Improving X-Codec-2.0 for Multi-Lingual Speech: 25 Hz Latent Rate and 24 kHz Sampling

Malaysia-AI

Abstract

X-Codec-2.0 has shown strong performance in neural audio compression and multilingual speech modeling, operating at a 50 Hz latent rate and a 16 kHz sampling rate using frozen HuBERT features. While effective, this configuration limits efficiency and audio fidelity. In this work, we explore a simple effective modification by introducing additional pooling and increasing the decoder hop size. This reduces the latent rate from 50 Hz to 25 Hz and simultaneously raises the output sampling rate from 16 kHz to 24 kHz, improving temporal efficiency and perceptual quality without altering the core architecture. Evaluated on the multilingual Common Voice 22 test set, the proposed configuration achieves a 0.5 MOS improvement over the original X-Codec-2.0 baseline. The source code, checkpoints and generation comparison released at xcodec2-25tps.github.io.

1 Introduction

Recent advances in neural audio tokenization have enabled end-to-end speech modeling through discrete latent representations. Among these, X-Codec-2.0 [1] has emerged as a powerful multilingual codec model, combining a HuBERT-based [2] semantic encoder and a transformer-based codec architecture to produce high-quality, language-agnostic audio tokens. Its design allows large language models (LLMs) and autoregressive decoders to handle speech as a discrete sequence prediction task.

A key appeal of X-Codec-2.0 lies in its simplicity: it employs a single codebook with a latent frame rate of 50 Hz, making its discrete token stream compact, stable, and easy to integrate into multimodal LLM pipelines. This property makes it particularly attractive for tasks such as text-to-speech (TTS) generation, speech-to-text modeling, or unified speech-language pretraining especially for streaming or low-latency scenarios where token-level autoregressive decoding is desirable.

However, X-Codec-2.0 operates at a 50 Hz latent frame rate with a 16 kHz target sampling rate. While effective, this configuration limits temporal resolution and upper-frequency fidelity, producing slightly muffled high-frequency content and longer token sequences for generation models. Moreover, as multilingual datasets expand to include more diverse acoustic and expressive conditions, the fixed 50 Hz resolution may underutilize the model's potential to capture fine-grained speech variation.

In this work, we propose a simple yet effective modification to X-Codec-2.0 that improves both efficiency and perceptual quality. By increasing the hop size to 960 samples and introducing a lightweight pooling layer before quantization, the codec operates at a reduced 25 Hz latent rate while simultaneously increasing the audio sampling rate to 24 kHz. All encoder components are kept frozen, and only the decoder is fine-tuned under this new temporal configuration. We find that this adjustment produces higher-quality reconstructions without additional parameters or training complexity.

Empirically, the proposed model achieves a +0.5 improvement in mean opinion score (MOS) as estimated by UTMOSv2 [3] on the multilingual Common Voice [4] 22 test set. The improvement is consistent across languages and demonstrates better high-frequency reconstruction and overall perceptual clarity. Our work shows that small architectural refinements by changing hop-size and

pooling adjustments can meaningfully improve codec quality while preserving the simplicity and modularity that make X-Codec-2.0 suitable for LLM-based speech modeling.

2 Method

The proposed modification to X-Codec-2.0 focuses on improving temporal efficiency and output audio fidelity through minimal architectural changes. The overall structure remains consistent with the original design, consisting of a frozen semantic encoder (HuBERT-based), a transformer codec encoder, and a vocoder-style decoder.

2.1 Temporal Pooling and Hop Size Adjustment

In the original X-Codec-2.0 configuration, the encoder operates at a latent rate of 50 Hz with a 16 kHz sampling rate, corresponding to a hop size of 320 samples. To achieve a lower latent rate and higher waveform sampling rate, we increased the hop size to 960 samples and introduced an additional average pooling layer:

$$\text{AvgPool1d}(k = 2,\ \text{stride} = 2).$$

This modification reduces the latent rate from 50 Hz to 25 Hz, halving the number of discrete tokens per second while maintaining temporal coherence. The pooling operation is applied before vector quantization, compressing the feature sequence by a factor of two.

2.2 Decoder Weight Interpolation

Changing the hop size also alters the dimensionality of the decoder's output layer. Rather than discarding the pretrained decoder weights, we applied one-dimensional linear interpolation to the output projection parameters of the generator head:

$$w'_i = (1 - \alpha_i)\, w_{\lfloor x_i \rfloor} + \alpha_i\, w_{\lceil x_i \rceil}, \quad x_i = \frac{L - 1}{L' - 1}\, i, \quad \alpha_i = x_i - \lfloor x_i \rfloor, \quad i = 0, \ldots, L' - 1, \quad (1)$$

Both the output weight and bias were interpolated to match the new hop size (960 samples). This procedure allows the decoder to retain the spectral characteristics of the original pretrained model while adapting smoothly to the new resolution. Although this interpolation strategy has not been empirically validated in isolation, it provided a straightforward way to transfer pretrained weights without retraining from scratch. We did not conduct ablation experiments with randomly initialized parameters, so it remains unclear whether interpolation contributes to faster convergence or improved stability during fine-tuning.

