IRIS: LOW-COMPLEXITY HIGH-EFFICIENCY NEURAL NETWORK CODEC FOR REAL-TIME AUDIO TRANSMISSION

Ziqian Wu*, Jiawei Jiang*, Kunpeng Lin*, He Wang*, Qingbo Huang*, Dejun Zhang*, Andong Li†

*ByteDance, China †Institute of Acoustics, Chinese Academy of Sciences, Beijing, China

1. Abstract

The 2025 Low-Resource Audio Codec (LRAC) Challenge aims to advance the development of low-resource neural audio codec. It requires participating systems to achieve high-quality audio compression under strict resource constraints. LRAC Challenge Track 1 focuses on transparent audio transmission under real-world noise and reverberation, while Track 2 integrates speech enhancement and transmission. This paper introduces IRIS (Internet Real-time Intelligent Streaming Codec): an end-to-end, low-complexity, low-latency neural audio codec. IRIS enables encoding and decoding at bitrates of 1 kbps and 6 kbps, with inherent robustness to noise and reverberation. We designed a lightweight codec architecture, incorporating multiple discriminators and loss functions to achieve high-quality reconstruction at extremely low complexity and bitrates. In the LRAC Challenge, IRIS ranked first in Track 1, demonstrating the effectiveness of the proposed codec.

2. Contribution

- **Audio Specification:** End-to-end support for encoding and decoding of 24kHz mono audio at extremely low bitrates of 1 kbps and 6 kbps.
- **Model Innovations:** Proposed an innovative residual encoder structure with **SnakeBeta** activation, and a highly efficient decoder based on Conv2Former blocks.
- **Low Complexity & Latency:** Achieved high-quality audio compression under strict resource limits. The total model complexity is **~695 MFLOPS** for Track 1 and **~2489 MFLOPS** for Track 2. The algorithmic system latency is merely **30 ms**.
- **Challenge Results:** Achieved **1st place** in Track 1 and **5th place** in Track 2 of the 2025 LRAC Challenge while strictly adhering to low-complexity constraints.

3. Datasets & Pre-processing

Datasets

- Strictly adhere to the competition requirements.
- Following the LRAC baseline system.

Training data processing methods

- Unifying duration.
- Pitch shifting.
- Adding noise or reverberation.
- Adding simultaneous speakers.
- Training target: raw signal for Track 1, clean signal for Track 2.

	Clean	Noisy	Reverb	Simultaneous Talkers
Track 1	8	5	5	2
Track 2	1	4	4	0

Weights of speech types in different tracks

4. Discriminators & Loss Functions

Discriminators

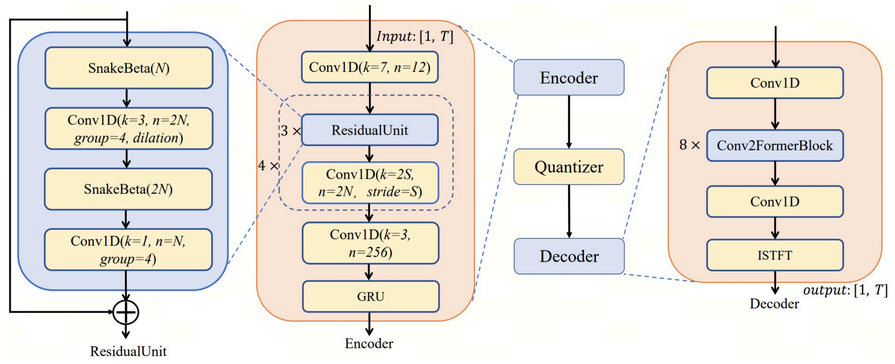
- Multi-period discriminators (MPD).
- Multi-scale STFT discriminators.
- Multi-scale subband STFT discriminators.
- Multi-length mel-spectrogram discriminators.
- All discriminators are updated at every training step.

Loss Functions

- Multi-scale STFT loss and multi-scale mel-spectrogram loss.
- Adversarial objectives use discriminator feature loss and generator loss.
- RVQ commitment and codebook losses.
- Perceptual PESQ loss.
- WaveFM-style optimized STFT loss.
- Loss functions comprehensively optimize waveform reconstruction, deep feature alignment, and perceptual quality.

5. Model Architecture

- End-to-end, asymmetric encoder–RVQ–decoder architecture.
- 24 kHz mono waveform input with 20 ms frame size.
- STFT magnitude and phase coefficients output.



Schematic diagram of the model architecture

Encoder

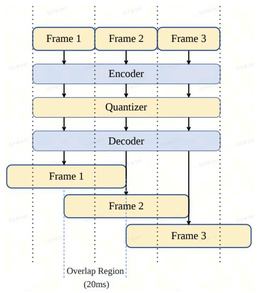
- An initial 1D convolution layer with kernel size 7 processes the input.
- Four encoding modules with strides [3, 4, 5, 8] to downsample features.
- Significant structural improvements:
  - Modified residual units use Conv1D for feature expansion and contraction.
  - Using **SnakeBeta(x) = x + (1/β)·sin²(αx)**, replacing the Snake activation used in DAC.
- A final Conv1D layer and GRU layer outputs encoded features.

Residual Vector Quantization (RVQ)

- DAC-style residual vector quantization design.
- 12 codebook layers are used, each with 1024 entries and dimension 8.
- Multi-bitrate support within a single codebook:
  - For the 1 kbps mode, only the first 2 codebook layers are activated.
  - For the 6 kbps mode, all 12 codebook layers are utilized.

