LOW RESOURCE AUDIO CODEC CHALLENGE TRACK2: DENOISING CODEC

Haoran Zhao, Zixiang Wan, Guochang Zhang, Runqiang Han, Jianqiang Wei

Anker Innovations, Beijing, China

ABSTRACT

We propose a frequency-domain *Denoising Codec* for the 2025 Low-Resource Audio Codec (LRAC) Challenge that jointly performs speech coding and noise suppression under strict constraints on complexity, latency, and bitrate. By integrating enhancement into the coding pipeline and employing residual vector quantization (RVQ), the system allocates bits to perceptually important speech components while reducing the noise. A three-stage training process combines spectral reconstruction with adversarial objectives to ensure stable optimization and high-quality output. Experiments across clean, noisy, and reverberant conditions demonstrate consistent improvements in both coding fidelity and robustness.

***Index Terms*—** speech codec, noise suppression, low resource, LRAC

1. INTRODUCTION

Neural audio codecs are emerging as powerful alternatives to traditional speech coders such as AMR-WB and Opus, delivering improved perceptual quality and flexible bitrate adaptation. Recent advances, SoundStream [1], Encodec [2], and DAC [3] employ autoencoder-based architectures with RVQ and adversarial training, achieving high-quality reconstruction at low bitrates.

In parallel, neural speech enhancement has advanced rapidly. Architectures such as U-Net [4], DCCRN [5], and DeepFilterNet [6] demonstrate robust noise suppression across diverse acoustic environments. Leveraging convolutional encoder–decoder backbones, recurrent layers, and attention mechanisms, these models effectively disentangle clean speech from noise and reverberation.

However, most codecs and enhancement systems are designed and optimized independently: codecs focus primarily on compression efficiency and reconstruction fidelity, while enhancement models target noise reduction and dereverberation. Under realistic constraints on complexity, latency, and bitrate, separating enhancement from coding can be suboptimal. A unified approach enables efficient bit allocation for perceptual speech quality and effective noise suppression.

The 2025 LRAC Challenge provides an ideal platform for such integrated solutions, emphasizing neural speech codecs operating under realistic noise and reverberation with strict limits on complexity, latency, and bitrate. It encourages unified designs that jointly address speech coding and enhancement within a low-resource framework. Motivated by this, we propose a frequency-domain *Denoising Codec* jointly optimized for noise suppression and speech coding.

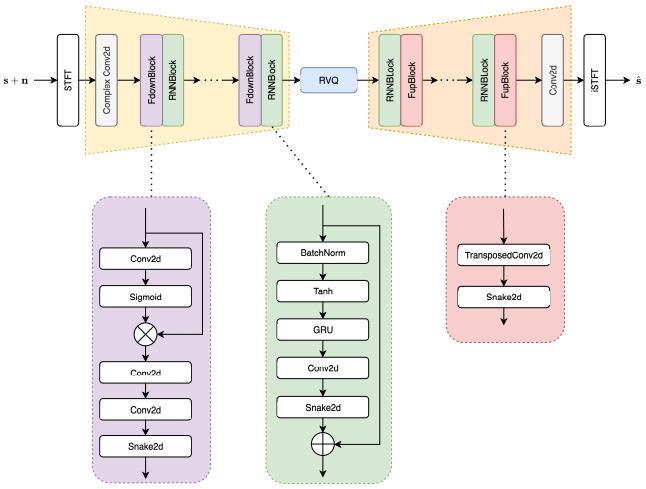

Fig. 1. The proposed model architecture.

2. METHOD

2.1. Architecture

The proposed end-to-end *Denoising Codec* is illustrated in Figure 1. It comprises an encoder, an RVQ module [1, 2], and a decoder, and operates entirely in the frequency domain. The noisy input signal is first transformed into a spectrogram via STFT, which is processed by the encoder to generate downsampled latent vectors; these vectors are quantized by RVQ and then reconstructed by the decoder through upsampling. The resulting spectrogram is finally converted back to the time domain using iSTFT. Noise suppression is implicitly achieved throughout the encoding–decoding process.

The encoder consists of a complex convolutional layer followed by 4 FdownBlocks and RNNBlocks, which perform downsampling, feature extraction, and implicit denoising.

Table 1. Evaluation results for different bitrates and acoustic conditions.