2.3 Parameter Freezing and Adaptation

All model parameters were frozen except for the decoder. The semantic encoder (frozen HuBERT) and codec encoder were reused directly from the pretrained X-Codec-2.0 checkpoint. The decoder was fine-tuned to accommodate the new hop size and temporal pooling. The resulting model generates 25 Hz discrete tokens and reconstructs 24 kHz audio with improved perceptual quality.

3 Experiments

3.1 Training Datasets

We train our model on a large-scale multilingual corpus totaling approximately 16,000 hours of speech. The dataset combines publicly available TTS and expressive speech corpora from Hugging Face, covering over 100 languages, including English, Mandarin, Malay, Japanese, Korean, Arabic (with various regional dialects), Hindi, Tamil, Bengali, and a wide range of Indic, European, and other low-resource languages. Each audio sample is uniformly resampled to 24 kHz and cropped into randomly selected 5-second segments. No text transcriptions are used during training. The full list of datasets is publicly available at malaysia-ai/Multilingual-TTS.

3.2 Model Initialization

We initialize from the official X-Codec-2.0 checkpoint provided by HKUSTAudio. As described in Section 2, the semantic encoder and codec encoder are frozen, while only the decoder is fine-tuned under the modified hop size (960 samples) and 25 Hz latent rate. Decoder head weights and biases are linearly interpolated from the pretrained model to match the new output dimensionality.

3.3 Training Configuration

All hyperparameters were kept consistent with the original X-Codec-2.0 implementation to ensure comparability. The model was trained on two NVIDIA RTX 3090 Ti GPUs using BF16 mixed precision for 3 million steps with a batch size of 20 per device. Both the generator and discriminator were optimized using the Adam optimizer ($\beta_1 = 0.8, \beta_2 = 0.9$) and a cyclic learning rate schedule with warmup and decay:

$$\text{max\_lr} = 1.0 \times 10^{-4}, \quad \text{min\_lr} = 2.0 \times 10^{-5}.$$

The loss function followed the same multi-objective formulation as X-Codec-2.0, combining mel-spectrogram, adversarial, and semantic losses except for feature matching:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{mel}}\mathcal{L}_{\text{mel}} + \lambda_{\text{adv}}\mathcal{L}_{\text{adv}} + \lambda_{\text{sem}}\mathcal{L}_{\text{sem}}.$$

Empirically, the following coefficients were retained from the original configuration: $\lambda_{\text{mel}} = 15$, $\lambda_{\text{adv}} = 1$, $\lambda_{\text{sem}} = 5$.

To align with the new sampling rate, the `MultiResolutionMelSpectrogramLoss` was configured for a 24 kHz sample rate. This ensures that spectral loss computation matches the decoder's output frequency resolution. Validation was performed every 4000 steps on 2560 held-out samples. Gradient clipping was set to 1.0 for both generator and discriminator to maintain stability. The overall setup reproduces the X-Codec-2.0 training dynamics while isolating the effect of the proposed hop-size and pooling modifications.

4 Evaluation

We evaluate perceptual quality using the UTMOSv2, a neural predictor trained to approximate human mean opinion scores (MOS) from speech spectrograms. Although UTMOSv2 is trained primarily on English speech, it has demonstrated strong correlation with subjective evaluations across diverse acoustic conditions and can be reasonably applied to multilingual speech quality assessment. During evaluation, we follow the inference procedure described in the original UTMOSv2 paper: for each audio sample, predictions are generated five times by randomly selecting different frame offsets of the input waveform, and the final score is obtained by averaging these predictions (i.e., test-time augmentation). This process improves robustness against temporal variability and ensures stable MOS estimation across datasets.

4.1 Evaluation Dataset

We use the multilingual Common Voice 22 test set, which covers 130 languages. All audio samples are preprocessed using the WebRTC Voice Activity Detection (VAD) to remove non-speech regions, including both leading, trailing, and intermediate silences, with a minimum silence duration threshold of 0.2 seconds. For each language, utterances shorter than 20 seconds after trimming are retained, then sorted in descending order of duration. The 500 longest samples per language are selected for evaluation, resulting in a total of 53,927 audio clips. The complete evaluation set is publicly available at malaysia-ai/xcodec2-25TPS-24k/test-set.

