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 17 test set, the proposed configuration achieves a 0.29 MOS improvement over the original X-Codec-2.0 baseline based on UTMOSv2. The source code, checkpoints and generation comparison released at huggingface.co/malaysia-ai/xcodec2-25TPS-24k.

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 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 in streaming or low-latency scenarios where token-level autoregressive decoding is desirable. Furthermore, the use of fused linear cross-entropy in modern LLM implementations can substantially reduce activation memory during training, enabling efficient extension of large language models to handle audio tokens without prohibitive computational overhead.

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 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 +0.29 improvement in mean opinion score (MOS) as estimated by UTMOSv2 [3] on the multilingual Common Voice [4] 17 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/2421a13e07226d96ac7009d5327d96a84672768c.

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.

4.1 Evaluation Dataset

We use the multilingual Common Voice 17 test set, which covers 116 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 48,489 audio clips. The complete evaluation set is publicly available at malaysia-ai/xcodec2-25TPS-24k/test-set.

4.2 Comparison with Other Models

We further benchmark our model against several recent neural audio codecs: DAC [5], DistilCodec [6], Encodec [7], Mimi [8], Neucodec [9], SNAC [10], SpeechTokenizer [11], UniCodec [12], and X-Codec-2.0 baseline.

Table 1: UTMOSv2 evaluation across eight representative languages from Common Voice 17.

Language	Model	Codebook Size	Nq	Token Rate (Hz)	UTMOSv2 (↑)
NL	DAC	1024	9	86	2.315
	DistilCodec	32768	1	93	2.416
	Encodec	1024	2	75	1.388
	Mimi	2048	32	12.5	2.336
	Neucodec	**65536**	**1**	**50**	**2.55**
	SNAC	4096	3	27	2.193
	SpeechTokenizer	1024	8	50	2.203
	UniCodec	16384	1	75	2.256
	Ours (25 Hz, 24 kHz)	65536	1	25	2.457
	X-Codec-2.0 (baseline)	65536	1	50	2.168
EN	DAC	1024	9	86	2.037
	DistilCodec	32768	1	93	2.13
	Encodec	1024	2	75	1.464
	Mimi	2048	32	12.5	2.032
	Neucodec	**65536**	**1**	**50**	**2.361**
	SNAC	4096	3	27	2.048
	SpeechTokenizer	1024	8	50	2.065
	UniCodec	16384	1	75	2.014
	Ours (25 Hz, 24 kHz)	65536	1	25	2.245
	X-Codec-2.0 (baseline)	65536	1	50	2.086
FR	DAC	1024	9	86	2.085
	DistilCodec	32768	1	93	2.141
	Encodec	1024	2	75	1.386
	Mimi	2048	32	12.5	2.081
	Neucodec	**65536**	**1**	**50**	**2.354**
	SNAC	4096	3	27	2.007
	SpeechTokenizer	1024	8	50	2.003
	UniCodec	16384	1	75	2.026
	Ours (25 Hz, 24 kHz)	65536	1	25	2.217
	X-Codec-2.0 (baseline)	65536	1	50	2.062
IT	DAC	1024	9	86	2.2
	DistilCodec	32768	1	93	2.261
	Encodec	1024	2	75	1.414
	Mimi	2048	32	12.5	2.152
	Neucodec	**65536**	**1**	**50**	**2.494**
	SNAC	4096	3	27	2.098
	SpeechTokenizer	1024	8	50	2.062
	UniCodec	16384	1	75	2.149
	Ours (25 Hz, 24 kHz)	65536	1	25	2.304
	X-Codec-2.0 (baseline)	65536	1	50	2.108
PL	DAC	1024	9	86	2.258
	DistilCodec	32768	1	93	2.312
	Encodec	1024	2	75	1.424
	Mimi	2048	32	12.5	2.296
	Neucodec	**65536**	**1**	**50**	**2.493**
	SNAC	4096	3	27	2.144
	SpeechTokenizer	1024	8	50	2.159
	UniCodec	16384	1	75	2.174
	Ours (25 Hz, 24 kHz)	65536	1	25	2.392
	X-Codec-2.0 (baseline)	65536	1	50	2.122

Language	Model	Codebook Size	Nq	Token Rate (Hz)	UTMOSv2 (↑)
PT	DAC	1024	9	86	2.231
	DistilCodec	32768	1	93	2.308
	Encodec	1024	2	75	1.405
	Mimi	2048	32	12.5	2.273
	Neucodec	**65536**	**1**	**50**	**2.491**
	SNAC	4096	3	27	2.187
	SpeechTokenizer	1024	8	50	2.202
	UniCodec	16384	1	75	2.221
	Ours (25 Hz, 24 kHz)	65536	1	25	2.386
	X-Codec-2.0 (baseline)	65536	1	50	2.136
ES	DAC	1024	9	86	2.204
	DistilCodec	32768	1	93	2.194
	Encodec	1024	2	75	1.369
	Mimi	2048	32	12.5	2.201
	Neucodec	**65536**	**1**	**50**	**2.504**
	SNAC	4096	3	27	2.117
	SpeechTokenizer	1024	8	50	2.096
	UniCodec	16384	1	75	2.111
	Ours (25 Hz, 24 kHz)	65536	1	25	2.314
	X-Codec-2.0 (baseline)	65536	1	50	2.139

We report UTMOSv2 results on eight representative languages, Dutch, English, French, German, Italian, Polish, Portuguese and Spanish, which collectively cover a range of phonetic and prosodic diversity among high-resource European languages. Results for the complete 116 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 based on UTMOSv2 across most languages.

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 predicts mean opinion scores (MOS) directly from audio representations. While this approach enables scalable and reproducible evaluation without the need for human raters, UTMOSv2 may not fully reflect subjective perceptual preferences. Moreover, as the model was primarily trained on English speech, its cross-lingual generalization remains uncertain. Although prior work has shown that UTMOSv2 maintains strong correlation with human judgments under diverse acoustic conditions, further validation on multilingual data is necessary to ensure fair and reliable assessment across languages.

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

Special thanks to my wife for her patience and for not getting too upset about the electricity bill caused by running two RTX 3090 Ti GPUs around the clock for 45 days.

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

[6] 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.

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

[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] Harry Julian, Rachel Beeson, Lohith Konathala, Johanna Ulin, and Jiameng Gao. Finite scalar quantization enables redundant and transmission-robust neural audio compression at low bit-rates, 2025.

[10] 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.

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

[12] 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.

A Full Multilingual Evaluation

In this section, we provide the full UTMOSv2 evaluation results across 116 languages.

Table 2: UTMOSv2 evaluation across 116 languages from Common Voice 17.

Language	Model	Codebook Size	Nq	Token Rate (Hz)	UTMOSv2 (↑)
AB	DAC	1024	9	86	2.33
	DistilCodec	32768	1	93	2.414
	Encodec	1024	2	75	1.268
	Mimi	2048	32	12.5	2.383
	Neucodec	**65536**	**1**	**50**	**2.638**
	SNAC	4096	3	27	2.256
	SpeechTokenizer	1024	8	50	2.181
	UniCodec	16384	1	75	2.392
	Ours (25 Hz, 24 kHz)	65536	1	25	2.409
	X-Codec-2.0 (baseline)	65536	1	50	2.16
AF	DAC	1024	9	86	2.397
	DistilCodec	32768	1	93	2.354
	Encodec	1024	2	75	1.431
	Mimi	2048	32	12.5	2.351
	Neucodec	**65536**	**1**	**50**	**2.416**
	SNAC	4096	3	27	2.215
	SpeechTokenizer	1024	8	50	2.286
	UniCodec	16384	1	75	2.248
	Ours (25 Hz, 24 kHz)	65536	1	25	2.401
	X-Codec-2.0 (baseline)	65536	1	50	2.145
AM	DAC	1024	9	86	2.174
	DistilCodec	32768	1	93	2.288
	Encodec	1024	2	75	1.388
	Mimi	2048	32	12.5	2.304
	Neucodec	**65536**	**1**	**50**	**2.519**
	SNAC	4096	3	27	2.037
	SpeechTokenizer	1024	8	50	2.037
	UniCodec	16384	1	75	2.198
	Ours (25 Hz, 24 kHz)	65536	1	25	2.412
	X-Codec-2.0 (baseline)	65536	1	50	2.02
AR	DAC	1024	9	86	2.239
	DistilCodec	32768	1	93	2.308
	Encodec	1024	2	75	1.294
	Mimi	2048	32	12.5	2.308
	Neucodec	**65536**	**1**	**50**	**2.588**
	SNAC	4096	3	27	2.069
	SpeechTokenizer	1024	8	50	2.152
	UniCodec	16384	1	75	2.298
	Ours (25 Hz, 24 kHz)	65536	1	25	2.428
	X-Codec-2.0 (baseline)	65536	1	50	2.115
AS	DAC	1024	9	86	2.245
	DistilCodec	32768	1	93	2.313
	Encodec	1024	2	75	1.288
	Mimi	2048	32	12.5	2.316
	Neucodec	**65536**	**1**	**50**	**2.602**
	SNAC	4096	3	27	2.112
	SpeechTokenizer	1024	8	50	2.073
	UniCodec	16384	1	75	2.177
	Ours (25 Hz, 24 kHz)	65536	1	25	2.416