Decoder

- Decoder takes quantized RVQ features as input.
- A Conv1D layer is followed by 8 stacked Conv2Former blocks.
- Outputs STFT magnitude and phase coefficients.
- ISTFT reconstructs 40 ms of audio: 10 ms past, 20 ms current, and 10 ms future context.



System latency description

Parameter/Metric	Track 1	Track 2
Encoder_dim	12	32
Encoder_group	4	8
Encoder_output_latent_dim	256	512
Conv2FormerBlock_input_dim	372	512
Conv2FormerBlock_hidden_dim	380	620
Encoder Complexity (MFLOPS)	385.5	1875.38
Quantizer Complexity (MFLOPS)	15.46	19.66
Decoder Complexity (MFLOPS)	294.02	594.26
Encoder Parameter Count (M)	0.973	5.145
Quantizer Parameter Count (M)	0.154	0.209
Decoder Parameter Count (M)	2.954	5.967

Model parameters, complexity and parameter count (Track 1 vs Track 2)

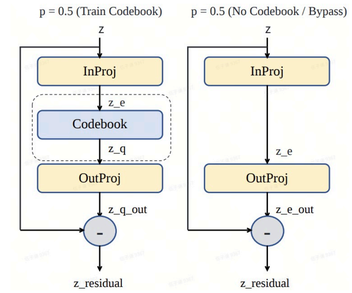
Contact Information

Email: wuziqian.0315@bytedance.com

ByteDance, China

6. Training Methods

- Randomly bypassing all codebook with a 50% probability during training allows direct gradient backpropagation between encoder and decoder.



Bypassing codebook training during training process

- Randomly dropping layers of codebook with a 50% probability to achieve multibitrate capability using one model and codebook.
- Reducing mel loss weight in stage 2 promotes the generation of mid-to-high frequency harmonics.

$$\begin{aligned}\text{Loss} = {} & \lambda_1 \text{Loss}_{\text{mel}} + \lambda_2 \text{Loss}_{\text{stft}} \\ & + \lambda_3 \text{Loss}_{\text{disc}} + \lambda_4 \text{Loss}_{\text{gen}} \\ & + \lambda_5 \text{Loss}_{\text{vqcommit}} + \lambda_6 \text{Loss}_{\text{vqcodebook}} \\ & + \lambda_7 \text{Loss}_{\text{pesq}} + \lambda_8 \text{Loss}_{\text{wavefm\_stft}}\end{aligned}$$

Weights	Stage-1	Stage-2
λ_1	15.0	1.0
λ_2	10.0	10.0
λ_3	2.0	2.0
λ_4	1.0	1.0
λ_5	0.25	0.25
λ_6	1.0	1.0
λ_7	5.0	5.0
λ_8	10.0	10.0

Loss weights (λ) for different training stages

7. Ablation Experiments and Challenge Results

Ablation Experiments

- Conv2Former-based decoders outperform ConvNeXt backbones under similar decoding complexity.
- PESQ-based loss yields objective gains.
- WavLM and MuQ feature losses failed to yield improvements in objective quality.
- Probabilistically dropping codebook training substantially improves generation quality.
- Two-stage fine-tuning with a lower mel loss weight further boosts PESQ scores.

Ablation Setup	Clean		Real world		Simultaneous talkers		DRT EN	
	1 kbps	6 kbps	1 kbps	6 kbps	1 kbps	6 kbps	1 kbps	6 kbps
Stage-1 200k steps.	1.83	3.00	1.77	2.98	1.48	2.60	1.90	2.98
ConvNeXt Block decoder, stage-1 200k steps.	1.74	2.46	1.70	2.38	1.43	2.10	1.75	2.44
Without PESQ loss, stage-1 200k steps.	1.66	2.70	1.62	2.72	1.38	2.40	1.74	2.72
With WavLM feature loss, stage-1 200k steps.	1.71	2.63	1.68	2.56	1.47	2.09	1.73	2.63
With MuQ feature loss, stage-1 200k steps.	1.68	2.90	1.66	2.96	1.45	2.65	1.71	2.89
Without dropping codebook training, stage-1 200k steps.	1.83	2.80	1.79	2.75	1.50	2.26	1.91	2.85
Stage-1 800k steps.	1.91	3.43	1.82	3.45	1.52	3.29	1.96	3.47
Stage-2 200k steps.	2.00	3.56	1.89	3.59	1.56	3.44	2.03	3.54

Ablation experiment

Challenge Results

- IRIS ranks **1st** in Track 1 and **5th** in Track 2 of the LRAC Challenge.
- A final aggregate score of **71.91** in Track 1.

Test Type	Clean speech		Real-world noise & reverb		Simultaneous talkers		Intelligibility (clean)	Aggregate Score	
Score / Rank	MUSHRA [0,100]	MUSHRA [0,100]	DMOS [1,5]	DMOS [1,5]	DMOS [1,5]	DMOS [1,5]	DRT [-100,100]	Final Score [0, 100]	Rank
Bitrate Mode / Weight	20% ULBR	20% LBR	20% ULBR	20% LBR	5% ULBR	5% LBR	10% ULBR	100%	
teamwzqaq (Proposed)	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

Results of LRAC Challenge Track 1