4.2 Evaluation with UTMOSv2 Across Checkpoints

To monitor model convergence, we compute UTMOSv2 scores periodically across training checkpoints. Scores are averaged over the full evaluation subset and plotted as a function of training steps. This analysis reveals how perceptual quality evolves during fine-tuning, indicating the point of diminishing returns and the stability of the interpolation-based adaptation process.

4.3 Comparison with Other Models

We further benchmark our model against several recent neural audio codecs: SNAC [5], Encodec [6], DAC [7], Mimi [8], BigCodec [9], SpeechTokenizer [10], WavTokenizer [11], BiCodec [12], Uni-Codec [13], DistilCodec [14], X-Codec-2.0 baseline.

Table 1: UTMOSv2 evaluation across eight representative languages from Common Voice 22.

Language	Model	Codebook Size	Nq	Token Rate (Hz)	UTMOSv2
English	Encodec	1024	1	50	–
	DAC	1024	1	50	–
	Mimi	2048	1	25	–
	BigCodec	8192	1	25	–
	SpeechTokenizer	65536	1	25	–
	WavTokenizer	65536	1	25	–
	X-Codec-2.0 (baseline)	65536	1	50	–
	Ours (25 Hz, 24 kHz)	**65536**	1	**25**	**–**
German	Encodec	1024	1	50	–
	DAC	1024	1	50	–
	Mimi	2048	1	25	–
	BigCodec	8192	1	25	–
	SpeechTokenizer	65536	1	25	–
	WavTokenizer	65536	1	25	–
	X-Codec-2.0 (baseline)	65536	1	50	–
	Ours (25 Hz, 24 kHz)	**65536**	1	**25**	**–**
Spanish	Encodec	1024	1	50	–
	DAC	1024	1	50	–
	Mimi	2048	1	25	–
	BigCodec	8192	1	25	–
	SpeechTokenizer	65536	1	25	–
	WavTokenizer	65536	1	25	–
	X-Codec-2.0 (baseline)	65536	1	50	–
	Ours (25 Hz, 24 kHz)	**65536**	1	**25**	**–**

We report UTMOSv2 results on eight representative languages, Spanish, German, English, French, Italian, Polish, Dutch, and Portuguese which collectively cover a range of phonetic and prosodic diversity among high-resource European languages. Results for the complete 130 languages evaluation are provided in Appendix A.

Overall, our model demonstrates consistent improvement over the X-Codec-2.0 baseline, achieving higher predicted MOS scores across most languages, particularly for high-frequency and expressive speech conditions.

5 Limitation

While the proposed modification improves the perceptual quality and efficiency of X-Codec-2.0, several limitations remain. First, the training data used in this study is primarily drawn from the Common Voice dataset, which is relatively clean and contains limited variation in background noise, speaker style, and expressiveness. As a result, the model does not generalize well to unseen languages or expressive speech domains such as animated or emotional voices. Continued fine-tuning with more diverse and expressive data is expected to alleviate this issue.

Second, all evaluations were conducted using the UTMOSv2 metric, which estimates mean opinion scores from audio features. While convenient for large-scale automatic evaluation, UTMOSv2 does not fully capture human perceptual preferences.

Finally, no downstream applications have been explored. With a vocabulary size of 65,536 and a 25 Hz token rate, each discrete token represents a larger amount of information. This increases the

effective prediction difficulty for autoregressive models that use these tokens, which may lead to higher perplexity.

6 Future Work

While this work focuses on a simple temporal modification introducing additional pooling and increasing the encoder hop size to reduce the latent rate from 50 Hz to 25 Hz in X-Codec-2.0 several directions remain to deepen the understanding and extend the capability of discrete audio representation models.

First, the semantic encoder in X-Codec-2.0, derived from a frozen HuBERT model, is known to capture cross-lingual acoustic and phonetic information effectively. However, its interaction with temporal resolution remains underexplored. Future work could systematically vary the latent frame rate (e.g., from 10 Hz to 100 Hz) and evaluate its impact on perceptual quality, token predictability, and downstream generative modeling. Such experiments would help characterize the decoder's temporal sensitivity and identify the optimal trade-off between information density and reconstruction quality. In particular, understanding how much semantic content is preserved when the latent rate decreases could inform the design of more compact and efficient tokenizers for both speech and general audio.

Second, reducing the latent rate introduces a natural compression bottleneck, potentially limiting the decoder's ability to recover fine-grained acoustic details. This raises an important design question regarding the balance between encoder compression strength and decoder capacity. Increasing the decoder depth, widening its receptive field, or incorporating attention-based upsampling mechanisms may help compensate for lost temporal detail when operating at lower frame rates. A systematic analysis of this trade-off to evaluate how decoder complexity scales with latent rate could provide new insights into optimal model scaling laws for discrete audio codecs.

