Large-Scale Room Impulse Response Dataset Compression With Neural Audio Codecs

Alessandro Ilic Mezza
DEIB
Politecnico di Milano
Milan, Italy
alessandroilic.mezza@polimi.it

Alberto Bernardini
DEIB
Politecnico di Milano
Milan, Italy
alberto.bernardini@polimi.it

Fabio Antonacci
DEIB
Politecnico di Milano
Milan, Italy
fabio.antonacci@polimi.it

***Abstract*—The virtualization of physical acoustic environments, essential for augmented reality and immersive spatial audio applications, typically requires the storage and transmission of a large quantity of room impulse responses (RIRs). Real-world RIRs often comprise tens of thousands of coefficients. As such, working with large databases of room acoustic measurements presents significant challenges in terms of memory and bandwidth requirements. To address this issue, we investigate neural audio codecs as a means to achieve lossy RIR data compression. In particular, by focusing on two publicly available datasets, we show that EnCodec, a recently proposed state-of-the-art neural audio codec with bitrate as low as 1.5 kbps, is able to achieve a compression ratio over two orders of magnitude larger than lossless coding. Objective metrics and a listening test reveal that EnCodec preserves perceptually relevant features of the decoded reverberation better than a traditional dimensionality reduction method based on singular value decomposition, encouraging further research on the topic of neural RIR coding.**

***Index Terms*—auralization, data compression, EnCodec, entropy coding, room acoustics, room impulse response, neural audio codec**

I. Introduction

Room impulse responses (RIRs) fully characterize the input-output relationship between an acoustic source and a receiver, capturing how sound propagates through an enclosed space and interacts with its boundaries. RIRs are central to many applications, including room acoustics analysis [1], room geometry inference [2], audio source separation [3], speech enhancement [4], and auralization [5].

A single RIR describes a single-input single-output (SISO) linear time-invariant system. Therefore, practical applications related to immersive audio and the auralization of navigable virtual spaces often entail a high spatial resolution of measurement points, i.e., source and receiver placements. Applications such as gaming and virtual reality, in turn, rely on the high-fidelity acoustic rendering of a large number of environments to create an immersive experience [6], and distributed wireless acoustic sensor networks [7] have the capacity to gather room acoustics data over a large area of interest.

At the same time, though, RIRs often consist of tens of thousands of coefficients at standard audio sampling rates. This poses considerable challenges when it comes to the management and storage of large databases of RIRs.

Moreover, when spatial audio processing is not performed by a centralized processing node but rather takes place on edge devices [8], RIRs must be transmitted over a telecommunication network, possibly in real-time. In this scenario, data size may also become a burden in terms of network bandwidth and throughput. Ultimately, high memory and network requirements create a strong need for effective lossy compression algorithms capable of reducing the dimensionality of RIRs as much as possible, all while preserving their acoustic qualities.

In this paper, we explore the application of EnCodec for large-scale RIR coding. EnCodec [9] is a recently proposed deep encoder-decoder architecture with latent Residual Vector Quantization (RVQ) that showed state-of-the-art performance in high-quality low-bitrate lossy compression of general audio signals, including speech, music, and environmental sounds, outperforming traditional digital signal processing based audio codecs such as Opus [10] and EVS [11].

While neural audio coding is an active and ever-growing field of research [9], [12]–[23], with some RVQ-based codecs also being used as building blocks for blind RIR estimation [24], to the best of our knowledge, the literature lacks a quantitative study of the impact of neural RIR coding on the perceived reverberation quality in relation to the data compression ratio.

We evaluate EnCodec against a recent RIR dimensionality reduction framework based on well-understood modal analysis principles [25]. To illustrate the methods' performance, we focus on two publicly available datasets of real-world RIRs, i.e., the MIT Acoustical Reverberation Scene Statistics Survey database [26] and HOMULA-RIR [27]. Perceptually-informed objective metrics indicate that EnCodec is capable of preserving temporal and timbral features significantly better than traditional SVD-based dimensionality reduction, while reducing the size of the two datasets by over two orders of magnitude with respect to FLAC encoding. These conclusions are confirmed by a MUSHRA listening test, suggesting that

This work has been funded by "REPERTORIUM project. Grant agreement number 101095065. Horizon Europe. Cluster II. Culture, Creativity and Inclusive Society. Call HORIZON-CL2-2022-HERITAGE-01-02."

This work was partially supported by the European Union under the Italian National Recovery and Resilience Plan (NRRP) of NextGenerationEU, partnership on "Telecommunications of the Future" (PE00000001—program "RESTART")

neural audio codecs represent a promising avenue of research for large-scale RIR data compression.

The remainder of the manuscript is organized as follows. In Section II, we present related work on RIR dimensionality reduction. In Section III, we focus on recent advances in neural audio coding. In Section IV, we outline the evaluation framework. In Section V, we discuss the experimental results. Finally, Section VI concludes this work.

II. Background on Dimensionality Reduction

When it comes to dimensionality reduction, autoencoders have been recently explored for compressing short RIR segments into small-size latent representations [28], [29].

Parametric methods such as artificial reverberators [30] can represent a low-cost alternative to full-scale convolution-based auralization [31], and can be regarded as reduced order models (ROMs) of the corresponding acoustic environments. Unless specifically optimized [32], [33], though, their low parameter count comes at the cost of limited controllability and realism.

