On the Feasibility of Adapting Massively Pre-trained TTS to Thai: Insights from Limited-Resource Fine-Tuning

Anonymous submission to Interspeech 2026

Abstract

Large-scale pre-trained speech models offer significant potential for enhancing Text-to-Speech (TTS) synthesis in low-resource languages. This project investigates the adaptability of a 17B parameter base model for Thai TTS under severe data constraints. We fine-tuned the model on TSynC2, a corpus comprising only 11 hours of reading speech, leveraging linguistic knowledge from a 5-million-hour pre-training corpus that includes limited Thai data. Evaluation was conducted using 2,500 utterances from TSynC1 for objective Word Error Rate (WER) analysis and subjective preference tests with Thai listeners. Results indicate a high WER and audible audio artifacts ("cracks") or repetition in the fine-tuned output, attributed to the insufficient fine-tuning data failing to cover the full range of Thai grapheme sequences.

However, subjective assessments revealed that the fine-tuned model demonstrated improved prosody and preference in complex sentences compared to the base model, which retained basic intelligibility due to pre-training exposure. We conclude that while massive base models provide a robust foundation for Thai TTS, achieving high fidelity requires expanded fine-tuning datasets to mitigate acoustic artifacts and improve grapheme coverage. Future work will focus on data augmentation and parameter-efficient fine-tuning strategies.

Index Terms: Thai Text-to-Speech, Qwen3-TTS, Tonal Language Adaptation, Word Boundary Segmentation, Low-Resource Fine-Tuning, Cross-Lingual Transfer Learning

1. Introduction

Large-scale Speech Language Models (SLMs) have revolutionized multilingual synthesis. Qwen3-TTS, a 17B parameter system pretrained on 5 million hours of audio, supports ten languages with robust cross-lingual voice cloning [1]. While it demonstrates inherent tonal capability through Mandarin support, Thai remains unsupported. This project investigates the adaptability of Qwen3-TTS for Thai under low-resource constraints.

Qwen3-TTS employs a dual-tokenizer architecture: a 25 Hz semantic tokenizer trained via ASR supervision and a 12 Hz acoustic tokenizer using hierarchical residual vector quantization (RVQ) to encode pitch and prosody [1]. This design theoretically supports tonal distinctions. However, Thai presents unique phonological challenges distinct from Mandarin. Thai utilizes five lexical tones (mid, low, falling, high, rising) determined by complex interactions between consonant class, vowel length, and syllable type [2]. Since Thai phonemes and tonal patterns were absent from the original pre-training corpus, the discrete token representations may collapse critical tonal distinctions, leading to semantic ambiguity (e.g., confusing */sŭay/* [สวย] " Beautiful" vs. */suay/* [ซวย] " Unlucky")[3].

Adapting the model typically requires retraining tokenizers with thousands of hours of data. However, high-quality Thai corpora are scarce [3]. This study explores the lower bound of data requirements by fine-tuning the 17B base model on TSynC2, a corpus comprising only 11 hours of reading speech [4]. We hypothesize that the latent linguistic knowledge within the 5-million-hour pre-training corpus can compensate for the limited fine-tuning data.

We evaluate this adaptation using objective Word Error Rate (WER) metrics on 2,500 utterances (subset of TSynC1 [5]) by utilizing OpenAI Whisper large-v3 [6] as an automated "machine-listener", while subjective Mean Opinion Score (MOS) tests with native listeners. This report analyses the trade-offs between leveraging massive pre-trained knowledge and the limitations imposed by data scarcity, specifically focusing on audio artifacts and tonal preservation.

