KD-Vocodec: A Low Complexity Model for Joint Speech Coding and Enhancement Using Knowledge Distillation

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

[1]*Key Laboratory of Modern Acoustics, Nanjing University*

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

Email: xuyang212518@gmail.com

Paper & Demo

LRAC 2025 CHALLENGE

ICASSP2026 Barcelona

INTRODUCTION

Background & Challenges

- Neural codecs enable efficient low-bitrate compression but often overlook **strict computational constraints**.
- **Perceptual quality degrades** significantly in real-world noisy and reverberant environments.

KD-Vocodec: A novel architecture **jointly optimizes speech coding and enhancement** within a unified architecture, which employs **feature-level knowledge distillation (KD)**.

Award: **2nd place** in Track 2 of the 2025 LRAC Challenge.

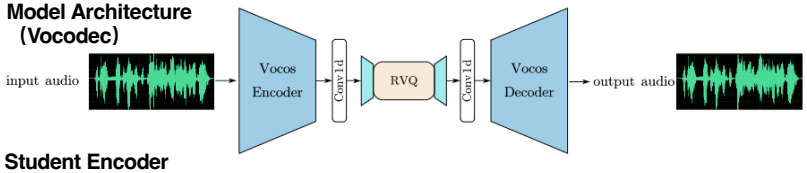
Ultra-low Latency: **30 ms**.

High Efficiency: **1.28G MACs/s** and **12.65M** parameters.

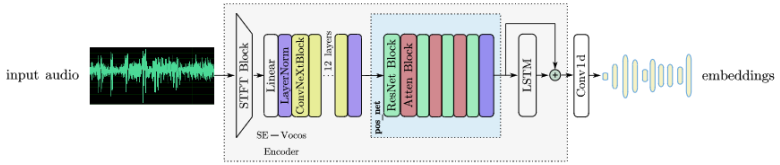
Flexible Bitrate: Scalable from **1 to 6 kbps**.

METHODS

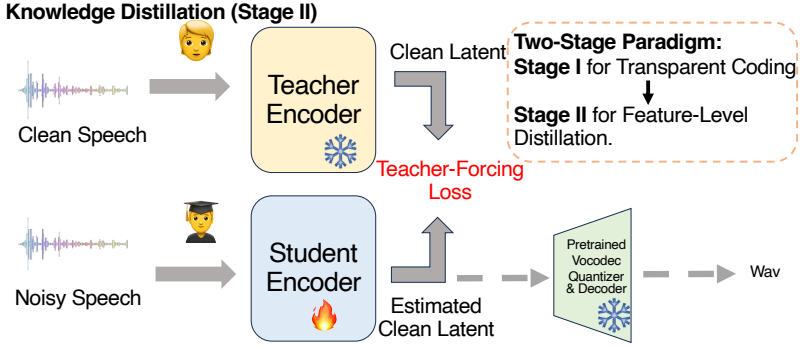
Model Architecture (Vocodec)



Student Encoder



Knowledge Distillation (Stage II)



EXPERIMENTS

Datasets:

Speech Datasets (24 kHz): Large-scale dataset curated from EARS, VCTK, LibriTTS, Common Voice, and DNS Challenge 5.

- **Teacher Training (Stage I):** Optimized exclusively on speech datasets。
- **Student Training (Stage II):** Clean speech mixed with VCTK, WHAM, FSD50K, and FMA noise (80% probability, SNR: -5 to 30 dB), augmented with Motus RIRs (50% probability).

Training Compute:

- Stage I: 8 x 3090 GPUs
- Stage II: 4 x 3090 GPUs

Training Pipeline:

Stage I: Employs the identical VQ-GAN training paradigm as **Vocodec (arXiv: 2601.13055)**.

Stage II: Achieves end-to-end knowledge distillation using MSE and Cosine distance losses.

Ablation Study On Student Encoder Architecture:

- The **Post-LSTM** configuration, when placed after contextual aggregation, stabilizes frame-level variations prior to quantization, proving crucial for robust latent representations.

Bitrate	Test Set	Model	ScoreQ-ref↓	UTMOS↑	Sheet-SSQA↑	PESQ↑	Audiobox AE-CE↑
6 kbps	Clean	Base	0.160	3.717	3.934	3.143	5.778
		Pre-LSTM	0.156	3.737	3.944	3.178	5.789
		ConvNeXt-only	0.160	3.725	3.963	3.170	5.780
		Post-LSTM	**0.153**	**3.749**	**3.982**	**3.219**	**5.794**
	Noisy	Base	0.430	3.316	3.585	2.171	5.358
		Pre-LSTM	0.409	3.354	3.641	2.178	5.405
		ConvNeXt-only	0.420	3.333	3.600	2.202	5.371
		Post-LSTM	**0.396**	**3.366**	**3.665**	**2.225**	**5.431**
	Reverb	Base	0.516	3.013	3.627	1.723	5.368
		Pre-LSTM	**0.468**	3.090	**3.677**	1.762	**5.429**
		ConvNeXt-only	0.499	3.060	3.624	1.769	5.377
		Post-LSTM	0.480	**3.096**	3.658	**1.802**	5.426
1 kbps	Clean	Base	0.393	3.254	3.619	1.909	5.490
		Pre-LSTM	0.386	3.248	3.614	1.916	**5.507**
		ConvNeXt-only	0.390	3.255	3.608	1.914	5.494
		Post-LSTM	**0.378**	**3.265**	**3.659**	**1.941**	5.506
	Noisy	Base	0.557	2.966	3.367	1.582	5.248
		Pre-LSTM	0.542	**3.007**	3.404	1.590	5.295
		ConvNeXt-only	0.543	2.989	3.397	1.603	5.255
		Post-LSTM	**0.530**	3.006	**3.434**	**1.609**	**5.302**
	Reverb	Base	0.648	2.685	3.329	1.390	5.154
		Pre-LSTM	0.622	2.734	3.360	1.407	5.181
		ConvNeXt-only	0.638	2.735	3.348	1.420	5.130
		Post-LSTM	**0.619**	**2.749**	**3.386**	**1.430**	**5.190**

RESULTS

- **Advantage:** Achieves exceptional clean speech coding and robust speech enhancement in noisy and reverberant environments across all bitrates.
- **Contribution:** A low-complexity, high-fidelity neural codec integrating robust speech enhancement with scalable bitrates.

- **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