Bitrate	Method	Clean					Noisy					Reverb				
		sheet ssqa	scoreq ref	audiobox AE_CE	utmos	pesq	sheet ssqa	scoreq ref	audiobox AE_CE	utmos	pesq	sheet ssqa	scoreq ref	audiobox AE_CE	utmos	pesq
1kbps	Baseline	2.07	1.01	3.96	1.37	1.21	1.95	1.15	3.70	1.35	1.18	2.43	1.12	3.55	1.32	1.15
	Proposed	3.44	0.43	5.23	3.18	2.07	3.11	0.61	4.9	2.94	1.8	2.27	0.95	4.32	2.05	1.38
6kbps	Baseline	3.55	0.43	5.25	2.97	2.13	2.92	0.75	4.60	2.56	1.73	2.67	0.92	4.25	1.79	1.29
	Proposed	4.22	0.18	5.62	3.80	3.34	3.78	0.41	5.17	3.47	2.41	2.96	0.74	4.62	2.30	1.61

Each FdownBlock includes two 1×1 convolutions and one downsampling convolution, incorporates a gating mechanism to enhance feature extraction, and adopts a Snake2D activation [7] to improve harmonic structure modeling. Each RNNBlock contains batch normalization, a GRU, and a 1×1 convolution, with residual connections to preserve gradient flow. The decoder mirrors the encoder with 4 RNNBlocks and FupBlocks for upsampling, followed by a final convolutional layer for spectrogram reconstruction. Due to computational constraints, each FupBlock contains only a transposed convolution and a Snake2D activation.

The model is trained using a loss function that combines complex spectrogram loss, multi-scale Mel-spectrogram loss, and a GAN-based loss, where Multi-Period (MPD) and Multi-Resolution (MRD) discriminators [8] are employed to capture both fine-grained temporal details and spectral characteristics.

2.2. Training Stages

We employ a three-stage training pipeline. (1) A quantizer-free encoder–decoder model is trained exclusively for the denoising task, optimized solely with a complex spectral loss to establish a clean and stable representation space for subsequent quantization. (2) Quantizer Integration: A quantizer is then incorporated into the pre-trained denoising model, and the entire system is jointly optimized for both denoising and codec objectives while still employing the complex spectral loss. This staged integration stabilizes training by initializing the quantizer within a well-structured representation space, thereby maintaining denoising performance while enabling effective quantization. (3) Perceptual Fine-tuning: Finally, the model is fine-tuned to enhance the perceptual audio quality by replacing the loss function with a multi-scale Mel-spectrogram reconstruction loss and introducing adversarial objectives via MPD and MRD. This combination further improves the naturalness and fidelity of the reconstructed audio.

3. EXPERIMENTS

3.1. Datasets

The 2025 LRAC Challenge provides training datasets comprising speech, noise, and room impulse response (RIR) subsets. All datasets are first resampled to 24 kHz, followed by curation to form finalized training subsets. Specifically for the noise dataset, we utilize a pre-trained audio understanding model to predict audio labels, and further filter out "dirty" data samples bearing speech labels.

To further enhance the generalization capability of the model, training data is synthesized online using randomly sampled parameters at each training step. The data augmentation is detailed as follows: Noisy conditions are simulated by mixing speech and noise at a probability of 0.75, using a signal-to-noise ratio (SNR) uniformly sampled from the range of -5 to 15 dB. Reverberation is simulated by convolving the speech signal with a RIR at a probability of 0.4. For the corresponding target speech, the RIR undergoes truncation commencing 1 ms after its peak amplitude prior to convolution.

3.2. Implementation Details

The proposed model has a total computational complexity of 2595M FLOPs and 3.9M parameters. Specifically, the encoder together with the RVQ module accounts for 1997M FLOPs and 2.5M parameters, while the decoder requires 598M FLOPs and 1.4M parameters. The system is designed for a sampling rate of 24kHz, with a frame length of 720 samples and a frame shift of 312 samples. No future frames are utilized, resulting in an algorithmic latency of only 30ms. The RVQ module contains 6 codebooks, each with a size of 8192 entries (equivalent to 13 bits) and a vector dimension of 16. During inference, the RVQ can dynamically select between using 1 to 6 codebooks, enabling bitrate scalability from 1kbps to 6kbps.