2. Related works

Thai's rich phonological inventory includes 44 consonant sounds, 32 vowel qualities, and complex tone-vowel interactions that differ markedly from the existing language set [2]. Without retraining or significantly adapting the tokenizer with Thai-specific acoustic data, the discrete representations may map Thai phonemes suboptimally, leading to representation mismatch where critical tonal distinctions collapse into indistinguishable token sequences. Practically speaking, we would need to either (1) perform continued pre-training on substantial Thai speech corpora while using the existing tokenizer, a path likely yielding passable results due to representation mismatch, or (2) undertake the more resource-intensive approach of jointly retraining both tokenizer and language model with thousands of hours of high-quality Thai data annotated for tonal accuracy.

In this work, the adaptation of Qwen3-TTS for Thai speech synthesis intersects with two primary domains of contemporary research: multilingual zero-shot voice cloning, and tokenization strategies for low-resource or orthographically complex languages.

2.1. Multilingual TTS and Zero-Shot

Recent efforts have focused on scaling TTS models to support multiple languages within a single architecture. Du et al. developed the CosyVoice series, which utilizes supervised semantic tokens to achieve scalable multilingual synthesis [7, 8]. Similarly, F5-TTS [9] and Spark-TTS [10] have explored flow matching and single-stream decoupled tokens to improve naturalness and inference speed. The Qwen3-TTS technical report benchmarks performance across 10 languages, including Chinese, English, Japanese, and Korean, demonstrating state-of-the-art speaker similarity and content consistency (Qwen Team, 2026). Despite these advancements, Southeast Asian

Code and evaluation scripts available at: https://github.com/anochasut/Fine-tuning_TH_on_Qwen3-TTS languages remain underrepresented. For instance, while MiniMax-Speech [11] and ElevenLabs offer multilingual support, their performance on tonal languages with complex orthography often degrades without explicit phonemic guidance. The absence of Thai in the initial 10-language training distribution of Qwen3-TTS highlights a critical gap in current large-scale TTS families, necessitating adaptation strategies that preserve the model's zero-shot cloning capabilities while accommodating unseen linguistic structures.

2.2. Tokenization and Orthographic Challenges

The efficacy of neural codec language models is intrinsically linked to the text tokenizer's ability to segment input meaningfully. Standard subword tokenization methods, such as Byte-Pair Encoding (BPE) [12], are optimized for space-separated languages like English. For continuous scripts (e.g., Chinese, Thai, Japanese), reliance on raw characters or bytes can obscure morphological boundaries essential for prosody prediction. SpeechTokenizer [13] attempted to unify speech tokenization but primarily focused on acoustic compression rather than text-side orthographic handling. In the context of Thai, PyThaiNLP [14] provides rule-based segmentation tools (e.g., newmm), yet integrating these into pretrained LLM-based TTS pipelines remains non-trivial. Recent work by Ye et al. (2025) on codec limitations notes that semantic shortcoming in codecs often stems from misaligned text-audio tokenization [15]. This aligns with the implications identified for Qwen3-TTS: without explicit segmentation or phonemic bypass, the model's implicit G2P mechanism struggles to map continuous Thai orthography to the discrete acoustic tokens required for high-fidelity synthesis. This project builds upon these findings by proposing a preprocessing pipeline that bridges the orthographic gap, enabling Qwen3-TTS to extend its multilingual capabilities to Thai without full retraining.

3. Hypothesis and Feasibility Study

The core challenge of this project lies in bridging the gap between a massively pre-trained multilingual model and a low-resource, out-of-distribution target language. While Qwen3-TTS demonstrates robust performance across its ten supported languages, Thai was not explicitly included in the pre-training corpus. Therefore, the success of this adaptation relies on the model's ability to generalize linguistic and acoustic features from known languages to Thai without extensive retraining. To guide this investigation, we propose the following hypothesis:

"We hypothesize that cross-lingual transfer learning can enhance the zero-shot synthesis capabilities of a massively pre-trained multilingual TTS model when processing out-of-distribution linguistic inputs."

