ChindaTTS Open: An Open-Weight Thai-English Text-to-Speech Model (Preview)

Technical Report

Sreyroth Phem, Chonlasit Sangbunlue, Smart Wattanapornmongkol, Krittapad Harnchang, Natthawat Chaimongkol, Thodsaporn Chay-intr, and Kobkrit Viriyayudhakorn
AI Division, iApp Technology Co., Ltd., Thailand

Abstract

We present ChindaTTS Open Preview, an open-weight Thai–English text-to-speech (TTS) model in our ChindaTTS series. The model is built upon a transformer-based architecture incorporating Multi-Scale Neural Audio Codec (SNAC) acoustic tokens as discrete representations to simplify the mapping between linguistic and acoustic domains. Training was performed in two stages: first, pre-training on 1,696 hours of diverse Thai–English corpora; then fine-tuning on 65.61 hours of high-fidelity speech data to improve prosody and intelligibility. ChindaTTS Open Preview achieves 10.56% CER on 1,000 randomly selected samples from the Common Voice 17 Thai Test Set, approaching the human-speech upper bound of 8.56% CER, while maintaining a real-time factor (RTF) of 1.59. The model is publicly available at `https://huggingface.co/iapp/chinda-tts-open-preview`.

1 Introduction

Recent progress in text-to-speech (TTS) technology has greatly improved multilingual speech quality. Thai, while no longer a low-resource language, still faces challenges in effective utilization of available data and limited accessibility to adaptable open models. Commercial systems such as Google Gemini TTS[1] produce highly natural Thai voices but remain closed to customization or research adaptation.

Several open initiatives have emerged, including MMS-TTS-THA [Pratap et al., 2023], F5-TTS-THAI [Chen et al., 2024], and Thai-native projects such as GoWajee [Chuangsuwanich et al., 2022]. Academic progress has also been made through Thai speech systems such as Thonburian Whisper [Aung et al., 2024], which enhance ASR robustness and support downstream Thai speech research. These efforts show strong advancement but remain limited in reproducibility or bilingual capability.

We present ChindaTTS Open Preview, a Thai–English bilingual model based on a transformer architecture and Multi-Scale Neural Audio Codec (SNAC) acoustic tokens [Siuzdak and colleagues, 2024]. The system models speech as discrete tokens, simplifying text-to-speech mapping and reducing reliance on mel-spectrogram and vocoder pipelines. The approach aligns conceptually with open TTS frameworks such as Orpheus TTS [Team, 2025], highlighting the potential of discrete token modeling for efficient speech synthesis.

ChindaTTS Open Preview serves as a bilingual baseline for Thai and English TTS, evaluated with ASR-as-a-Judge and real-time factor (RTF) metrics. The model and methodology are partially released to support reproducible research and strengthen Thailand's open speech technology ecosystem.

[1] `https://cloud.google.com/text-to-speech`

2 Approach

ChindaTTS Open Preview treats speech synthesis as sequential generation over discrete acoustic tokens. Instead of a conventional acoustic-model + vocoder stack, the system integrates text encoding, token generation, and neural-codec decoding in a single generative pipeline.

2.1 Model Architecture

Backbone. The model utilizes a transformer-based large language model architecture conceptually similar to the LLaMA 3 family [Touvron et al., 2023, 2024]. Input text is tokenized and encoded into contextual embeddings that capture linguistic structure and cues relevant to prosody and intelligibility.

Discrete speech token generation. Rather than emitting subword text tokens, the decoder predicts streams of discrete acoustic tokens that represent quantized speech features. These streams correspond to multiple codebook levels and temporal rates, enabling direct modeling of temporal dynamics without an intermediate acoustic model. This formulation follows recent open TTS frameworks that generate codec tokens end-to-end [Team, 2025].

Neural-codec decoding (SNAC). Predicted tokens are converted to waveform using a Multi-Scale Neural Audio Codec (SNAC) [Siuzdak and colleagues, 2024]. SNAC extends residual vector quantization with quantizers operating at different temporal resolutions, providing high-fidelity reconstruction at 24 kHz while maintaining a low bitrate. We use the public `hubertsiuzdak/snac_24khz` checkpoint for encoding/decoding of discrete acoustic tokens.