Language	Model	Codebook Size	Nq	Token Rate (Hz)	UTMOSv2 (↑)
	X-Codec-2.0 (baseline)	65536	1	50	2.046

Language	Model	Codebook Size	Nq	Token Rate (Hz)	UTMOSv2 (↑)
AST	DAC	1024	9	86	2.023
	DistilCodec	32768	1	93	2.106
	Encodec	1024	2	75	1.378
	Mimi	2048	32	12.5	1.965
	Neucodec	**65536**	**1**	**50**	**2.297**
	SNAC	4096	3	27	1.776
	SpeechTokenizer	1024	8	50	1.883
	UniCodec	16384	1	75	1.932
	Ours (25 Hz, 24 kHz)	65536	1	25	1.928
	X-Codec-2.0 (baseline)	65536	1	50	1.895
AZ	DAC	1024	9	86	2.262
	DistilCodec	32768	1	93	2.376
	Encodec	1024	2	75	1.397
	Mimi	2048	32	12.5	2.431
	Neucodec	**65536**	**1**	**50**	**2.556**
	SNAC	4096	3	27	2.267
	SpeechTokenizer	1024	8	50	2.051
	UniCodec	16384	1	75	2.292
	Ours (25 Hz, 24 kHz)	65536	1	25	2.429
	X-Codec-2.0 (baseline)	65536	1	50	2.216
BA	DAC	1024	9	86	2.138
	DistilCodec	32768	1	93	2.201
	Encodec	1024	2	75	1.166
	Mimi	2048	32	12.5	2.171
	Neucodec	**65536**	**1**	**50**	**2.486**
	SNAC	4096	3	27	1.962
	SpeechTokenizer	1024	8	50	1.947
	UniCodec	16384	1	75	2.185
	Ours (25 Hz, 24 kHz)	65536	1	25	2.329
	X-Codec-2.0 (baseline)	65536	1	50	2.062
BAS	DAC	1024	9	86	1.886
	DistilCodec	32768	1	93	2.019
	Encodec	1024	2	75	1.222
	Mimi	2048	32	12.5	2.039
	Neucodec	**65536**	**1**	**50**	**2.437**
	SNAC	4096	3	27	1.786
	SpeechTokenizer	1024	8	50	1.826
	UniCodec	16384	1	75	1.842
	Ours (25 Hz, 24 kHz)	65536	1	25	2.187
	X-Codec-2.0 (baseline)	65536	1	50	1.856
BE	DAC	1024	9	86	2.236
	DistilCodec	32768	1	93	2.339
	Encodec	1024	2	75	1.321
	Mimi	2048	32	12.5	2.295
	Neucodec	**65536**	**1**	**50**	**2.586**
	SNAC	4096	3	27	2.148
	SpeechTokenizer	1024	8	50	2.154
	UniCodec	16384	1	75	2.347
	Ours (25 Hz, 24 kHz)	65536	1	25	2.437
	X-Codec-2.0 (baseline)	65536	1	50	2.166

Language	Model	Codebook Size	Nq	Token Rate (Hz)	UTMOSv2 (↑)
BG	DAC	1024	9	86	2.341
	DistilCodec	32768	1	93	2.427
	Encodec	1024	2	75	1.402
	Mimi	2048	32	12.5	2.39
	Neucodec	**65536**	**1**	**50**	**2.557**
	SNAC	4096	3	27	2.219
	SpeechTokenizer	1024	8	50	2.191
	UniCodec	16384	1	75	2.284
	Ours (25 Hz, 24 kHz)	65536	1	25	2.493
	X-Codec-2.0 (baseline)	65536	1	50	2.159
BN	DAC	1024	9	86	2.126
	DistilCodec	32768	1	93	2.144
	Encodec	1024	2	75	1.25
	Mimi	2048	32	12.5	2.219
	Neucodec	**65536**	**1**	**50**	**2.501**
	SNAC	4096	3	27	1.954
	SpeechTokenizer	1024	8	50	1.981
	UniCodec	16384	1	75	2.21
	Ours (25 Hz, 24 kHz)	65536	1	25	2.327
	X-Codec-2.0 (baseline)	65536	1	50	2.031
BR	DAC	1024	9	86	2.325
	DistilCodec	32768	1	93	2.4
	Encodec	1024	2	75	1.421
	Mimi	2048	32	12.5	2.33
	Neucodec	**65536**	**1**	**50**	**2.656**
	SNAC	4096	3	27	2.229
	SpeechTokenizer	1024	8	50	2.163
	UniCodec	16384	1	75	2.222
	Ours (25 Hz, 24 kHz)	65536	1	25	2.522
	X-Codec-2.0 (baseline)	65536	1	50	2.153
CA	DAC	1024	9	86	2.138
	DistilCodec	32768	1	93	2.307
	Encodec	1024	2	75	1.331
	Mimi	2048	32	12.5	2.245
	Neucodec	**65536**	**1**	**50**	**2.538**
	SNAC	4096	3	27	2.074
	SpeechTokenizer	1024	8	50	2.131
	UniCodec	16384	1	75	2.222
	Ours (25 Hz, 24 kHz)	65536	1	25	2.37
	X-Codec-2.0 (baseline)	65536	1	50	2.158
CKB	DAC	1024	9	86	2.337
	DistilCodec	32768	1	93	2.473
	Encodec	1024	2	75	1.393
	Mimi	2048	32	12.5	2.419
	Neucodec	**65536**	**1**	**50**	**2.645**
	SNAC	4096	3	27	2.208
	SpeechTokenizer	1024	8	50	2.328
	UniCodec	16384	1	75	2.414
	Ours (25 Hz, 24 kHz)	65536	1	25	2.555
	X-Codec-2.0 (baseline)	65536	1	50	2.349

Language	Model	Codebook Size	Nq	Token Rate (Hz)	UTMOSv2 (↑)
CNH	DAC	1024	9	86	2.45
	DistilCodec	32768	1	93	2.515
	Encodec	1024	2	75	1.403
	Mimi	2048	32	12.5	2.519
	Neucodec	**65536**	**1**	**50**	**2.601**
	SNAC	4096	3	27	2.306
	SpeechTokenizer	1024	8	50	2.328
	UniCodec	16384	1	75	2.286
	Ours (25 Hz, 24 kHz)	65536	1	25	2.552
	X-Codec-2.0 (baseline)	65536	1	50	2.581
CS	DAC	1024	9	86	2.342
	DistilCodec	32768	1	93	2.437
	Encodec	1024	2	75	1.377
	Mimi	2048	32	12.5	2.374
	Neucodec	**65536**	**1**	**50**	**2.588**
	SNAC	4096	3	27	2.241
	SpeechTokenizer	1024	8	50	2.231
	UniCodec	16384	1	75	2.335
	Ours (25 Hz, 24 kHz)	65536	1	25	2.457
	X-Codec-2.0 (baseline)	65536	1	50	2.186
CV	DAC	1024	9	86	2.187
	DistilCodec	32768	1	93	2.28
	Encodec	1024	2	75	1.279
	Mimi	2048	32	12.5	2.221
	Neucodec	**65536**	**1**	**50**	**2.519**
	SNAC	4096	3	27	2.071
	SpeechTokenizer	1024	8	50	2.015
	UniCodec	16384	1	75	2.221
	Ours (25 Hz, 24 kHz)	65536	1	25	2.339
	X-Codec-2.0 (baseline)	65536	1	50	1.987
CY	DAC	1024	9	86	2.33
	DistilCodec	32768	1	93	2.454
	Encodec	1024	2	75	1.345
	Mimi	2048	32	12.5	2.403
	Neucodec	**65536**	**1**	**50**	**2.675**
	SNAC	4096	3	27	2.254
	SpeechTokenizer	1024	8	50	2.261
	UniCodec	16384	1	75	2.436
	Ours (25 Hz, 24 kHz)	65536	1	25	2.572
	X-Codec-2.0 (baseline)	65536	1	50	2.268
DA	DAC	1024	9	86	2.359
	DistilCodec	32768	1	93	2.53
	Encodec	1024	2	75	1.409
	Mimi	2048	32	12.5	2.375
	Neucodec	**65536**	**1**	**50**	**2.617**
	SNAC	4096	3	27	2.256
	SpeechTokenizer	1024	8	50	2.285
	UniCodec	16384	1	75	2.361
	Ours (25 Hz, 24 kHz)	65536	1	25	2.557
	X-Codec-2.0 (baseline)	65536	1	50	2.31