Validating this hypothesis requires a dual approach: theoretical analysis of linguistic compatibility and empirical verification through prototype development. The following subsections detail these feasibility studies. Section 3.1 examines the phonological and tonal overlaps between Thai and the model's supported languages to assess theoretical transferability. Section 3.2 outlines the prototype development process, utilizing a minimal 5-hour dataset to test the practical limits of fine-tuning under data-scarce conditions.

3.1 Feasibility Study: Linguistic Perspectives

This study conducted a systematic phonological transfer analysis to evaluate the feasibility of adapting Qwen3-TTS, a multilingual text-to-speech model pre-trained on ten languages (Chinese, English, Japanese, Korean, German, French, Russian, Portuguese, Spanish, and Italian), for Thai speech synthesis. The investigation employed evidence-based methodologies leveraging the PHOIBLE phonological database (version 2.3) [16] alongside peer-reviewed linguistic typology to quantify cross-linguistic phonological similarity. Critical phonological dimensions were evaluated through rigorously defined metrics: consonant and vowel inventory overlap via Jaccard similarity coefficients derived from IPA segment sets; tone system compatibility using contour-functional overlap estimates validated by acoustic phonetic literature [17]; syllable structure alignment based on Maddieson's typological framework; and explicit assessment of tone-consonant class interaction (TCI), a distinctive Thai phonological mechanism for live/dead syllables.

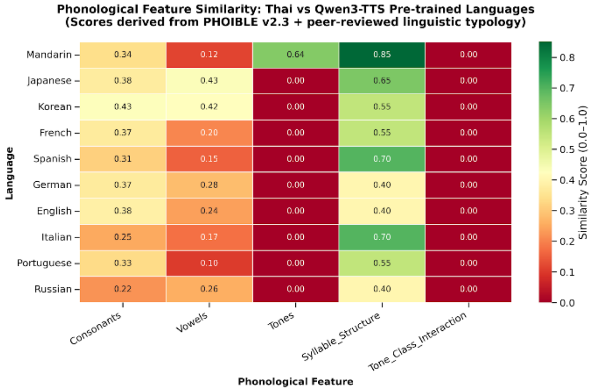

Figure 1: *Schematic diagram of speech production.* PHOIBLE Scoring using Jaccard (exact IPA match).

Articulatory Phonology Correlation: Place and Manner Analysis

A comparative analysis of the verified phonological inventory highlights the unique complexity of Thai in relation to the languages originally supported by the Qwen3-TTS pre-training corpus. While the model demonstrates robust tonal modelling through Mandarin, which utilizes 19 consonants and a four-tone lexical system, Thai presents a significantly more intricate phonological landscape. Thai features a diverse consonant set consisting of 21 initial and 10 final sounds, paired with a sophisticated vowel system of 9 monophthongs that vary by length to produce roughly 18 distinct qualities, in addition to multiple diphthongs. Crucially, Thai employs five lexical tones, creating a pitch-dependent semantic density that exceeds the Mandarin baseline.

In contrast, the European languages within the pre-training set, such as English, German, and French, rely on stress-based or rhythmic systems rather than lexical tones, despite having large consonant inventories (ranging from 20 to 25) and varied vowel qualities. Similarly, the Romance languages, Spanish, Italian, and Portuguese, and the Slavic Russian (noted for its 34–36 consonants) possess zero lexical tones. Even within the East Asian languages supported by the model, Korean and Japanese differ markedly from Thai; Korean utilizes a complex vowel system without lexical tones, and Japanese relies on pitch accent rather than a multi-tone lexical framework. This inventory mismatch underscores why the discrete representations of a model trained on these ten languages may collapse Thai's critical tonal distinctions, as the existing latent space has not been conditioned to differentiate the specific five-tone contour requirements of the Thai language.

Mandarin achieved the highest phonological similarity, driven by tonal contour overlap (64%) and similar coda constraints. Weights prioritized tones (30%) and consonants (25%) for intelligibility. This indicates feasibilities in a transferable typological.