2.2 Advantages over conventional TTS

Traditional text-to-speech pipelines (e.g., Tacotron2 + HiFi-GAN) rely on mel-spectrogram generation and a separate vocoder stage. In contrast, discrete token modeling with a neural codec simplifies the architecture and enhances robustness:

- **End-to-end mapping:** The model learns direct conversion from text to audio in the discrete domain, eliminating the need for hand-crafted intermediate features.
- **Simplified inference:** Removing a dedicated vocoder stage reduces system complexity and potential error propagation.
- **Temporal stability:** Hierarchical token rates in SNAC maintain prosodic continuity and minimize artifacts compared to mel-spectrogram-based approaches.

2.3 Training

Table 1: Summary of datasets used for pre-training and fine-tuning.

Training Stage	Language	Dataset	Hours
Pre-training	Thai	Common Voice 17.0 (Train) [Mozilla Foundation, 2024] Porjai (Central Thai) [CMKL University, 2024] WanJuan-Thai [OpenDataLab, 2024]	37 700 200
	English	Elise [MrDragonFox, 2024a] Emilia Yodas [MrDragonFox, 2024b] Expresso (tagged by ylacombe) [ylacombe, 2024] LibriSpeech ASR (cleaned train) [OpenSLR, 2015]	3 616 40 100
		Total Pre-training Hours	**1,696**
Fine-tuning	Thai English	High-fidelity Thai set High-fidelity English set	30 35.61
		Total Fine-tuning Hours	**65.61**
Total			**1,761.61**

The model was trained in two stages. The first stage involved large-scale pre-training on bilingual Thai–English speech corpora to enhance generalization across speakers, accents, and recording environments. The second stage performed fine-tuning on a smaller, high-fidelity dataset to refine prosody and intelligibility.

Pre-training covered 1,696 hours of bilingual data, while fine-tuning added 65.61 hours of curated high-fidelity recordings (30 hours Thai and 35.61 hours English). This two-stage process provided broad linguistic coverage and tonal robustness, followed by refinement for prosody and naturalness.

3 Evaluation

We evaluate ChindaTTS Open Preview on two dimensions: intelligibility using the *ASR-as-a-Judge* protocol, and synthesis efficiency measured by real-time factor (RTF).

3.1 ASR-as-a-Judge

To assess intelligibility objectively, we use an automatic speech recognition (ASR) model to transcribe both human and synthesized audio, then compare transcriptions with reference text via character error rate (CER). This provides a reproducible and scalable proxy for pronunciation accuracy and clarity. Two pipelines are evaluated under identical ASR conditions using the Whisper Large V3 model [Radford et al., 2022][2].

- **Upper bound (human speech)** — Original recordings from the Common Voice 17 Thai test set are transcribed to establish the CER representing natural human intelligibility.
- **TTS-generated (synthesized speech)** — The same reference texts are synthesized by ChindaTTS Open Preview, transcribed by the same ASR, and compared with the ground truth to compute TTS CER.

We randomly sample 1,000 utterances from the Common Voice 17.0 Thai test split, which balances statistical reliability with computational cost. Table 2 summarizes the decoding and evaluation configuration used throughout all experiments.

Table 2: Decoding and evaluation parameters.

Parameter	Setting
Sampling rate	24 kHz
Max output length	15 s per utterance
ASR model	Whisper Large V3
Decoding method	Beam search (width = 5)
Language	Thai (explicitly forced)
Evaluation samples	1,000
Hardware	NVIDIA H100 (80 GB)

The key metric is the gap between human and synthesized CERs—the smaller the gap, the closer the generated speech is to human-level intelligibility. This protocol is reproducible (public data and ASR), automated (no human raters), and interpretable (CER directly measures intelligibility and pronunciation accuracy).

3.2 Real-time factor (RTF)

Efficiency is measured by the real-time factor,

$$\mathrm{RTF} = \frac{t_{\text{process}}}{t_{\text{audio}}}, \tag{1}$$

where t_{process} is wall-clock synthesis time and t_{audio} is the duration of the generated waveform. An $\mathrm{RTF} < 1$ indicates faster-than-real-time generation suitable for streaming, while $\mathrm{RTF} > 1$ denotes slower, but acceptable, performance for offline or batch applications.

4 Results

We report the intelligibility and synthesis efficiency results of ChindaTTS Open Preview, evaluated using the ASR-as-a-Judge framework and real-time factor (RTF) metrics.

[2] https://huggingface.co/openai/whisper-large-v3

4.1 Intelligibility

Table 3 presents the Character Error Rate (CER) comparison between human speech (upper bound) and ChindaTTS-generated speech on the Common Voice 17.0 Thai test set.

Table 3: Character Error Rate (CER) comparison on the Common Voice 17.0 Thai test set (1,000 randomly selected samples).