Language	Model	Codebook Size	Nq	Token Rate (Hz)	UTMOSv2 (↑)
DE	DAC	1024	9	86	2.148
	DistilCodec	32768	1	93	2.224
	Encodec	1024	2	75	1.451
	Mimi	2048	32	12.5	2.17
	Neucodec	**65536**	**1**	**50**	**2.459**
	SNAC	4096	3	27	2.095
	SpeechTokenizer	1024	8	50	2.112
	UniCodec	16384	1	75	2.089
	Ours (25 Hz, 24 kHz)	65536	1	25	2.321
	X-Codec-2.0 (baseline)	65536	1	50	2.124
DV	DAC	1024	9	86	2.174
	DistilCodec	32768	1	93	2.268
	Encodec	1024	2	75	1.227
	Mimi	2048	32	12.5	2.303
	Neucodec	**65536**	**1**	**50**	**2.583**
	SNAC	4096	3	27	2.031
	SpeechTokenizer	1024	8	50	2.084
	UniCodec	16384	1	75	2.313
	Ours (25 Hz, 24 kHz)	65536	1	25	2.389
	X-Codec-2.0 (baseline)	65536	1	50	2.109
DYU	DAC	1024	9	86	1.975
	DistilCodec	32768	1	93	1.984
	Encodec	1024	2	75	1.246
	Mimi	2048	32	12.5	2.057
	Neucodec	**65536**	**1**	**50**	**2.5**
	SNAC	4096	3	27	1.866
	SpeechTokenizer	1024	8	50	1.842
	UniCodec	16384	1	75	2.209
	Ours (25 Hz, 24 kHz)	65536	1	25	2.289
	X-Codec-2.0 (baseline)	65536	1	50	1.958
EL	DAC	1024	9	86	2.406
	DistilCodec	32768	1	93	2.472
	Encodec	1024	2	75	1.445
	Mimi	2048	32	12.5	2.385
	Neucodec	**65536**	**1**	**50**	**2.601**
	SNAC	4096	3	27	2.256
	SpeechTokenizer	1024	8	50	2.254
	UniCodec	16384	1	75	2.269
	Ours (25 Hz, 24 kHz)	65536	1	25	2.528
	X-Codec-2.0 (baseline)	65536	1	50	2.224
EN	DAC	1024	9	86	2.037
	DistilCodec	32768	1	93	2.13
	Encodec	1024	2	75	1.464
	Mimi	2048	32	12.5	2.032
	Neucodec	**65536**	**1**	**50**	**2.361**
	SNAC	4096	3	27	2.048
	SpeechTokenizer	1024	8	50	2.065
	UniCodec	16384	1	75	2.014
	Ours (25 Hz, 24 kHz)	65536	1	25	2.245
	X-Codec-2.0 (baseline)	65536	1	50	2.086

Language	Model	Codebook Size	Nq	Token Rate (Hz)	UTMOSv2 (↑)
EO	DAC	1024	9	86	2.245
	DistilCodec	32768	1	93	2.388
	Encodec	1024	2	75	1.265
	Mimi	2048	32	12.5	2.35
	Neucodec	**65536**	**1**	**50**	**2.615**
	SNAC	4096	3	27	2.251
	SpeechTokenizer	1024	8	50	2.106
	UniCodec	16384	1	75	2.316
	Ours (25 Hz, 24 kHz)	65536	1	25	2.399
	X-Codec-2.0 (baseline)	65536	1	50	2.03
ES	DAC	1024	9	86	2.204
	DistilCodec	32768	1	93	2.194
	Encodec	1024	2	75	1.369
	Mimi	2048	32	12.5	2.201
	Neucodec	**65536**	**1**	**50**	**2.504**
	SNAC	4096	3	27	2.117
	SpeechTokenizer	1024	8	50	2.096
	UniCodec	16384	1	75	2.111
	Ours (25 Hz, 24 kHz)	65536	1	25	2.314
	X-Codec-2.0 (baseline)	65536	1	50	2.139
ET	DAC	1024	9	86	2.421
	DistilCodec	32768	1	93	2.541
	Encodec	1024	2	75	1.311
	Mimi	2048	32	12.5	2.488
	Neucodec	**65536**	**1**	**50**	**2.656**
	SNAC	4096	3	27	2.355
	SpeechTokenizer	1024	8	50	2.263
	UniCodec	16384	1	75	2.509
	Ours (25 Hz, 24 kHz)	65536	1	25	2.51
	X-Codec-2.0 (baseline)	65536	1	50	2.084
EU	DAC	1024	9	86	2.215
	DistilCodec	32768	1	93	2.369
	Encodec	1024	2	75	1.279
	Mimi	2048	32	12.5	2.281
	Neucodec	**65536**	**1**	**50**	**2.601**
	SNAC	4096	3	27	2.134
	SpeechTokenizer	1024	8	50	2.125
	UniCodec	16384	1	75	2.339
	Ours (25 Hz, 24 kHz)	65536	1	25	2.411
	X-Codec-2.0 (baseline)	65536	1	50	2.14
FA	DAC	1024	9	86	2.145
	DistilCodec	32768	1	93	2.207
	Encodec	1024	2	75	1.323
	Mimi	2048	32	12.5	2.2
	Neucodec	**65536**	**1**	**50**	**2.564**
	SNAC	4096	3	27	1.996
	SpeechTokenizer	1024	8	50	2.072
	UniCodec	16384	1	75	2.21
	Ours (25 Hz, 24 kHz)	65536	1	25	2.39
	X-Codec-2.0 (baseline)	65536	1	50	2.088

Language	Model	Codebook Size	Nq	Token Rate (Hz)	UTMOSv2 (↑)
FI	DAC	1024	9	86	2.339
	DistilCodec	32768	1	93	2.47
	Encodec	1024	2	75	1.34
	Mimi	2048	32	12.5	2.413
	Neucodec	**65536**	**1**	**50**	**2.659**
	SNAC	4096	3	27	2.26
	SpeechTokenizer	1024	8	50	2.271
	UniCodec	16384	1	75	2.435
	Ours (25 Hz, 24 kHz)	65536	1	25	2.523
	X-Codec-2.0 (baseline)	65536	1	50	2.238
FR	DAC	1024	9	86	2.085
	DistilCodec	32768	1	93	2.141
	Encodec	1024	2	75	1.386
	Mimi	2048	32	12.5	2.081
	Neucodec	**65536**	**1**	**50**	**2.354**
	SNAC	4096	3	27	2.007
	SpeechTokenizer	1024	8	50	2.003
	UniCodec	16384	1	75	2.026
	Ours (25 Hz, 24 kHz)	65536	1	25	2.217
	X-Codec-2.0 (baseline)	65536	1	50	2.062
FY-NL	DAC	1024	9	86	2.236
	DistilCodec	32768	1	93	2.355
	Encodec	1024	2	75	1.336
	Mimi	2048	32	12.5	2.301
	Neucodec	**65536**	**1**	**50**	**2.601**
	SNAC	4096	3	27	2.143
	SpeechTokenizer	1024	8	50	2.155
	UniCodec	16384	1	75	2.322
	Ours (25 Hz, 24 kHz)	65536	1	25	2.471
	X-Codec-2.0 (baseline)	65536	1	50	2.109
GA-IE	DAC	1024	9	86	2.376
	DistilCodec	32768	1	93	2.497
	Encodec	1024	2	75	1.379
	Mimi	2048	32	12.5	2.47
	Neucodec	**65536**	**1**	**50**	**2.687**
	SNAC	4096	3	27	2.29
	SpeechTokenizer	1024	8	50	2.298
	UniCodec	16384	1	75	2.299
	Ours (25 Hz, 24 kHz)	65536	1	25	2.532
	X-Codec-2.0 (baseline)	65536	1	50	2.175
GL	DAC	1024	9	86	2.345
	DistilCodec	32768	1	93	2.48
	Encodec	1024	2	75	1.39
	Mimi	2048	32	12.5	2.385
	Neucodec	**65536**	**1**	**50**	**2.649**
	SNAC	4096	3	27	2.291
	SpeechTokenizer	1024	8	50	2.224
	UniCodec	16384	1	75	2.381
	Ours (25 Hz, 24 kHz)	65536	1	25	2.538
	X-Codec-2.0 (baseline)	65536	1	50	2.276