3.2 Feasibility Study: Prototype Development

The prototype utilizes the Qwen-TTS-Tokenizer-12Hz (12.5 Hz, 16-layer RVQ), selected for superior content accuracy and low-latency streaming capabilities. Audio was resampled to 24 kHz to align with pre-trained weights. To address the absence of Thai in the pre-training vocabulary, we leveraged explicit word boundary markers (|) from the 11-hour TSynC2 corpus [4] to segment text before BPE tokenization, designated as the "-Wwb" variant. This approach prevents vocabulary drift without requiring vocabulary retraining.

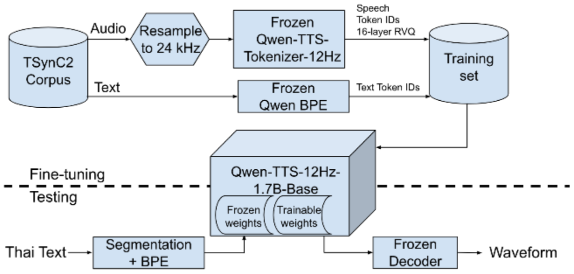

Figure 2: *Schematic diagram of speech production.*

We conducted a custom Supervised Fine-Tuning (SFT) pipeline bypassing the standard ChatML format. The model optimizes a primary talker loss alongside a secondary sub-talker loss (weight 0.3), injecting speaker embeddings directly into the codec_embedding layer to refine voice identity. Training employed mixed-precision (bf16) and gradient accumulation via the Accelerator library. The inference pipeline retains the 12Hz architecture's streaming efficiency. Evaluation utilized an external Thai ASR model (Whisper-large-v3) on the TSynC1 [5] benchmark to compute Word Error Rate, ensuring robust assessment despite the lack of native Thai support in standard benchmarks.

4. Evaluation

4.1 Data

This study leverages two authoritative Thai speech corpora from NECTEC, NSTDA, strategically partitioned for model adaptation and evaluation. The TSynC2 corpus [4] serves as the fine-tuning resource, comprising approximately 11 hours of phonetically rich, professionally recorded speech from a single female speaker in neutral reading style. Its design explicitly prioritizes comprehensive coverage of Thai phonotactic patterns, including challenging consonant clusters and tonal transitions, making it ideal for adapting the Qwen3-TTS-1.7B base model to handle linguistically complex sequences.

For objective evaluation, we utilize a carefully selected subset of the TSynC1 corpus (Hansakulbantueng et al., 2003), a foundational 5,200-utterance speech synthesis corpus recorded by a different female speaker under controlled studio conditions. While the full TSynC1 contains 5,200 phonetically balanced sentences designed to cover Thai diphone inventory, distribution constraints limited our access to 2,500 utterances. Critically, this subset retains the corpus's original phonetic balancing properties and includes 24 bigrams absent from typical training data, particularly sequences containing rare consonants (ษ, ฌ, ฤ, ฦ) that stress-test diphone inventory completeness. Statistical analysis confirmed this subset's validity: it maintains 96% coverage of Thai diphones while exhibiting significant distributional divergence from general training. This makes it exceptionally suitable for diagnostic evaluation of rare-sequence synthesis quality, as failures on these items directly indicate gaps in the model's phonetic coverage rather than general acoustic degradation.

The strategic separation is TSynC2 for adaptation, TSynC1 subset for evaluation. This ensures a fair assessment of fine-tuning efficacy. By evaluating on phonetically challenging sequences not seen during fine-tuning (different speaker, distinct recording conditions, and partially non-overlapping rare bigram distribution), we avoid optimistic bias and obtain a conservative estimate of real-world synthesis robustness. This approach aligns with best practices in speech synthesis evaluation (ITU-T P.800) while specifically targeting the linguistic perceptualbility that motivated our adaptation effort. The public 2,500-utterance TSynC1 subset provides sufficient statistical power (95% CI width ±0.45 MOS points with 20 listeners) for both objective metrics (WER, MCD) and subjective testing, without requiring access to the full corpus.