Condition	**CER (%)**	**ΔCER**
Upper Bound (human speech)	8.56	—
ChindaTTS Open Preview	10.56	+2.00

ChindaTTS Open Preview achieves a CER of 10.56% on Thai speech, only 2 percentage points higher than the human upper bound of 8.56%. This small margin indicates that the synthesized output approaches human-level intelligibility, with Whisper Large V3 accurately transcribing the vast majority of generated utterances.

4.2 Synthesis Efficiency

The real-time factor (RTF) was measured on an NVIDIA H100 GPU using the decoding configuration in Table 2. ChindaTTS Open Preview attains an RTF of 1.59, meaning speech generation takes approximately 1.6 times the duration of the produced audio. Although not faster than real-time ($\text{RTF} < 1$), this speed is suitable for offline or semi-interactive applications such as content narration, accessibility tools, and language learning systems. Future optimization techniques—such as model distillation, quantization, and architectural refinement—can further reduce latency.

4.3 Qualitative Observations

In addition to quantitative evaluation, qualitative inspection highlights the following:

- **Prosody and Naturalness**: SNAC-based tokenization yields stable rhythm and intonation with fewer artifacts than mel-spectrogram-based models. The 24 kHz reconstruction provides clear, natural audio.
- **Cross-lingual Consistency**: The model handles Thai–English code-switching smoothly, maintaining consistent timbre across language boundaries.
- **Pronunciation Stability**: Tonal accuracy remains strong across most utterances, though occasional errors occur with rare or compound words.
- **Speaker Consistency**: The preview release employs a single-speaker voice for reliability; multi-speaker or expressive variants are planned for future versions.

5 Discussion

ChindaTTS Open Preview demonstrates competitive intelligibility and acceptable synthesis efficiency for Thai speech generation. The model attains 10.56% CER, only 2.00 percentage points above the human upper bound (8.56%), showing that discrete-token modeling with SNAC effectively captures Thai tonal structure while maintaining clarity. The measured RTF of 1.59 indicates practical usability for offline and semi-interactive applications.

5.1 ASR-as-a-Judge Validation

The ASR-as-a-Judge protocol yields consistent and reproducible results. The small CER gap confirms that Whisper Large V3 can reliably evaluate both human and synthesized Thai speech, validating CER as an objective proxy for intelligibility. However, CER alone does not reflect prosody, expressiveness, or naturalness. Future evaluations should incorporate perceptual metrics such as Mean Opinion Score (MOS) or listener preference tests for a fuller quality assessment.

5.2 Synthesis Efficiency

The current RTF reflects a trade-off between model quality and computational cost. Generation time is influenced by long token sequences, the transformer backbone size, and SNAC decoding latency. Future work will explore sequence compression, faster attention mechanisms, and model distillation to improve speed without sacrificing intelligibility.

5.3 Role in the Thai Speech Ecosystem

ChindaTTS Open Preview expands the Thai speech technology ecosystem by offering an open-weight, bilingual foundation for research and development. It supports fine-tuning, style adaptation, and multilingual exploration, providing a reproducible platform that complements commercial and academic efforts in Thai TTS.

6 Limitations

While ChindaTTS Open Preview demonstrates strong intelligibility and stability, several limitations remain:

Expressive control: The current model lacks explicit mechanisms for emotion, speaking rate, or prosodic variation. Future extensions may incorporate style tokens or conditioning vectors for expressive and controllable speech synthesis.

Speaker diversity: This release is trained on a single Thai speaker, limiting personalization and voice variety. Multi-speaker modeling or speaker embedding adaptation will be required for broader coverage.

Evaluation scope: Assessment currently focuses on ASR-based intelligibility metrics. Comprehensive evaluation of prosody, naturalness, and speaker similarity will require human listening tests such as Mean Opinion Score (MOS) and preference studies.

Synthesis speed: An RTF of 1.59 limits real-time applications. Model compression, faster decoding strategies, and architectural optimization could reduce latency for interactive or streaming use cases.

Language coverage: Performance is optimized for Thai–English synthesis. Code-switching beyond these two languages or adaptation to other regional dialects may degrade pronunciation stability.

Rare word handling: Long compound words, technical terms, and proper nouns occasionally yield pronunciation errors, especially when underrepresented in the training corpus.

7 Conclusion

We presented ChindaTTS Open Preview, a Thai–English text-to-speech model built on transformer architecture and Multi-Scale Neural Audio Codec (SNAC) acoustic tokens. Using the ASR-as-a-Judge evaluation protocol, the model achieved 10.56% CER on Thai speech—within 2 percentage points of the human upper bound (8.56%)—and an RTF of 1.59, demonstrating strong intelligibility and practical synthesis efficiency.