Language	Model	Codebook Size	Nq	Token Rate (Hz)	UTMOSv2 (↑)
GN	DAC	1024	9	86	2.172
	DistilCodec	32768	1	93	2.238
	Encodec	1024	2	75	1.204
	Mimi	2048	32	12.5	2.285
	Neucodec	**65536**	**1**	**50**	**2.597**
	SNAC	4096	3	27	2.024
	SpeechTokenizer	1024	8	50	2.02
	UniCodec	16384	1	75	2.199
	Ours (25 Hz, 24 kHz)	65536	1	25	2.375
	X-Codec-2.0 (baseline)	65536	1	50	2.036
HA	DAC	1024	9	86	2.158
	DistilCodec	32768	1	93	2.276
	Encodec	1024	2	75	1.293
	Mimi	2048	32	12.5	2.205
	Neucodec	**65536**	**1**	**50**	**2.584**
	SNAC	4096	3	27	2.013
	SpeechTokenizer	1024	8	50	2.023
	UniCodec	16384	1	75	2.078
	Ours (25 Hz, 24 kHz)	65536	1	25	2.341
	X-Codec-2.0 (baseline)	65536	1	50	1.985
HE	DAC	1024	9	86	2.388
	DistilCodec	32768	1	93	2.517
	Encodec	1024	2	75	1.609
	Mimi	2048	32	12.5	2.347
	Neucodec	**65536**	**1**	**50**	**2.577**
	SNAC	4096	3	27	2.1
	SpeechTokenizer	1024	8	50	2.24
	UniCodec	16384	1	75	2.207
	Ours (25 Hz, 24 kHz)	65536	1	25	2.524
	X-Codec-2.0 (baseline)	65536	1	50	2.246
HI	DAC	1024	9	86	2.296
	DistilCodec	32768	1	93	2.397
	Encodec	1024	2	75	1.302
	Mimi	2048	32	12.5	2.346
	Neucodec	**65536**	**1**	**50**	**2.629**
	SNAC	4096	3	27	2.238
	SpeechTokenizer	1024	8	50	2.145
	UniCodec	16384	1	75	2.312
	Ours (25 Hz, 24 kHz)	65536	1	25	2.463
	X-Codec-2.0 (baseline)	65536	1	50	2.2
HSB	DAC	1024	9	86	2.497
	DistilCodec	32768	1	93	2.7
	Encodec	1024	2	75	1.466
	Mimi	2048	32	12.5	2.631
	Neucodec	**65536**	**1**	**50**	**2.803**
	SNAC	4096	3	27	2.47
	SpeechTokenizer	1024	8	50	2.46
	UniCodec	16384	1	75	2.462
	Ours (25 Hz, 24 kHz)	65536	1	25	2.693
	X-Codec-2.0 (baseline)	65536	1	50	2.304

Language	Model	Codebook Size	Nq	Token Rate (Hz)	UTMOSv2 (↑)
HU	DAC	1024	9	86	2.365
	DistilCodec	32768	1	93	2.508
	Encodec	1024	2	75	1.368
	Mimi	2048	32	12.5	2.424
	Neucodec	**65536**	**1**	**50**	**2.658**
	SNAC	4096	3	27	2.303
	SpeechTokenizer	1024	8	50	2.27
	UniCodec	16384	1	75	2.427
	Ours (25 Hz, 24 kHz)	65536	1	25	2.553
	X-Codec-2.0 (baseline)	65536	1	50	2.28
HY-AM	DAC	1024	9	86	2.423
	DistilCodec	32768	1	93	2.585
	Encodec	1024	2	75	1.281
	Mimi	2048	32	12.5	2.478
	Neucodec	**65536**	**1**	**50**	**2.694**
	SNAC	4096	3	27	2.413
	SpeechTokenizer	1024	8	50	2.242
	UniCodec	16384	1	75	2.576
	Ours (25 Hz, 24 kHz)	65536	1	25	2.559
	X-Codec-2.0 (baseline)	65536	1	50	2.282
IA	DAC	1024	9	86	2.522
	DistilCodec	32768	1	93	2.612
	Encodec	1024	2	75	1.353
	Mimi	2048	32	12.5	2.647
	Neucodec	**65536**	**1**	**50**	**2.813**
	SNAC	4096	3	27	2.456
	SpeechTokenizer	1024	8	50	2.413
	UniCodec	16384	1	75	2.5
	Ours (25 Hz, 24 kHz)	65536	1	25	2.649
	X-Codec-2.0 (baseline)	65536	1	50	2.352
ID	DAC	1024	9	86	2.159
	DistilCodec	32768	1	93	2.207
	Encodec	1024	2	75	1.33
	Mimi	2048	32	12.5	2.232
	Neucodec	**65536**	**1**	**50**	**2.499**
	SNAC	4096	3	27	2.084
	SpeechTokenizer	1024	8	50	2.042
	UniCodec	16384	1	75	2.116
	Ours (25 Hz, 24 kHz)	65536	1	25	2.343
	X-Codec-2.0 (baseline)	65536	1	50	2.058
IG	DAC	1024	9	86	2.053
	DistilCodec	32768	1	93	2.313
	Encodec	1024	2	75	1.096
	Mimi	2048	32	12.5	2.38
	Neucodec	**65536**	**1**	**50**	**2.88**
	SNAC	4096	3	27	1.998
	SpeechTokenizer	1024	8	50	2.06
	UniCodec	16384	1	75	2.592
	Ours (25 Hz, 24 kHz)	65536	1	25	2.572
	X-Codec-2.0 (baseline)	65536	1	50	2.127

Language	Model	Codebook Size	Nq	Token Rate (Hz)	UTMOSv2 (↑)
IT	DAC	1024	9	86	2.2
	DistilCodec	32768	1	93	2.261
	Encodec	1024	2	75	1.414
	Mimi	2048	32	12.5	2.152
	Neucodec	**65536**	**1**	**50**	**2.494**
	SNAC	4096	3	27	2.098
	SpeechTokenizer	1024	8	50	2.062
	UniCodec	16384	1	75	2.149
	Ours (25 Hz, 24 kHz)	65536	1	25	2.304
	X-Codec-2.0 (baseline)	65536	1	50	2.108
JA	DAC	1024	9	86	2.464
	DistilCodec	32768	1	93	2.645
	Encodec	1024	2	75	1.348
	Mimi	2048	32	12.5	2.537
	Neucodec	**65536**	**1**	**50**	**2.685**
	SNAC	4096	3	27	2.517
	SpeechTokenizer	1024	8	50	2.268
	UniCodec	16384	1	75	2.577
	Ours (25 Hz, 24 kHz)	65536	1	25	2.612
	X-Codec-2.0 (baseline)	65536	1	50	2.33
KA	DAC	1024	9	86	2.425
	DistilCodec	32768	1	93	2.535
	Encodec	1024	2	75	1.387
	Mimi	2048	32	12.5	2.463
	Neucodec	**65536**	**1**	**50**	**2.655**
	SNAC	4096	3	27	2.375
	SpeechTokenizer	1024	8	50	2.274
	UniCodec	16384	1	75	2.584
	Ours (25 Hz, 24 kHz)	65536	1	25	2.565
	X-Codec-2.0 (baseline)	65536	1	50	2.252
KAB	DAC	1024	9	86	2.068
	DistilCodec	32768	1	93	2.169
	Encodec	1024	2	75	1.381
	Mimi	2048	32	12.5	2.123
	Neucodec	**65536**	**1**	**50**	**2.406**
	SNAC	4096	3	27	2.028
	SpeechTokenizer	1024	8	50	2.006
	UniCodec	16384	1	75	2.054
	Ours (25 Hz, 24 kHz)	65536	1	25	2.272
	X-Codec-2.0 (baseline)	65536	1	50	1.972
KK	DAC	1024	9	86	2.168
	DistilCodec	32768	1	93	2.264
	Encodec	1024	2	75	1.288
	Mimi	2048	32	12.5	2.195
	Neucodec	**65536**	**1**	**50**	**2.457**
	SNAC	4096	3	27	2.069
	SpeechTokenizer	1024	8	50	2.053
	UniCodec	16384	1	75	2.133
	Ours (25 Hz, 24 kHz)	65536	1	25	2.302
	X-Codec-2.0 (baseline)	65536	1	50	2.085

