ENHANCE-NANOCODEC: ENHANCEMENT NEURALAUDIO CODEC FOR THE LRAC 2025 CHALLENGE TRACK 2

Lingling Dai[⋆∗], *Zhe Han*[†], *Andong Li*[⋆∗], *Yiqing Guo*[†], *Linping Xu*[†],
Hua Gao[†], Xiaodong Li[⋆∗], Chengshi Zheng[⋆∗]

⋆ Institute of Acoustics, Chinese Academy of Sciences, Beijing, China
†ByteDance, China
∗University of Chinese Academy of Sciences, Beijing, China

ABSTRACT

This paper presents Enhance-NanoCodec, which is designed to perform codec transmission in conjunction with simultaneous denoising and dereverberation under the constraints of low complexity, low bitrate and real-time processing. Our architecture operates in the time-frequency (T-F) domain, where we discard the phase and only encode the magnitude features on the encoder side—both the magnitude and phase are estimated on the receiver side. To scientifically allocate the complexity ratio of the model between the encoder and decoder, and to utilize the codebook more efficiently, we designed a multi-stage training scheme, which excellently accomplishes the joint task of speech enhancement and coding. In addition, we propose an efficient convolution-style attention block as the core modeling unit. Enhance-NanoCodec achieves a total latency of 50 ms and a computational complexity of 1.86 GFlops (0.58 for the decoder), and is submitted to the LRAC Challenge Track 2.

Index Terms— Neural audio codec, speech enhancement, low-complexity, low bitrate, real-time

1. INTRODUCTION

Audio codec technologies are foundational to on-demand streaming. End-to-end Neural audio codecs (NACs) with learnable encoders, including SoundStream [1] and DAC [2], have attracted significant research interest. They stand out for high-quality audio at very low bitrates, a performance target conventional audio coding struggles to achieve. However, several critical issues persist as key focus areas for advancing the practical deployment of NACs in real-world transmission scenarios, including high computational cost, strict causality constraints, non-negligible algorithmic delay, and the ongoing challenge of ensuring clear speech transmission amid complex background noise.

The objective of Track 2 in the LRAC 2025 Challenge [1] is to achieve the integration of speech enhancement and coding under the joint constraints of low latency, low computational complexity, real-time processing, and high quality. To this end, we propose the Enhance-NanoCodec architecture. This system is engineered to fulfill the challenge constraints while maintaining robust performance, and its capabilities are fully optimized through a multi-stage training scheme for high-quality speech enhancement and coding.

First, Enhance-NanoCodec operates in the time-frequency (T-F) domain for high-fidelity spectral detail reconstruction. As target coding or estimation in the time domain becomes especially challenging when computational resources are limited, we disregard the phase and utilize only the magnitude for feature encoding, with both magnitude and phase reconstructed in the decoder, leveraging a Fourier prior to ease the learning process. Second, we adopt a convolution-style attention block for spectral modeling. It uses large convolution kernels to generate the attention distribution, effectively aggregating contextual information. Third, joint magnitude and phase estimation under limited resources remains an open challenge. Following [3], we use an omnidirectional phase loss for phase optimization, which captures differential relations between center and neighboring phase bins. We further extend this to the spectrum's real and imaginary (RI) parts, proposing an omnidirectional RI loss. Finally, inspired by [4], we design a multi-stage training scheme to further enhance the codebook's efficiency in leveraging clean speech data within Track 2, while optimizing task allocation between the encoder and decoder. This comprehensive training strategy enables the model to accomplish the dual objectives of high-quality speech enhancement and reconstruction, all while fully complying with the challenge requirements.

[1] https://crowdsourcing.cisco.com/lrac-challenge/2025/

2. METHOD ILLUSTRATIONS

2.1. Overall Architecture

The overall structure of the proposed Enhance-NanoCodec is presented in Fig. 1(a). Given the input waveform $x \in \mathbb{R}^L$, we first transform it into the time-frequency (T-F) domain using the short-time Fourier transform (STFT), obtaining the complex spectrogram $X \in \mathbb{C}^{F\times T}$, where F and T denote the number of frequency bins and time frames, respectively. For the encoder input, we drop the phase counterpart and use the normalized magnitude spectrogram $|X| \in \mathbb{R}^{F\times T}$ along with the spectral energy, which is extracted via the energy-content decoupling (ECD) layer. Then the encoder extracts the frequency information and obtains highly compressed hidden representations, which are matched with a sequence of discrete codes $C \in \mathbb{R}^{N_q\times D\times T}$ through residual vector quantization (RVQ), where N_q is the codebook number and D is the feature dimension. The decoder takes the quantized codes as input and reconstructs both the magnitude spectrogram and the phase spectrogram. Finally, we recover the enhanced waveform $\hat{x} \in \mathbb{R}^L$ by applying the inverse STFT (iSTFT). Both encoder and decoder share the same modeling unit, which is composed of a stack of Large Kernel Convolution-Style Attention Block (LKCAB) as shown in Fig. 1(b). The proposed LKCAB uses large convolution kernels to generate the attention distribution, effectively aggregating contextual information.