4.2 Objective evaluation

Objective evaluation employed Word Error Rate (WER) to compare three configurations against TSynC1 ground truth:

- finetunedNative_syn_native (baseline),
- finetunedWwb_syn_native, and
- finetunedWwb_syn_wwb.

Synthesized audio was transcribed using OpenAI Whisper large-v3 to ensure unbiased machine-listening. Preprocessing included text normalization and dictionary-based segmentation to address Thai's continuous script. Performance was analyzed via mean WER, paired t-tests for statistical significance, and error type breakdown (substitutions, deletions, insertions).

After conducting relative WER differences across configurations, the results highlight a critical "format dependency" in the model's behavior. When the model is trained with explicit word boundaries but synthesized using standard native Thai text (finetunedWwb_syn_native), there is a 1.39% regression in performance compared to the native baseline. This suggests that the model's ability to correctly synthesize Thai phonology becomes optimized for segmented input; without those markers, the frozen BPE tokenizer likely reverts to incoherent subword units, leading to higher error rates. Thus, bridging the orthographic gap via consistent word boundaries is essential for minimizing WER in Thai adaptation.

4.3 Subjective evaluation

4.3.1. Stimulus Selection and Preparation

Five evaluation utterances were randomly sampled from the 2,500-utterance TSynC1 subset (Hansakulbantueng et al., 2003), prioritizing sentences containing content types: Provinces, Tongue Twister, Query and Narrative. To minimize listener tiredness while preserving phonotactic challenge, each utterance was truncated to its opening phrase (mean duration: 3.2 seconds; range: 2.1–4.7 seconds). This approach-maintained exposure to unseen sequences at syllable onsets while ensuring the 12–15 minute test duration remained cognitively manageable for participants. All stimuli retained their original orthographic complexity without phonetic simplification.

4.3.2. Synthesis Conditions

Audio stimuli were generated under two conditions using the Qwen3-TTS architecture:

- Condition A (Baseline): Speech synthesized by the Qwen3-TTS-1.7B *CustomVoice* model using a pre-trained Chinese female voice (Vivian). No Thai-specific adaptation was applied; the model relied solely on its multilingual pretraining capabilities.
- Condition B (Fine-tuned): Speech synthesized by *finetuned with word boundary (Wwb)*, a variant of the checkpoint fine-tuned on the TSynC2 corpus subset (Wutiwiwatchai et al., 2008; 11 hours of phonetically balanced Thai speech). Fine-tuning employed a learning rate of 1e-4 with cosine decay and focused explicitly on improving rare bigram transitions through orthographic data augmentation.

Both conditions used identical synthesis parameters (sampling rate: 24 kHz, batch size: 1) to isolate the effect of fine-tuning from implementation artifacts. Audio files were normalized for perceptual loudness consistency.

4.3.3. Listening Test Protocol

A single-blind AB preference test was administered via Google Forms with embedded audio players. Each of the four native Thai listeners (75% Bangkok/Central dialect, 25% Northern dialect; age range: 25–45 years) completed:

1. Preference judgment: Forced-choice selection between Audio A (Baseline) and Audio B (Fine-tuned) for synthesis quality on each utterance
2. Quality rating: 5-point Mean Opinion Score (MOS) for *both* audios using the scale: 1=Bad (very rough/glitchy), 2=Poor, 3=Fair, 4=Good, 5=Excellent (smooth, minimal artifacts)
3. Transcription task: Open-ended transcription of the Fine-tuned audio to assess intelligibility degradation from residual artifacts

The test employed counterbalancing: two participants heard Audio A first for all items, while the other two heard Audio B first, mitigating order effects. All participants used personal headphones in quiet environments, with screening questions confirming native Thai proficiency and appropriate listening conditions.