Other noteworthy model order reduction techniques rely on modal analysis. Rooted in linear algebra, these methods exploit singular value decomposition (SVD) as a means to extract the most prominent physical modes from measurement data and achieve dimensionality reduction. In this family of methods, we find, e.g., the well-known Eigensystem Realization Algorithm (ERA) [34], that aims to identify a state-space filter from impulse response data. Likewise, Dynamic Mode Decomposition (DMD) [35] characterizes measurements as the superposition of sinusoidal modes, and is able to project high-dimensional vectors onto low-dimensional representations by capturing the underlying system's dynamics.

Adopting the latter approach, Huang et al. recently proposed an SVD-based RIR dimensionality reduction technique and explored its application for system identification [25]. In the following, we provide an overview of the method.

Given an L-sample RIR $h[n]$, we can define the vector

$$\mathbf{h} = \left[h[0], h[1], ..., h[L-1]\right]^T , \tag{1}$$

where $(\cdot)^T$ denotes the transpose operator. Then, a dataset of N RIRs can be expressed as the matrix

$$\mathbf{H} = \begin{bmatrix} | & | & & | \\ \mathbf{h}_1 & \mathbf{h}_2 & \cdots & \mathbf{h}_N \\ | & | & & | \end{bmatrix} . \tag{2}$$

In likely case of length mismatch, to avoid truncation errors when constructing the matrix $\mathbf{H}$, every RIR where $L < L_{\text{max}}$ can be padded with trailing zeros, such that $\mathbf{H} \in \mathbb{R}^{L_{\max} \times N}$.

The economy-sized SVD of the data matrix yields

$$\mathbf{H} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T, \tag{3}$$

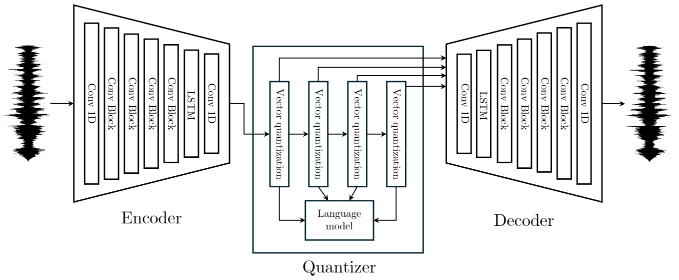


Fig. 1: EnCodec architecture [9].

where $\mathbf{U}$ is the semi-unitary matrix containing the left singular vectors of $\mathbf{H}$. We thus define the *signature matrix* $\mathbf{Q} \in \mathbb{R}^{L_{\max} \times R}$ as the matrix containing the columns of $\mathbf{U}$ associated with the R largest singular values in $\mathbf{\Sigma}$. As long as $R \ll \min\{L_{\text{max}}, N\}$, dimensionality reduction is achieved by linearly projecting a RIR onto the column space of $\mathbf{Q}$, i.e.,

$$\mathbf{z} = \mathbf{Q}^T \mathbf{h} , \tag{4}$$

where $\mathbf{z} \in \mathbb{R}^R$ is the reduced representation of $\mathbf{h} \in \mathbb{R}^{L_{\max}}$.

The vector $\mathbf{z}$ and the signature matrix $\mathbf{Q}$ are then either stored or transmitted. This way, decoding $\mathbf{z}$ amounts to

$$\hat{\mathbf{h}} = \mathbf{Q}\mathbf{z} . \tag{5}$$

Given a dataset of N RIRs of length $L \leq L_{\text{max}}$, the total number of real-valued coefficients in the reduced dataset is

$$\mathcal{C}_R(N) = L_{\text{max}}R + NR, \tag{6}$$

which has a sizeable constant part due to the high dimensionality of the (dataset-specific) signature matrix, and a relatively small cost for storing a single RIR, i.e., R coefficients. Therefore, this approach is better suited for databases so large as to offset the constant factor $L_{\text{max}}R$. Indeed, virtually all methods inspired by modal analysis require storing auxiliary data structures which, in the case of [25], have the same dimensionality of a RIR, i.e., tens of thousands of coefficients. This approach fundamentally differs from that of neural audio codecs, which will be discussed in the next section.

III. Neural Audio Codecs

At its core, an audio codec is a system designed to transform an audio signal into a compact sequence of discrete codes which, ideally, contains enough information to reconstruct the input signal with negligible distortion. A typical audio codec comprises three modules: an encoder, a quantizer, and a decoder. The bitrate of the quantized codes is usually much lower than that of the input signals, such that codeword indices are stored and transmitted at a lower cost.

While conventional codecs exploit digital signal processing, psychoacoustics principles, and sound production models, neural audio codecs [12]–[23] tackle the problem by parameterizing the encoder and the decoder as deep neural networks, and apply vector quantization with learnable codebooks to the innermost latent representations.

In general terms, a Vector Quantizer (VQ) compares the encoder output with each codeword stored in a learnable codebook matrix, and returns the index of the most similar according to some suitable metric. The decoder, in turn, is tasked with reconstructing the input signal from the embedding retrieved from the codebook via table lookup. Such a direct approach, however, would result in a exceedingly large codebook when using a single quantizer layer. Pioneered by [13], residual vector quantization (RVQ) offers an elegant solution, employing a cascade of smaller VQs that progressively encode the residual of the preceding stage.

Depicted in Fig. 1, EnCodec [9] features an RVQ with a variable number of codebooks depending on the target bitrate. Two main variants of EnCodec exist: a 48 kHz non-causal model trained with stereophonic music, and a 24 kHz causal model trained with a variety of monophonic audio signals, including speech, music, and sound events. In this work, we focus on the latter for two reasons. First, we limit our study to SISO RIRs.[1] Second, causality is a desirable property as it enables low-latency data streaming.

