A LOW-LATENCY VQ-GAN-BASED CODEC WITH KNOWLEDGE DISTILLATION FOR JOINT SPEECH CODING AND ENHANCEMENT

Yang Xu[1,2,†], *Ronghui Hu*[1,2,†], *Leyan Yang*[1,2,†], *Jing Lu*[1,2,*]

[1]Key Laboratory of Modern Acoustics, Nanjing University, Nanjing, 210093, Jiangsu, China
[2]NJU-Horizon Intelligent Audio Lab, Horizon Robotics, Beijing, 100094, Beijing, China

ABSTRACT

The advancement of speech interfaces operating in resource-constrained environments drives the need for neural speech codecs that achieve a critical balance among computational efficiency, minimized bitrate, and low latency. These codecs must also maintain high speech quality under challenging acoustic conditions, integrating robust enhancement capabilities to counteract real-world noise and reverberation. To address these challenges, we present KD-Vocodec, an efficient knowledge distillation (KD) framework for joint speech coding and enhancement. The proposed system achieves superior performance by training a student model to replicate the intermediate representations of a high-fidelity teacher model through feature-level knowledge distillation, thereby delivering high-quality audio at a latency of 30 ms and scalable bitrates from 1 to 6 kbps. Rigorous evaluation on a public test set confirms the superior capability of KD-Vocodec.

***Index Terms*—** neural speech codec, knowledge distillation, speech enhancement

1. INTRODUCTION

The deployment of neural speech codecs on devices with constrained resources requires balancing critical trade-offs between bitrate, computational complexity, latency, and robustness to acoustic noise. The 2025 Low-Resource Audio Codec (LRAC) Challenge focuses on this problem, calling for codecs that perform effectively under realistic and noisy conditions. Motivated by this challenge, a novel framework called KD-Vocodec is proposed in this paper for joint speech coding and enhancement. Its key innovation is a feature-level knowledge distillation technique, which enables the system to learn compact and noise-invariant representations. The resulting codec achieves a low algorithmic latency of 30 ms and supports variable bitrates, delivering enhanced performance without a significant increase in computational complexity.

2. PROPOSED METHOD

Our proposed framework leverages feature-level knowledge distillation to achieve joint speech coding and enhancement under strict latency and bitrate constraints. The system architecture, depicted in Fig.1, is built upon a VQ-GAN-based [1] clean teacher codec. The overarching design employs a teacher-student paradigm wherein a student encoder is trained to replicate the intermediate representations of a pre-trained teacher encoder, facilitating the learning of clean features. However, the final system retains the original decoder weights without fine-tuning.

2.1. Teacher codec architecture

We adopt Vocos [2] as the backbone of our teacher codec architecture, due to its superior performance in speech synthesis. Specifically, a mirrored variant of the Vocos structure is employed as the encoder–decoder backbone. The input waveform is first converted into a time–frequency representation via STFT. The complex spectrogram is split into magnitude and phase components, which are concatenated along the frequency dimension and fed into the network. This combined input is projected into a latent space with dimension D via a linear layer. The encoder consists of multiple convolutional blocks inspired by ConvNeXt [3], aiming to extract deep hierarchical features. Each block contains a 1D depthwise convolution with weight normalization, followed by a pointwise convolution. To ensure strict causality and avoid algorithmic delay, all temporal padding is causal. To enhance sequence modeling, ResNet blocks are incorporated. Inspired by WavTokenizer [4], a causal self-attention mechanism is inserted after the second convolutional block. The resulting features are passed to the quantizer, which uses a Residual Vector Quantizer (RVQ) [5] with 6 layers and gradient-based codebook updates. Linear layers before and after quantization map features between the quantization dimension and a lower-dimensional space.

The encoder configuration is as follows: STFT window size is 720 samples with a hop size of 180; hidden dimension D is 256; the encoder stack contains 12 ConvNeXt layers, each with an expansion channel size of 896. The decoder mirrors the encoder's structure but with reduced capacity to meet receiver-side computational constraints: D is 252, the number of ConvNeXt layers is 4, and the expansion channel size is 256. The projection dimension for RVQ is set to 8.