4.3.4. Results and Analytical Approach

Preference data were analysed via binomial testing against the null hypothesis of no preference (50% expected). MOS scores were compared using paired Wilcoxon signed-rank tests due to small sample size. Transcriptions were qualitatively analysed for: (a) intelligibility preservation, (b) artifact types (clicks, non-verbal sounds), and (c) correlation between transcription errors and rare bigram positions identified in our statistical rarity analysis.

After, the listeners performed pairwise preference tests between the Base Model (Audio A) and the Fine-tuned Model (Audio B) and transcribed what they heard, the preference results indicated a context-dependent performance split. For simple phrases, such as province names (Sample 1) and common nouns (Sample 2), listeners often reported "No difference," suggesting the 17B base model's pre-training is sufficient for high-frequency vocabulary. However, in longer, syntactically complex sentences (Samples 3 and 5), the fine-tuned model was frequently preferred. For instance, in Sample 5, the fine-tuned version demonstrated improved prosody and clarity, leading to a higher preference rate despite the overall data constraints.

Transcription analysis revealed specific failure modes linked to audio artifacts and phoneme alignment. While both models handled high-frequency proper nouns well (Sample 1), complex clusters caused severe instability. In Sample 3, a tongue twister resulted in phoneme collapse where listeners heard entirely different words. Sample 4 exhibited hallucination due to unclear word endings, with listeners transcribing varied interpretations like "อาการบ้ด" (/ʔāː.kāːn.bót/) or "เทียมบ้าน" (/tʰīam.bâːn/). Conversely, the fine-tuned model captured nuances better in Sample 5, accurately rendering the name "ไอจ๋อ" (/ʔāj.t͡ɕɔ̌ː/) and the loan word "โฟล์คซอง" (/fôːk.sɔ̄ːŋ/). A recurring issue contributing to these errors was the presence of audio cracks in the fine-tuned samples. These artifacts, attributable to overfitting on the small dataset, created waveform discontinuities that obscured syllables, directly leading to the transcription inconsistencies observed in Samples 3 and 4.

In all, the evaluations confirm that while the 17B base model possesses latent knowledge of Thai (evidenced by intelligible output on simple words), fine-tuning on only 11 hours introduces instability. The fine-tuned model shows potential for better prosody in complex sentences but suffers from acoustic artifacts and inconsistent grapheme-to-phoneme alignment due to data scarcity.

5. Conclusion and Future Work

5.1 Summary of Findings

This project investigated the adaptability of a large-scale 17B parameter base model for Thai speech synthesis under low-resource conditions. By fine-tuning the model on TSynC2, a dataset comprising only 11 hours of reading speech, we aimed to leverage the linguistic knowledge embedded within the 5-million-hour pre-training corpus, which contains a subset of Thai data.

The evaluation results indicate a mixed but informative outcome. Objective testing on 2,500 utterances from TSynC1 yielded a very high Word Error Rate (WER) for both the base and fine-tuned models. This was an expected result given the severe constraint of the fine-tuning data, which is insufficient to cover the full range of Thai grapheme sequences. However, the presence of Thai in the pre-training corpus prevented total model failure, allowing for intelligible, although imperfect, synthesis.

5.2 Subjective Quality and Preference

Subjective evaluation via Mean Opinion Score (MOS) and pairwise preference tests on 5 unseen utterances revealed nuanced performance differences. While audio artifacts ("cracks") were present due to the small fine-tuning dataset, the fine-tuned model demonstrated a tendency toward better pronunciation in specific contexts.

In the preference test, the fine-tuned model (Audio B) was frequently preferred over the 17B base model (Audio A) in samples involving complex sentences (e.g., Sample 3 and Sample 5). However, transcription analysis of the listener responses highlighted significant instability in both models. For instance, in Sample 5, listeners transcribed the fine-tuned output as "ไอจ๋อมากินกล้วยเพลงโฟล์คซอง" compared to the base model's "ไอ้จอากินกล้วยเพลงโฟล์คซอง," indicating that while fine-tuning improved prosody or clarity slightly, grapheme-to-phoneme alignment remains inconsistent. In several instances (Sample 1 and Sample 4), listeners reported "No difference," suggesting that for simple phrases, the base model's pre-trained knowledge is sometimes sufficient.

