LOW RESOURCE AUDIO CODEC CHALLENGE TRACK1: TRANSPARENCY CODEC

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

Anker Innovations, Beijing, China

ABSTRACT

We propose a frequency–time domain fusion audio codec for the 2025 Low-Resource Audio Coding (LRAC) Challenge, designed to meet strict constraints on complexity, latency, and bitrate while ensuring high quality and robustness. The system achieves 698 M FLOPs, 1.48 M parameters, and sub-30 ms latency, combining a frequency-domain encoder, Residual Vector Quantization (RVQ), and a time-domain decoder. Multi-Period and Multi-Resolution GANs jointly refine temporal and spectral fidelity. A multi-stage training process combines spectral reconstruction with adversarial objectives and noise-reduction strategies to ensure stable optimization and high-quality output. Evaluations at 1 kbps and 6 kbps in clean, noisy, and reverberant settings show consistent and significant gains over the baseline.

***Index Terms*—** speech codec, frequency–time domain fusion, low resource

1. INTRODUCTION

Speech interfaces have become essential in embedded systems, mobile devices, and other platforms with limited computational power or energy budgets. In such low-resource environments, speech codecs must deliver real-time processing while balancing complexity, bitrate, and latency, and still preserve high audio quality under noise and reverberation. While end-to-end neural audio coding has improved quality and compression efficiency, simultaneously achieving low complexity, low latency, low bitrate, and robustness in real acoustic conditions remains a major challenge.

The 2025 Low-Resource Audio Coding (LRAC) Challenge provides a stringent benchmark for this problem, with strict limits on complexity, latency, and bitrate, and a requirement for real-world operation. It serves both as a test of engineering capability and a driver for advances in integrated low-resource speech coding.

To address these demands, we propose a frequency–time domain fusion end-to-end audio codec for high-fidelity speech reconstruction under extreme resource constraints. The system combines frequency-domain encoding and time-domain decoding, augmented by a multi-stage training process, and noise-reduction techniques. These components jointly enhance transmission quality and fine-detail reproduction within tight computational and storage budgets, meeting LRAC's requirements for low latency, low bitrate, and high intelligibility, and delivering superior performance across diverse evaluation scenarios.

2. METHOD

2.1. Architecture

We propose a frequency–time domain fusion end-to-end audio codec that achieves high-quality speech transmission under strict resource constraints. The overall architecture, illustrated in Fig. 1, consists of a frequency-domain encoder, a residual vector quantizer (RVQ) [1, 2], and a time-domain decoder. The input audio is first transformed into an amplitude spectrogram via short-time Fourier transform (STFT). The frequency-domain encoder, built upon SpecTokenizer [3], employs a complex convolution layer followed by four cascaded FdownBlocks and RNNBlocks to extract and compress spectral features. Each FdownBlock combines a 2D convolution with Snake2D activation to enhance harmonic structure modeling, while each RNNBlock integrates FLNorm, Tanh, GRU, 2D convolution, and Snake2D activation, with residual connections to maintain stable gradient flow and preserve feature fidelity.

The latent representation is subsequently quantized by the RVQ module and passed to a BigCodec-based time-domain decoder [4]. This decoder comprises a 1D convolution, a unidirectional LSTM with residual connections, four sequential DecoderBlocks, Snake1D activation, an output 1D convolution, and Tanh activation. Each DecoderBlock contains Snake1D activation [5], a 1D transposed convolution for upsampling, and several ResidualBlocks. Each ResidualBlock consists of two 1D convolutions with different kernel sizes and Snake1D activations, coupled with a residual connection at the end, thereby improving high-frequency detail restoration and spatial perceptual quality in waveform reconstruction.

Model training adopts a multi-objective loss function, including multi-scale mel-spectrogram loss, VQ quantization loss, and GAN-based adversarial loss. During adversarial training, a Multi-Period Discriminator (MPD) and Multi-

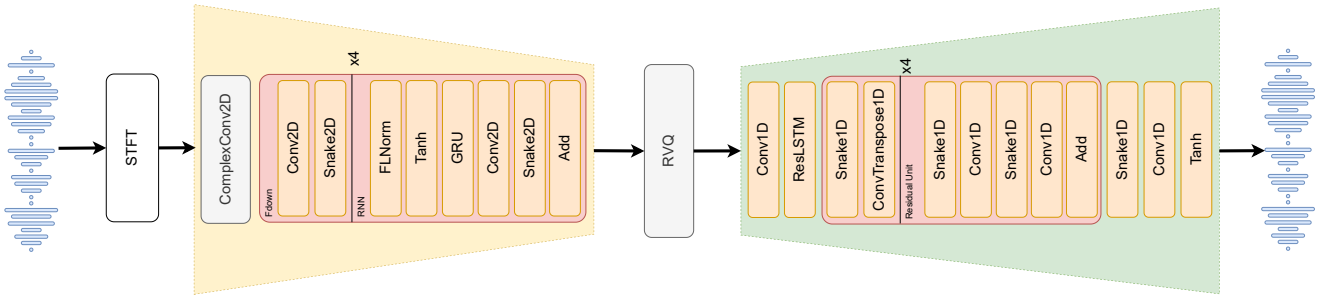

Fig. 1. The proposed model architecture.

