H audio—the new standard for universal spatial/3D audio coding," *J. Audio Eng. Soc.*, vol. 62, no. 12, pp. 821–830, 2014.

[21] V. Pulkki, "Spatial sound reproduction with directional audio coding," *J. Audio Eng. Soc.*, vol. 55, no. 6, pp. 503–516, 2007.

[22] J. Vilkamo, T. Lokki, and V. Pulkki, "Directional audio coding: Virtual microphone-based synthesis and subjective evaluation," *J. Audio Eng. Soc.*, vol. 57, no. 9, pp. 709–724, 2009.

[23] T. Hirvonen, J. Ahonen, and V. Pulkki, "Perceptual compression methods for metadata in directional audio coding applied to audiovisual teleconference," in *Proc. 126th Audio Eng. Soc. Conv.*, 2009, Convention Paper 7708.

[24] C. Hold, L. McCormack, A. Politis, and V. Pulkki, "Perceptually-motivated spatial audio codec for higher-order ambisonics compression," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2024, pp. 1121–1125.

[25] C. Hold, V. Pulkki, A. Politis, and L. McCormack, "Compression of higher-order ambisonic signals using directional audio coding," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 32, pp. 651–665, 2024.

[26] G. Fuchs, F. Ghido, D. Weckbecker, and O. Thiergart, "A first-order DirAC-based parametric ambisonic coder for immersive communications," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2025, pp. 12 937–12 941.

[27] D. McGrath, S. Bruhn, H. Purnhagen, M. Eckert, J. Torres, S. Brown, and D. Darcy, "Immersive audio coding for virtual reality using a metadata-assisted extension of the 3GPP EVS codec," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 730–734.

[28] A. Daniel, R. Nicol, and S. McAdams, "Parametric spatial audio coding based on spatial auditory blurring," in *Proc. AES 45th Int. Conf. Appl. Time-Freq. Process. Audio*, 2012, pp. 231–240.

[29] B.-H. Juang and A. H. Gray, Jr., "Multiple stage vector quantization for speech coding," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 7, 1982, pp. 597–600.

[30] S. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inf. Theory*, vol. 28, no. 2, pp. 129–137, 1982.

[31] A. Politis, K. Shimada, P. Sudarsanam, S. Adavanne, D. Krause, Y. Koyama, N. Takahashi, S. Takahashi, Y. Mitsufuji, and T. Virtanen, "STARSS22: A dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events," in *Proc. 7th Workshop Detection Classification Acoust. Scenes Events*, 2022, pp. 125–129.

[32] K. Shimada, A. Politis, P. Sudarsanam, D. A. Krause, K. Uchida, S. Adavanne, A. Hakala, Y. Koyama, N. Takahashi, S. Takahashi, T. Virtanen, and Y. Mitsufuji, "STARSS23: An audio-visual dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events," *Adv. Neural Inf. Process. Syst.*, vol. 36, pp. 72 931–72 957, 2023.

[33] B. R. Glasberg and B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hear. Res.*, vol. 47, no. 1–2, pp. 103–138, 1990.

[34] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A Python package for audio room simulation and array processing algorithms," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 351–355.

[35] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943–950, 1979.

[36] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 5206–5210.

[37] K. J. Piczak, "ESC: Dataset for environmental sound classification," in *Proc. 23rd ACM Int. Conf. Multimedia*, 2015, pp. 1015–1018.

[38] M. Yasuda, B. T. Nguyen, N. Harada, R. Serizel, M. Mishra, M. Delcroix, S. Araki, D. Takeuchi, D. Niizumi, Y. Ohishi, T. Nakatani, T. Kawamura, and N. Ono, "Description and discussion on DCASE 2025 Challenge Task 4: Spatial semantic segmentation of sound scenes," *arXiv preprint arXiv:2506.10676*, 2025.

[39] J. Yamagishi, C. Veaux, and K. MacDonald, "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92)," University of Edinburgh, Centre for Speech Technology Research, 2019.

[40] M. Yasuda, Y. Ohishi, and S. Saito, "Echo-aware adaptation of sound event localization and detection in unknown environments," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2022, pp. 226–230.

[41] C. J. Steinmetz and J. D. Reiss, "auraloss: Audio-focused loss functions in PyTorch," in *Proc. Digit. Music Res. Netw. One-Day Workshop (DMRN+15)*, 2020.

[42] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proc. 40th Int. Conf. Mach. Learn.*, ser. Proc. Mach. Learn. Res., vol. 202, 2023, pp. 28 492–28 518.

[43] K. Shimada, Y. Koyama, S. Takahashi, N. Takahashi, E. Tsunoo, and Y. Mitsufuji, "Multi-ACCDOA: Localizing and detecting overlapping sounds from the same class with auxiliary duplicating permutation invariant training," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2022, pp. 316–320.

[44] S. Adavanne, "DCASE 2023: Sound event localization and detection evaluated in real spatial sound scenes," 2023, Accessed: Sep. 27, 2026. [Online]. Available: https://github.com/sharathadavanne/seld-dcase2023.