Language	Model	Codebook Size	Nq	Token Rate (Hz)	UTMOSv2 (↑)
KMR	DAC	1024	9	86	2.153
	DistilCodec	32768	1	93	2.259
	Encodec	1024	2	75	1.293
	Mimi	2048	32	12.5	2.249
	Neucodec	**65536**	**1**	**50**	**2.54**
	SNAC	4096	3	27	2.046
	SpeechTokenizer	1024	8	50	2.149
	UniCodec	16384	1	75	2.207
	Ours (25 Hz, 24 kHz)	65536	1	25	2.402
	X-Codec-2.0 (baseline)	65536	1	50	2.128
KO	DAC	1024	9	86	2.274
	DistilCodec	32768	1	93	2.413
	Encodec	1024	2	75	1.307
	Mimi	2048	32	12.5	2.325
	Neucodec	**65536**	**1**	**50**	**2.596**
	SNAC	4096	3	27	2.251
	SpeechTokenizer	1024	8	50	2.123
	UniCodec	16384	1	75	2.269
	Ours (25 Hz, 24 kHz)	65536	1	25	2.345
	X-Codec-2.0 (baseline)	65536	1	50	2.097
KY	DAC	1024	9	86	2.042
	DistilCodec	32768	1	93	2.134
	Encodec	1024	2	75	1.229
	Mimi	2048	32	12.5	2.113
	Neucodec	**65536**	**1**	**50**	**2.386**
	SNAC	4096	3	27	1.976
	SpeechTokenizer	1024	8	50	1.912
	UniCodec	16384	1	75	2.049
	Ours (25 Hz, 24 kHz)	65536	1	25	2.228
	X-Codec-2.0 (baseline)	65536	1	50	1.939
LG	DAC	1024	9	86	1.999
	DistilCodec	32768	1	93	2.007
	Encodec	1024	2	75	1.173
	Mimi	2048	32	12.5	2.105
	Neucodec	**65536**	**1**	**50**	**2.388**
	SNAC	4096	3	27	1.814
	SpeechTokenizer	1024	8	50	1.848
	UniCodec	16384	1	75	2.081
	Ours (25 Hz, 24 kHz)	65536	1	25	2.192
	X-Codec-2.0 (baseline)	65536	1	50	1.856
LIJ	DAC	1024	9	86	2.483
	DistilCodec	32768	1	93	2.578
	Encodec	1024	2	75	1.33
	Mimi	2048	32	12.5	2.57
	Neucodec	**65536**	**1**	**50**	**2.766**
	SNAC	4096	3	27	2.302
	SpeechTokenizer	1024	8	50	2.36
	UniCodec	16384	1	75	2.474
	Ours (25 Hz, 24 kHz)	65536	1	25	2.605
	X-Codec-2.0 (baseline)	65536	1	50	2.345

Language	Model	Codebook Size	Nq	Token Rate (Hz)	UTMOSv2 (↑)
LO	DAC	1024	9	86	2.366
	DistilCodec	32768	1	93	2.379
	Encodec	1024	2	75	1.267
	Mimi	2048	32	12.5	2.408
	Neucodec	**65536**	**1**	**50**	**2.542**
	SNAC	4096	3	27	2.15
	SpeechTokenizer	1024	8	50	2.278
	UniCodec	16384	1	75	2.214
	Ours (25 Hz, 24 kHz)	65536	1	25	2.347
	X-Codec-2.0 (baseline)	65536	1	50	2.148
LT	DAC	1024	9	86	2.303
	DistilCodec	32768	1	93	2.41
	Encodec	1024	2	75	1.303
	Mimi	2048	32	12.5	2.401
	Neucodec	**65536**	**1**	**50**	**2.601**
	SNAC	4096	3	27	2.179
	SpeechTokenizer	1024	8	50	2.225
	UniCodec	16384	1	75	2.367
	Ours (25 Hz, 24 kHz)	65536	1	25	2.446
	X-Codec-2.0 (baseline)	65536	1	50	2.147
LTG	DAC	1024	9	86	2.47
	DistilCodec	32768	1	93	2.603
	Encodec	1024	2	75	1.355
	Mimi	2048	32	12.5	2.542
	Neucodec	**65536**	**1**	**50**	**2.688**
	SNAC	4096	3	27	2.394
	SpeechTokenizer	1024	8	50	2.365
	UniCodec	16384	1	75	2.514
	Ours (25 Hz, 24 kHz)	65536	1	25	2.666
	X-Codec-2.0 (baseline)	65536	1	50	2.345
LV	DAC	1024	9	86	2.343
	DistilCodec	32768	1	93	2.506
	Encodec	1024	2	75	1.326
	Mimi	2048	32	12.5	2.44
	Neucodec	**65536**	**1**	**50**	**2.699**
	SNAC	4096	3	27	2.298
	SpeechTokenizer	1024	8	50	2.233
	UniCodec	16384	1	75	2.447
	Ours (25 Hz, 24 kHz)	65536	1	25	2.56
	X-Codec-2.0 (baseline)	65536	1	50	2.269
MDF	DAC	1024	9	86	2.246
	DistilCodec	32768	1	93	2.264
	Encodec	1024	2	75	1.31
	Mimi	2048	32	12.5	2.163
	Neucodec	**65536**	**1**	**50**	**2.465**
	SNAC	4096	3	27	2.165
	SpeechTokenizer	1024	8	50	2.072
	UniCodec	16384	1	75	2.329
	Ours (25 Hz, 24 kHz)	65536	1	25	2.405
	X-Codec-2.0 (baseline)	65536	1	50	2.229

Language	Model	Codebook Size	Nq	Token Rate (Hz)	UTMOSv2 (↑)
MHR	DAC	1024	9	86	2.297
	DistilCodec	32768	1	93	2.382
	Encodec	1024	2	75	1.239
	Mimi	2048	32	12.5	2.355
	Neucodec	**65536**	**1**	**50**	**2.661**
	SNAC	4096	3	27	2.232
	SpeechTokenizer	1024	8	50	2.126
	UniCodec	16384	1	75	2.423
	Ours (25 Hz, 24 kHz)	65536	1	25	2.541
	X-Codec-2.0 (baseline)	65536	1	50	2.102
MK	DAC	1024	9	86	2.37
	DistilCodec	32768	1	93	2.467
	Encodec	1024	2	75	1.298
	Mimi	2048	32	12.5	2.366
	Neucodec	**65536**	**1**	**50**	**2.647**
	SNAC	4096	3	27	2.182
	SpeechTokenizer	1024	8	50	2.182
	UniCodec	16384	1	75	2.371
	Ours (25 Hz, 24 kHz)	65536	1	25	2.467
	X-Codec-2.0 (baseline)	65536	1	50	2.193
ML	DAC	1024	9	86	2.468
	DistilCodec	32768	1	93	2.573
	Encodec	1024	2	75	1.336
	Mimi	2048	32	12.5	2.541
	Neucodec	**65536**	**1**	**50**	**2.795**
	SNAC	4096	3	27	2.361
	SpeechTokenizer	1024	8	50	2.338
	UniCodec	16384	1	75	2.36
	Ours (25 Hz, 24 kHz)	65536	1	25	2.645
	X-Codec-2.0 (baseline)	65536	1	50	2.222
MN	DAC	1024	9	86	2.186
	DistilCodec	32768	1	93	2.244
	Encodec	1024	2	75	1.453
	Mimi	2048	32	12.5	2.174
	Neucodec	**65536**	**1**	**50**	**2.307**
	SNAC	4096	3	27	2.046
	SpeechTokenizer	1024	8	50	2.044
	UniCodec	16384	1	75	2.018
	Ours (25 Hz, 24 kHz)	65536	1	25	2.264
	X-Codec-2.0 (baseline)	65536	1	50	2.161
MR	DAC	1024	9	86	2.1
	DistilCodec	32768	1	93	2.121
	Encodec	1024	2	75	1.119
	Mimi	2048	32	12.5	2.217
	Neucodec	**65536**	**1**	**50**	**2.587**
	SNAC	4096	3	27	2.012
	SpeechTokenizer	1024	8	50	1.843
	UniCodec	16384	1	75	2.285
	Ours (25 Hz, 24 kHz)	65536	1	25	2.364
	X-Codec-2.0 (baseline)	65536	1	50	1.908

Language	Model	Codebook Size	Nq	Token Rate (Hz)	UTMOSv2 (↑)
MRJ	DAC	1024	9	86	2.179
	DistilCodec	32768	1	93	2.252
	Encodec	1024	2	75	1.145
	Mimi	2048	32	12.5	2.263
	Neucodec	**65536**	**1**	**50**	**2.523**
	SNAC	4096	3	27	1.976
	SpeechTokenizer	1024	8	50	2.037
	UniCodec	16384	1	75	2.279
	Ours (25 Hz, 24 kHz)	65536	1	25	2.451
	X-Codec-2.0 (baseline)	65536	1	50	1.979
MT	DAC	1024	9	86	2.424
	DistilCodec	32768	1	93	2.518
	Encodec	1024	2	75	1.409
	Mimi	2048	32	12.5	2.444
	Neucodec	**65536**	**1**	**50**	**2.639**
	SNAC	4096	3	27	2.321
	SpeechTokenizer	1024	8	50	2.235
	UniCodec	16384	1	75	2.337
	Ours (25 Hz, 24 kHz)	65536	1	25	2.568
	X-Codec-2.0 (baseline)	65536	1	50	2.235
MYV	DAC	1024	9	86	2.253
	DistilCodec	32768	1	93	2.334
	Encodec	1024	2	75	1.242
	Mimi	2048	32	12.5	2.371
	Neucodec	**65536**	**1**	**50**	**2.738**
	SNAC	4096	3	27	2.106
	SpeechTokenizer	1024	8	50	2.146
	UniCodec	16384	1	75	2.347
	Ours (25 Hz, 24 kHz)	65536	1	25	2.549
	X-Codec-2.0 (baseline)	65536	1	50	2.164
NAN-TW	DAC	1024	9	86	2.208
	DistilCodec	32768	1	93	2.358
	Encodec	1024	2	75	1.357
	Mimi	2048	32	12.5	2.278
	Neucodec	**65536**	**1**	**50**	**2.473**
	SNAC	4096	3	27	2.148
	SpeechTokenizer	1024	8	50	2.178
	UniCodec	16384	1	75	2.121
	Ours (25 Hz, 24 kHz)	65536	1	25	2.372
	X-Codec-2.0 (baseline)	65536	1	50	2.185
NE-NP	DAC	1024	9	86	2.422
	DistilCodec	32768	1	93	2.45
	Encodec	1024	2	75	1.453
	Mimi	2048	32	12.5	2.461
	Neucodec	**65536**	**1**	**50**	**2.637**
	SNAC	4096	3	27	2.4
	SpeechTokenizer	1024	8	50	2.298
	UniCodec	16384	1	75	2.365
	Ours (25 Hz, 24 kHz)	65536	1	25	2.554
	X-Codec-2.0 (baseline)	65536	1	50	2.279