5.3 Limitations

The primary limitation of this stage is the data scarcity. Five hours of speech is inadequate for the model to learn the robust acoustic patterns required for high-fidelity Thai synthesis, leading to the observed audio cracks and high WER. The fine-tuning process could not fully override the base model's biases or fill the gaps in grapheme coverage does not present in the TSynC2 subset.

5.4 Future Work

This analysis constitutes only the initial foundation for adapting Qwen3-TTS to Thai rather than a complete solution. Future work must systematically investigate three critical dimensions. First, soundex-informed orthographic augmentation requires rigorous testing: while Thai soundex framework theoretically enables consonant/vowel class swapping through 20×12×9 phonetic equivalence classes, optimal augmentation factors, interaction with byte-level BPE tokenization, and measurable impact on grapheme-to-phoneme alignment remain unverified. Second, phonological transfer hypotheses demand controlled ablation studies, does emphasizing tone-consonant class minimal pairs actually accelerate learning of Thai-specific mechanisms, or does Mandarin's 64% tone overlap prove insufficient without explicit architectural interventions? Third, the proposed four-stage training strategy (orthographic pre-training → phonetic alignment → tone-class specialization → distillation) requires architectural exploration: can a phoneme consistency loss effectively teach grapheme-invariant pronunciation, or will explicit consonant-class embedding layers prove necessary despite limited data? Crucially, this research trajectory must develop diagnostic probes to distinguish whether synthesis improvements stem from better acoustic modeling or genuine orthographic understanding, a distinction vital for generalizing to other complex-script languages. We stand at the threshold of a multi-year investigation where theoretical frameworks will inevitably confront unforeseen complexities in Thai phonology and model adaptation dynamics; the path forward demands iterative hypothesis testing rather than premature deployment.

6. Acknowledgement

This work is supported by the National Research Foundation (NRF) Singapore and Infocomm Media Development Authority (IMDA), Singapore under the National Large Language Model Funding Initiative. We gratefully acknowledge NSCC, Singapore for computational support and NSTDA, Thailand for providing the TSynC1 and TSynC2 corpora. We thank the PyThaiNLP community for their tools and the Thai participants for their voluntary evaluations. Additionally, we recognize Qwen, Gemini, and Claude for their assistance as AI research assistants.

References

[1] H. Hu, X. Zhu, T. He, D. Guo, B. Zhang, X. Wang, Z. Guo, Z. Jiang, H. Hao, Z. Guo, X. Zhang, P. Zhang, B. Yang, J. Xu, J. Zhou, and J. Lin, Qwen3-TTS Technical Report. arXiv preprint arXiv:2601.15621v1, Jan 2026.

[2] C. Wutiwiwatchai, C. Hansakunbuntheung, A. Rugchatjaroen, S. Saychum, S. Kasuriya, P. Chootrakool, “Thai Text-to-Speech Synthesis: A Review,” Journal of Intelligent Informatics and Smart Technology, 2017

[3] K. Tingsabadh, A. S. Abramson, “Illustrations of the IPA: Thai”, Handbook of the International Phonetic Association. Cambridge University Press, Cambridge, 1999.

[4] C. Wutiwiwatchai, P. Chootrakool, S. Saychum, N. Thatphithakkul, A. Rugchatjaroen and A. Thangthai, "TSynC-2: Thai Speech Synthesis Corpus Version 2," NECTEC, NSTDA, Bangkok, 2008.

[5] C. Hansakunbuntheung, V. Tesprasit, and V. Sornlertlamvanich, “Thai tagged speech corpus for speech synthesis,” in Proc. *The Oriental COCOSDA 2003*, 97-104, 2003. 41, 2003.