Resolution Discriminator (MRD) [6] are employed jointly to constrain both time-domain details and spectral textures, significantly enhancing mid-to-high frequency energy reproduction and naturalness. As a result, the proposed system delivers high-fidelity speech reconstruction that combines audio quality and intelligibility under low-latency and low-bitrate conditions.

2.2. Training Stages

We first trained the codec without noise-reduction to obtain a performance-stable baseline model, and then introduced a noise-reduction stage after convergence. Although the competition rules explicitly state that noise-reduction features neither yield additional credit nor incur penalties in evaluation, our experiments show that incorporating this stage significantly improves speech quality in real acoustic environments. Consequently, we consider noise-reduction training an essential component of system optimization.

The codec training process consists of two parts: a Mel stage and a GAN stage. In the Mel stage, only the multi-scale mel-spectrogram loss is used for optimization. The model converges rapidly in this stage and achieves excellent reconstruction in the low-frequency range (0–1.5 kHz), with correspondingly high objective scores. However, because the mel loss provides insufficient constraint in the mid-to-high frequency range, the generated audio above 1.5 kHz often exhibits blurred spectral detail, energy attenuation, and slight mechanical artifacts, affecting subjective naturalness. To address this issue, we switch to the GAN stage after Mel-stage convergence, leveraging both the Multi-Period Discriminator (MPD) and Multi-Resolution Discriminator (MRD) for adversarial training. This significantly enhances mid-to-high frequency detail restoration, produces spectral energy distributions closer to natural speech, and effectively reduces mechanical noise. While some objective metrics (e.g., PESQ and Scoreq) degrade slightly in this stage, subjective ratings improve markedly, with richer spatial perception and more natural fine detail.

During training, we observed an interesting phenomenon: after several cycles in the GAN stage, returning to the Mel stage for further optimization causes objective scores not only to recover but to exceed the best results of the initial Mel stage. This may be because the GAN stage encourages the generator to explore a broader solution space, providing the mel loss with a better optimization starting point and helping the model escape local minima.

In the noise-reduction training stage, the input data comprise a random mix of clean, noisy, and reverberant speech, with the target output being the corresponding clean speech. The loss functions and hyperparameters are kept identical to those in codec training, and adversarial learning is again applied to further improve the realism and richness of generated audio. The discriminator configuration follows a staged policy: MPD alone in the early phase to strengthen time-domain periodicity discrimination; MPD plus MRD in the mid phase to impose multi-resolution spectral constraints; and MRD alone in the late phase to focus optimization on spectral detail restoration. Subjective listening tests indicate that this configuration yields the best improvements in mid-to-high frequency clarity, spectral extension, and overall intelligibility, producing speech more closely resembling real recordings.

3. EXPERIMENTS

3.1. Datasets

All training data in this study are sourced from the official LRAC2025 dataset and underwent rigorous filtering and preprocessing prior to use. For noise data, labels were predicted using a pre-trained audio understanding model, and any non-pure noise samples containing speech were removed to ensure clean noise content. For reverberation data, room impulse responses (RIRs) were truncated before convolution, retaining only the 1 ms segment following the peak. This reduces long-tail decay that can impair speech clarity while preserving spatial characteristics.

Based on this, we applied a data augmentation strategy by mixing clean, noisy, and reverberant speech in a 1:1:1 ratio. In noise mixing, the signal-to-noise ratio (SNR) was uniformly sampled within the range of 10–30 dB to increase acoustic

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	1.84	1.15	3.90	1.44	1.15	1.72	1.29	3.40	1.33	1.11	1.85	1.36	2.94	1.26	1.07
	Proposed	3.79	0.35	5.31	3.42	2.09	3.65	0.38	5.18	3.32	1.92	2.80	0.59	4.53	2.58	1.46
6kbps	Baseline	3.84	0.35	5.28	3.23	2.67	3.12	0.82	4.37	2.70	1.81	2.22	1.13	3.43	1.32	1.18
	Proposed	4.17	0.18	5.62	3.77	2.98	3.99	0.30	5.45	3.64	2.50	3.14	0.53	4.75	2.74	1.62

Table 2. Latency breakdown of the proposed system.

Source	**Samples**	**Notes**
STFT hopsize	192 @ 16kHz	Frame shift
Decoder Residual Units	272 @ 16kHz	$64 \times 3 + 16 \times 4 + 4 \times 4 + 1 \times 5$
Final decoder convolution	3 @ 16kHz	Kernel size = 7
Resampling delay	8 @ 24kHz	Maximum group delay of the IIR filter
Total (24kHz)	716 @ 24kHz (29.83 ms)	$472 \times \frac{3}{2} + 8$

diversity.

