End-to-End Thai Text-to-Speech with Linguistic Unit

Kontawat Wisetpaitoon
Kasikorn Labs Co. Ltd.
Nonthaburi, Thailand
o_kontawat.w@kbtg.tech

Sattaya Singkul
Kasikorn Labs Co. Ltd.
Nonthaburi, Thailand
sattaya.s@kbtg.tech

Theerat Sakdejayont
Kasikorn Labs Co. Ltd.
Nonthaburi, Thailand
theerat.s@kbtg.tech

Tawunrat Chalothorn
Kasikorn Labs Co. Ltd.
Nonthaburi, Thailand
tawunrat.c@kbtg.tech

ABSTRACT

In this study, we explore the influence of Thai Linguistic Units (TH-LUs) and speech trimming on the state-of-the-art Thai Text-to-Speech (TTS) systems. We propose an end-to-end Thai TTS framework that emphasizes phonemes, syllables, and words, essential for accurate text pronunciation. To thoroughly investigate these aspects, we designed two main experiments: the TH-LU factor experiment and the TH-LU with speech trimming factor experiment. Our assessment targeted speaker tone and pronunciation accuracy. VITS model demonstrated a standout performer in tonal accuracy, which is evaluated by the Speaker Encoder Cosine Similarity (SECS) method, across different TH-LUs in both trim and non-trim speech training data. For pronunciation accuracy, we integrated a Thai speech-to-text model to evaluate. Our results indicate that VITS with the word linguistic unit outperforms all baselines in overall performance, excelling in both speaker tone and pronunciation accuracy. This research significantly advances the field of TTS, particularly for the Thai language, by highlighting the importance of diverse TH-LU and speech trimming in TTS model development and underlining the need for evaluation methods that account for both tonal accuracy and pronunciation quality.

CCS CONCEPTS

• **Computing methodologies** → **Neural networks**; • **Theory of computation** → *Models of learning*.

KEYWORDS

Thai Text-to-Speech; Linguistic Unit Analysis; Speech Synthesis; Tonal Accuracy; Speech Trimming

ICMR '24, June 10–14, 2024, Phuket, Thailand

ACM ISBN 979-8-4007-0619-6/24/06
https://doi.org/10.1145/3652583.3658029

ACM Reference Format:
Kontawat Wisetpaitoon, Sattaya Singkul, Theerat Sakdejayont, and Tawunrat Chalothorn. 2024. End-to-End Thai Text-to-Speech with Linguistic Unit. In *Proceedings of the 2024 International Conference on Multimedia Retrieval (ICMR '24), June 10–14, 2024, Phuket, Thailand*. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3652583.3658029

1 INTRODUCTION

Text-to-Speech (TTS) [11] synthesis emerges as an indispensable technology with diverse applications, ranging from aiding visually impaired individuals to augmenting human-computer interaction (HCI) across digital platforms. TTS fundamentally transforms written text into spoken language, acting as a conduit between text-based information and audible communication. The quality and naturalness of TTS systems, in general, rely greatly on the underlying linguistic units (LU) employed during the synthesis process.

The foundation of any TTS system lies in the segmentation of text into appropriate linguistic units (LU) [21], such as phonemes [14, 26], syllables [15, 16, 20], or words [20]. These units dictate how text is pronounced in varying segments. The selection of the linguistic unit exerts a profound influence on the overall quality, comprehensibility, and naturalness of the synthesized speech. In this paper, we focus on Thai TTS. The Thai language [5], distinguished by its intricate phonological structure and tonal complexities, presents a distinct challenge within the domain of TTS synthesis. While the Thai language's phonological structure, which refers to its system of sounds and their organization, is notably complex for TTS synthesis due to its tonal nature, where a word's tone alters its meaning, and the intricate combination of consonants and vowels forming syllables. Phonologically, Thai [18, 19, 23] lacks explicit markers for word or sentence boundaries, making the segmentation into such units nontrivial. Additionally, Thai [27] is a tonal language featuring four explicit tone marks but exhibiting five tonal sounds. Tonal coarticulation is prevalent, while slight deviations in tonal pronunciation are particularly noticeable to native listeners. These characteristics collectively contribute to the challenges inherent in Thai TTS.

Addressing these challenges necessitates the selection of appropriate LUs and boundary markers in Thai TTS. In this study, we

investigate the representation of Thai linguistic units (TH-LUs) in state-of-the-art TTS models (SOTA-TTS), which is extracted based on text or transcription. We examine how phonemes, syllables, and words, as TH-LUs, influence speech pronunciation. These units are renowned for their role in TH-LUs [14, 15, 26]. A comparative analysis of these three LUs has been conducted to assess their effectiveness in Thai TTS. Subsequently, the study explores the consequences of opting for one distinct LU over the alternatives.

To evaluate the efficacy of Thai TTS, each of our TH-LUs is integrated into three SOTA-TTS models: Tacontron2-DDC [8, 17], VITS [10], and YourTTS [4]. Our objective is to explore the impact of TH-LU on critical dimensions of synthesized speech, encompassing pronunciation accuracy, prosody, and overall auditory quality, that evaluated by Tsync1 [9] and Tsync2 [26] datasets. Through this comprehensive analysis, we aim to provide critical insights that will contribute to the development of more efficient and naturally expressive Thai TTS systems.

Our main contributions of this paper are as follows: (i) the end-to-end Thai text-to-speech framework with linguistic unit was proposed. (ii) The study of the Thai TH-LU factor, which includes phoneme, syllable, and word information, and TH-LU with speech trimming factor were investigated and make the SOTA-TTS model efficient in Thai. (iii) A comparative study between speaker tone and pronunciation performance in our Thai text-to-speech framework was proposed.

2 RELATED WORK

The fusion of SOTA-TTS models with linguistic unit (LU) analysis [15, 16, 20, 21, 26] has resulted in significant advancements in the field of speech synthesis. This section provides an in-depth review of the synergy between SOTA-TTS and LUs, highlighting the pivotal contributions that have shaped this evolving landscape.