The combination of discrete acoustic token modeling and modern LLM-based decoding provides an effective framework for Thai speech synthesis and a foundation for future multilingual research. While not all resources are released, model weights and evaluation recipes are publicly available to support reproducibility and further development within the Thai AI community.

Future work will focus on expanding to multi-speaker training, improving tonal accuracy, enabling emotional and prosodic control, and enhancing overall speech quality. ChindaTTS Open Preview aims to promote more effective and efficient use of Thai speech resources and strengthen accessibility to high-quality bilingual TTS technologies through open-weight research.

Acknowledgments

This work was supported by Thammasat University's Research Fund, the Center of Excellence in Intelligent Informatics, Speech and Language Technology and Service Innovation (CILS), the Intelligent Informatics and Service Innovation (IISI) Research Center, and the Thailand Research Fund (Grant No. RTA6080013). We thank all members of iApp Technology Co., Ltd. for their contributions, the open-source community for providing datasets and tools such as Common Voice, LibriSpeech, SNAC, Whisper, and OrpheusTTS, and SiamAI Cloud[3] for computational resources.

References

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. Scaling speech technology to 1,000+ languages. *arXiv preprint arXiv:2301.02111*, 2023. URL https://arxiv.org/abs/2301.02111.

[3] https://siam.ai/

Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, Jian Zhao, Kai Yu, and Xie Chen. F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching. *arXiv preprint arXiv:2410.06885*, 2024. URL https://arxiv.org/abs/2410.06885.

Ekapol Chuangsuwanich et al. Gowajee: Thai speech ai platform with text-to-speech and asr capabilities. https://www.gowajee.ai/, 2022. Accessed: 2025-11-01.

Zaw Htet Aung, Thanachot Thavornmongkol, Atirut Boribalburephan, Vittavas Tangsriworakan, Knot Pipatsrisawat, and Titipat Achakulvisut. Thonburian whisper: Robust fine-tuned and distilled whisper for Thai. In Mourad Abbas and Abed Alhakim Freihat, editors, *Proceedings of the 7th International Conference on Natural Language and Speech Processing (ICNLSP 2024)*, pages 149–156, Trento, October 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.icnlsp-1.17/.

Hubert Siuzdak and colleagues. Snac: Multi-scale neural audio codec for high-fidelity speech generation. *arXiv preprint arXiv:2403.01799*, 2024. URL https://arxiv.org/abs/2403.01799.

Orpheus TTS Team. Orpheus tts: An open neural codec-based text-to-speech framework. https://github.com/canopyai/Orpheus-TTS, 2025. Accessed: 2025-11-01.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. URL https://arxiv.org/abs/2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yacine Jernite, Thomas Wang, Sai Prasanna, Ross Taylor, Nicolas Usunier, Thomas Wolf, Guillaume Lample, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

Mozilla Foundation. Common voice 17.0 – thai train split (37 hours). https://huggingface.co/datasets/mozilla-foundation/common_voice_17_0/tree/main, 2024. Accessed: 2025-10-30.

CMKL University. Porjai thai voice dataset (central thai) – 700 hours. https://huggingface.co/datasets/CMKL/Porjai-Thai-voice-dataset-central, 2024. Accessed: 2025-10-30.

OpenDataLab. Wanjuan-thai corpus – 200 hours. https://opendatalab.com/OpenDataLab/WanJuan-Thai/tree/main, 2024. Accessed: 2025-10-30.

MrDragonFox. Elise english speech dataset – 3 hours. https://huggingface.co/datasets/MrDragonFox/Elise, 2024a. Accessed: 2025-10-30.

MrDragonFox. Emilia yodas english speech dataset – 616 hours. https://huggingface.co/datasets/MrDragonFox/EN_Emilia_Yodas_616h, 2024b. Accessed: 2025-10-30.

ylacombe. Expresso-tagged english speech dataset – 40 hours. https://huggingface.co/datasets/ylacombe/expresso-tagged, 2024. Accessed: 2025-10-30.

OpenSLR. Librispeech asr corpus – cleaned train split (100 hours). https://huggingface.co/datasets/openslr/librispeech_asr/viewer/clean, 2015. Accessed: 2025-10-30.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*, 2022. URL https://arxiv.org/abs/2212.04356. Model reference: Whisper Large V3, https://huggingface.co/openai/whisper-large-v3.