Language	Model	Codebook Size	Nq	Token Rate (Hz)	UTMOSv2 (↑)
NHI	DAC	1024	9	86	2.517
	DistilCodec	32768	1	93	2.409
	Encodec	1024	2	75	1.148
	Mimi	2048	32	12.5	2.134
	Neucodec	65536	1	50	2.53
	SNAC	4096	3	27	2.561
	SpeechTokenizer	1024	8	50	2.461
	UniCodec	**16384**	**1**	**75**	**2.87**
	Ours (25 Hz, 24 kHz)	65536	1	25	2.722
	X-Codec-2.0 (baseline)	65536	1	50	2.431
NL	DAC	1024	9	86	2.315
	DistilCodec	32768	1	93	2.416
	Encodec	1024	2	75	1.388
	Mimi	2048	32	12.5	2.336
	Neucodec	**65536**	**1**	**50**	**2.55**
	SNAC	4096	3	27	2.193
	SpeechTokenizer	1024	8	50	2.203
	UniCodec	16384	1	75	2.256
	Ours (25 Hz, 24 kHz)	65536	1	25	2.457
	X-Codec-2.0 (baseline)	65536	1	50	2.168
NN-NO	DAC	1024	9	86	2.534
	DistilCodec	32768	1	93	2.718
	Encodec	1024	2	75	1.437
	Mimi	2048	32	12.5	2.627
	Neucodec	**65536**	**1**	**50**	**2.761**
	SNAC	4096	3	27	2.495
	SpeechTokenizer	1024	8	50	2.407
	UniCodec	16384	1	75	2.536
	Ours (25 Hz, 24 kHz)	65536	1	25	2.727
	X-Codec-2.0 (baseline)	65536	1	50	2.369
OC	DAC	1024	9	86	2.349
	DistilCodec	32768	1	93	2.441
	Encodec	1024	2	75	1.296
	Mimi	2048	32	12.5	2.39
	Neucodec	**65536**	**1**	**50**	**2.681**
	SNAC	4096	3	27	2.287
	SpeechTokenizer	1024	8	50	2.187
	UniCodec	16384	1	75	2.398
	Ours (25 Hz, 24 kHz)	65536	1	25	2.517
	X-Codec-2.0 (baseline)	65536	1	50	2.208
OR	DAC	1024	9	86	2.443
	DistilCodec	32768	1	93	2.599
	Encodec	1024	2	75	1.27
	Mimi	2048	32	12.5	2.491
	Neucodec	**65536**	**1**	**50**	**2.738**
	SNAC	4096	3	27	2.252
	SpeechTokenizer	1024	8	50	2.261
	UniCodec	16384	1	75	2.391
	Ours (25 Hz, 24 kHz)	65536	1	25	2.624
	X-Codec-2.0 (baseline)	65536	1	50	2.194

Language	Model	Codebook Size	Nq	Token Rate (Hz)	UTMOSv2 (↑)
OS	DAC	1024	9	86	2.145
	DistilCodec	32768	1	93	2.104
	Encodec	1024	2	75	1.194
	Mimi	2048	32	12.5	2.008
	Neucodec	65536	1	50	2.325
	SNAC	4096	3	27	1.723
	SpeechTokenizer	1024	8	50	1.959
	UniCodec	16384	1	75	2.034
	Ours (25 Hz, 24 kHz)	**65536**	**1**	**25**	**2.344**
	X-Codec-2.0 (baseline)	65536	1	50	2.082
PA-IN	DAC	1024	9	86	2.391
	DistilCodec	32768	1	93	2.508
	Encodec	1024	2	75	1.316
	Mimi	2048	32	12.5	2.539
	Neucodec	**65536**	**1**	**50**	**2.708**
	SNAC	4096	3	27	2.394
	SpeechTokenizer	1024	8	50	2.311
	UniCodec	16384	1	75	2.487
	Ours (25 Hz, 24 kHz)	65536	1	25	2.631
	X-Codec-2.0 (baseline)	65536	1	50	2.37
PL	DAC	1024	9	86	2.258
	DistilCodec	32768	1	93	2.312
	Encodec	1024	2	75	1.424
	Mimi	2048	32	12.5	2.296
	Neucodec	**65536**	**1**	**50**	**2.493**
	SNAC	4096	3	27	2.144
	SpeechTokenizer	1024	8	50	2.159
	UniCodec	16384	1	75	2.174
	Ours (25 Hz, 24 kHz)	65536	1	25	2.392
	X-Codec-2.0 (baseline)	65536	1	50	2.122
PS	DAC	1024	9	86	1.876
	DistilCodec	32768	1	93	1.996
	Encodec	1024	2	75	1.146
	Mimi	2048	32	12.5	1.985
	Neucodec	**65536**	**1**	**50**	**2.619**
	SNAC	4096	3	27	1.859
	SpeechTokenizer	1024	8	50	1.731
	UniCodec	16384	1	75	2.026
	Ours (25 Hz, 24 kHz)	65536	1	25	2.177
	X-Codec-2.0 (baseline)	65536	1	50	1.932
PT	DAC	1024	9	86	2.231
	DistilCodec	32768	1	93	2.308
	Encodec	1024	2	75	1.405
	Mimi	2048	32	12.5	2.273
	Neucodec	**65536**	**1**	**50**	**2.491**
	SNAC	4096	3	27	2.187
	SpeechTokenizer	1024	8	50	2.202
	UniCodec	16384	1	75	2.221
	Ours (25 Hz, 24 kHz)	65536	1	25	2.386
	X-Codec-2.0 (baseline)	65536	1	50	2.136

Language	Model	Codebook Size	Nq	Token Rate (Hz)	UTMOSv2 (↑)
RM-SURSILV	DAC	1024	9	86	2.286
	DistilCodec	32768	1	93	2.517
	Encodec	1024	2	75	1.291
	Mimi	2048	32	12.5	2.49
	Neucodec	**65536**	**1**	**50**	**2.744**
	SNAC	4096	3	27	2.43
	SpeechTokenizer	1024	8	50	2.23
	UniCodec	16384	1	75	2.451
	Ours (25 Hz, 24 kHz)	65536	1	25	2.598
	X-Codec-2.0 (baseline)	65536	1	50	2.326
RM-VALLADER	DAC	1024	9	86	2.304
	DistilCodec	32768	1	93	2.481
	Encodec	1024	2	75	1.241
	Mimi	2048	32	12.5	2.392
	Neucodec	**65536**	**1**	**50**	**2.697**
	SNAC	4096	3	27	2.344
	SpeechTokenizer	1024	8	50	2.148
	UniCodec	16384	1	75	2.451
	Ours (25 Hz, 24 kHz)	65536	1	25	2.627
	X-Codec-2.0 (baseline)	65536	1	50	2.387
RO	DAC	1024	9	86	2.318
	DistilCodec	32768	1	93	2.481
	Encodec	1024	2	75	1.381
	Mimi	2048	32	12.5	2.413
	Neucodec	**65536**	**1**	**50**	**2.594**
	SNAC	4096	3	27	2.272
	SpeechTokenizer	1024	8	50	2.244
	UniCodec	16384	1	75	2.362
	Ours (25 Hz, 24 kHz)	65536	1	25	2.526
	X-Codec-2.0 (baseline)	65536	1	50	2.191
RU	DAC	1024	9	86	2.288
	DistilCodec	32768	1	93	2.361
	Encodec	1024	2	75	1.38
	Mimi	2048	32	12.5	2.302
	Neucodec	**65536**	**1**	**50**	**2.594**
	SNAC	4096	3	27	2.159
	SpeechTokenizer	1024	8	50	2.176
	UniCodec	16384	1	75	2.275
	Ours (25 Hz, 24 kHz)	65536	1	25	2.368
	X-Codec-2.0 (baseline)	65536	1	50	2.134
RW	DAC	1024	9	86	1.774
	DistilCodec	32768	1	93	1.759
	Encodec	1024	2	75	1.148
	Mimi	2048	32	12.5	1.8
	Neucodec	**65536**	**1**	**50**	**2.212**
	SNAC	4096	3	27	1.706
	SpeechTokenizer	1024	8	50	1.633
	UniCodec	16384	1	75	1.87
	Ours (25 Hz, 24 kHz)	65536	1	25	1.954
	X-Codec-2.0 (baseline)	65536	1	50	1.645