SOTA-TTS models have revolutionized speech synthesis by leveraging deep learning architectures to generate more natural and expressive speech. Tacotron [24], a well-known TTS model, is a sequence-to-sequence model that directly predicts mel spectrograms from input text. Tacotron's integration of attention mechanisms improved the alignment between text and spectrogram, enhancing the quality and coherence of synthesized speech. Further evolution brought forth a refined Tacotron2 [17], incorporating a more robust encoder-decoder architecture. The model's ability to generate high-quality spectrograms represented more human-like speech output.

Tacotron2 proposes to compute multiple non-overlapping output frames by the decoder. However, Tacontron2 requires "reduction rate" to set the number of output frames per decoder step. Larger the reduction rate, fewer the number of decoder steps required for the model to produce the same length output. Thereby, the model achieves faster training convergence and easier attention alignment. However, larger reduction rate values also produce smoother output frames and therefore, reduce the frame-level details. To solve this problem, the double decoder consistency (DDC) [8] is presented for Tacontron2 as Tacontron2-DDC. The DDC of Tacontron2-DDC [8, 17] is based on two decoders working simultaneously with different reduction factors. One decoder works with a large, and the other decoder (fine) works with a small reduction factor. These provide steady attention alignment and provide a choice in a spectrum of quality and speed switching between the fine and the coarse decoders at inference.

On the other hand, Tacontron2-DDC has a high effort for feature and architecture engineering. To reduce this effort like end-to-end learning, VITS [10] introduces a sequence-to-sequence model that embeds an auto-alignment mechanism within it, thus addressing the temporal duration of synthesized speech results. VITS utilizes speech embedding from a pre-trained model, enhancing the model's robustness compared to conventional approaches. Moreover, to support multi-speaker and multi-language capabilities, YourTTS [4] is developed based on the VITS architecture, incorporating language embeddings into the model input. Notably, YourTTS introduces the concept of speaker consistency loss (SCL) [28] to enhance multi-speaker synthesis within each language. These advancements collectively position YourTTS as an outperformer compared to traditional models, particularly evident in zero-shot prediction scenarios.

The interplay between SOTA-TTS and LUs introduces a new dimension to speech synthesis. Traditional TTS systems [11, 21] often segmented input text into phonemes, syllables, or words, affecting the naturalness and intelligibility of the synthesized speech. The integration of linguistic units within SOTA-TTS addresses this challenge, leading to more coherent prosody and improved phonetic accuracy. Researchers have explored the selection of linguistic units for optimal TTS synthesis. Syllable-based segmentation [15, 16] has gained prominence, particularly in tonal languages like Thai, where capturing tonal variations is crucial. Moreover, the incorporation of boundary markers enhances the robustness of synthesized speech, enabling smoother transitions between different linguistic units.

In summary, the integration of SOTA-TTS models with linguistic units marks a paradigm shift in speech synthesis. The evolution from phoneme-based methods to comprehensive unit representations has enhanced the authenticity and naturalness of synthesized

speech. This fusion not only offers a deeper understanding of linguistic nuances but also opens avenues for more effective and expressive TTS systems, enriching human-computer interaction and accessibility for diverse users. Our research highlights crucial areas for enhancement within Thai TTS systems, with the objective of improving naturalness and efficiency. We propose an end-to-end Thai TTS framework that incorporates TH-LUs, designed to streamline the process and elucidate the role of TH-LUs in TTS synthesis. Our findings emphasize the critical importance of selecting appropriate TH-LUs, as this choice markedly impacts the quality of the synthesized speech. Through a detailed comparative analysis of each TH-LU within SOTA-TTS models, using both trimmed and non-trimmed speech training data, we focus on speaker tone and pronunciation accuracy.

3 PROPOSED IDEA

Our research focuses on improving Thai TTS synthesis by incorporating and examining TH-LU representations within the SOTA-TTS model. We propose an end-to-end Thai TTS framework that includes data input processing, feature extraction, and the end-to-end TTS model, as shown in Fig. 1. The TH-LU is integrated in this framework to find the proper LU in Thai. In addition, we delve into the role of speech trimming in enhancing TTS performance, examining how it influences the pronunciation of TH-LUs.

3.1 Data Preprocessing

Our framework employs the Tsync1 and Tsync2 datasets, as detailed in Section 4.3. The data preprocessing is methodically divided into two distinct phases. Initially, text data undergoes a cleaning and normalization process. Special characters and extraneous whitespace are filtered out from the Thai text. Subsequently, text in languages other than Thai is converted using a Thai transliteration model[1]. Numerals and abbreviations are then fully expanded into their Thai word counterparts. The final step in this phase involves employing the "newmm-tokenizer"[2] to accurately pre-compute Thai word boundaries, which is essential for extracting the TH-LU information. The subsequent phase is focused on speech data preprocessing. Here, we ensure that all speech files are consistently resampled to a uniform rate of 16kHz and formatted to 16-bit PCM resolution, establishing a standard quality across the audio dataset.

3.2 Thai Linguistic Unit Extraction

In our framework, TH-LUs play a crucial role in determining pronunciation for TTS synthesis. Thai's linguistic structure is complex, with a phonetic system comprising 44 consonants and 21 vowel symbols that create five distinct tonal variations. This tonality introduces intricate phonological challenges. The Thai language also poses specific obstacles for TTS technologies, particularly in word segmentation. For instance, the phrase "นอนตากลมอยู่ริมทะเล" could be segmented as "นอน | ตาก | ลม | อยู่ | ริม | ทะเล" or "นอน | ตา | กลม | อยู่ | ริม | ทะเล," with each segmentation altering the meaning substantially [18, 19, 29]. Such ambiguities significantly impact the development of precise TTS models. Besides, tonal nuances in Thai can also affect meaning, often introducing sarcasm or entirely different interpretations based on tonal delivery, which complicates TTS synthesis. Words with multiple tonal readings can confuse TTS algorithms, making accurate tonal pronunciation a complex task. Homographs, which are words with identical spellings but different meanings depending on context and pronunciation, add another layer of complexity to TTS conversion processes. Such words require careful consideration, as a single written form may represent different meanings. Given these complexities, we study three primary TH-LU for TTS synthesis: phonemes, syllables, and words. Each unit is essential for capturing the subtleties of Thai pronunciation and intonation. We show examples of TH-LUs utilizing the Tsync1 and Tsync2 datasets, with specific instances detailed in Table 1.

