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, China
[2] NJU-Horizon Intelligent Audio Lab, Horizon Robotics, Beijing 100094, China

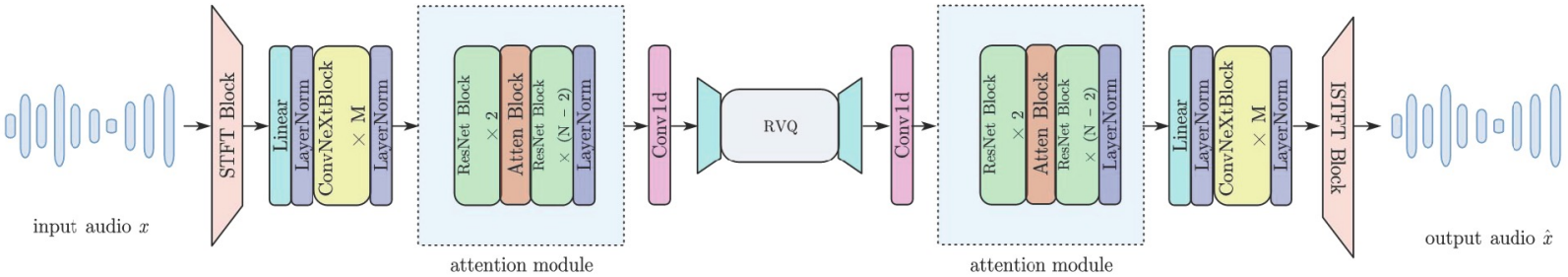
* Presenter

Contents

1. **Method**
   - 1.1 Overall architecture
   - 1.2 Design principles
   - 1.3 Progressive refinement training
2. **Experiments**
   - 2.1 Experimental setup
   - 2.2 Results on track1
   - 2.3 Results on track2
3. **Conclusion**
   - 3.1 Contributions
   - 3.2 Limitations
   - 3.3 Future work

1. Method

➢ Overall Architecture: VoCodec



- **Generator:** Causal Vocos backbone
- **Discriminator:** Multi-scale STFT discriminator (MSSTFTD)
- **Quantization:** 6-layer RVQ with 1kbps per layer at 100Hz

- **Training Loss for Track 1 (codec for clean speech):**
  - For the generator: $\mathcal{L}_{\text{generator}} = 15\mathcal{L}_{\text{rec}} + \mathcal{L}_g + 2\mathcal{L}_{\text{feat}} + \mathcal{L}_{\text{code}} + 0.25\mathcal{L}_{\text{c}}$
  - For the discriminator: $\mathcal{L}_d = \|1 - D(x)\|_2^2 + \|D(\hat{x})\|_2^2$

$$\mathcal{L}_{rec} = \|\log(\mathcal{M}(x)) - \log(\mathcal{M}(\hat{x}))\|_1$$

$$\mathcal{L}_g = \|1 - D(\hat{x})\|_2^2$$

$$\mathcal{L}_{feat} = 2\sum_l \left\|D^l(x) - D^l(\hat{x})\right\|_1$$

$$\mathcal{L}_{code} = \|\text{sg}[\mathbf{z}_e] - \mathbf{e}_k\|_2^2$$

$$\mathcal{L}_c = \|\mathbf{z}_e - \text{sg}[\mathbf{e}_k]\|_2^2$$

1. Method

➢ Design Principles

- **Operating in time-frequency (T-F) domain:**
  - Clear harmonic structure of speech in the T-F domain.
  - The temporal dimension is downsampled before entering the encoder, thereby reducing the computational complexity.
- **Causal Vocos backbone:**
  - Vocos backbone with **causal** convolution.
  - **Masked** attention module.
- **Discriminator Choices:**
  - Why MSSTFTD: The generator directly operates in the T-F domain.
  - Why only MSSTFTD:
    Preliminary experiments indicate that **incorporating other discriminators (e.g., multi-period discriminator) leads to a significant performance degradation.**

Our findings suggest that aligning the operational domains of the generator and the discriminator is crucial for achieving optimal performance.

Model	Epochs	PESQ
VoCodec (only MSSTFTD)	200	2.77
VoCodec (MSSTFTD + MPD)	200	2.26