†Equal contribution.
*Corresponding author: lujing@nju.edu.cn

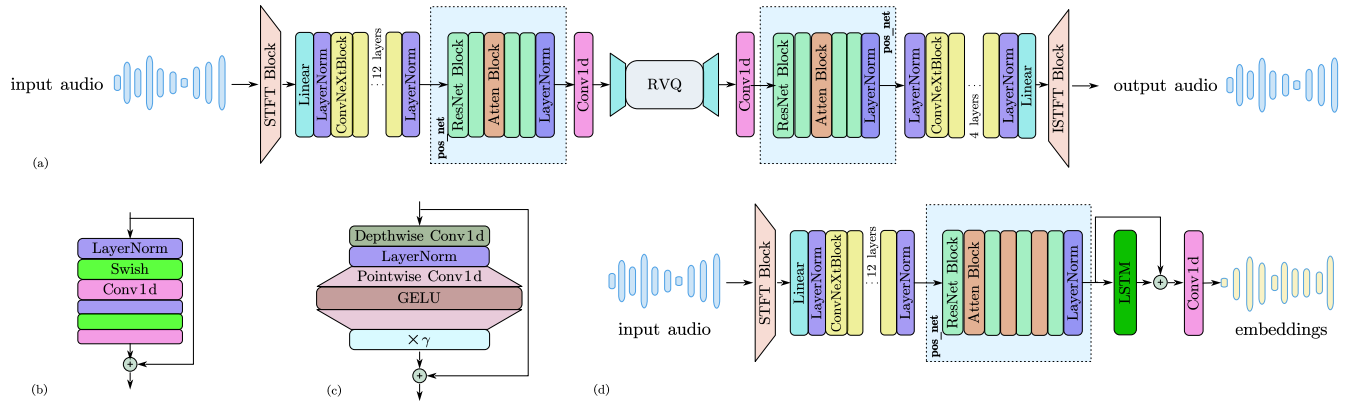

Fig. 1. Architecture of the proposed KD-Vocodec framework. (a) Overall pipeline of the teacher codec; (b) Detailed structure of the ResNet Block; (c) Detailed structure of the ConvNeXt Block; (d) Architecture of the student encoder.

2.2. Student encoder

The student encoder is designed by augmenting the encoder with several key components. This design is motivated by the hypothesis that these augmentations will enable a more robust derivation of clean embeddings from distorted speech inputs. Specifically, causal self-attention modules are incorporated after each ResNet block, except for the final one, to capture long-range contextual dependencies under causal constraints. Furthermore, a two-layer LSTM layer is introduced immediately preceding the final convolutional layer to enhance temporal sequence modeling. A skip connection is also employed between the input and output of this LSTM to facilitate gradient flow and preserve fine-grained temporal information.

2.3. Discriminator

Given that the input to our model is derived from the STFT time-frequency representation, it is advantageous to employ a Multi-Scale STFT Discriminator (MSSTFTD) [6] to assess the reconstruction quality directly in the spectral domain. A set of window lengths [128, 256, 512, 1024, 2048] is used, and the hop length is fixed to one-fourth of the window length. Accordingly, we introduce adversarial training solely using the MSSTFTD to refine the output of the teacher codec.

2.4. Loss function

The training of the teacher codec utilizes a composite loss function within a GAN framework. The total generator loss, $L_{generator}$, is a weighted sum of multiple components: the multi-scale mel-spectrogram reconstruction loss L_{rec} [7], the generator adversarial loss L_g, the feature matching loss L_{feat} applied to the discriminator's features, the codebook loss L_{code}, and the commitment loss L_c. It is formulated as:

$$L_{rec} = \|\mathcal{M}(x) - \mathcal{M}(\hat{x})\|_1 \tag{1}$$

$$L_g = \|1 - D(\hat{x})\|_2^2 \tag{2}$$

$$L_{feat} = 2\sum_l \left\|D^l(x) - D^l(\hat{x})\right\|_1 \tag{3}$$

$$\begin{aligned} L_{\text{generator}} =& \lambda_{\text{rec}} L_{\text{rec}} + \lambda_g L_g + \lambda_{\text{feat}} L_{\text{feat}} \\ &+ \lambda_{\text{code}} \underbrace{\|\text{sg}[\mathbf{z}_e] - \mathbf{e}_k\|_2^2}_{L_{\text{code}}} + \lambda_c \underbrace{\|\mathbf{z}_e - \text{sg}[\mathbf{e}_k]\|_2^2}_{L_c} \end{aligned} \tag{4}$$

