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*

[2]*NJU-Horizon Intelligent Audio Lab, Horizon Robotics*

Email: ronghui.hu@smail.nju.edu.cn

Demo

LRAC 2025 CHALLENGE

ICASSP2026 Barcelona

INTRODUCTION

Background & Challenges

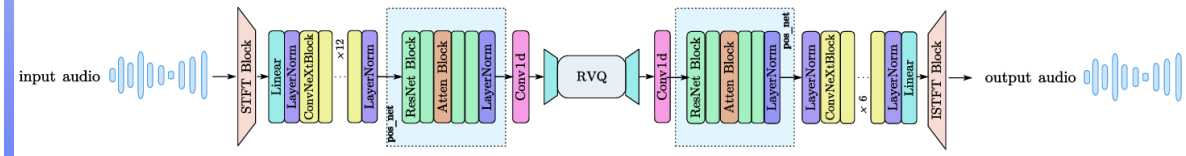
- ⚠️ **Resource Constraints:** On-device neural speech coding requires real-time processing under strict bitrate and compute limits.
- 🌪️ **Acoustic Interference:** Real-world noise and reverberation severely degrade perceptual quality, demanding integrated Speech Enhancement (SE).

PR-Vocodec

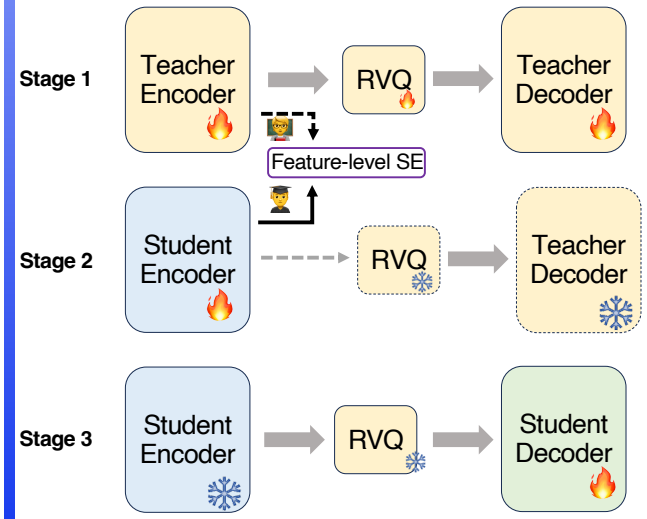
- ✨ **Unified Framework:** A highly efficient (**1.25 G MACs/s** computational complexity and **12.37 M** parameters), low-latency (**50 ms**) neural codec supporting scalable bitrates (**1~6 kbps**).
- 📈 **Progressive Refinement (PR):** Proposes a novel three-stage collaborative training strategy.
- 🏆 **1st Place** in Track 2 (Speech Enhancement Codecs) of the 2025 LRAC Challenge!

METHODS

Model Architecture (Vocodec, arXiv: 2601.13055):



Progressive Refinement (PR) Strategy:



Teacher Pre-training: Train a high-fidelity codec on **clean speech** to establish optimal latent targets

Encoder Alignment: Force student encoder to map **noisy/reverberant inputs** to clean teacher embeddings.

Decoder Refinement: Retrain student decoder via adversarial **fine-tuning**

EXPERIMRNTS & RESULTS

Datasets & Augmentation:

Trained on 24 kHz clean speech (EARS, VCTK, LibriTTS, Common Voice, and DNS Challenge 5) , with Stages 2 and 3 dynamically augmented by diverse noise (80% probability, SNR -5 to 30 dB) and Motus RIR reverberation (50% probability).

Optimization Objectives:

- **Stage 1 (Teacher):** Trained via a **GAN** framework with a composite loss including multi-scale mel-spectrogram reconstruction, adversarial, feature matching, and RVQ codebook losses.
- **Stage 2 (Encoder):** Optimized using a combination of **Mean Squared Error (MSE)** and **cosine distance** to align student and teacher embeddings.
- **Stage 3 (Decoder): Jointly trained** with embedding alignment loss and the GAN-based reconstruction loss to effectively remove artifacts.

Ablation Study for PR Strategy:

Bitrate	Model	Data	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	**0.147**	**3.815**	**3.938**	**3.305**	**5.813**
		Noisy	**0.306**	**3.667**	**3.772**	**2.482**	**5.619**
		Reverb	**0.329**	**3.623**	**3.890**	**2.142**	**5.672**
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	**0.347**	**3.353**	**3.643**	**2.015**	**5.515**
		Noisy	**0.454**	**3.280**	3.532	**1.804**	**5.411**
		Reverb	**0.458**	**3.233**	3.610	**1.686**	**5.435**

Stage 2: Negligible degradation on clean speech coding.

Stage 3: PESQ boosted by >0.09 (6 kbps) via decoder retraining.

The PR strategy seamlessly integrates robust speech enhancement with insignificant degradation to clean speech coding.

- **Subjective Leaderboard** (crowdsourced listening tests conducted by LRAC challenge):

Test Type	Clean speech		Real-world speech in noise		Real-world speech reverb		Intelligibility in clean	Intelligibility in noise	Aggregate Score	
Scale	MUSHRA [0, 100]		MOS [1, 5]		MOS [1, 5]		DRT score [-100, 100]	DRT score [-100, 100]	Weighted sum of normalized test mean scores [0, 100]	
Weight	10%	15%	10%	20%	10%	20%	5%	10%	100%	
Bitrate Mode	ULBR	LBR	ULBR	LBR	ULBR	LBR	ULBR	LBR	Final Score	Overall Rank
nju-aalab	67.30	87.47	2.66	3.42	3.04	3.80	80.61	72.09	68.32	1
xuyang	63.66	87.10	2.33	3.18	2.57	3.57	84.46	72.01	63.64	2
nano-codec	56.32	85.76	2.16	3.08	2.74	3.85	67.69	68.34	63.01	3