Language	Model	Codebook Size	Nq	Token Rate (Hz)	UTMOSv2 (↑)
SAH	DAC	1024	9	86	2.235
	DistilCodec	32768	1	93	2.426
	Encodec	1024	2	75	1.245
	Mimi	2048	32	12.5	2.288
	Neucodec	**65536**	**1**	**50**	**2.536**
	SNAC	4096	3	27	2.153
	SpeechTokenizer	1024	8	50	2.141
	UniCodec	16384	1	75	2.288
	Ours (25 Hz, 24 kHz)	65536	1	25	2.404
	X-Codec-2.0 (baseline)	65536	1	50	2.188
SAT	DAC	1024	9	86	2.318
	DistilCodec	32768	1	93	2.367
	Encodec	1024	2	75	1.194
	Mimi	2048	32	12.5	2.427
	Neucodec	**65536**	**1**	**50**	**2.625**
	SNAC	4096	3	27	2.008
	SpeechTokenizer	1024	8	50	2.183
	UniCodec	16384	1	75	2.197
	Ours (25 Hz, 24 kHz)	65536	1	25	2.429
	X-Codec-2.0 (baseline)	65536	1	50	2.123
SC	DAC	1024	9	86	2.168
	DistilCodec	32768	1	93	2.262
	Encodec	1024	2	75	1.183
	Mimi	2048	32	12.5	2.254
	Neucodec	**65536**	**1**	**50**	**2.622**
	SNAC	4096	3	27	2.113
	SpeechTokenizer	1024	8	50	2.106
	UniCodec	16384	1	75	2.203
	Ours (25 Hz, 24 kHz)	65536	1	25	2.334
	X-Codec-2.0 (baseline)	65536	1	50	1.965
SK	DAC	1024	9	86	2.276
	DistilCodec	32768	1	93	2.341
	Encodec	1024	2	75	1.427
	Mimi	2048	32	12.5	2.33
	Neucodec	**65536**	**1**	**50**	**2.524**
	SNAC	4096	3	27	2.186
	SpeechTokenizer	1024	8	50	2.196
	UniCodec	16384	1	75	2.198
	Ours (25 Hz, 24 kHz)	65536	1	25	2.453
	X-Codec-2.0 (baseline)	65536	1	50	2.143
SKR	DAC	1024	9	86	2.079
	DistilCodec	32768	1	93	2.111
	Encodec	1024	2	75	1.23
	Mimi	2048	32	12.5	2.176
	Neucodec	**65536**	**1**	**50**	**2.522**
	SNAC	4096	3	27	1.83
	SpeechTokenizer	1024	8	50	1.969
	UniCodec	16384	1	75	2.127
	Ours (25 Hz, 24 kHz)	65536	1	25	2.3
	X-Codec-2.0 (baseline)	65536	1	50	1.965

Language	Model	Codebook Size	Nq	Token Rate (Hz)	UTMOSv2 (↑)
SL	DAC	1024	9	86	2.436
	DistilCodec	32768	1	93	2.594
	Encodec	1024	2	75	1.449
	Mimi	2048	32	12.5	2.476
	Neucodec	**65536**	**1**	**50**	**2.725**
	SNAC	4096	3	27	2.392
	SpeechTokenizer	1024	8	50	2.394
	UniCodec	16384	1	75	2.383
	Ours (25 Hz, 24 kHz)	65536	1	25	2.639
	X-Codec-2.0 (baseline)	65536	1	50	2.334
SQ	DAC	1024	9	86	2.331
	DistilCodec	32768	1	93	2.501
	Encodec	1024	2	75	1.36
	Mimi	2048	32	12.5	2.343
	Neucodec	**65536**	**1**	**50**	**2.602**
	SNAC	4096	3	27	2.261
	SpeechTokenizer	1024	8	50	2.263
	UniCodec	16384	1	75	2.398
	Ours (25 Hz, 24 kHz)	65536	1	25	2.538
	X-Codec-2.0 (baseline)	65536	1	50	2.252
SR	DAC	1024	9	86	2.391
	DistilCodec	32768	1	93	2.47
	Encodec	1024	2	75	1.462
	Mimi	2048	32	12.5	2.394
	Neucodec	**65536**	**1**	**50**	**2.607**
	SNAC	4096	3	27	2.259
	SpeechTokenizer	1024	8	50	2.318
	UniCodec	16384	1	75	2.258
	Ours (25 Hz, 24 kHz)	65536	1	25	2.53
	X-Codec-2.0 (baseline)	65536	1	50	2.286
SV-SE	DAC	1024	9	86	2.291
	DistilCodec	32768	1	93	2.431
	Encodec	1024	2	75	1.385
	Mimi	2048	32	12.5	2.341
	Neucodec	**65536**	**1**	**50**	**2.556**
	SNAC	4096	3	27	2.179
	SpeechTokenizer	1024	8	50	2.22
	UniCodec	16384	1	75	2.308
	Ours (25 Hz, 24 kHz)	65536	1	25	2.463
	X-Codec-2.0 (baseline)	65536	1	50	2.202
SW	DAC	1024	9	86	1.936
	DistilCodec	32768	1	93	1.929
	Encodec	1024	2	75	1.157
	Mimi	2048	32	12.5	2.005
	Neucodec	**65536**	**1**	**50**	**2.371**
	SNAC	4096	3	27	1.845
	SpeechTokenizer	1024	8	50	1.811
	UniCodec	16384	1	75	2.029
	Ours (25 Hz, 24 kHz)	65536	1	25	2.181
	X-Codec-2.0 (baseline)	65536	1	50	1.777

Language	Model	Codebook Size	Nq	Token Rate (Hz)	UTMOSv2 (↑)
TA	DAC	1024	9	86	2.222
	DistilCodec	32768	1	93	2.357
	Encodec	1024	2	75	1.197
	Mimi	2048	32	12.5	2.327
	Neucodec	**65536**	**1**	**50**	**2.673**
	SNAC	4096	3	27	2.11
	SpeechTokenizer	1024	8	50	2.042
	UniCodec	16384	1	75	2.331
	Ours (25 Hz, 24 kHz)	65536	1	25	2.379
	X-Codec-2.0 (baseline)	65536	1	50	2.027
TE	DAC	1024	9	86	2.304
	DistilCodec	32768	1	93	2.476
	Encodec	1024	2	75	1.309
	Mimi	2048	32	12.5	2.327
	Neucodec	**65536**	**1**	**50**	**2.63**
	SNAC	4096	3	27	2.385
	SpeechTokenizer	1024	8	50	2.237
	UniCodec	16384	1	75	2.306
	Ours (25 Hz, 24 kHz)	65536	1	25	2.534
	X-Codec-2.0 (baseline)	65536	1	50	2.273
TH	DAC	1024	9	86	2.318
	DistilCodec	32768	1	93	2.473
	Encodec	1024	2	75	1.251
	Mimi	2048	32	12.5	2.362
	Neucodec	**65536**	**1**	**50**	**2.575**
	SNAC	4096	3	27	2.344
	SpeechTokenizer	1024	8	50	2.223
	UniCodec	16384	1	75	2.399
	Ours (25 Hz, 24 kHz)	65536	1	25	2.489
	X-Codec-2.0 (baseline)	65536	1	50	2.317
TI	DAC	1024	9	86	2.102
	DistilCodec	32768	1	93	2.27
	Encodec	1024	2	75	1.389
	Mimi	2048	32	12.5	2.245
	Neucodec	65536	1	50	2.171
	SNAC	4096	3	27	1.867
	SpeechTokenizer	1024	8	50	1.939
	UniCodec	16384	1	75	2.338
	Ours (25 Hz, 24 kHz)	**65536**	**1**	**25**	**2.525**
	X-Codec-2.0 (baseline)	65536	1	50	2.216
TIG	DAC	1024	9	86	2.467
	DistilCodec	32768	1	93	2.619
	Encodec	1024	2	75	1.449
	Mimi	2048	32	12.5	2.535
	Neucodec	**65536**	**1**	**50**	**2.752**
	SNAC	4096	3	27	2.447
	SpeechTokenizer	1024	8	50	2.447
	UniCodec	16384	1	75	2.308
	Ours (25 Hz, 24 kHz)	65536	1	25	2.573
	X-Codec-2.0 (baseline)	65536	1	50	2.423