3.2.1 Thai Phoneme-based Linguistic Unit. In the phoneme-based linguistic unit of our framework, we delve into the intricate rules of phonetic expression, such as those codified by the International Phonetic Alphabet (IPA). The IPA is globally recognized and extensively employed, especially in cross-lingual TTS applications, due to its precise representation of speech sounds. We utilize the IPA to examine its complex phonetic and phonological structure comprehensively.

Thai consonants span a broad array, including voiceless stops like /p, t, k/ and voiced stops /b, d, g/. The language also encompasses nasals /m, n, ŋ/, fricatives /f, s, h/, approximants /j, w/, and affricates /ʨ, ʥ/, with each class adding to the rich phonemic diversity. For example, the voiceless alveolar plosive /t/ is articulated as [t] in IPA, as in "ต้น" (ton), meaning "tree," where the /t/ sound occurs in the initial position of the syllable. Besides, for vowels, the Thai vowel system is equally varied with monophthongs /i, e, ɛ, a, ɔ, o, u/ and diphthongs /ia, ua, ai, au, ei, ou/, contributing to the language's vocalic richness. The monophthong /a/, for instance, is noted as [a] and is evident in "ราตรี" (raatrii), meaning "night," the /a/ sound is pronounced in the initial syllable.

Tonal differentiation is fundamental to Thai, with five distinct tones—high, mid, low, falling, and rising—each altering a word's meaning. Tone markers such as mai ek (่), mai tho (้), mai tri (๊), and mai chattawa (๋) indicate these variations, as seen in the difference between "ขาว" (khao), "white," and "ข้าว" (khâao), "rice,"

[1] https://github.com/cakimpei/wunsen.git
[2] https://github.com/wisesight/newmm-tokenizer

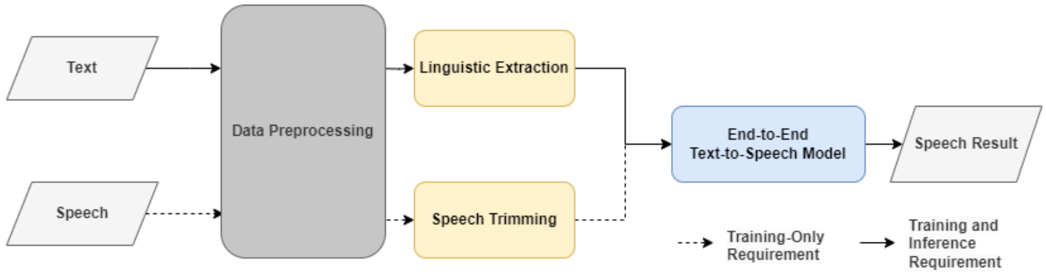

Figure 1: Our end-to-end Thai text-to-speech framework with linguistic units. Note that the speech pathway is applied only during the training phase of the model.

distinguished by mid and high tones, respectively. However, the syllable structure in Thai generally adheres to consonant-vowel-consonant (CVC) or consonant-vowel (CV) patterns, avoiding complex clusters and promoting a smooth flow. This structure, in synergy with tone, influences the language's prosody, exemplified by "หมา" (maa), "dog," following a CVC pattern. This tone of the syllable is influenced by the combination of the consonant and vowel. Lastly, in Thai tonal coarticulation and sandhi, where adjacent tones affect each other, introduce dynamic changes to the tonal landscape. This is seen in the phrase "คุณเป็นอะไร" (khun pen à-rai), "What are you?" where the tones of "คุณ" (khun) and "เป็น" (pen) interact.

The detailed Thai phoneme-based IPA thus lays a solid foundation for our SOTA-TTS model, offering a nuanced understanding of the language's consonants, vowels, tones, and syllable structures, essential for developing refined, natural-sounding Thai speech synthesis. In our work, we utilize CharisuG2P [30] to extract phonemes, which is based on the IPA concept.

3.2.2 *Thai Syllable-based Linguistic Unit.*

Syllables in Thai are foundational to the language's tonal and phonetic identity, playing a pivotal role in the articulation and meaning of words. They are not merely clusters of sounds but are laden with semantic significance, contributing to the language's rhythm and meter. Thai syllables are intricate constructs that are vital to pronunciation and linguistic expression. Each syllable is a composite of consonants, vowels, and tones, orchestrating the formation of words and sentences. Unlike English, which is non-tonal, or Mandarin, which has a limited tonal spectrum, Thai syllables embody a broader array of tonal possibilities, each bearing substantial semantic value as discussed in section 3.2.1. This comparative perspective illuminates the unique phonetic and tonal landscape of Thai. The syllables are not just phonetic building blocks but are imbued with rich tonal, phonetic, and rhythmic qualities that distinguish Thai's linguistic essence. It is this intricate fusion of tonal complexity and syllabic structure that forms the bedrock of Thai's phonetic identity and enriches its linguistic tapestry. We utilize PMSeg[5] to extract syllables, which is based on this concept.