1. Method

LRAC 2025 CHALLENGE

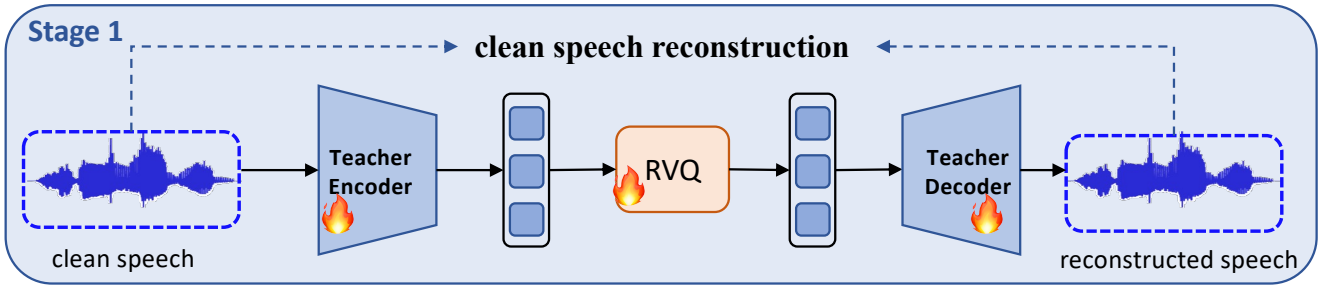
➢ Progressive Refinement (PR) Training

- **Motivation**
  - Preliminary experiments indicate that the joint coding and enhancement model trained **in an end-to-end manner** yields sub-optimal performance.
  - The presence of noise and reverberation interferes with the learning process of the clean speech codebook.

- **Stage 1**

✓ Train a speech codec model with high reconstruction quality.

✓ Obtain codebooks for clean speech.



1. Method

➢ Progressive Refinement (PR) Training

- **Stage2**

✓ Performing enhancement at the **latent embedding level.**

✓ Preventing the codebook from being corrupted.

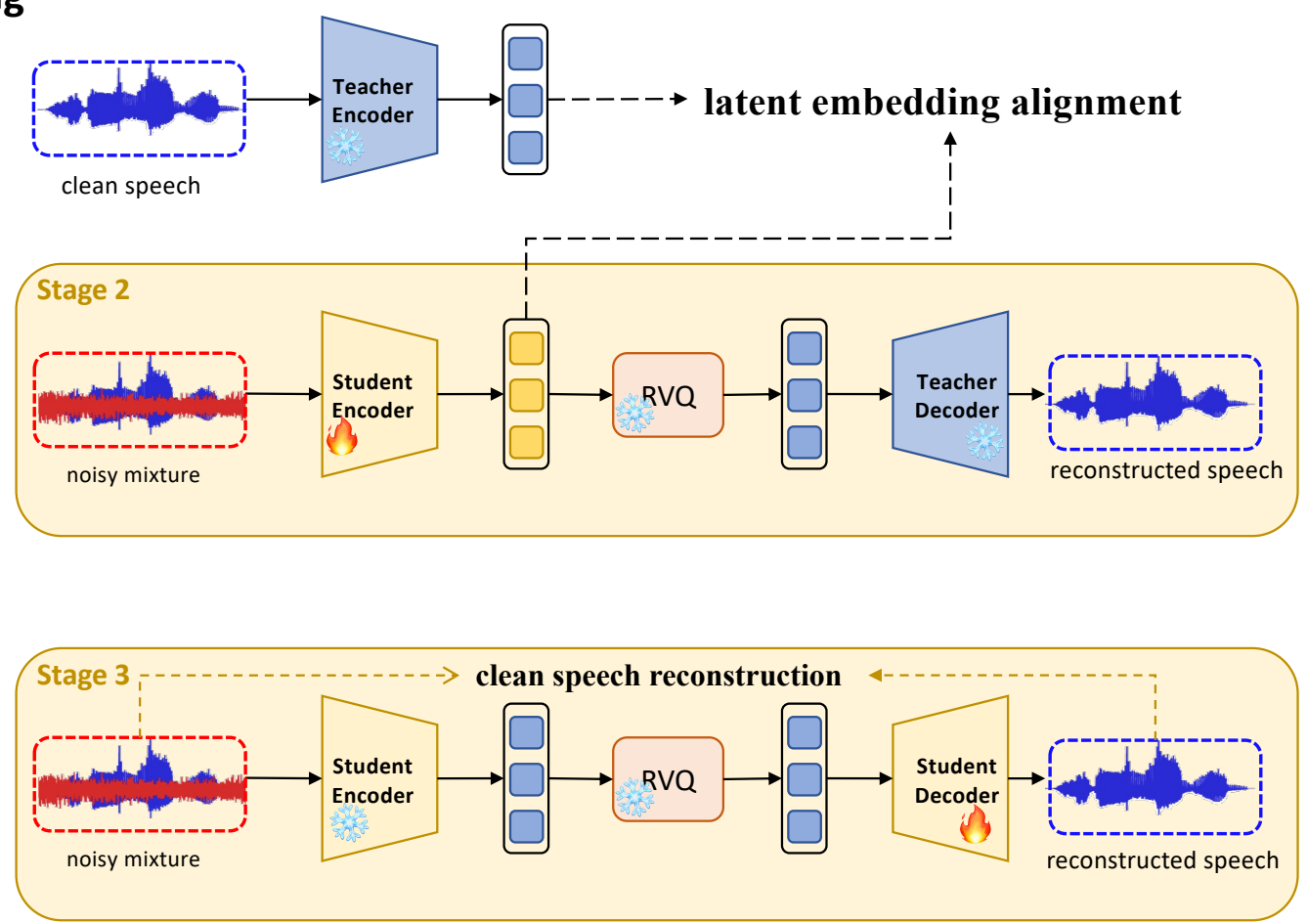
Training Loss:

$$L_{\text{MSE}} = \frac{1}{N}\sum_{i=1}^{N}\left\|\mathbf{z}_i^{(T)} - \mathbf{z}_i^{(S)}\right\|_2^2$$

$$L_{\cos} = 1 - \frac{\mathbf{z}^{(T)} \cdot \mathbf{z}^{(S)}}{\left\|\mathbf{z}^{(T)}\right\|_2 \left\|\mathbf{z}^{(S)}\right\|_2}$$

- **Stage 3**

✓ Strengthening the decoder's resilience to imperfect latent embedding



2. Experiments

LRAC 2025 CHALLENGE

➢ Experimental Setup

- **Training Data Preparation**

Dataset Type	Dataset Name	Original Sample Rate [kHz]	Duration [hours]	Language
Speech	LibriTTS reading speech	8-24	191	EN
	LibriVox data from DNS5 challenge	8-48	314	EN
	VCTK reading speech	48	79	EN with accents
	EARS	48	87	EN
	Globe (CommonVoices)	48	535	EN with 164 accents
	Multilingual Librispeech	8-48	450	DE, ES, FR

Dataset Type	Dataset Name	Condition	Original Sample Rate [kHz]	Duration [hours]
Noise	VCTK Noises	Acoustic background & babble noises	TBA	TBA
	Audioset+FreeSound noises in DNS5 challenge	Crowdsourced + youtube	8-48	~180
	WHAM! noise	4 Urban environments	48	~70
	FSD50K (human voice filtered)	Crowdsourced	8-48	~100
	Free Music Archive	Free Music Archive (directed by WFMU)	8-44	200

Dataset Type	Dataset Name	Condition	Original Sample Rate [kHz]	Duration [# examples]	License
RIR	Open SLR 28	Real & simulated RIRs	48	~60k samples	CC BY 4.0
	Motus	Real RIRs	48	3k samples	CC BY 4.0

- **All above training data follow the cleaning and preprocessing procedures defined by the challenge.**
- For track1, 3-second clean speech segments are extracted during training.
- For track2, each clean speech segment is **mixed with noise with 80% probability**, with the **SNR uniformly distributed in the range of [-5, 30] dB; reveberation is introduced with 50% probability.**

2. Experiments

➢ Experimental Setup

- **Configuration**
- **Track1**

Model	Params (M)	MACs/s (M)	Latency (ms)
Encoder	1.95	194.56	10*
Quantizer	0.02	1.96	0
Decoder	1.50	144.82	20**
VoCodec	3.47	349.29	30

- Trained on 8 NVIDIA RTX 4090 GPUs, **with the batch size set to 24 per GPU (192 in total).**
- Training procedure lasts for 1000 epochs, with 500 iterations per epoch.

STFT & iSTFT: **win_length = 720 (30ms), hop_length = 240 (10ms),** n_fft = 720

* The buffering latency of 10ms is introduced by the STFT operation.

** The algorithmic latency of 20 ms is introduced by the iSTFT operation.

2. Experiments

LRAC 2025 CHALLENGE

➢ Experimental Setup

- **Configuration**
- **Track2**

Model	Params (M)	MACs/s (G)	Latency (ms)
Encoder	9.54	9.49	10 + 20***
Quantizer	0.02	0.002	0
Decoder	2.81	2.81	20
Student	12.37	1.25	50

- **Using VoCodec for track1 as the teacher model.**
- Stage2: trained for 500 epochs, with a total batch size of 40.
- Stage3: trained for 200 epochs, with a total batch size of 192.

*** The additional algorithmic latency of 20 ms is introduced by the lookahead requirements in the encoder.

2. Experiments

LRAC 2025 CHALLENGE

➢ Results on Track1

- **Results on the open test set**

Bitrate	Model	Condition	ScoreQ-ref ↓	UTMOS ↑	Sheet-SSQA ↑	PESQ ↑	Audiobox AE-CE ↑
6 kbps	Baseline	Clean	0.35	3.23	3.84	2.67	5.28
		Noisy	0.82	2.76	3.12	1.81	4.37
		Reverb	1.13	1.32	2.22	1.18	3.43
	VoCodec	Clean	**0.17**	**3.73**	**4.22**	**3.20**	**5.66**
		Noisy	**0.70**	**3.10**	**3.43**	**2.03**	**4.82**
		Reverb	**0.94**	**1.55**	**2.80**	**1.21**	**3.98**
1 kbps	Baseline	Clean	1.15	1.44	1.84	1.15	3.90
		Noisy	1.29	1.33	1.72	1.11	3.40
		Reverb	1.36	1.26	1.85	1.07	2.94
	VoCodec	Clean	**0.40**	**3.24**	**3.55**	**1.95**	**5.31**
		Noisy	**0.83**	**2.67**	**2.93**	**1.56**	**4.43**
		Reverb	**1.10**	**1.48**	**2.19**	**1.17**	**3.59**

- Outperforming the official baseline **across all metrics**.