In the above equations, x and $\hat{x}$ denote the target and reconstructed speech, respectively, $\mathcal{M}(\cdot)$ is the mel-spectrogram transform, $D(\cdot)$ is the discriminator output, $D^l(\cdot)$ represents the feature map of the l-th discriminator layer, $\mathbf{z}_e$ is the quantizer output, and $\mathbf{e}_k$ is the codebook vector. The multi-scale mel-spectrogram loss L_{rec} is computed using window length samples [32, 64, 128, 256, 512, 1024, 2048], with the hop length fixed at $1/4$ of each window length. Each scale uses different mel bins of [5, 10, 20, 40, 80, 160, 320]. Loss weights are set as: $\lambda_{\text{rec}} = 15$, $\lambda_g = 2$, $\lambda_{\text{feat}} = 1$, $\lambda_{\text{code}} = 1$, $\lambda_c = 0.25$. The discriminator is trained separately with the adversarial loss L_d.

$$L_d = \|1 - D(x)\|_2^2 + \|D(\hat{x})\|_2^2 \tag{5}$$

For knowledge distillation in the student encoder, the loss combines the MSE and cosine distance between the teacher and student embeddings, weighted by 1.0 and 0.1, respectively.

3. EXPERIMENTAL SETUP

3.1. Training data preparation

We trained our model on a large-scale speech dataset curated from high-quality speech samples obtained from the EARS, VCTK, Common Voice, LibriTTS, Multilingual LibriSpeech datasets, and DNS Challenge 5 dataset. All speech signals are resampled to 24 kHz. To extend the noise suppression and dereverberation capabilities of the model, we further constructed a noise data set that includes noise from the VCTK, WHAM, FSD50K, and FMA datasets, encompassing various noise types. During training, each speech sample is combined with background noise with an 80% probability, where signal-to-noise ratios (SNRs) are uniformly distributed between -5 dB and 30 dB. For reverberation, we use room impulse responses (RIRs) from the Motus dataset, and each sample is

Table 1. Objective Performance Comparison on the Open Test Set

Bitrate	Model	Condition	ScoreQ-ref	UTMOS	Sheet-SSQA	PESQ	Audiobox AE-CE
6 kbps	Baseline	Clean	0.43	2.97	3.55	2.13	5.25
		Noisy	0.75	2.56	2.92	1.73	4.6
		Reverb	0.92	1.79	2.67	1.29	4.25
	Proposed	Clean	**0.15**	**3.74**	**4.26**	**3.22**	**5.69**
		Noisy	**0.40**	**3.36**	**3.73**	**2.23**	**5.29**
		Reverb	**0.48**	**3.08**	**3.51**	**1.80**	**5.27**
1 kbps	Baseline	Clean	1.01	1.37	2.07	1.21	3.96
		Noisy	1.15	1.35	1.95	1.18	3.7
		Reverb	1.12	1.32	2.43	1.15	3.55
	Proposed	Clean	**0.38**	**3.26**	**3.60**	**1.94**	**5.37**
		Noisy	**0.53**	**3.00**	**3.30**	**1.61**	**5.14**
		Reverb	**0.62**	**2.74**	**3.06**	**1.43**	**5.01**

augmented with reverberation with a probability of 50% during training. All training data follow the cleaning and preprocessing procedures defined in the baseline[1].

3.2. Implementation Details

Notably, our approach avoids using any pre-trained models throughout the training and inference pipeline. The training procedure consists of two distinct stages. The first stage involves training the teacher codec using a GAN-based reconstruction objective. This model is trained for 1000 epochs with a batch size of 128, using the AdamW optimizer with a cosine annealing learning rate scheduler. In the subsequent distillation stage, the student encoder is trained to replicate the teacher's embeddings. This stage runs for 500 epochs with a batch size of 384, optimized by RAdam with an exponential decay scheduler.