The streamable EnCodec model can operate at five bitrates, i.e., 1.5, 3, 6, 12, 24 kbps, corresponding to 2, 4, 8, 16, 32 codebooks, respectively, each consisting of 1024 codewords. It implements causal convolutions and frames the input using a sliding window with a stride of 13 ms. This way, the encoder produces $f_\mathrm{r} = 75$ latent codes per second at $f_\mathrm{s} = 24$ kHz. This means that, using K codebooks, a dataset of N RIRs can be represented using $\mathcal{C}_K(N)$ unsigned integer values ranging from 0 to 1023, where

$$\mathcal{C}_K(N) = N \cdot \left(\frac{f_\mathrm{r}}{f_\mathrm{s}} LK \right). \tag{7}$$

It is worth pointing out that, while (6) is expressed in terms of floating-point numbers, the coefficients here can be represented with as few as 10 bits each.

On top of that, EnCodec uses a small language model (LM) to estimate token probabilities and apply range-based entropy coding (EC) [36]. This further compresses the representation by up to 40% at the cost of an increased computational load at encoding time. In the following, we evaluate EnCodec both with and without LM-based EC.

IV. Evaluation

As a lossless reference, we consider the total disk space in kilobytes occupied by each dataset when stored in FLAC. Moreover, we consider the baseline method presented in [25] for $R = 4, 8, 16, 32, 64, 96, 128$. Whereas increasing the free parameter R may further increase the quality of the reconstruction, we observed that the memory requirements already exceeded those of lossless coding. Finally, we evaluate the causal EnCodec model for all the available bitrates. It is worth noting that EnCodec was trained on general audio, and not with RIRs. The main goal here is thus to assess whether an off-the-shelf neural audio codec can generalize and achieve satisfactory coding results on unseen room acoustics measurements.

A. Room Impulse Response Datasets

To illustrate the capabilities of the methods under scrutiny, we consider two datasets: the MIT Acoustical Reverberation Scene Statistics Survey database [26], or "MIT Survey" for short, and the recently released HOMULA-RIR [27].

The choice of these datasets is to cover two different scenarios. First, MIT Survey contains 270 impulse responses (IRs), both of indoor and outdoor spaces. Since each recording took place in a different environment, the respective IRs exhibit a wide range of reverberation times (T_{60}) and are likely to be little correlated with one another.

Second, HOMULA-RIR comprises multi-channel RIRs obtained for two sources using 25 higher-order microphones (HOMs) with eight capsules each, as well as a uniform linear array (ULA) with 64 sensors, totaling 528 SISO RIRs. Being measured in the same furnished seminar room of the Politecnico di Milano, Milan, Italy, all RIRs in HOMULA-RIR are affected by the same room geometry and are characterized by much more homogeneous recording conditions compared to those in the MIT Survey dataset.

As a data preprocessing step, we resample and normalize each RIR in order to avoid level mismatch.

B. Objective Metrics

In recent work [37], [38], the *normalized misalignment* was used as a metric to evaluate RIR estimation algorithms

$$\mathcal{M} := 20 \log_{10} \left(\frac{\|\mathbf{h} - \hat{\mathbf{h}}\|_2}{\|\mathbf{h}\|_2} \right), \tag{8}$$

where $\mathbf{h}$ is the reference and $\hat{\mathbf{h}}$ is the decoded RIR.

Targeting low-bitrate lossy compression, though, point-wise errors may not adequately reflect the ability to maintain the perceptual qualities of the target RIRs. Instead, we use the *normalized EDC misalignment* to compare different temporal behaviors. Furthermore, we assess the spectral coloration of the compressed and reference RIRs using the *normalized magnitude misalignment*.

The Energy Decay Curve (EDC) of an L-sample impulse response $h[n]$ can be defined through Schroeder's backward integration [39]

$$\varepsilon[n] = \sum_{\tau=n}^{L-1} h^2[\tau]. \tag{9}$$

It is well-understood that the EDC closely relates to the reverberation time T_{60}, as well as other widespread metrics such as early decay time, clarity C_{80}, and definition D_{50} [1]. Thus, we define the normalized EDC misalignment as

$$\mathcal{M}_\mathrm{EDC} := 20 \log_{10} \left(\frac{\|\boldsymbol{\varepsilon} - \hat{\boldsymbol{\varepsilon}}\|_2}{\|\boldsymbol{\varepsilon}\|_2} \right), \tag{10}$$

where $\boldsymbol{\varepsilon} = [\varepsilon[0], ..., \varepsilon[L-1]]^T$.

Similarly, we define the normalized magnitude misalignment as

$$\mathcal{M}_\mathrm{mag} := 20 \log_{10} \left(\frac{\big\| \, |H[k]| - |\hat{H}[k]| \, \big\|_2}{\big\| \, |H[k]| \, \big\|_2} \right), \tag{11}$$

[1] It is worth mentioning that binaural room impulse responses (BRIRs) are inherently stereophonic. The investigation of the non-streamable model for BRIRs compression is left for future work.