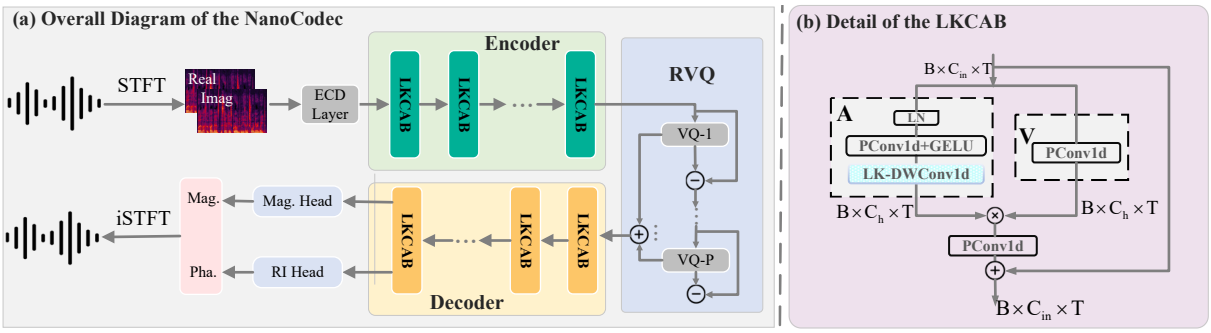

Fig. 1. (a) Overall structure of the proposed NanoCodec; (b) Internal structure of the adopted LKCAB.

2.2. Multi-stage Training Scheme

In Track 2 of the LRAC 2025 Challenge, the computational budget is intentionally biased toward the transmitter side, with a substantially higher complexity allocation compared to the receiver. Since the track specifically focuses on enhanced speech, we introduce a multi-stage training strategy aimed at improving the codebook's efficiency in representing clean speech while achieving a more balanced computational distribution between the encoder and decoder. The detailed training procedure is elaborated as follows.

2.2.1. Stage 1: Training codebook of clean speech

To ensure that all information in the codebook is dedicated to transmitting valid clean speech, only clean speech is used during the training process. It is important to note that the codebook for both 1 kbps and 6 kbps were finalized in this stage. To better guide the encoder's performance and avoid being constrained by the decoding bottleneck of the decoder, a decoder with a complexity exceeding that required by the challenge is employed for speech encoding during this stage.

2.2.2. Stage 2: Training a speech-enhancement encoder

In Stage 2, an encoder with noise reduction capability is trained. At this stage, various speech augmentations were applied to the data, including the addition of noise and reverberation, as well as other augmentations mentioned in 3.3. During this phase, the codebook and decoder learned in Stage 1 were fixed, with only the encoder undergoing training.

2.2.3. Stage 3: Training a low complexity decoder

In Stage 3, the objective was to train a decoder that meets the computational complexity requirements and is compatible with the encoder and codebook obtained in the previous two stages. During this stage, both the codebook and the encoder were fixed.

3. MISCELLANEOUS CONFIGURATIONS

3.1. Network Setups

For both STFT and iSTFT, the window length is set to 50 ms with a hop size of 12.5 ms. No auxiliary look-ahead nor algorithmic delay is introduced, resulting in a total system latency of 50 ms. The number of LKCABs used in the encoder is set to 12 with a hidden dimension of 600, while the decoder uses 10 LKCABs with a hidden dimension of 420. The number of codebooks is set to 1 with a codebook size of 5792 for the 1 kbps transmission rate. For the 6 kbps transmission rate, we reuse the codebook from the 1 kbps setup. Additionally, we introduce the grouped RVQ, where the codebooks are divided into two groups, with each group containing 3 codebooks and a codebook size of 1024. The theoretical transmission rate is 1.00 kbps for 1 kbps transmission and 5.80 kbps for 6 kbps transmission. The total trainable parameter count for Enhance-NanoCodec is 11.51 M, and the total computational complexity is 1.86 GFlops, where the decoder accounts for 0.58 GFlops. The detailed model parameters and computational complexity of each module are presented in Table 1.

Module	Para. (M)	Complexity (MFlops)
Encoder	7.84	1266.66
Quantizer	0.08	16.32
Decoder	3.59	578.63

Table 1. Model parameter and computational complexity.

3.2. Loss Setups

We use both reconstruction and adversarial losses during Stage 1 and Stage 2 training. The reconstruction loss consists of multi-resolution STFT loss, multi-resolution Mel loss, as well as our proposed omni-directional phase loss, which captures differential relations between center and neighboring phase bins. For adversarial training, we employ a multi-period discriminator (MPD), multi-resolution STFT discriminator (MRSTFTD), and multi-band discriminator (MBD), along with a feature matching loss. In Stage 3, to further improve the performance of the low-complexity decoder, we additionally incorporate PESQ loss, UTMOS loss, as well as our proposed omnidirectional RI loss for optimization, where the former two provide perceptual supervision and the latter enables finer joint magnitude-and-phase reconstruction.