3.3. Computational complexity

The teacher codec operates with 1.11G multiply–accumulate operations per second (MACs) computational complexity (with the decoder accounting for 281.57M MACs) and contains 11.07M parameters. By integrating the student encoder (979.18M MACs), the complete system achieves a complexity of 1.28G MACs with 12.65M parameters. The system maintains strict causality without look-ahead. Consequently, the algorithmic latency is determined solely by the 30-ms STFT analysis window at a 24 kHz sampling rate. The system supports variable bitrates via its RVQ module, where each quantizer layer provides approximately 1 kbps (using a 1024-codebook at 100 fps), allowing operational modes of 1 kbps (1-layer) and 6 kbps (6-layers).

3.4. Checkpoint selection strategy

We employ a systematic strategy for selecting the final model checkpoint. The validation objective metrics are evaluated at regular intervals during training. Should a consistent and pronounced degradation in these metrics be observed, the early stopping strategy is triggered, and the checkpoint with the best performance up to that point is selected. Otherwise, the model checkpoint achieving the lowest training loss at the end of the training process was chosen as the final model.

[1] https://github.com/cisco-open/lrac_data_generation

4. EVALUATION RESULTS

The proposed approach is evaluated using the official challenge metrics and compared against the official baseline system [8]. As shown in Table 1, the KD-Vocodec framework demonstrates consistent performance improvements at both operational bitrates of 1 kbps and 6 kbps. The evaluation employs five objective metrics—ScoreQ_ref [9], UTMOS [10], Sheet-SSQA [11], PESQ [12], and Audiobox Aesthetics_CE [13]—selected for their high correlation with subjective quality assessments, as confirmed by Pearson correlation analysis, thereby providing a reliable measure of decompressed speech quality.

5. CONCLUSION

This paper introduces KD-Vocodec, our submission to Track 2 of the 2025 Low-Resource Audio Codec (LRAC) Challenge. Experimental evaluations demonstrate that KD-Vocodec delivers superior performance over the baseline under diverse acoustic conditions. The system provides an effective solution for real-world speech coding applications that require efficient processing on resource-constrained devices.

6. REFERENCES

[1] Patrick Esser, Robin Rombach, and Bjorn Ommer, "Taming transformers for high-resolution image synthesis," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 12873–12883.

[2] Hubert Siuzdak, "Vocos: Closing the gap between time-domain and fourier-based neural vocoders

for high-quality audio synthesis," *arXiv preprint arXiv:2306.00814*, 2023.

[3] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie, "A convnet for the 2020s," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 11976–11986.

[4] Shengpeng Ji, Ziyue Jiang, Wen Wang, Yifu Chen, Minghui Fang, Jialong Zuo, Qian Yang, Xize Cheng, Zehan Wang, Ruiqi Li, et al., "Wavtokenizer: an efficient acoustic discrete codec tokenizer for audio language modeling," *arXiv preprint arXiv:2408.16532*, 2024.

[5] Biing-Hwang Juang and A Gray, "Multiple stage vector quantization for speech coding," in *ICASSP'82. IEEE International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 1982, vol. 7, pp. 597–600.

[6] Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi, "High fidelity neural audio compression," *arXiv preprint arXiv:2210.13438*, 2022.

[7] 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, pp. 27980–27993, 2023.

[8] Yusuf Ziya Isik and Rafał Łaganowski, "Low resource audio codec challenge baseline systems," *arXiv preprint arXiv:2510.00264*, 2025.

[9] Alessandro Ragano, Jan Skoglund, and Andrew Hines, "Scoreq: Speech quality assessment with contrastive regression," *Advances in Neural Information Processing Systems*, vol. 37, pp. 105702–105729, 2024.

[10] Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari, "Utmos: Utokyo-sarulab system for voicemos challenge 2022," *arXiv preprint arXiv:2204.02152*, 2022.

[11] Wen-Chin Huang, Erica Cooper, and Tomoki Toda, "Mos-bench: Benchmarking generalization abilities of subjective speech quality assessment models," *arXiv preprint arXiv:2411.03715*, 2024.

[12] Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*. IEEE, 2001, vol. 2, pp. 749–752.

[13] Andros Tjandra, Yi-Chiao Wu, Baishan Guo, John Hoffman, Brian Ellis, Apoorv Vyas, Bowen Shi, Sanyuan Chen, Matt Le, Nick Zacharov, et al., "Meta audiobox aesthetics: Unified automatic quality assessment for speech, music, and sound," *arXiv preprint arXiv:2502.05139*, 2025.