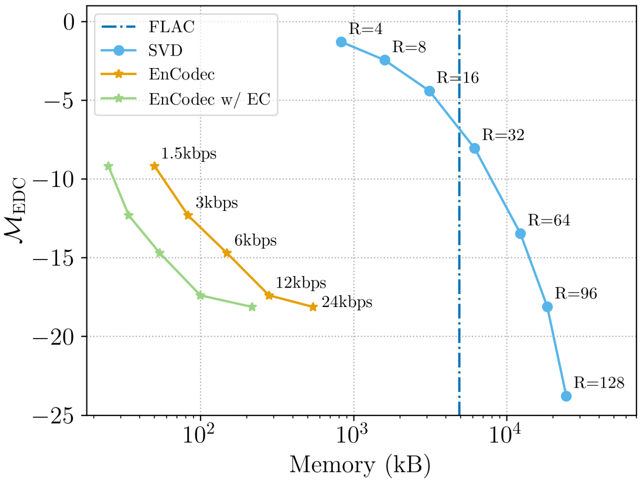


Fig. 2: **MIT Acoustical Reverberation Scene Statistics Survey**: Normalized EDC misalignment as a function of the required disk space. The x-axis is in logarithmic scale.

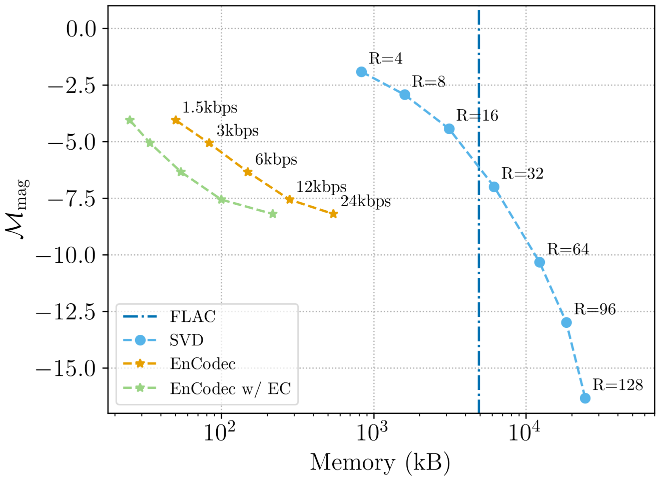


Fig. 4: **MIT Acoustical Reverberation Scene Statistics Survey**: Normalized magnitude misalignment as a function of the required disk space. The x-axis is in logarithmic scale.

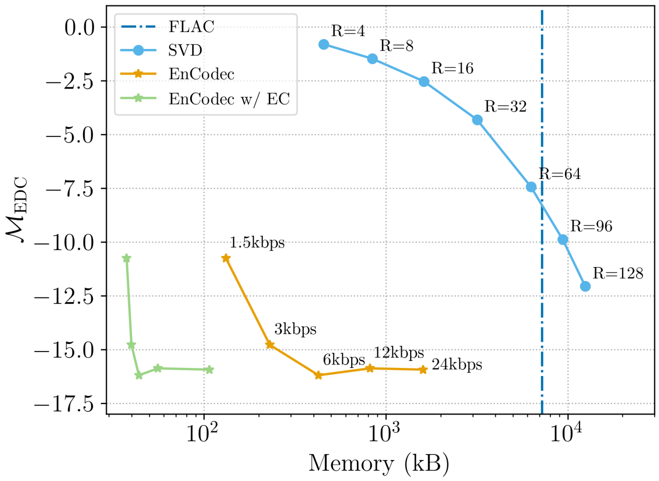


Fig. 3: **HOMULA-RIR**: Normalized EDC misalignment as a function of the required disk space. The x-axis is in logarithmic scale.

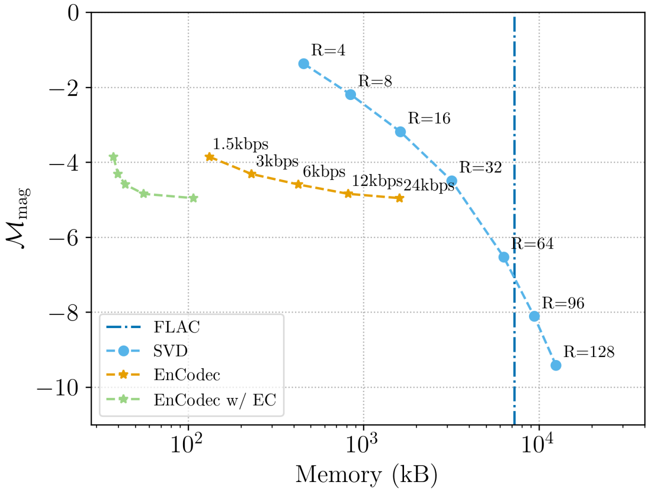


Fig. 5: **HOMULA-RIR**: Normalized magnitude misalignment as a function of the required disk space. The x-axis is in logarithmic scale.

where $|H[k]|$ and $|\hat{H}[k]|$ are the magnitude of the Fast Fourier Transform (FFT) of $h[n]$ and $\hat{h}[n]$, respectively.

The perceptual relevance of the metrics presented above is confirmed by the listening test detailed in the next section.

C. Listening Test

We conducted a Multiple Stimuli with Hidden Reference and Anchor (MUSHRA) test where six anechoic signals (two male speech, two female speech, and two music signals) were convolved with RIRs at different compression levels.

The speech signals were taken from VCTK [40], while the music signals (sax and cello) came from AVAD-VR [41]. In a DAW, we stitched together different utterances from the same speaker and manually extracted a complete musical phrase from each music clip. This resulted in audio files with duration ranging from 15 to 18 seconds,

For each file, we randomly selected a RIR from either dataset, such that one male speech, one female speech, and one music signal were associated with RIRs from MIT Survey, while the remaining ones were paired with RIRs from HOMULA-RIR. The pairings are reported in Table I.