Due to computational constraints, the channel configurations of the encoder and decoder are asymmetric. The FdownBlocks in the encoder have channel sizes of $[48, 144, 192, 288]$ with strides $[6, 5, 4, 3]$, while the upsampling layers in the decoder have channel sizes of $[24, 48, 124, 288]$.

The MPD employs period settings of $[2, 3, 5, 7, 11]$, while the MRD adopts $[3072, 1536, 768, 384, 206, 126, 78]$ as window sizes [9]. Additionally, The generator is trained with a learning rate of 3×10^{-4}, while the discriminator uses 1×10^{-4}. The Adam optimizer is employed throughout all training stages.

3.3. Results

Our evaluation employs Versa [10], the official evaluation toolkit recommended by the 2025 LRAC Challenge, which provides standardized implementations of metrics such as *sheet_ssqa*, *score_ref*, *audiobox AE_CE*, *UTMO*, and *PESQ*. Experiments are conducted under three acoustic conditions: clean, noisy, and reverberant. Leveraging the RVQ module, which supports variable-bitrate operation by selectively discarding codebooks during inference, we further assess the model's performance at 1 kbps and 6 kbps.

The evaluation results are summarized in Table 1, where the proposed method demonstrates significant improvements over the baseline across all metrics under all three acoustic conditions and at both 1 kbps and 6 kbps bitrates. The evaluation across different acoustic conditions reveals distinct characteristics of *Denoising Codec*. In clean scenarios, the model demonstrates superior performance in speech compression and reconstruction. For noisy and reverberant conditions, it exhibits strong robustness by effectively suppressing background noise and reverberation. However, the audio quality under reverberant conditions is somewhat compromised compared to other scenarios.

4. CONCLUSION

We presented a unified *Denoising Codec* that integrates speech coding and noise suppression in the frequency domain, enabling scalable bitrate via RVQ and delivering high perceptual quality across diverse acoustic conditions. The staged training strategy stabilizes optimization and enhances overall performance while meeting strict low-resource constraints. Future work will focus on further improving reconstruction quality under strict low-resource constraints.

5. REFERENCES

[1] Neil Zeghidour, Anatoly Luebs, Mohammad Omran, Jan Skoglund, and Marco Tagliasacchi, "Soundstream: An end-to-end neural audio codec," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021, vol. 30, pp. 495–507.

[2] Alexandre Défossez, Neil Zeghidour, Nicolas Usunier, and Gabriel Synnaeve, "High fidelity neural audio compression," *arXiv preprint arXiv:2210.13438*, 2022.

[3] Fabian Mentzer, Eirikur Agustsson, Michael Tschannen, Radu Timofte, Tim Salimans, and Marco Tagliasacchi, "Fully neural audio coding using variational autoencoders," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 14266–14276.

[4] Andreas Jansson, Eric Humphrey, Nicola Montecchio, Rachel Bittner, Aparna Kumar, and Till Weyde, "Singing voice separation with deep u-net convolutional networks," in *18th International Society for Music Information Retrieval Conference (ISMIR)*, 2017.

[5] Yongxin Hu, Yue Wang, Yao, et al., "Dccrn: Deep complex convolution recurrent network for phase-aware speech enhancement," in *INTERSPEECH*, 2020, pp. 2472–2476.

[6] Hendrik Schröter, Tim Gburrek, Thomas Appel, and Andreas Maier, "Deepfilternet: Lightweight speech enhancement for full-band audio," *arXiv preprint arXiv:2205.07847*, 2022.

[7] Sang gil Lee, Wei Ping, Boris Ginsburg, Bryan Catanzaro, and Sungroh Yoon, "Bigvgan: A universal neural vocoder with large-scale training," 2023.

[8] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," 2020.

[9] Julian D Parker, Anton Smirnov, Jordi Pons, CJ Carr, Zack Zukowski, Zach Evans, and Xubo Liu, "Scaling transformers for low-bitrate high-quality speech coding," 2024.

[10] Jiatong Shi, Hye jin Shim, Jinchuan Tian, Siddhant Arora, Haibin Wu, Darius Petermann, Jia Qi Yip, You Zhang, Yuxun Tang, Wangyou Zhang, Dareen Safar Alharthi, Yichen Huang, Koichi Saito, Jionghao Han, Yiwen Zhao, Chris Donahue, and Shinji Watanabe, "Versa: A versatile evaluation toolkit for speech, audio, and music," 2025.