3.2.3 *Thai Word-based Linguistic Unit.*

Building on our exploration of phonemes and syllables, Thai words represent a more upper level of linguistic structure. Composed of syllables, they form the smallest meaningful units of sound in the language. Thai words often embody complex semantics, with meanings deeply rooted in the cultural and contextual intricacies of the language. Consider the Thai word "แม่น้ำ" (mae nam), translating to "river" in English. This compound word is formed by combining two syllables: "แม่" (mae), meaning "mother", and "น้ำ" (nam), meaning "water". Individually, these syllables convey distinct meanings, but together, they create a contextually enriched term within the Thai lexicon. Furthermore, as emphasized in section 3.2.1, Thai's tonal aspect adds another dimension to word formation. The tonal variation in each syllable can significantly influence the overall word meaning, underscoring the language's tonal sensitivity, which affects both pronunciation and interpretation.

A comprehensive analysis of word-based linguistic units in Thai is crucial for fully understanding the language. The construction of Thai words, shaped by tonal variations, morphological processes, and semantic richness, presents distinct challenges and opportunities in linguistic and language technology research. Grasping these complexities is fundamental for effectively engaging with the nuanced Thai linguistic landscape.

Table 1: The example of Thai linguistic unit results when applying our concept to the Tsync1 and Tsync2 datasets.

<table>
<tr><th>Linguistic Unit</th><th>Tsync1 Result</th><th>Tsync2 Result</th></tr>
<tr><td>Original</td><td>สามารถยืดอายุแบตเตอรี่ได้ (can extend battery life.)</td><td>เพราะตะกั่วถือเป็นโลหะหนัก (because lead is considered a heavy metal.)</td></tr>
<tr><td>Phoneme</td><td>saː˩˩˥.maː˧.rot˥˥ | jɯːt˥˩.ʔaː˧.juʔ˥˥ | bɛːt˨˩.tɤː˧.riː˥˩ | daj˥˩</td><td>pʰrɔʔ˥˥ | ta˨˩.kua̯˨˩ | tʰɯː˩˩˥ | pen˧ | loː˧.haʔ˨˩ | nak˨˩</td></tr>
<tr><td>Syllable</td><td>สา | มารถ | ยืด | อา | ยุ | แบต | เตอ | รี่ | ได้</td><td>เพราะ | ตะ | กั่ว | ถือ | เป็น | โล | หะ | หนัก</td></tr>
<tr><td>Word</td><td>สามารถ (can) | ยืดอายุ (extend life) | แบตเตอรี่ (battery) | ได้ (get)</td><td>เพราะ (because) | ตะกั่ว (lead) | ถือ (hold) | เป็น (is) | โลหะ (metal) | หนัก (heavy)</td></tr>
</table>

[5] https://pypi.org/project/basicthainlp

3.3 Speech Trimming

In the speech trimming phase, we hypothesize that the trimming process will mitigate noise and enhance model performance. Trim speech data refers to audio samples where silence at the beginning and end of recordings has been removed. For this crucial step, we employ the 'librosa' library, specifically the 'effects.trim' function with a threshold set at 30 decibels. This function meticulously trims leading and trailing silences from the audio files, which is anticipated to refine the clarity and efficacy of the synthesized speech. It is important to note that during audio processing with libraries such as 'librosa', audio samples are conventionally normalized to floating-point values within the range of -1.0 to +1.0. This normalization standardizes the audio signal, streamlining various audio processing tasks, and remains consistent regardless of the audio's original bit depth. This step is essential as it ensures that the speech trimming and subsequent processing are performed on a uniform scale, thereby facilitating more accurate and reliable noise reduction.

3.4 End-to-End Text-to-Speech Model

In our pursuit of SOTA-TTS capabilities, we have chosen Tacotron2-DDC, VITS, and YourTTS models for their distinctive strengths and functionalities. Tacotron2-DDC[24] is selected for its traditional TTS approach with Mel feature extraction, VITS [10] for its innovative model incorporating encoded features, and YourTTS[4] for its integration of language identifier features within the VITS model. Tacotron2-DDC is an end-to-end model notable for processing character sequences alongside their corresponding speech waveforms. A pivotal component of Tacotron2-DDC is the inclusion of "WaveNet," essential for extracting features from waveforms and converting them into mel spectrograms. These spectrograms provide a lower-level acoustic representation, allowing Tacotron2-DDC to capture the intricate dynamics of speech and making it a valuable TTS performance.

Following Tacotron, VITS emerged as a new benchmark in TTS technology. It introduces an encoder component to preprocess both text and audio signals. In VITS, phonemes are utilized as input features for the text encoder, and audio signals are processed through a posterior encoder, yielding vectors representing audio signals. This architecture facilitates a more nuanced synthesis of speech. YourTTS builds upon the VITS architecture, enhancing it with multi-speaker and multi-language capabilities. It achieves this by incorporating language embeddings into the model's input, allowing for greater versatility and adaptability. These three models—Tacotron2-DDC, VITS, and YourTTS—represent the SOTA-TTS in our TTS study, each contributing uniquely to the development of advanced TTS systems.

4 EXPERIMENTS

In this paper, we introduce an end-to-end framework for Thai TTS that incorporates linguistic units to achieve optimal performance in Thai language synthesis. We have tailored the SOTA-TTS model to fit within this framework. Our experimental analysis concentrates on two primary factors that influence performance: the Thai Linguistic Unit (TH-LU) factor and the combined TH-LU with speech trimming factor. These factors were assessed using the Tsync1 and Tsync2 datasets. The experiments were conducted on a desktop computer running Ubuntu OS, equipped with an AMD Ryzen 9 5950X processor, 32 GB of RAM, and an NVIDIA RTX A5000 GPU with 24 GB of memory.

4.1 Datasets

In our experimental framework, we leveraged two primary corpora: Tsync1 and Tsync2, both of which are specifically tailored for Thai TTS applications [25]. These datasets, Tsync1 and Tsync2, are comprehensive and encompass a wide range of Thai speech variations, which contains speech and transcription. Together, they consist of approximately 7,868 utterances, totaling 12.36 hours of audio recordings [22]. These recordings were made by a professional female speaker, known for her clear articulation, and delivered in a reading style that reflects the standard Thai accent. This extensive collection is crucial for training robust TTS models, as it ensures a broad representation of Thai phonetics and intonation patterns.