The test was conducted using webMUSHRA [42], a Web Audio API-based software compliant to the ITU-R Rec. BS.1534 [43]. References were obtained by convolving the audio signals with the uncompressed RIRs, whereas clean, non-reverberant clips were used as anchors. After an initial training page, participants were tasked to rate the similarity of each item with the reference on a scale of 0 to 100. On each

Stimulus	VCTK speaker ID	AVAD-VR ID	MIT Survey ID	HOMULA-RIR ID	T_{60} [s]
1	p300	–	h060_Office_ConferenceRoom_3txts	–	1.42
2	p227	–	h001_Bedroom_65txts	–	0.43
3	–	DontMeanAthing_Sax	h052_Gym_WeightRoom_3txts	–	1.22
4	p225	–	–	rir-S1-R1-HOM1 (0)	0.83
5	p292	–	–	rir-S2-R2-HOM4 (1)	0.91
6	–	Canon_Cello	–	rir-S1-R3-HOM2 (2)	0.92

TABLE I: Signal-RIR pairings of all MUSHRA test stimuli.

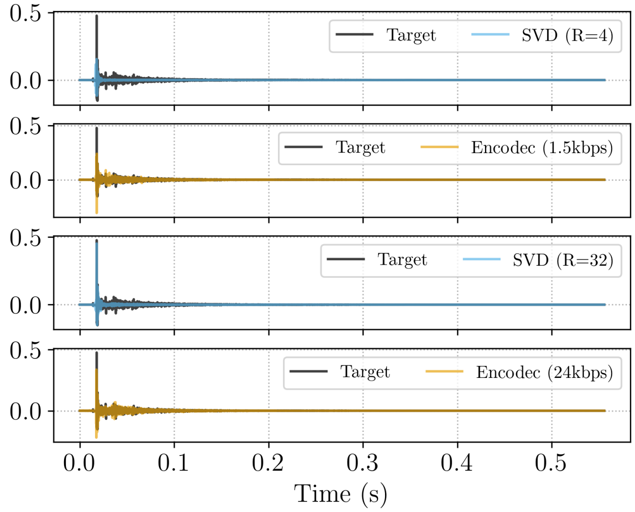


Fig. 6: **MIT Acoustical Reverberation Scene Statistics Survey**: Room impulse responses (h002_Bedroom_62txts).

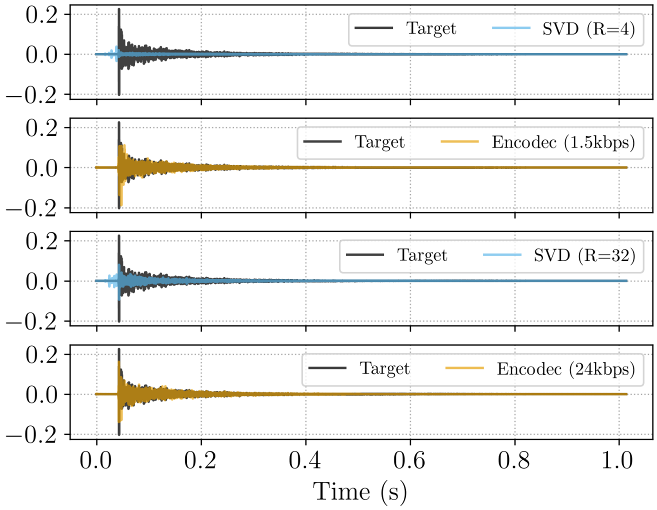


Fig. 7: **HOMULA-RIR**: Room impulse responses (S1-R5-HOM5).

page, ten conditions were assessed, including the hidden reference and anchor; we evaluate EnCodec at all bitrates (1.5, 3, 6, 12, 24 kbps), as well as the baseline with $R \in \{4, 8, 16\}$, i.e., configurations in which the compression ratio was significantly larger than FLAC. The loudness of each file was normalized to -24 LUFS according to ITU-R Rec. BS.1770-4 [44] using pyloudnorm [45]. Volume adjustments were allowed during the training phase. Then, subjects were asked to keep the level constant for the duration of the test. A total of 14 participants took part in the experiment, with age ranging from 25 to 38, none of whom reported hearing impairments. According to the post-screening guidelines [43], one subject was excluded. The remaining participants were students or members of the Image and Sound Processing Lab (ISPL) at Politecnico di Milano, and had previous experience with MUSHRA tests.

V. Results and Discussion

A. *Objective Evaluation*

Fig. 2 and Fig. 3 show the normalized EDC misalignment as a function of the memory required to store MIT Survey and HOMULA-RIR, respectively. Fig. 4 and Fig. 5, in turn, depict the normalized magnitude misalignment.