Third, extending the experiments to larger and more diverse datasets, including noisy or expressive speech (e.g., emotional, conversational, or singing voice data), would reveal the model's robustness and generalization ability. The current results, based on primarily reflect performance on clean and monotonic speech, future work could investigate whether adjusting the latent frame rate interacts differently with noisy or highly variable input conditions.

Finally, downstream evaluation remains an open area. One promising direction is to assess how the discrete tokens perform in text-to-speech (TTS) or speech-language modeling pipelines, especially under large vocabulary sizes (e.g., 65,536 tokens) and shorter temporal sequences. Analyzing token-level perplexity, attention behavior, and reconstruction fidelity in LLM-based TTS finetuning could help determine whether lower frame rates hinder or help generative quality.

7 Acknowledgement

References

[1] Zhen Ye, Peiwen Sun, Jiahe Lei, Hongzhan Lin, Xu Tan, Zheqi Dai, Qiuqiang Kong, Jianyi Chen, Jiahao Pan, Qifeng Liu, Yike Guo, and Wei Xue. Codec does matter: Exploring the semantic shortcoming of codec for audio language model, 2024.

[2] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units, 2021.

[3] Kaito Baba, Wataru Nakata, Yuki Saito, and Hiroshi Saruwatari. The t05 system for the voicemos challenge 2024: Transfer learning from deep image classifier to naturalness mos prediction of high-quality synthetic speech, 2024.

[4] Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus, 2020.

[5] Hubert Siuzdak, Florian Grötschla, and Luca A Lanzendörfer. Snac: Multi-scale neural audio codec. In *Audio Imagination: NeurIPS 2024 Workshop AI-Driven Speech, Music, and Sound Generation*, 2024.

[6] Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression, 2022.

[7] Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar. High-fidelity audio compression with improved rvqgan, 2023.

[8] Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. Moshi: a speech-text foundation model for real-time dialogue, 2024.

[9] Detai Xin, Xu Tan, Shinnosuke Takamichi, and Hiroshi Saruwatari. Bigcodec: Pushing the limits of low-bitrate neural speech codec, 2024.

[10] Xin Zhang, Dong Zhang, Shimin Li, Yaqian Zhou, and Xipeng Qiu. Speechtokenizer: Unified speech tokenizer for speech large language models, 2024.

[11] Shengpeng Ji, Ziyue Jiang, Wen Wang, Yifu Chen, Minghui Fang, Jialong Zuo, Qian Yang, Xize Cheng, Zehan Wang, Ruiqi Li, Ziang Zhang, Xiaoda Yang, Rongjie Huang, Yidi Jiang, Qian Chen, Siqi Zheng, and Zhou Zhao. Wavtokenizer: an efficient acoustic discrete codec tokenizer for audio language modeling, 2025.

[12] Xinsheng Wang, Mingqi Jiang, Ziyang Ma, Ziyu Zhang, Songxiang Liu, Linqin Li, Zheng Liang, Qixi Zheng, Rui Wang, Xiaoqin Feng, Weizhen Bian, Zhen Ye, Sitong Cheng, Ruibin Yuan, Zhixian Zhao, Xinfa Zhu, Jiahao Pan, Liumeng Xue, Pengcheng Zhu, Yunlin Chen, Zhifei Li, Xie Chen, Lei Xie, Yike Guo, and Wei Xue. Spark-tts: An efficient llm-based text-to-speech model with single-stream decoupled speech tokens, 2025.

[13] Yidi Jiang, Qian Chen, Shengpeng Ji, Yu Xi, Wen Wang, Chong Zhang, Xianghu Yue, ShiLiang Zhang, and Haizhou Li. Unicodec: Unified audio codec with single domain-adaptive codebook, 2025.

[14] Rui Wang, Qianguo Sun, Tianrong Chen, Zhiyun Zeng, Junlong Wu, and Jiaxing Zhang. Unitts: An end-to-end tts system without decoupling of acoustic and semantic information, 2025.

A Full Multilingual Evaluation

In this section, we provide the full UTMOSv2 evaluation results across 130 languages.

Model	Codebook Size	Nq	Token Rate (Hz)	Language	UTMOSv2
Encodec	1024	1	50	English	4.12
DAC	1024	1	50	German	3.95
Mimi	2048	1	25	Spanish	4.20
BigCodec	8192	1	25	French	4.35
SpeechTokenizer	65536	1	25	Italian	4.28
WavTokenizer	65536	1	25	Polish	4.22
X-Codec-2.0 (baseline)	65536	1	50	Dutch	4.05
Ours (25 Hz, 24 kHz)	**65536**	1	**25**	**Portuguese**	**4.56**