PROGRESSIVE REFINEMENT TRAINING FOR LOW-RESOURCE NEURAL SPEECH CODING AND ENHANCEMENT

Ronghui Hu[1,2,†], *Leyan Yang*[1,2,†], *Yang Xu*[1,2,†], *Qinwen Hu*[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, China

ABSTRACT

Speech codec is a key challenge in hands-free communication systems, where on-device deployment requires real-time processing under strict constraints on bitrate and computational complexity. Meanwhile, real-world acoustic conditions demand integrated speech enhancement (SE). In this paper, we propose a novel Progressive Refinement (PR) strategy to build a high-performance codec for joint speech coding and enhancement. With this strategy, we introduce PR-Vocodec, a low-latency, high-fidelity, and low-bitrate codec, which can perform noise reduction and dereverberation simultaneously with low computational overhead. Experimental results demonstrate that the PR-Vocodec delivers superior performance across multiple evaluation metrics.

Index Terms— progressive refinement, audio neural codec, speech enhancement.

1. INTRODUCTION

The 2025 Low-Resource Audio Codec (LRAC) Challenge focuses on codecs with low computational complexity, low latency, and low transmission bandwidth, as well as multi-task codecs coupled with front-end enhancement tasks. In this paper, we introduce PR-Vocodec, our system submitted to the Challenge. The system is built upon the Vocos architecture [1] and employs a six-layer Residual Vector Quantizer (RVQ) [2] in the quantization module, supporting both 1 kbps and 6 kbps bitrates. The training follows a three-stage progressive refinement (PR) strategy. Stage 1 focuses on constructing a high-fidelity teacher model. Stages 2 and 3 progressively train the student model, enhancing its noise suppression and dereverberation capabilities. This progressive refinement framework not only preserves the quality of the codebooks but also significantly improves the speech enhancement performance and generalization of the student model under low-bitrate constraints.

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

2. PROPOSED METHOD

2.1. Codec architecture

As illustrated in Fig.1, we design the backbone architecture based on the Vocos [1] framework and employ it as the decoder. The decoder consists of six 1D ConvNeXt [3] blocks with a hidden dimension of 558, followed by a post-processing network comprising four ResNet blocks and a causal self-attention module [4]. The encoder is constructed as a mirror-symmetric counterpart of the decoder, performing feature extraction of the input speech at the transmitting end through a reversed information flow. Since the encoding stage is coupled with the SE task, we adopt an asymmetric parameter configuration to enhance the encoder's feature extraction and multi-task processing capabilities. Specifically, the encoder consists of twelve 1D ConvNeXt blocks with the hidden dimension increased to 1096. In addition, we employ an RVQ module to encode the embeddings extracted by the encoder. The RVQ consists of six quantization layers, each with a codebook size of 1024.

2.2. PR training strategy

The PR strategy enables the model to achieve high-fidelity audio coding while simultaneously performing high-quality speech enhancement, including noise suppression and dereverberation. As illustrated in Fig. 2, the training process consists of three progressive stages.

In Stage 1, the model follows the standard audio codec training paradigm to obtain a low-bitrate, high-fidelity codec, which serves as the teacher model. The training process adopts a generative adversarial network (GAN) framework, where a multi-scale short-time Fourier transform discriminator (MS-STFTD) [5] is employed to impose multi-scale time–frequency constraints on the reconstructed audio, thereby enhancing the accuracy in frequency band reconstruction.

In Stage 2, the encoder of the student model is trained from scratch to perform joint coding and enhancement. Specifically, the clean speech is fed into the encoder of the teacher model to generate target embeddings, while

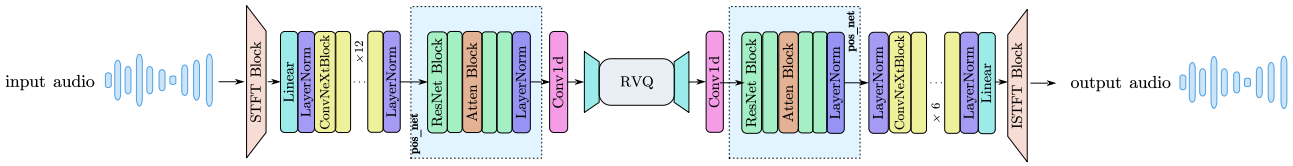

Fig. 1. An overview of the proposed PR-Vocodec backbone.

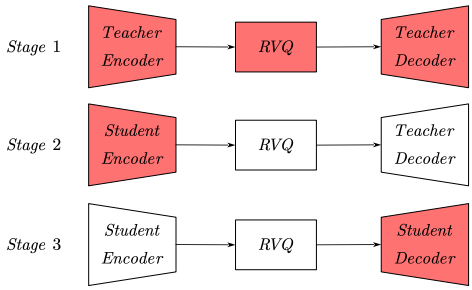

Fig. 2. The Schematic of the PR training strategy. The red blocks are updated during the training stage, while the white blocks are frozen.

the noisy and reverberant speech is passed through the student encoder. During training, the two sets of embeddings are aligned, guiding the student encoder to produce representations that closely match those of the teacher model when processing clean inputs. This alignment effectively implements noise and reverberation suppression within the encoder module. Crucially, the codebook remains frozen throughout this stage, ensuring that the decoder's input space remains consistent with that of the teacher model.

Stage 3 is the dual process of Stage 2, aiming to enhance the robustness of the student decoder and thereby improve the system's generalization to noisy or reverberant inputs. During this stage, the encoders and RVQ modules of both teacher and student models are frozen. Clean and noisy speech pairs are processed in parallel, aligning the decoder outputs and the reconstructed waveforms to promote consistent decoding behavior. This process ensures the output of the student decoder closely approximates the output of the teacher decoder for clean speech, enhancing the robustness against variations in encoder output. Furthermore, adversarial training is incorporated in this stage by employing the discriminator pre-trained during Stage 1 to refine the decoder outputs of the student model. To balance the training progress between the decoder and the discriminator, we update the decoder five times for each discriminator update to ensure balanced convergence and maintain stable adversarial training.