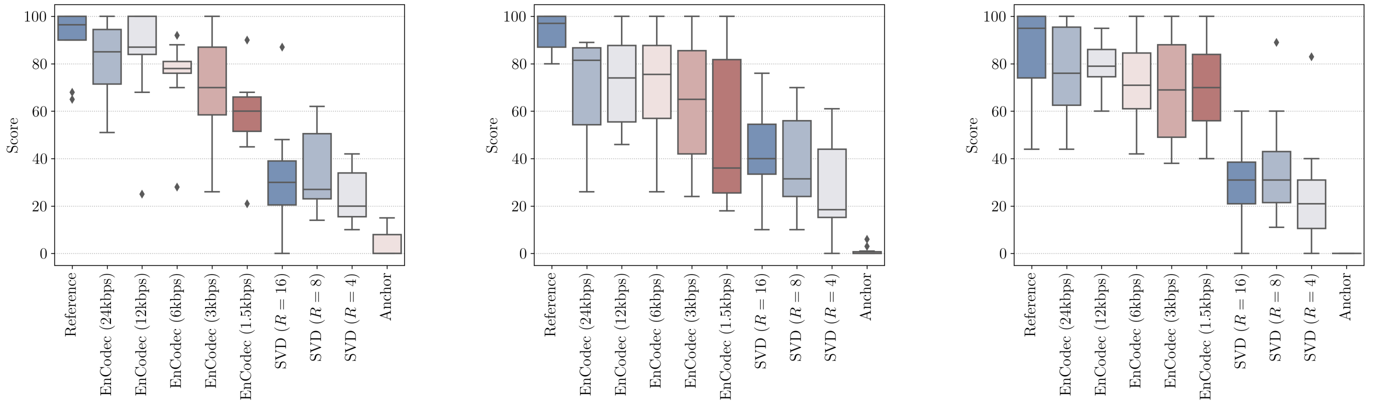
EnCodec with LM-based EC achieves the highest compression gain, with a memory footprint two orders of magnitude smaller than FLAC. Comparatively, whereas the baseline method can sometimes achieve better metrics than EnCodec, this happens only for data sizes that either approach or exceed the threshold for lossless compression, i.e., failing to provide any substantial saving. Overall, EnCodec is shown to outperform the baseline method by over 15 dB in terms of $\mathcal{M}_{\text{EDC}}$, and by more than 3 dB in terms of $\mathcal{M}_{\text{mag}}$. This suggests that EnCodec is better suited for encoding temporal and spectral features despite the higher compression gain.

This is also noticeable in the RIRs depicted in Fig. 6 and Fig. 7. In these examples, we compare the baseline method with $R = 4$ and $R = 32$ (in blue), and EnCodec with a bitrate of 1.5 and 24 kbps (in orange). In both Fig. 6 and Fig. 7, EnCodec appears able to match the target RIRs more closely than the baseline, which is instead characterized by a steep energy decay. Additionally, Fig. 7 also shows that the baseline method can sometimes produce pre-ringing artifacts on HOMULA-RIR data, while EnCodec does a better job at preserving the temporal location of the onset.

B. *Subjective Evaluation*

The outcomes of the MUSHRA test appear to validate the conclusions drawn in the previous section.

In Fig. 8 and Fig. 9, each box represents the interquartile range (IQR) which spans from the first quartile (Q1) to the third quartile (Q3). The line inside the boxes marks the median score. The whiskers extend from Q1 and Q3 to the smallest and largest values within 1.5 times the IQR, indicating the

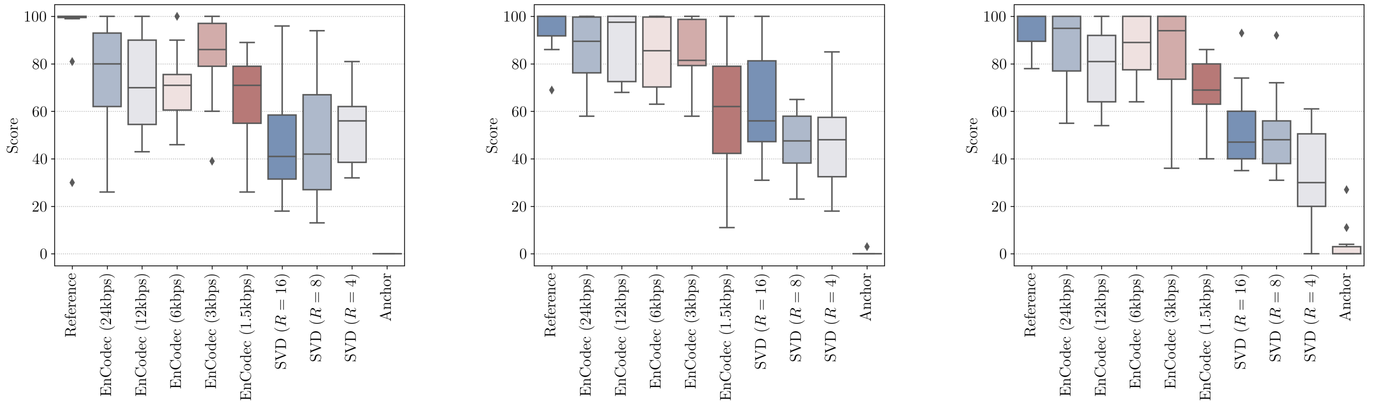


(a) Female speech (`p300`) in a conference room (`h060_Office_ConferenceRoom_3txts`).

(b) Male speech (`p227`) in a bedroom (`h001_Bedroom_65txts`).

(c) Sax (`DontMeanAthing`) in a weight room (`h052_Gym_WeightRoom_3txts`).

Fig. 8: **MIT Acoustical Reverberation Scene Statistics Survey**: MUSHRA test results.



(a) Female speech (`p225`) with the first capsule of `rir-S1-R1-HOM1`.

(b) Male speech (`p292`) with the second capsule of `rir-S2-R2-HOM4`.

(c) Cello (`Canon`) with the third capsule of `rir-S1-R3-HOM2`

Fig. 9: **HOMULA-RIR**: MUSHRA test results.

range of the bulk of the data. Any rating scores outside the whiskers are considered outliers and plotted separately.