For the Tsync1 and Tsync2 datasets, each are allocated into three-part, 80% of the data for training, 10% for validation, and the remaining 10% for testing. This distribution is designed to optimize the performance of the TTS models. The training set, comprising the majority of the data, provides a rich and diverse array of speech samples essential for the comprehensive training of the TTS algorithms. The validation set, plays a crucial role in fine-tuning the model parameters and preventing overfitting. Lastly, the testing set is vital for evaluating the model's capability in synthesizing natural-sounding Thai speech under varied and realistic conditions. This structured approach ensures a well-balanced dataset, catering to all aspects of model training, validation, and testing.

4.2 Experimental Setup

In our study, we operated under two fundamental assumptions impacting our TTS framework. Firstly, we posited that the TH-LU is a critical factor influencing performance variation. Secondly, we hypothesized that speech trimming serves as another factor contributing to performance differences. To test these assumptions, we designed two main experiments: the TH-LU factor experiment and the TH-LU with speech trimming factor experiment.

In the TH-LU factor experiment, we varied the TH-LU (including phoneme, syllable, and word levels) while maintaining a constant none speech trimming setting across each SOTA-TTS models, as shown in Figure 2. Conversely, in the TH-LU with speech trimming factor experiment, we followed a similar procedure as the first experiment but incorporated varying speech trimming parameters, as shown in Figure 3. These experiments align with our initial assumptions and are designed to validate the impact of these factors on our TTS framework.

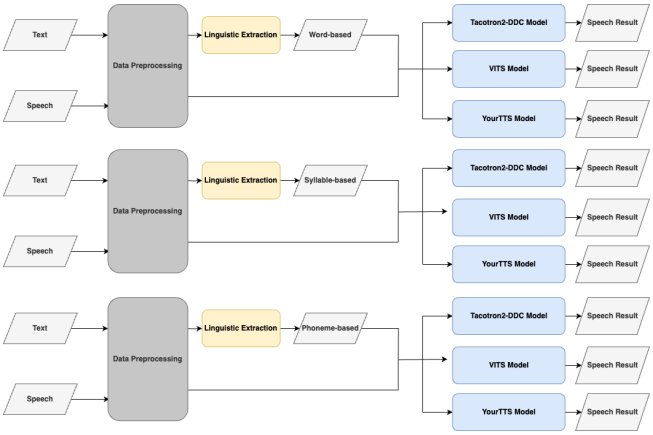

Figure 2: A comparative experiment based on the TH-LU factor experiment setup. Note that each of the processed units is employed in an individual end-to-end TTS model (blue block).

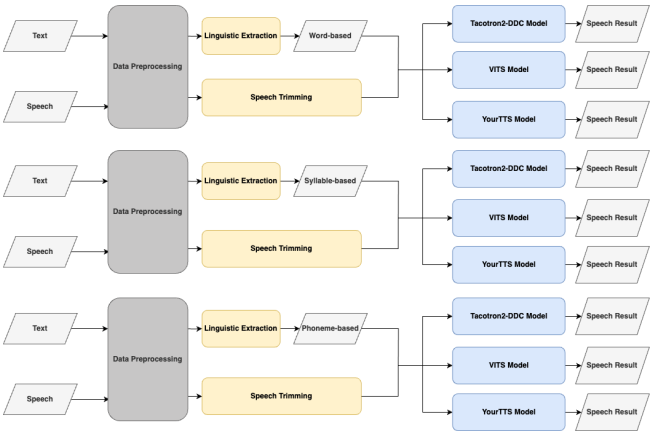

Figure 3: A comparative experiment based on the TH-LU with speech trimming factor experiment setup. Note that each of the processed units is employed in an individual end-to-end TTS model (blue block).

4.3 Evaluation

Our framework is evaluated based on two objectives including speaker tone and pronunciation. For the speaker tone objective, we utilized the Speaker Encoder Cosine Similarity (SECS) metric [3, 4, 12] to assess the resemblance between the synthesized speech and the original speaker's speech. This method involves calculating the cosine similarity between the speaker embeddings derived from two speech samples, using a speaker encoder. This approach effectively measures how closely the synthesized voice matches the characteristics of the original speaker's voice, providing a quantitative evaluation of the synthesis quality. We utilize the Coqui speaker encoder [2, 7], trained on the large dataset of VoxCeleb1 [13], Voxceleb2 [6], and all language CommonVoice [1] datasets. This encoder ensures broad generalizability in our SECS evaluations. In addition, for pronunciation, we utilized a Thai speech-to-text model[3]. The underlying assumption is that high-quality synthesized speech should yield similar speech-to-text results as the original speech. This method allows us to gauge the accuracy of pronunciation in the synthesized speech. Our evaluation script and the utilized speaker encoder model are available for public access[4], facilitating replication and further research in the field.

4.4 Metric Results and Discussions

In this study, we set two primary evaluation objectives—speaker tone and pronunciation—to gauge the performance of Thai TTS frameworks, as detailed in our experimental setup (see Subsection 4.2). Utilizing the SECS metric, we discerned the tonal accuracy of various TTS models, as shown in Table 2. VITS showcased exemplary performance, with SECS scores of 98.04 for Tsync1 (trim) and 96.49 for Tsync2 (non-trim) speech data in the word unit, and SECS scores of 98.82 for Tsync1 (trim) and 96.58 for Tsync2 (non-trim) in the syllable unit, indicating a high fidelity to the target speaker's tone. This trend continued over the phoneme unit , where VITS scored impressively, signifying its adeptness at capturing the essential tonal variations required for accurate Thai pronunciation.

While VITS stood out in tonal accuracy, Tacontron2-DDC and YourTTS lagged with lower SECS scores, indicating room for improvement in tonal representation. However, it's crucial to note that the SECS metric, though indicative of tonal precision, does not encompass pronunciation accuracy—a pivotal aspect in tonal languages like Thai, where the correct articulation of sounds is crucial for meaning conveyance. Despite the high SECS scores suggesting excellent tonal replication, VITS's capabilities in pronunciation clarity required further investigation.