3.3. Dataset Setups

The training corpus employed in this study is sourced from the LRAC 2025 Challenge. For speech, we use the speech clips from LibriSpeech [5], LibriVox [6], VCTK [7], EARS [8] and Multilingual Librispeech [9]. For noise set, we include Audioset [10], Freesound [11](from the DNS5 challenge[2]), FMA [12],

[2]https://github.com/microsoft/DNS-Challenge

WHAM! [13] and FSD50K [14]. For reverberation generation, we include the room impulse responses (RIRs) from Open SLR 28[3] and Motus [15]. Further refinement was performed by excluding audio segments that were excessively short or exhibited abnormally low energy, thereby ensuring the quality and consistency of the training samples.

During model training after stage 1, noisy and reverberant signals were synthesized on-the-fly via random sampling from the speech, RIR, and noise datasets. Specifically, under noisy speech conditions, the signal-to-noise ratio (SNR) was set to range from -5 dB to 20 dB. To enhance model generalization, we applied additional data augmentation to 20% of the training corpus, implementing specific techniques including bandwidth limitation, amplitude clipping, and packet loss concealment (PLC).

3.4. Evaluation Metrics

In this study, model performance is initially evaluated using both the non-intrusive metric UTMOS [16] and the intrusive metric PESQ [17], which facilitated rapid evaluation and informed iterative adjustments to the model architecture and training procedures. For the final selection of the model, comprehensive human listening tests were conducted to ensure robust perceptual quality.

3.5. Training Settings

We optimized the model using AdamW optimizer [18] with its default betas (0.8, 0.99) and an initial learning rate of 0.0002. The learning rate is scheduled using an ExponentialLR scheduler with a gamma of 0.999998 per epoch. Additionally, we set the batch size to 16 and the duration of each sample to 5 seconds. For each training stage, the number of training steps was set to a range of 500,000 to 1,000,000, depending on the convergence of the evaluation metrics.

4. REFERENCES

[1] Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi, "SoundStream: An End-to-End Neural Audio Codec," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 495–507, 2021.

[2] Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar, "High-fidelity audio compression with improved rvqgan," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[3] Andong Li, Tong Lei, Zhihang Sun, Rilin Chen, Erwei Yin, Xiaodong Li, and Chengshi Zheng, "Learning Neural Vocoder from Range-Null Space Decomposition," *arXiv preprint arXiv:2507.20731*, 2025.

[4] Yunlong Liu, Tao Huang, Weisheng Dong, Fangfang Wu, Xin Li, and Guangming Shi, "Low-light image enhancement with multi-stage residue quantization and brightness-aware attention," in *Proc. ICCV*, 2023, pp. 12140–12149.

[5] Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu, "Libritts: A corpus derived from librispeech for text-to-speech," *arXiv preprint arXiv:1904.02882*, 2019.

[6] Jodi Kearns, "Librivox: Free public domain audiobooks," 2014.

[7] Christophe; MacDonald Kirsten Yamagishi, Junichi; Veaux, "CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit,," .

[8] Julius Richter, Yi-Chiao Wu, Steven Krenn, Simon Welker, Bunlong Lay, Shinji Watanabe, Alexander Richard, and Timo Gerkmann, "EARS: An anechoic fullband speech dataset benchmarked for speech enhancement and dereverberation," *arXiv preprint arXiv:2406.06185*, 2024.

[9] Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert, "Mls: A large-scale multilingual dataset for speech research," *arXiv preprint arXiv:2012.03411*, 2020.

[10] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. ICASSP*. IEEE, 2017, pp. 776–780.

[11] Eduardo Fonseca, Jordi Pons Puig, Xavier Favory, Frederic Font Corbera, Dmitry Bogdanov, Andres Ferraro, Sergio Oramas, Alastair Porter, and Xavier Serra, "Freesound datasets: a platform for the creation of open audio datasets," 2017.

[12] Michaël Defferrard, Kirell Benzi, Pierre Vandergheynst, and Xavier Bresson, "FMA: A dataset for music analysis," *arXiv preprint arXiv:1612.01840*, 2016.

[13] Gordon Wichern, Joe Antognini, Michael Flynn, Licheng Richard Zhu, Emmett McQuinn, Dwight Crow, Ethan Manilow, and Jonathan Le Roux, "Wham!: Extending speech separation to noisy environments," *arXiv preprint arXiv:1907.01160*, 2019.

[14] Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra, "Fsd50k: an open dataset of human-labeled sound events," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 829–852, 2021.

[15] Georg Götz, Sebastian J Schlecht, and Ville Pulkki, "A dataset of higher-order ambisonic room impulse responses and 3d models measured in a room with varying furniture," in *2021 Immersive and 3D Audio: from Architecture to Automotive (I3DA)*. IEEE, 2021, pp. 1–8.

[16] Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari, "UTMOS: UTokyo-SaruLab System for the VoiceMOS Challenge 2022," in *Proc. Interspeech*, 2022, pp. 4521–4525.

[17] A.W. Rix, J.G. Beerends, M.P. Hollier, and A.P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. ICASSP*, 2001, vol. 2, pp. 749–752 vol.2.

[18] Ilya Loshchilov and Frank Hutter, "Decoupled Weight Decay Regularization," in *International Conference on Learning Representations*, 2019.

[3] https://www.openslr.org/28/