In all six cases, the baseline method received the lowest ratings, whereas most EnCodec variants showcase whiskers approaching 100, suggesting that several participants could not detect any significant difference between the test conditions and the reference. This aligns with the fact that the hidden reference was not consistently identified, resulting in a broader rating distribution for five out of six reference signals.

Overall, little to no dependency on reverberation time was observed. In fact, the ratings in Fig. 8a, corresponding to the longest RIR ($T_{60} \approx 1.42$), follow a trend similar to those in Fig. 8b, which instead pertains to the shortest RIR of the six ($T_{60} \approx 0.43$). The latter, however, exhibit a larger spread.

In summary, EnCodec variants operating at 3, 6, 12, and 24 kbps show a comparable perceptual performance, with average ratings of 75.8, 75.9, 78.6, and 78.8, respectively. However, a noticeable drop is observed for the 1.5 kbps model, which received an average score of 62.3. Meanwhile, the baseline method achieved average ratings of 45.2, 42.3, and 37.0 for $R = 16$, 8, and 4, respectively. This reveals that, on average, the neural audio codec outperforms SVD-based dimensionality reduction by over 30 points.

VI. Conclusions

In this paper, we examined the application of EnCodec, a general-purpose pretrained RVQ-based neural audio codec, for compressing large-scale datasets of RIRs. Our analysis, both objective and subjective, demonstrates that EnCodec is effective for very low-bitrate lossy compression. It significantly reduces data size while maintaining the perceptual qualities of the RIRs, enabling more manageable storage and efficient transmission of large room acoustics datasets. While this work indicates off-the-shelf neural audio codecs as a viable strategy to achieve low data rate realistic-sounding reverberation, future work will investigate their ability to encode spatial cues, such as sound directivity, diffuseness, and spatial coherence. Future research could thus explore the effectiveness of space-time processing using encoded RIRs. This includes applications such as auralization, sound source localization, and room geometry inference.

References

[1] "Acoustics – measurement of room acoustic parameters. Part 1: performance spaces," ISO 3382-1:2009, International Organization for Standardization, Geneva, Switzerland, June 2009.

[2] C. Tuna, A. Canclini, F. Borra, P. Götz, F. Antonacci, A. Walther, A. Sarti, and E. A. P. Habets, "3D room geometry inference using a linear loudspeaker array and a single microphone," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1729–1744, 2020.

[3] E. Vincent, T. Virtanen, and S. Gannot, *Audio source separation and speech enhancement*. John Wiley & Sons, 2018.

[4] J. Benesty, S. Makino, and J. Chen, *Speech enhancement*. Springer Science & Business Media, 2006.

[5] M. Kleiner, B.-I. Dalenbäck, and P. Svensson, "Auralization—an overview," *Journal of the Audio Engineering Society*, vol. 41, pp. 861–875, november 1993.

[6] T. Potter, Z. Cvetković, and E. De Sena, "On the relative importance of visual and spatial audio rendering on vr immersion," *Frontiers in Signal Processing*, vol. 2, p. 904866, 2022.

[7] A. Bertrand, "Applications and trends in wireless acoustic sensor networks: A signal processing perspective," in *2011 18th IEEE Symposium on Communications and Vehicular Technology in the Benelux (SCVT)*, 2011, pp. 1–6.

[8] F. Martusciello, C. Centofanti, C. Rinaldi, and A. Marotta, "Edge-enabled spatial audio service: Implementation and performance analysis on a MEC 5G infrastructure," in *2023 4th International Symposium on the Internet of Sounds*, 2023, pp. 1–8.

[9] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, "High fidelity neural audio compression," *arXiv preprint arXiv:2210.13438*, 2022.

[10] J.-M. Valin, G. Maxwell, T. B. Terriberry, and K. Vos, "High-quality, low-delay music coding in the Opus codec," *Journal of the Audio Engineering Society*, no. 8942, october 2013.

[11] M. Dietz, M. Multrus, V. Eksler, V. Malenovsky, E. Norvell, H. Pobloth, L. Miao, Z. Wang, L. Laaksonen, A. Vasilache *et al.*, "Overview of the EVS codec architecture," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5698–5702.

[12] S. Kankanahalli, "End-to-end optimized speech coding with deep neural networks," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 2521–2525.

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

[14] T. Jayashankar, T. Koehler, K. Kalgaonkar, Z. Xiu, J. Wu, J. Lin, P. Agrawal, and Q. He, "Architecture for variable bitrate neural speech codec with configurable computation complexity," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 861–865.

[15] X. Jiang, X. Peng, C. Zheng, H. Xue, Y. Zhang, and Y. Lu, "End-to-end neural speech coding for real-time communications," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 866–870.

[16] D. Yang, S. Liu, R. Huang, J. Tian, C. Weng, and Y. Zou, "HiFi-Codec: Group-residual vector quantization for high fidelity audio codec," *arXiv preprint arXiv:2305.02765*, 2023.

[17] X. Zhang, D. Zhang, S. Li, Y. Zhou, and X. Qiu, "SpeechTokenizer: Unified speech tokenizer for speech large language models," *arXiv preprint arXiv:2308.16692*, 2023.

[18] Z. Huang, C. Meng, and T. Ko, "RepCodec: A speech representation codec for speech tokenization," *arXiv preprint arXiv:2309.00169*, 2023.