[3]https://huggingface.co/biodatlab/whisper-th-medium-combined
[4]https://github.com/JoesSattes/Thai-TTS-Evaluation.git

To that end, we employed a Thai speech-to-text model[3] for a more rounded assessment of pronunciation performance (Table 3), which revealed that VITS maintained a strong lead, notably in the word linguistic unit, with the lowest Word Error Rate (WER) and Character Error Rate (CER) in both trim and non-trim scenarios. These metrics, in conjunction with SECS, present a holistic view of each model's strengths and highlight VITS's nuanced understanding of the Thai language.

Our discussion must also address the noticeable decline in SECS performance from trim to non-trim conditions for VITS, albeit less pronounced than in other models. This observation suggests potential areas for optimization, particularly in handling unprocessed speech data. Furthermore, the pronounced drop in YourTTS's phoneme-level performance, with high WER and CER, accentuates the inherent challenges TTS systems face with smaller LUs, especially in non-trim conditions.

The overarching conclusion from our analysis is that the word linguistic unit is the most effective for TTS synthesis in Thai, with VITS as the standout model, adeptly balancing clarity and naturalness in speech synthesis. This finding underscores the necessity to tailor TTS systems to the unique challenges of tonal languages and posits the integration of larger linguistic contexts as a path forward to refine speech synthesis quality. Our research contributes significantly to Thai TTS technology, paving the way for future advancements that enhance the natural expressiveness of TTS systems, potentially shaping the landscape of audio-focused machine learning research.

Table 2: The speaker tone performance of Thai end-to-end text-to-speech frameworks with the linguistic unit when evaluated on SECS.

Model	Linguistic Unit	Trim		Non-trim	
		Tsync1	Tsync2	Tsync1	Tsync2
Tacontron2-DDC	Word	76.48	79.44	74.51	77.35
VITS	Word	**98.04**	97.31	95.90	**96.49**
YourTTS	Word	92.18	85.27	80.14	80.98
Tacontron2-DDC	Syllable	77.30	77.11	74.45	76.93
VITS	Syllable	97.29	**98.82**	**96.58**	94.16
YourTTS	Syllable	91.10	84.15	80.24	81.35
Tacontron2-DDC	Phoneme	74.27	78.63	75.34	78.04
VITS	Phoneme	92.08	97.41	96.16	94.75
YourTTS	Phoneme	62.26	49.79	78.56	79.93

4.5 Mel Spectrogram Results and Discussions

Our experimental analysis of the Thai TTS systems is visually encapsulated in the Mel spectrograms, as shown in Figure 4. The VITS is chosen from the best performance previously mentioned. These

Table 3: The pronunciation performance of Thai end-to-end text-to-speech framework with linguistic unit when evaluated on WER and CER.

Model	Linguistic Unit	WER (%)				CER (%)			
		Trim		Non-trim		Trim		Non-trim	
		Tsync1	Tsync2	Tsync1	Tsync2	Tsync1	Tsync2	Tsync1	Tsync2
Tacontron2-DDC	Word	30.12	55.35	29.37	28.89	12.55	35.14	9.65	10.13
VITS	Word	**26.61**	54.16	**26.53**	**23.12**	**9.96**	31.37	**8.03**	**9.84**
YourTTS	Word	29.18	33.35	25.77	35.34	13.26	13.77	10.21	14.28
Tacontron2-DDC	Syllable	31.98	31.17	32.85	39.33	14.10	13.58	15.89	15.11
VITS	Syllable	28.44	**26.24**	30.55	37.08	12.13	**11.15**	15.52	14.04
YourTTS	Syllable	34.88	38.82	37.74	42.36	17.35	16.55	16.49	17.24
Tacontron2-DDC	Phoneme	71.89	40.15	117.50	35.67	54.01	20.28	81.88	14.31
VITS	Phoneme	71.66	39.33	104.08	34.69	52.52	19.11	81.25	13.95
YourTTS	Phoneme	251.38	379.35	584.82	291.57	107.32	225.30	520.88	216.71

spectrograms illustrate the frequency response over time for synthesized speech, providing a clear depiction of the model's ability to generate the melodic contours that are characteristic of natural speech. The vertical axis represents the Mel-frequency channels, indicative of the perceived pitch, while the horizontal axis represents the time frame of the spoken audio.

In the spectrograms, we observe well-defined patterns of harmonics, which are represented by the brighter yellow bands. These harmonics are essential in replicating the natural resonance and timbre of human speech. The darker areas represent lower energy, which typically corresponds to the absence of vocalization, such as pauses or unvoiced consonants.

The consistency in the patterns across the different spectrograms suggests a stable performance by the TTS model in generating speech with clear and consistent tonal quality. Notably, the model appears to exhibit proficiency in capturing the essential tonal variations required for accurate Thai pronunciation, which is critical given the tonal nature of the language.

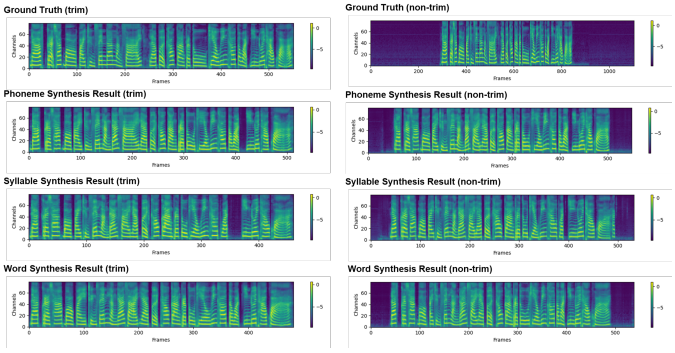

Figure 4: The Mel spectrogram results from the VITS model of our end-to-end with TH-LUs framework.

