PHOENIXCODEC: TAMING NEURAL SPEECH CODING FOR EXTREME LOW-RESOURCE SCENARIOS

Zixiang Wan[1], Haoran Zhao[2], Guochang Zhang[2], Runqiang Han[2], Jianqiang Wei[2], Yuexian Zou[1]

[1]Peking University, Shenzhen [2]Anker Inc, Beijing

1. MOTIVATION & KEY CHALLENGE

Background

- Neural speech codecs demand high-quality audio over low-bandwidth networks
- Existing methods face trade-offs under extreme constraints

LRAC 2025 Challenge Constraints

- ✓ Dual bitrates: **1 kbps & 6 kbps**
- ✓ Computation: $\leq$ **700 MFLOPs**
- ✓ Latency: $\leq$ **30 ms** (real-time)

Problem: How to achieve optimal balance of efficiency, quality, and adaptability?

2. KEY CONTRIBUTIONS

Three Innovations:

(1) Frequency–Time Asymmetric Architecture

- Combines F-domain encoding + T-domain decoding
- Solves "resource-scattering" bottleneck
- Improves efficiency at comparable complexity

(2) Cyclical Calibration & Refinement (CCR)

- Periodically suspends adversarial losses
- Dynamic regularization with reconstruction loss
- Enhances stability & extends performance ceiling

(3) Noise-Invariant Fine-Tuning (NIFT)

- Trains on noisy samples with clean targets
- Learns noise-invariant representations
- Improves bit allocation efficiency

3-1. METHOD: ARCHITECTURE

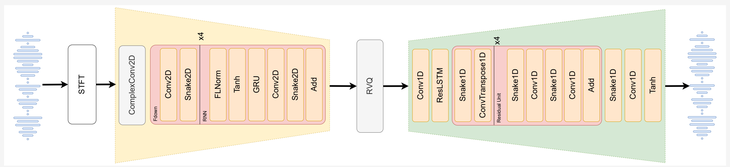
Why Frequency-Time (F-T) Design?

Key Insight: Encoder determines information preservation (critical for WER), Decoder determines audio quality (critical for MOS score)

Problem with F-domain Decoder:

- Fixed frequency grouping $\rightarrow$ resource-scattering bottleneck
- Insufficient capacity per sub-band
- Severely limits performance under extreme computational budgets

Solution: Asymmetric F-T Architecture



- Frequency-domain encoder: preserves rich information (benefits WER)
- Time-domain decoder: avoids resource-scattering, improves signal fidelity
- GAN-based training enhances perceptual quality
- Achieves optimal balance of information preservation, signal fidelity, and perceptual quality under strict constraints

3-2. NETWORK STRUCTURE DETAILS

Frequency-domain Encoder:

- **Input:** STFT representation (hop size: 288 for 24khz, approx. 83Hz)
- **Architecture:** 4 residual Conv2D + RNN blocks with downsampling
- **RVQ:** 6 codebooks, each with 4096 entries and 8 dimensions, support 1-6 kbps

Time-domain Decoder:

- **Input:** Quantized latent codes from RVQ
- **Architecture:** 4 transposed Conv1D + Conv1D residual unit blocks with upsampling
- **Output:** Raw waveform (24kHz)

4. METHOD: CCR TRAINING STRATEGY

Optimization Objective:

$$\mathcal{L}_G = \mathcal{L}_{\text{mel}} + \lambda_{\text{vq}}\mathcal{L}_{\text{vq}} + \lambda_{\text{fm}}\mathcal{L}_{\text{fm}} + \lambda_{\text{adv}}\mathcal{L}_{\text{adv}}$$

where $\mathcal{L}_{\text{mel}}$: multi-resolution mel-spectrogram reconstruction loss, $\mathcal{L}_{\text{vq}}$: vector quantization loss, $\mathcal{L}_{\text{fm}}$: feature matching loss, $\mathcal{L}_{\text{adv}}$: adversarial loss

Three-Stage Process:

Stage I: Pre-training ($\lambda_{\text{fm}} = 0, \lambda_{\text{adv}} = 0$)
$\rightarrow$ Stable spectral representation
$\rightarrow$ $\sim$200k steps, LR: 8×10^{-4}

$\Downarrow$

Stage II: Joint Optimization (All losses)
$\rightarrow$ Enhance perceptual quality
$\rightarrow$ $\sim$200k steps, LR: 1×10^{-4}

$\Downarrow$

Stage III: Cyclical Refinement
$\circlearrowleft$ Calibration: $\lambda_{\text{fm}} = 0, \lambda_{\text{adv}} = 0$ (stabilize)
$\circlearrowleft$ Refinement: All losses (enhance)
$\rightarrow$ $\sim$10k steps/phase, LR: 1×10^{-5}

Why CCR Works?