Model evaluation was conducted on an open test set from the same source, with inference performed directly on the original official data without additional processing, and performance tested at both 1 kbps and 6 kbps bitrates.

3.2. Implementation Details

The proposed model has an overall computational complexity of 698 M FLOPs and 1.48 M parameters, with the encoder and RVQ module accounting for 399 M FLOPs and 1.17 M parameters, and the decoder for 299 M FLOPs and 0.32 M parameters. The system operates at a sampling rate of 24 kHz, with a frame length of 720 samples and a frame shift of 288 samples (approximately 83 Hz frame rate). In the STFT computation, only frequency bins 0–240 (0–8kHz) are used, effectively yielding a 24kHz to 16 kHz downsampling without introducing additional latency.

The encoder employs convolution kernels and strides of 1, introducing no additional latency. The decoder primarily uses causal convolutions and causal transposed convolutions, but non-causal convolutions are applied in specific positions to enhance reconstruction quality: the first convolution layer in the decoder (kernel = 1, stride = 1), the first convolution layer within repeated ResidualBlocks (kernel sizes = [7, 9, 9, 11], stride = 1), and the final convolution layer in the decoder (kernel = 7, stride = 1). These designs significantly improve mid-to-high frequency detail within the latency budget. The end-to-end latency is determined by both the STFT window length and the non-causal convolutions, and is kept within 30 ms overall.

To convert the 16 kHz audio output of the decoder to 24 kHz without noticeably increasing latency, we use a fractional-rate resampling strategy. First, the signal is upsampled by a factor of three using zero-insertion. Next, the spectral images introduced by zero-insertion are removed with an 11th-order IIR Butterworth low-pass filter with an 8 kHz cutoff frequency. Finally, the signal is downsampled by a factor of two to reach the target sampling rate. Compared to an FIR-based approach, this IIR design exhibits a maximum passband group delay of only 8 samples near 8 kHz, making it well-suited for real-time applications. The latency breakdown is shown in Table 2.

The RVQ module consists of six codebooks, each containing 4096 entries (indexed with 12-bit codes) and a vector dimension of 8. During inference, either 1 codebook (for 1 kbps) or all 6 codebooks (for 6 kbps) can be selected, enabling operation at two different bitrates. The encoder channel configuration is [32, 32, 32, 128, 335], with time-axis kernel sizes and strides of [1, 1, 1, 1] and frequency-axis kernel sizes and strides of [5, 4, 4, 3]. The decoder channels are [117, 58, 29, 14, 7], with upsampling rates of [3, 4, 4, 4]. For the discriminators, the MPD uses periods [2, 3, 5, 7, 11], and the MRD operates with window sizes [128, 256, 512, 1024, 2048].

For optimization, both the generator and discriminators use an initial learning rate of 8×10^{-4} during the Mel stage and 1×10^{-4} during the GAN stage, gradually reduced to 1×10^{-5}. Adam is used throughout all training stages.

Checkpoint Selection Strategy: For system submission, we

performed subjective listening evaluations on multiple models from different training stages using the open test set, selecting the checkpoint that yielded the best combination of audio quality and fine-detail reproduction as the final competition version.

3.3. Results

Our evaluation uses Versa [7], the official toolkit recommended by the 2025 LRAC Challenge, which provides standardized implementations of multiple metrics, including *sheet_ssqa*, *scoreq_ref*, *audiobox AE_CE*, *UTMOS*, and *PESQ*. Experiments are conducted under three acoustic conditions: clean, noisy, and reverberant. Using the RVQ module's ability to achieve variable bitrate by selectively dropping codebooks during inference, we further evaluate the model at 1 kbps and 6 kbps.

The evaluation results are summarized in Table 1. Under all three acoustic conditions and both bitrates, the proposed method outperforms the baseline system across all metrics.

4. CONCLUSION

We propose a frequency–time domain fusion end-to-end codec for low-resource audio coding, combining iterative optimization with noise-reduction to enhance quality and robustness across diverse acoustic conditions and bitrates. Exploiting the complementarity of frequency-domain encoding and time-domain decoding, the system achieves high-fidelity speech reconstruction within strict complexity and latency limits. Experiments demonstrate consistent gains over the baseline in clean, noisy, and reverberant settings, confirming the effectiveness of the approach and its potential for more complex scenarios.

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] Zixiang Wan, Guochang Zhang, Yifeng He, and Jianqiang Wei, "SpecTokenizer: A lightweight streaming codec in the compressed spectrum domain," in *Proc. Interspeech 2025*, 2025, pp. 599–603.

[4] Detai Xin, Xu Tan, Shinnosuke Takamichi, and Hiroshi Saruwatari, "BigCodec: Pushing the limits of low-bitrate neural speech codec," *arXiv preprint arXiv:2409.05377*, 2024.

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

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

[7] 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.