4.5.1 Thai Linguistic Unit Analysis based on Mel Spectrogram Results. The analysis of our VITS in Thai SOTA-TTS using Mel spectrogram results reveals critical insights across different TH-LUs - phonemes, syllables, and words:

Phonemes: Spectrograms indicate precise articulation and clear transitions between individual sounds, showcasing the model's ability to accurately reproduce the unique tonal characteristics essential for speech intelligibility and Thai phonology nuances.

Syllables: Observations reveal a cohesive structure with smoothly blending harmonics, mirroring the syllabic integrity vital for the Thai language's tonal and rhythmic patterns. This suggests the model's advanced handling of coarticulation phenomena and its capability to maintain consistent energy levels and appropriate attenuation at syllabic boundaries.

Words: The model effectively preserves prosodic features of longer linguistic constructs, demonstrating a natural fluctuation of intonation that aligns with fluent Thai speech's prosodic contours. This highlights the model's adeptness in simulating native Thai speech patterns, maintaining word clarity amidst the complex interplay of tone and stress.

The analysis underscores the word-based TH-LU as providing the most comprehensive context for leveraging the model's synthesis capabilities. This approach allows for better prediction and generation of speech with enhanced prosody and naturalness, crucial for tonal languages.

4.5.2 Thai Speech Trimming Analysis. The analysis on speech trimming in our VITS, as part of SOTA-TTS, demonstrates a clear impact on model performance between trimmed and non-trimmed speech data. Removing silences from the beginning and end of recordings (trimmed data) enhances tonal and prosodic clarity, leading to higher quality synthesis. This is evidenced by improved SECS scores in Table 2 and spectrogram pattern in Figure 4. This reflects a closer resemblance to the target speaker's tone, though it does not necessarily guarantee pronunciation accuracy.

In contrast, non-trimmed data, retaining these silences, presents more challenges for TTS models. Despite this, the VITS model showcases robust performance in processing such raw inputs, an essential feature for applications in real-world settings where speech inputs may not be preprocessed. Based on these findings, we recommend speech trimming to enhance model performance, especially for TTS systems that may not possess the same level of robustness as our VITS model, to ensure the highest quality of synthesized speech.

5 CONCLUSION

Our research offers a thorough examination of Thai TTS systems, centering in the influence of TH-LUs and speech trimming on the performance of SOTA-TTS models. We propose an end-to-end TTS framework that incorporates TH-LUs to enhance the clarity of process and experimentation in Thai. This framework focuses on three units: phonemes, syllables, and words, which are crucial in dictating the pronunciation of text in varying segments. In evaluating these systems, we concentrated on assessing speaker tone and pronunciation performance. Our methodical experiments have yielded significant insights. VITS model emerged as a leader in tonal accuracy according to the SECS evaluations, displaying exceptional performance across various TH-LUs in both trim and non-trim speech training data. Notably, in speaker tone evaluation, VITS with the syllable linguistic unit surpassed baselines in the Tsync1 dataset without speech trimming and the Tsync2 dataset with speech trimming. Moreover, VITS with the word linguistic unit excelled in the Tsync1 dataset with speech trimming and the Tsync2 dataset without speech trimming, underscoring the impact of speech trimming on these results. A pivotal observation from our study is the limitation of the SECS metric, which predominantly focuses on tonal aspects, thus neglecting pronunciation accuracy. This aspect is crucial in a tonal language like Thai, where pronunciation intricacies significantly influence meaning. While the high SECS scores of VITS indicate effective tonal replication, they do not guarantee pronunciation accuracy. This highlights the need for a more comprehensive evaluation framework. To address this, we incorporated a Thai speech-to-text model to evaluate pronunciation performance, presenting a more holistic assessment of TTS models. This method assesses both tonal accuracy and the clarity and correctness of pronunciation, ensuring a more accurate representation of a TTS model's effectiveness in replicating human-like speech. Our results indicate that VITS with the word linguistic unit outperforms all baselines overall. In terms of both speaker tone and pronunciation accuracy, the VITS with the word linguistic unit proves to be superior.

Thus, our study makes a significant contribution to the field of TTS, particularly for Thai language synthesis. It underscores the importance of considering various TH-LUs and speech trimming in SOTA-TTS model development and highlights the necessity of comprehensive evaluation methods that encompass both tonal accuracy and pronunciation quality. We believe our findings and methodologies will pave the way for further advancements in the development of more efficient and expressively natural Thai TTS systems, potentially impacting future audio-focused machine learning research.

ACKNOWLEDGEMENTS

This work was supported by Kasikorn Business-Technology Group (KBTG).

REFERENCES

[1] Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common Voice: A Massively-Multilingual Speech Corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*. 4218–4222.

[2] Edresson Casanova, Arnaldo Candido Junior, Christopher Shulby, Frederico Santos de Oliveira, Lucas Rafael Stefanel Gris, Hamilton Pereira da Silva, Sandra Maria Aluísio, and Moacir Antonelli Ponti. 2021. Speech2Phone: a novel and efficient method for training speaker recognition models. In *Brazilian Conference on Intelligent Systems*. Springer, 572–585.

[3] Edresson Casanova, Christopher Shulby, Eren Gölge, Nicolas Michael Müller, Frederico Santos de Oliveira, Arnaldo Candido Junior, Anderson da Silva Soares, Sandra Maria Aluisio, and Moacir Antonelli Ponti. 2021. Sc-glowtts: an efficient zero-shot multi-speaker text-to-speech model. *arXiv preprint arXiv:2104.05557* (2021).

[4] Edresson Casanova, Julian Weber, Christopher D Shulby, Arnaldo Candido Junior, Eren Gölge, and Moacir A Ponti. 2022. Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone. In *International Conference on Machine Learning*. PMLR, 2709–2720.

