VoCodec: An Efficient Lightweight Low-Bitrate Speech Codec

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

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

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

Demo

Paper

LRAC 2025 CHALLENGE

ICASSP2026 Barcelona

INTRODUCTION

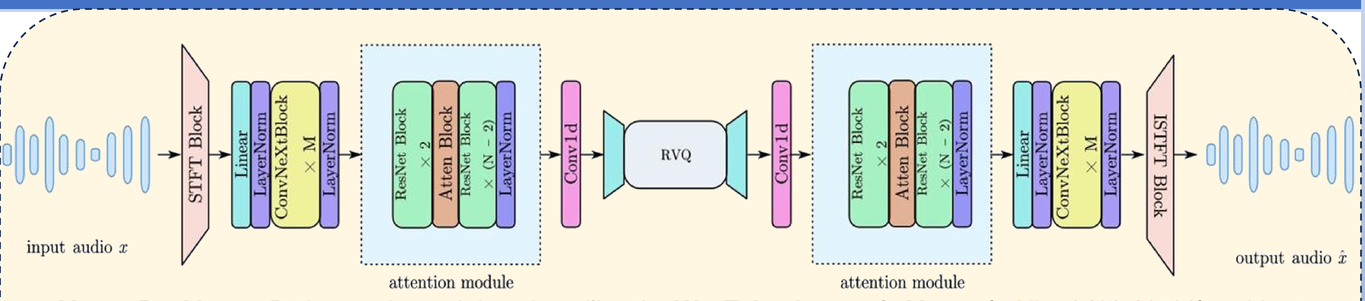
Background & Challenges

- **Real-Time Communication Bottlenecks:** Existing high-performance neural speech codecs often suffer from high computational complexity and high latency.
- **Environmental Disturbances:** Real-world noise and reverberation cause severe interference to real-time communication.

VoCodec: A novel **VQ-GAN-based** codec model simultaneously featuring **low computational complexity**, **low latency**, and support for **1 and 6 kbps transmission**.

- **Ultra-Low Overhead:** Total computational complexity is only **349.29M MACs/s** (receiver-side: **144.82M MACs/s**), with a low latency of **30 ms** under **24kHz sampling rate**.
- **Speech Enhancement Scalability for Track 2:** Optionally cascading a ultra-lightweight speech enhancement network (**extended UL-UNAS, 935.36M MACs/s**) at the front end for joint noise reduction and dereverberation.

METHODS

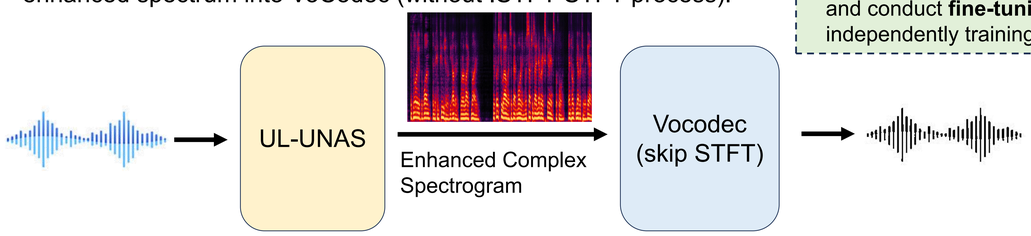


- **Vocos Backbone:** Both encoder and decoder utilize the **WavTokenizer-style Vocos (arXiv: 2408.16532)** architecture.
- **Strict Frame-wise Causality: Causal convolutions** and **masked attention**
- **Quantization Strategy:** An improved **6-layer RVQ** featuring factorized codes and L2-normalization, supporting a total bitrate of up to 6 kbps (**1 kbps per layer at 100 Hz with 1024 codewords**)
- **Discriminator:** Only a **multi-scale STFT discriminator**

Track 2 Extension: Front-End Speech Enhancement (SE) Cascade:

- Cascade an extended UL-UNAS (a time-frequency domain SE network) at the front end. It seamlessly feeds the enhanced spectrum into VoCodec (without ISTFT-STFT process).

⋆ we freeze the parameters of UL-UNAS and conduct **fine-tuning** of VoCodec after independently training each one.



UL-UNAS Paper

EXPERIMENTS

Datasets: Follow the official LRAC Challenge pipeline.

- **Augmentation for Track 2:** Dynamically mix samples with background noise (50% probability, SNR uniformly distributed between -5 dB and 30 dB) and reverberation (50% probability).

Training Compute: 8 x 4090 GPUs

Training Loss: The UL-UNAS is optimized via negative SI-SNR and power-compressed spectrum losses , whereas the VoCodec combines multi-scale mel-spectrogram reconstruction, adversarial, feature matching, and quantization losses.

RESULTS

Objective Performance Comparison on the Open Test Set of Track1

- **Superiority Across All Metrics:**

VoCodec outperforms the official baseline in all objective metrics across clean, noisy, and reverberant conditions.

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**

Subjective Performance Comparison on the LRAC Blind Test Set

(i). Track 1

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

(ii). Track 2 (with UL-UNAS)

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

Conclusions

- **Outstanding High-Fidelity:** Achieve the highest MUSHRA score (89.19 at 6 kbps and 65.20 at 1kbps) on the Track 1 clean speech set.
- **Competitive Joint System (Track 2):** Cascading the SE model provides robust interference suppression and competitive multi-metric performance.