- **Results on the blind test set**

Test Type	Clean speech		Real-world light noise and reverb		Simultaneous talkers		Intelligibility in clean	Aggregate Score	
Scale	MUSHRA [0, 100]		DMOS [1, 5]		DMOS [1, 5]		DRT score [-100, 100]	Weighted sum of normalized test mean scores [0, 100]	
Weight	20%	20%	20%	20%	5%	5%	10%	100%	
Bitrate Mode	ULBR	LBR	ULBR	LBR	ULBR	LBR	ULBR	Final Score	Overall Rank
teamwzqaq	62.65	81.75	3.02	4.44	2.82	4.35	85.43	71.91	1
nano-codec	59.23	81.17	3.13	4.44	2.60	4.22	78.12	70.86	2
aitd-go	60.90	80.69	3.40	4.16	2.08	2.98	85.57	69.22	3
nju-aalab	65.20	89.19	2.74	4.12	1.70	2.82	82.98	67.48	4
boya-audio	35.22	77.24	2.21	4.30	2.03	4.26	80.29	59.42	5
pdura7	42.75	62.56	2.30	3.29	1.68	2.05	75.34	49.94	6
lrac-challenge	17.92	74.28	1.31	3.35	1.26	2.20	75.90	42.36	7

- Ranking **4th** on the final blind test set.
- Achieving **the highest MUSHRA scores** on the clean speech set.

2. Experiments

➢ Results on Track2

- **Results on the open test set**

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**

- Outperforming the official baseline **across all metrics**.
- Demonstrating **the effectiveness of PR strategy.**

- **Results on the blind test set**

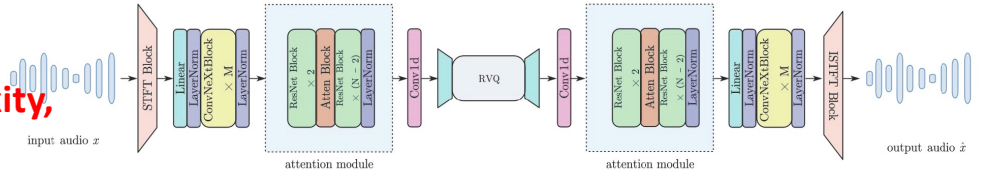
Test Type	Clean speech		Real-world speech in noise		Real-world speech reverb		Intelligibility in clean	Intelligibility in noise	Aggregate Score	
Scale	MUSHRA [0, 100]		MOS [1, 5]		MOS [1, 5]		DRT score [-100, 100]	DRT score [-100, 100]	Weighted sum of normalized test mean scores [0, 100]	
Weight	10%	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
aitd-go	63.41	82.07	2.48	3.37	2.39	3.36	78.72	73.47	62.62	4
teamwzqaq	60.47	81.09	1.94	2.86	2.43	3.45	81.78	63.78	58.42	5
boya-audio	49.74	78.17	1.92	2.91	2.38	3.66	70.43	70.39	58.08	6
leyan	58.63	75.96	2.09	2.89	2.15	3.05	76.21	66.35	55.28	7
parslog	41.34	71.40	1.90	2.57	2.06	2.76	66.95	50.68	48.10	8
pdura7	38.86	61.72	1.80	2.36	2.12	2.87	73.90	67.19	46.80	9
lrac-challenge	21.26	60.06	1.30	2.25	1.39	2.32	75.86	68.21	38.52	10

- Ranking **1st** on the final blind test set.

3. Conclusion

➢ Contributions

- VoCodec: a codec model that **simultaneously achieves high reconstruction quality, low computational complexity, low bitrate, and low latency.**
- PR strategy: a novel training strategy for **codec models with enhancement capabilities.**



➢ Limitations

- The reconstruction quality at ultra-low bitrates (e.g., 1 kbps) leaves room for improvement.
- Performance in real-world scenarios (involving noise, reverberation, and multi-speaker conditions) remains sub-optimal.

➢ Future Work

- Further refining the network architecture or training strategy to boost performance at ultra-low bitrates.
- Investigating methods to further enhance the model's effectiveness in real-world scenarios.

Thanks for your attention!