2.3. 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} [6], 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. $\text{sg}[\cdot]$ denotes the stop-gradient operation, indicating that its gradients are detached from the computation graph and do not participate in backpropagation. 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 with adversarial loss L_d, which is formulated as:

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

In Stage 2, the loss function $L_{PR-encoder}$ combines the mean squared error (MSE) and cosine distance between the teacher and student embeddings, weighted by 1.0 and 0.2, respectively.

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	0.435	2.972	3.548	2.126	5.381
		Noisy	0.753	2.562	3.122	1.723	4.754
		Reverb	0.913	1.803	3.273	1.295	4.381
	Stage 2	Clean	0.164	3.790	3.917	3.215	5.786
		Noisy	0.348	3.594	3.706	2.428	5.540
		Reverb	0.364	3.517	3.883	2.092	5.597
	Stage 3	Clean	**0.158**	**3.785**	**3.929**	**3.244**	**5.795**
		Noisy	**0.317**	**3.613**	**3.755**	**2.444**	**5.592**
		Reverb	**0.340**	**3.560**	**3.890**	**2.116**	**5.659**
1 kbps	Baseline	Clean	1.008	1.371	2.079	1.207	4.163
		Noisy	1.150	1.351	2.520	1.180	3.918
		Reverb	1.117	1.323	3.065	1.153	3.723
	Stage 2	Clean	0.386	**3.306**	3.609	1.959	5.470
		Noisy	0.470	3.236	3.537	1.753	5.370
		Reverb	0.466	3.202	**3.666**	1.657	**5.392**
	Stage 3	Clean	**0.364**	3.305	**3.648**	**1.991**	**5.490**
		Noisy	**0.463**	**3.242**	**3.541**	**1.786**	**5.383**
		Reverb	**0.465**	**3.211**	3.626	**1.674**	5.391

In Stage 3, the outputs of the decoder's final hidden layer are optimized using the same loss function as in Stage 2, denoted as $L_{PR-decoder}$. Meanwhile, the decoded speech is trained with the same loss formulation as in Stage 1, denoted as $L_{generator}$. The overall training objective for the decoder at this stage is therefore given by:

$$L_{stage-3} = L_{PR-decoder} + L_{generator}. \quad (6)$$

3. EXPERIMENTAL SETUP

3.1. Training data preparation

In Stage 1, the teacher model is trained on the EARS, VCTK, Common Voice, LibriTTS, Multilingual LibriSpeech, and DNS Challenge 5 datasets. All speech data are resampled to 24 kHz. In Stages 2 and 3, we extend the student model's capability in noise suppression and dereverberation by constructing an additional noise dataset derived from VCTK, WHAM, FSD50K, and FMA, covering a diverse range of noise types. During training, each clean speech sample is mixed with background noise with a probability of 80%, where the signal-to-noise ratio (SNR) is uniformly sampled between -5 dB and 30 dB. To simulate reverberant conditions, room impulse responses (RIRs) from the Motus dataset are applied, with each sample augmented with reverberation at a probability of 50%. All training data are processed following the cleaning and preprocessing procedures specified in the official baseline[1]

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

3.2. Implementation Details

In Stage 1, the teacher model is trained for 1000 epochs with a batch size of 192, using the AdamW optimizer with a cosine annealing learning rate scheduler. In Stage 2, the student encoder is trained to replicate the teacher model's embeddings. This stage runs for 500 epochs with a batch size of 40, optimized by RAdam with an exponential decay scheduler. In Stage 3, the student decoder is trained for 200 epochs with a batch size of 192 and optimized by the AdamW optimizer with an exponential learning rate decay scheduler.

3.3. Computational complexity and latency

The computational complexity of the teacher model is 349.29M multiply–accumulate operations per second (MACs/s)[2] (with the decoder accounting for 281.57M MACs/s) and the model contains 3.47M parameters. The overall student model comprises 12.37M parameters and operates with a computational complexity of 1.25G MACs/s (with the decoder accounting for 281.29M MACs)

The teacher model incurs an algorithmic latency of 30 ms due to the 720-point STFT. In contrast, the student model has a total latency of 50 ms, comprising the same 30 ms algorithmic latency and an additional 20 ms buffering latency introduced in the encoder.

[2] The computational complexity is calculated by ptflops: `https://github.com/tel-0s/ptflops`.

4. RESULTS

Evaluation on the open test set is conducted using the official metrics provided by the challenge, which contain five objective metrics: ScoreQ_ref [7], UTMOS [8], Sheet-SSQA [9], PESQ [10], and Audiobox Aesthetics_CE [11].[3] Experimental results are summarized in Table 1. The results show that PR-Vocodec significantly outperforms the baseline across all scenarios at both bitrates, particularly demonstrating strong robustness and generalization for reverberant data. Furthermore, the comparison between Stage 2 and Stage 3 shows that the decoder retraining enhances the model's adaptability and consistency, thereby validating the effectiveness of the PR training strategy in achieving high-fidelity speech coding with strong enhancement capability.

5. CONCLUSION

This paper introduces our proposed PR training strategy designed for joint speech coding and enhancement tasks, and our PR-Vocodec model submitted to the LRAC Challenge. The proposed approach achieves competitive performance in the LRAC challenge, surpassing the baseline by a large margin across different bitrates and input conditions.

6. REFERENCES

[1] 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.

[2] 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.

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

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[3] Sheet-SSQA and Audiobox AE-CE scores show some deviations from the official results provided by the LRAC challenge.