[19] R. Kumar, P. Seetharaman, A. Luebs, I. Kumar, and K. Kumar, "High-fidelity audio compression with improved RVQGAN," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[20] Y. Zheng, W. Tu, L. Xiao, and X. Xu, "Srcodec: Split-residual vector quantization for neural speech codec," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 451–455.

[21] Y. Ai, X.-H. Jiang, Y.-X. Lu, H.-P. Du, and Z.-H. Ling, "APCodec: A neural audio codec with parallel amplitude and phase spectrum encoding and decoding," *arXiv preprint arXiv:2402.10533*, 2024.

[22] H. Liu, X. Xu, Y. Yuan, M. Wu, W. Wang, and M. D. Plumbley, "SemantiCodec: An ultra low bitrate semantic audio codec for general sound," *arXiv preprint arXiv:2405.00233*, 2024.

[23] Y. Ren, T. Wang, J. Yi, L. Xu, J. Tao, C. Y. Zhang, and J. Zhou, "Fewer-token neural speech codec with time-invariant codes," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 12 737–12 741.

[24] S. Lee, H.-S. Choi, and K. Lee, "Yet another generative model for room impulse response estimation," in *2023 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2023, pp. 1–5.

[25] G. Huang, J. Benesty, and J. Chen, "Dimensionality reduction of room acoustic impulse responses and applications to system identification," *IEEE Signal Processing Letters*, 2023.

[26] J. Traer and J. H. McDermott, "Statistics of natural reverberation enable perceptual separation of sound and space," *Proceedings of the National Academy of Sciences*, vol. 113, no. 48, pp. E7856–E7865, 2016.

[27] F. Miotello, P. Ostan, M. Pezzoli, L. Comanducci, A. Bernardini, F. Antonacci, and A. Sarti, "HOMULA-RIR: A room impulse response dataset for teleconferencing and spatial audio applications acquired through higher-order microphones and uniform linear microphone arrays," in *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops*, 2024, pp. 795–799.

[28] D. A. Sanaguano-Moreno, J. F. Lucio-Naranjo, R. A. Tenenbaum, L. Bravo-Moncayo, and G. B. Regattiere-Sampaio, "A deep learning approach for the generation of room impulse responses," in *2022 Third International Conference on Information Systems and Software Technologies*, 2022, pp. 64–71.

[29] M. Eineborg, F. Pind, and S. Thrastarson, "Generating impulse responses using autoencoders," in *Proceedings of the 10th Convention of the European Acoustics Association Forum Acusticum 2023*, Jan. 2024, p. 3177–3180.

[30] V. Valimaki, J. D. Parker, L. Savioja, J. O. Smith, and J. S. Abel, "Fifty years of artificial reverberation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 5, pp. 1421–1448, 2012.

[31] E. De Sena, H. Hacihabiboglu, and Z. Cvetkovic, "Scattering delay network: An interactive reverberator for computer games," *Journal of the Audio Engineering society*, no. 3-1, february 2011.

[32] A. I. Mezza, R. Giampiccolo, E. De Sena, and A. Bernardini, "Data-driven room acoustic modeling via differentiable feedback delay networks with learnable delay lines," *arXiv preprint arXiv:2404.00082*, 2024.

[33] A. I. Mezza, R. Giampiccolo, and A. Bernardini, "Modeling the frequency-dependent sound energy decay of acoustic environments with differentiable feedback delay networks," in *Proceedings of the 27th International Conference on Digital Audio Effects*, 2024.

[34] J.-N. Juang and R. S. Pappa, "An eigensystem realization algorithm for modal parameter identification and model reduction," *Journal of guidance, control, and dynamics*, vol. 8, no. 5, pp. 620–627, 1985.

[35] J. N. Kutz, S. L. Brunton, B. W. Brunton, and J. L. Proctor, *Dynamic mode decomposition: data-driven modeling of complex systems*. SIAM, 2016.

[36] J. Rissanen and G. Langdon, "Universal modeling and coding," *IEEE Transactions on Information Theory*, vol. 27, no. 1, pp. 12–23, 1981.

[37] L.-M. Dogariu, J. Benesty, C. Paleologu, and S. Ciochină, "Identification of room acoustic impulse responses via kronecker product decompositions," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 2828–2841, 2022.

[38] M. Jälmby, F. Elvander, and T. v. Waterschoot, "Low-rank room impulse response estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 957–969, 2023.

[39] M. R. Schroeder, "New method of measuring reverberation time," *Journal of the Acoustical Society of America*, vol. 37, no. 6, pp. 1187–1188, 1965.

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

[41] D. Thery and B. F. Katz, “Anechoic audio and 3D-video content database of small ensemble performances for virtual concerts,” in *International Congress on Acoustics (ICA)*, 2019.

[42] M. Schoeffler, S. Bartoschek, F.-R. Stöter, M. Roess, S. Westphal, B. Edler, and J. Herre, “webMUSHRA—a comprehensive framework for web-based listening tests,” *Journal of Open Research Software*, Feb 2018.

[43] “Method for the subjective assessment of intermediate quality level of audio systems,” Rec. ITU-R BS.1534-3, International Telecommunications Union, Geneva, Switzerland, Jun. 2021.

[44] “Algorithms to measure audio programme loudness and true-peak audio level,” Rec. ITU-R BS.1770-4, International Telecommunications Union, Geneva, Switzerland, Nov. 2023.

[45] C. J. Steinmetz and J. D. Reiss, “pyloudnorm: A simple yet flexible loudness meter in Python,” *Journal of the Audio Engineering Society*, no. 10483, may 2021.