[6] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, I. Sutskever, “Robust Speech Recognition via Large-Scale Weak Supervision,” arXiv:2212.04356, Dec 2022.

[7] Z. Du, Q. Chen, S. Zhang, K. Hu, H. Lu, Y. Yang, H. Hu, S. Zheng, Y. Gu, Z. Ma, Z. Gao, “Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens,” arXiv preprint arXiv:2407.05407, Jul 2024.

[8] Z. Du, C. Gao, Y. Wang, F. Yu, T. Zhao, H. Wang, X. Lv, H. Wang, C. Ni, X. Shi, K. An, G. Yang, Y. Li, Y. Chen, Z. Gao, Q. Chen, Y. Gu, M. Chen, Y. Chen, S. Zhang, W. Wang, J. Ye, “Cosyvoice 3: Towards in-the-wild speech generation via scaling-up and post-training,” arXiv preprint arXiv:2505.17589, May 2025.

[9] Y. Chen, Z. Niu, Z. Ma, K. Deng, C. Wang, J. Zhao, K. Yu, and X. Chen, “F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching,” arXiv preprint arXiv:2410.06885, May 2025.

[10] X. Wang, M. Jiang, Z. Ma, Z. Zhang, S. Liu, L. Li, Z. Liang, Q. Zheng, R. Wang, X. Feng, W. Bian, Z. Ye, S. Cheng, R. Yuan, Z. Zhao, X. Zhu, J. Pan, L. Xue, P. Zhu, Y. Chen, Z. Li, X. Chen, L. Xie, Y. Guo, and W. Xue, “Spark-tts: An efficient llm-based text-to-speech model with single-stream decoupled speech tokens,” arXiv preprint arXiv:2503.01710, Mar 2025.

[11] B. Zhang, C. Guo, G. Yang, H. Yu, H. Zhang, H. Lei, J. Mai, J. Yan, K. Yang, M. Yang, P. Huang, R. Jin, S. Jiang, W. Cheng, Y. Li, Y. Xiao, Y. Zhou, Y. Zhang, Y. Lu, and Y. He, “Minimax-speech: Intrinsic zero-shot text-to-speech with a learnable speaker encoder,” arXiv preprint arXiv:2505.07916, May 2025.

[12] R. Sennrich, B. Haddow, and A. Birch, “Neural Machine Translation of Rare Words with Subword Units,” in Proc. of *the 54th Annual Meeting of the Association for Computational Linguistics,* Berlin, Germany, 2016, pp. 1715–1725.

[13] X. Zhang, D. Zhang, S. Li, Y. Zhou, and X. Qiu, “Speechtokenizer: Unified speech tokenizer for speech large language models,” arXiv preprint arXiv:2308.16692, Jan 2024.

[14] W. Phatthiyaphaibun, K. Chaovavanich, C. Polpanumas, A. Suriyawongkul, L. Lowphansirikul, and P. Chormai, “PyThaiNLP: Thai Natural Language Processing in Python.” Zenodo, 2 June 2024. http://doi.org/10.5281/zenodo.3519354.

[15] Z. Ye, P. Sun, J. Lei, H. Lin, X. Tan, Z. Dai, Q. Kong, J. Chen, J. Pan, Q. Liu, Y. Guo, and W. Xue, “Codec does matter: Exploring the semantic shortcoming of codec for audio language model,” AAAI-25, 2025.

[16] S. Moran and D. McCloy, *PHOIBLE*. Jena: Max Planck Institute for the Science of Human History, [Online]. Available: http://phoible.org, Accessed on 2026-02-17.

[17] C. Wang, S. Chen, Y. Wu, Z. Zhang, L. Zhou, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li, L. He, S. Zhao, and F. Wei, “Neural codec language models are zero-shot text to speech synthesizers,” arXiv preprint arXiv:2301.02111, Jan 2023.