Language	Model	Codebook Size	Nq	Token Rate (Hz)	UTMOSv2 (↑)
TK	DAC	1024	9	86	2.164
	DistilCodec	32768	1	93	2.266
	Encodec	1024	2	75	1.249
	Mimi	2048	32	12.5	2.268
	Neucodec	**65536**	**1**	**50**	**2.517**
	SNAC	4096	3	27	2.005
	SpeechTokenizer	1024	8	50	2.05
	UniCodec	16384	1	75	2.131
	Ours (25 Hz, 24 kHz)	65536	1	25	2.299
	X-Codec-2.0 (baseline)	65536	1	50	2.073
TOK	DAC	1024	9	86	2.389
	DistilCodec	32768	1	93	2.572
	Encodec	1024	2	75	1.374
	Mimi	2048	32	12.5	2.476
	Neucodec	**65536**	**1**	**50**	**2.714**
	SNAC	4096	3	27	2.444
	SpeechTokenizer	1024	8	50	2.306
	UniCodec	16384	1	75	2.449
	Ours (25 Hz, 24 kHz)	65536	1	25	2.609
	X-Codec-2.0 (baseline)	65536	1	50	2.32
TR	DAC	1024	9	86	2.283
	DistilCodec	32768	1	93	2.386
	Encodec	1024	2	75	1.415
	Mimi	2048	32	12.5	2.345
	Neucodec	**65536**	**1**	**50**	**2.56**
	SNAC	4096	3	27	2.189
	SpeechTokenizer	1024	8	50	2.194
	UniCodec	16384	1	75	2.302
	Ours (25 Hz, 24 kHz)	65536	1	25	2.45
	X-Codec-2.0 (baseline)	65536	1	50	2.217
TT	DAC	1024	9	86	2.133
	DistilCodec	32768	1	93	2.233
	Encodec	1024	2	75	1.196
	Mimi	2048	32	12.5	2.218
	Neucodec	**65536**	**1**	**50**	**2.521**
	SNAC	4096	3	27	2.045
	SpeechTokenizer	1024	8	50	1.979
	UniCodec	16384	1	75	2.204
	Ours (25 Hz, 24 kHz)	65536	1	25	2.327
	X-Codec-2.0 (baseline)	65536	1	50	1.924
TW	DAC	1024	9	86	2.169
	DistilCodec	32768	1	93	2.486
	Encodec	1024	2	75	1.224
	Mimi	2048	32	12.5	2.369
	Neucodec	**65536**	**1**	**50**	**2.678**
	SNAC	4096	3	27	1.97
	SpeechTokenizer	1024	8	50	2.331
	UniCodec	16384	1	75	2.193
	Ours (25 Hz, 24 kHz)	65536	1	25	2.514
	X-Codec-2.0 (baseline)	65536	1	50	2.359

Language	Model	Codebook Size	Nq	Token Rate (Hz)	UTMOSv2 (↑)
UG	DAC	1024	9	86	2.348
	DistilCodec	32768	1	93	2.502
	Encodec	1024	2	75	1.323
	Mimi	2048	32	12.5	2.381
	Neucodec	**65536**	**1**	**50**	**2.584**
	SNAC	4096	3	27	2.303
	SpeechTokenizer	1024	8	50	2.263
	UniCodec	16384	1	75	2.406
	Ours (25 Hz, 24 kHz)	65536	1	25	2.529
	X-Codec-2.0 (baseline)	65536	1	50	2.349
UK	DAC	1024	9	86	2.337
	DistilCodec	32768	1	93	2.449
	Encodec	1024	2	75	1.316
	Mimi	2048	32	12.5	2.402
	Neucodec	**65536**	**1**	**50**	**2.674**
	SNAC	4096	3	27	2.254
	SpeechTokenizer	1024	8	50	2.218
	UniCodec	16384	1	75	2.426
	Ours (25 Hz, 24 kHz)	65536	1	25	2.531
	X-Codec-2.0 (baseline)	65536	1	50	2.155
UR	DAC	1024	9	86	2.298
	DistilCodec	32768	1	93	2.409
	Encodec	1024	2	75	1.266
	Mimi	2048	32	12.5	2.418
	Neucodec	**65536**	**1**	**50**	**2.692**
	SNAC	4096	3	27	2.374
	SpeechTokenizer	1024	8	50	2.152
	UniCodec	16384	1	75	2.496
	Ours (25 Hz, 24 kHz)	65536	1	25	2.484
	X-Codec-2.0 (baseline)	65536	1	50	2.13
UZ	DAC	1024	9	86	2.09
	DistilCodec	32768	1	93	2.19
	Encodec	1024	2	75	1.284
	Mimi	2048	32	12.5	2.201
	Neucodec	**65536**	**1**	**50**	**2.462**
	SNAC	4096	3	27	1.966
	SpeechTokenizer	1024	8	50	2.024
	UniCodec	16384	1	75	2.16
	Ours (25 Hz, 24 kHz)	65536	1	25	2.312
	X-Codec-2.0 (baseline)	65536	1	50	2.037
VI	DAC	1024	9	86	2.194
	DistilCodec	32768	1	93	2.25
	Encodec	1024	2	75	1.249
	Mimi	2048	32	12.5	2.209
	Neucodec	**65536**	**1**	**50**	**2.467**
	SNAC	4096	3	27	2.168
	SpeechTokenizer	1024	8	50	2.063
	UniCodec	16384	1	75	2.087
	Ours (25 Hz, 24 kHz)	65536	1	25	2.252
	X-Codec-2.0 (baseline)	65536	1	50	2.064

Language	Model	Codebook Size	Nq	Token Rate (Hz)	UTMOSv2 (↑)
VOT	DAC	1024	9	86	2.001
	DistilCodec	32768	1	93	2.409
	Encodec	1024	2	75	1.371
	Mimi	2048	32	12.5	1.892
	Neucodec	**65536**	**1**	**50**	**2.538**
	SNAC	4096	3	27	2.323
	SpeechTokenizer	1024	8	50	1.799
	UniCodec	16384	1	75	2.105
	Ours (25 Hz, 24 kHz)	65536	1	25	1.897
	X-Codec-2.0 (baseline)	65536	1	50	2.14
YI	DAC	1024	9	86	2.398
	DistilCodec	32768	1	93	2.702
	Encodec	1024	2	75	1.348
	Mimi	2048	32	12.5	2.39
	Neucodec	**65536**	**1**	**50**	**2.863**
	SNAC	4096	3	27	2.323
	SpeechTokenizer	1024	8	50	2.138
	UniCodec	16384	1	75	2.506
	Ours (25 Hz, 24 kHz)	65536	1	25	2.485
	X-Codec-2.0 (baseline)	65536	1	50	2.483
YO	DAC	1024	9	86	2.096
	DistilCodec	32768	1	93	2.125
	Encodec	1024	2	75	1.21
	Mimi	2048	32	12.5	2.182
	Neucodec	**65536**	**1**	**50**	**2.414**
	SNAC	4096	3	27	1.974
	SpeechTokenizer	1024	8	50	1.932
	UniCodec	16384	1	75	2.105
	Ours (25 Hz, 24 kHz)	65536	1	25	2.304
	X-Codec-2.0 (baseline)	65536	1	50	2.022
YUE	DAC	1024	9	86	2.42
	DistilCodec	32768	1	93	2.536
	Encodec	1024	2	75	1.325
	Mimi	2048	32	12.5	2.476
	Neucodec	**65536**	**1**	**50**	**2.647**
	SNAC	4096	3	27	2.374
	SpeechTokenizer	1024	8	50	2.339
	UniCodec	16384	1	75	2.474
	Ours (25 Hz, 24 kHz)	65536	1	25	2.59
	X-Codec-2.0 (baseline)	65536	1	50	2.434
ZGH	DAC	1024	9	86	2.337
	DistilCodec	32768	1	93	2.517
	Encodec	1024	2	75	1.426
	Mimi	2048	32	12.5	2.437
	Neucodec	**65536**	**1**	**50**	**2.673**
	SNAC	4096	3	27	2.24
	SpeechTokenizer	1024	8	50	2.245
	UniCodec	16384	1	75	2.223
	Ours (25 Hz, 24 kHz)	65536	1	25	2.607
	X-Codec-2.0 (baseline)	65536	1	50	2.176

Language	Model	Codebook Size	Nq	Token Rate (Hz)	UTMOSv2 (↑)
ZH-CN	DAC	1024	9	86	2.181
	DistilCodec	32768	1	93	2.252
	Encodec	1024	2	75	1.304
	Mimi	2048	32	12.5	2.184
	Neucodec	**65536**	**1**	**50**	**2.471**
	SNAC	4096	3	27	2.071
	SpeechTokenizer	1024	8	50	2.008
	UniCodec	16384	1	75	2.155
	Ours (25 Hz, 24 kHz)	65536	1	25	2.288
	X-Codec-2.0 (baseline)	65536	1	50	2.057