[5] Suphattharachai Chomphan. 2011. Analysis of Decision Trees in Context Clustering of Hidden Markov Model Based Thai Speech Synthesis. *Journal of Computer Science* 7, 3 (Mar 2011), 359–365. https://doi.org/10.3844/jcssp.2011.359.365

[6] J Chung, A Nagrani, and A Zisserman. 2018. VoxCeleb2: Deep speaker recognition. *Interspeech 2018* (2018).

[7] Joon Son Chung, Jaesung Huh, Seongkyu Mun, Minjae Lee, Hee Soo Heo, Soyeon Choe, Chiheon Ham, Sunghwan Jung, Bong-Jin Lee, and Icksang Han. 2020. In defence of metric learning for speaker recognition. In *Proc. Interspeech*.

[8] Eren Gölge. 2020. Solving attention problems of tts models with double decoder consistency.

[9] Chatchawarn Hansakunbuntheung, Virongrong Tesprasit, and Virach Sornlertlamvanich. 2003. Thai tagged speech corpus for speech synthesis. *The Oriental COCOSDA 2003* (2003), 97–104.

[10] Jaehyeon Kim, Jungil Kong, and Juhee Son. 2021. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *International Conference on Machine Learning*. PMLR, 5530–5540.

[11] Yogesh Kumar, Apeksha Koul, and Chamkaur Singh. 2023. A deep learning approaches in text-to-speech system: A systematic review and recent research perspective. *Multimedia Tools and Applications* 82, 10 (2023), 15171–15197.

[12] Chao Li, Xiaokong Ma, Bing Jiang, Xiangang Li, Xuewei Zhang, Xiao Liu, Ying Cao, Ajay Kannan, and Zhenyao Zhu. 2017. Deep speaker: an end-to-end neural speaker embedding system. *arXiv preprint arXiv:1705.02304* (2017).

[13] Arsha Nagrani, Joon Son Chung, Weidi Xie, and Andrew Zisserman. 2020. Voxceleb: Large-scale speaker verification in the wild. *Computer Speech & Language* 60 (2020), 101027.

[14] Wannaphong Phatthiyaphaibun. 2020. thai-g2p. https://github.com/wannaphong/thai-g2p/.

[15] Anocha Rugchatjaroen, Sittipong Saychum, Sarawoot Kongyoung, Patcharika Chootrakool, Sawit Kasuriya, and Chai Wutiwiwatchai. 2019. Efficient two-stage processing for joint sequence model-based Thai grapheme-to-phoneme conversion. *Speech Communication* 106 (2019), 105–111.

[16] Sittipong Saychum, Anocha Rugchatjaroen, and Chai Wutiwiwatchai. 2019. A great reduction of wer by syllable toneme prediction for thai grapheme to phoneme conversion. In *2019 22nd Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*. IEEE, 1–5.

[17] Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. 2018. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 4779–4783.

[18] Sattaya Singkul, Borirat Khampingyot, Nattasit Maharattamalai, Supawat Taerungruang, and Tawunrat Chalothorn. 2019. Parsing thai social data: A new challenge for thai nlp. In *2019 14th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)*. IEEE, 1–7.

[19] Sattaya Singkul and Kuntpong Woraratpanya. 2019. Thai dependency parsing with character embedding. In *2019 11th International Conference on Information Technology and Electrical Engineering (ICITEE)*. IEEE, 1–5.

[20] Sunayana Sitaram, Sukhada Palkar, Yun-Nung Chen, Alok Parlikar, and Alan W Black. 2013. Bootstrapping text-to-speech for speech processing in languages without an orthography. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 7992–7996.

[21] Paul Taylor. 2009. *Text-to-speech synthesis*. Cambridge university press.

[22] Ausdang Thangthai, Sumonmas Thatphithakkul, Kwanchiva Thangthai, and Arnon Namsanit. 2020. TSynC-3miti: Audiovisual Speech Synthesis Database from Found Data. In *2020 23rd Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*. 77–82. https://doi.org/10.1109/O-COCOSDA50338.2020.9295001

[23] Nipon Theera-Umpon, Suppakarn Chansareewittaya, and Sansanee Auephanwiriyakul. 2011. Phoneme and tonal accent recognition for Thai speech. *Expert Systems with Applications* 38, 10 (2011), 13254–13259.

[24] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. 2017. Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135* (2017).

[25] Chai Wutiwiwatchai, Patcharika Chootrakool, Sittipong Saychum, Nattanun Thatphithakkul, Anocha Rugchatjaroen, and Ausdang Thangthai. [n. d.]. TSynC-2: Thai Speech Synthesis Corpus Version 2. ([n. d.]).

[26] Chai Wutiwiwatchai, Patcharika Chootrakool, Sittipong Saychum, Nattanun Thatphithakkul, Anocha Rugchatjaroen, and Ausdang Thangthai. 2008. TSynC-2: Thai Speech Synthesis Corpus Version 2 TSynC-2. (2008).

[27] Chai Wutiwiwatchai, Chatchawarn Hansakunbuntheung, Anocha Rugchatjaroen, Sittipong Saychum, Sawit Kasuriya, and Patcharika Chootrakool. 2017. Thai text-to-speech synthesis: a review. *Journal of Intelligent Informatics and Smart Technology* 2, 2 (2017), 1–8.

[28] Detai Xin, Yuki Saito, Shinnosuke Takamichi, Tomoki Koriyama, and Hiroshi Saruwatari. 2021. Cross-Lingual Speaker Adaptation Using Domain Adaptation and Speaker Consistency Loss for Text-To-Speech Synthesis.. In *Interspeech*. 1614–1618.

[29] Sukanya Yimngam, Wichian Premchaisawadi, and Worapoj Kreesuradej. 2008. State of the Art Review on Thai Text-to-Speech System. In *2008 International Conference on Computer Science and Information Technology*. 194–198. https://doi.org/10.1109/ICCSIT.2008.158

[30] Jian Zhu, Cong Zhang, and David Jurgens. 2022. ByT5 model for massively multilingual grapheme-to-phoneme conversion. *arXiv preprint arXiv:2204.03067* (2022).