- Alternates between perceptual quality & signal fidelity
- Adversarial losses explore acoustic details while reconstruction loss stabilizes generation

5. METHOD: NOISE-INVARIANT FINE-TUNING

Motivation: At 1 kbps, the bottleneck is **bit allocation efficiency**

Two-Stage Process:

- **Baseline:** Train on clean speech (s), optimize $d(s, \mathcal{C}(s;\theta))$
- **NIFT:**
  Input: Clean/Noisy/Reverberant (1:1:1)
  Target: Always clean speech
  Optimize: $d(s, \mathcal{C}(x;\theta))$ for all x

where s: clean speech, x: input (clean/noisy/reverberant), $\mathcal{C}(\cdot;\theta)$: codec, $d(\cdot,\cdot)$: distance metric

Benefits:

- ✓ Learn noise/reverberation-invariant representations
- ✓ Improve bit allocation for speech features

6. EXPERIMENTS

Model Settings

- **Total Parameters:** 1.48M
- **Computational Cost:** 699 MFLOPs
- **Latency:** 29.83 ms (real-time compliant)

LRAC 2025 Challenge Results

System	Clean MUSHRA		Noisy DMOS		Multi-talkers DMOS		Intel. DRT
	1 kbps	6 kbps	1 kbps	6 kbps	1 kbps	6 kbps	1 kbps
Baseline	17.92	74.28	1.31	3.35	1.26	2.20	75.90
1st Place	62.65	81.75	3.02	4.44	2.82	4.35	85.43
Proposed	60.90	80.69	**3.40**	4.16	2.08	2.98	**85.57**

Key Achievements:

- ✓ **3rd Place Overall** in Track 1
- ✓ **1st Place at 1 kbps** in Noisy (DMOS: 3.40)
- ✓ **1st Place at 1 kbps** in Intelligibility (DRT: 85.57)

Ablation Studies

A. Architecture Comparison

Method	WER↓	PESQ_c ↑	UTMOS_c ↑	Scoreq_c ↓	PESQ_n ↑	UTMOS_n ↑	Scoreq_n ↓
F-F	**7.41**	1.73	2.32	0.68	1.42	1.96	0.94
T-T	10.87	1.92	2.94	0.50	1.49	2.55	0.83
F-T	7.47	**1.98**	**2.99**	**0.49**	**1.53**	**2.61**	**0.81**

Finding: F-T design combines frequency encoder's complete information preservation (WER 3.40% lower than T-T) with time decoder's enhanced audio quality (UTMOS 0.67 higher than F-F), validating the asymmetric approach for extreme low-bitrate scenarios

B. CCR Training Strategy

Method	WER↓	PESQ_c ↑	UTMOS_c ↑	Scoreq_c ↓	PESQ_n ↑	UTMOS_n ↑	Scoreq_n ↓
Joint-opt	16.49	1.72	2.87	0.55	1.37	2.29	0.93
CCR_{s1}	**7.47**	1.98	2.99	0.49	1.53	2.61	0.81
CCR_{s2}	8.86	1.94	3.29	0.40	1.50	2.67	0.81
$\text{CCR}_{s3\text{-}c}$	8.03	**2.13**	**3.44**	**0.38**	**1.60**	**2.89**	**0.76**

Finding: CCR reduces WER 8.46% and improves clean UTMOS by 0.57, proving cyclical refinement effectiveness over direct joint training

C. Noise-Invariant Fine-Tuning

Method	WER↓	PESQ_c ↑	UTMOS_c ↑	Scoreq_c ↓	PESQ_n ↑	UTMOS_n ↑	Scoreq_n ↓
w/o NIFT	**5.89**	**1.80**	2.13	0.73	1.46	1.76	0.98
w/ NIFT	8.52	1.78	**2.39**	**0.67**	**1.60**	**2.29**	**0.73**

Finding: NIFT boosts noisy UTMOS by 0.53 (from 1.76 to 2.29) with minimal clean quality loss, critical for robust 1 kbps deployment in real-world noisy environments

7. CONCLUSION

Summary: PhoenixCodec addresses extreme low-resource neural speech coding through three synergistic innovations:

- **F-T Architecture** solves resource-scattering with asymmetric frequency-time design
- **CCR Training** enables stable optimization and extended performance ceiling
- **NIFT** improves bit allocation efficiency and robustness to noise

Impact: 3rd place overall, 1st place at 1 kbps in noisy and intelligibility
Contribution: State-of-the-art ultra-low-bitrate codec under strict real-time constraints