



Multi-Speaker Thai Speech Synthesis using Transfer Learning

Kittikan Charoenrattana¹, Pusadee Seresangtakul² and Pongsathon Janyoi³

ABSTRACT

This paper investigates transfer learning for developing a multi-speaker Thai text-to-speech (TTS) system under low-resource conditions, with a focus on cross-lingual knowledge transfer from source languages with different phonological and prosodic characteristics. The model is pre-trained on speech data from Thai, Mandarin Chinese, and English, and subsequently fine-tuned using Thai speech from three target speakers, each with only 15–30 minutes of data, covering both male and female speakers. Both objective and subjective evaluations consistently demonstrate that Thai pre-training achieves the best overall performance. Among the cross-lingual models, transfer learning from Mandarin Chinese outperforms transfer from English, yielding lower F0 RMSE (20.79 vs. 21.65) and higher MOS scores (3.55 vs. 3.44). In addition, speaker-dependent analysis indicates that speaker gender has a noticeable influence on synthesis quality under limited data conditions, suggesting that acoustic similarity between pre-training data and target speakers can affect the effectiveness of knowledge transfer. However, this factor is secondary to linguistic and prosodic similarity. Natural speech achieves a MOS of 4.93, while the Thai, Mandarin Chinese, and English-pre-trained models obtain MOS scores of 3.65, 3.55, and 3.44, respectively. These results highlight the importance of linguistic proximity in cross-lingual multi-speaker TTS, particularly tonal and prosodic similarity between source and target languages. Overall, the study confirms that transfer learning is an effective approach for low-resource Thai TTS and that tonal source languages provide more beneficial knowledge transfer than non-tonal, accent-based languages.

Article information:

Keywords: Text-to-Speech, Speech Synthesis, Transfer Learning, Multi-speaker

Article history:

Received: January 24, 2026

Revised: March 26, 2026

Accepted: April 23, 2026

Published: April 30, 2026

(Online)

DOI: 10.37936/ecti-cit.2026202.265851

1. INTRODUCTION

Text-to-speech (TTS) technology is a remarkable innovation that enables computers to convert text into speech. The early implementations relied on statistical parametric modeling [1], which involved three main stages: text analysis, acoustic modeling, and speech synthesis. These stages significantly increased system complexity. In recent years, end-to-end neural networks (E2E) for speech synthesis [2], [3] have simplified this process by generating speech directly from text and speech pairs without an explicit text analysis step. This design reduces processing time at this stage. E2E systems are categorized into two approaches: autoregressive and non-autoregressive, based on their output sequence generation mechanisms. Autoregressive models generate

acoustic features sequentially, where each prediction depends on previous outputs. This sequential generation process effectively models temporal dependencies and dynamic speech structures, often resulting in high-quality synthesized speech. However, step-by-step generation inherently results in high inference latency and limited parallelization efficiency. Non-autoregressive models [4] generate multiple acoustic frames simultaneously, without dependence on outputs from previous time steps. This design enables fully parallelized computation, as demonstrated by models such as FastSpeech [5] and FastSpeech2 [6], resulting in a significant improvement in speech synthesis speed.

Multi-speaker speech synthesis systems have attracted considerable research attention in recent years

^{1,2,3}The authors are with the Natural Language and Speech Processing Laboratory, Department of Computer Science, College of Computing, Khon Kaen University, Thailand, Email: ckittikan@kku.ac.th, pusadee@kku.ac.th and pongsathon@kku.ac.th

²Corresponding author: pusadee@kku.ac.th

[7]-[9]. These systems aim to generate speech from multiple speakers within a shared architecture. Such models leverage speech data from diverse speakers to learn variations in vocal characteristics, speaking styles, and phonetic attributes. Speaker-related information is commonly incorporated into the model, such as speaker identifiers [7] or reference speech signals [10], [11]. A key advantage of multi-speaker TTS systems lies in their flexibility in collecting speech data from heterogeneous sources, as well as their ability to reduce the amount of training data required per speaker. Nevertheless, substantial inter-speaker variability makes learning effective multi-speaker models from limited data particularly challenging, especially for achieving high-quality speech synthesis while preserving speaker-specific characteristics.

Despite ongoing efforts to develop modern TTS systems, many languages still lack high-quality speech corpora, particularly in low-resource languages [12]. This limitation often results in reduced correctness and naturalness in synthesized speech. Monolingual training typically requires large amounts of data to achieve high-quality speech synthesis, an approach that is often impractical or insufficient for low-resource languages. To address this issue, numerous studies have adopted transfer learning [13]-[15], in which TTS models are first pre-trained from high-resource languages and subsequently fine-tuned on target languages or domains with limited data. This approach enables models to reuse previously learned representations, such as phonetic, prosodic, and acoustic features, effectively reducing the amount of training data required for the target language. Moreover, in multi-speaker TTS scenarios, transfer learning has been shown to preserve speech quality while maintaining speaker identity, even when only a small amount of speaker-specific data is available.

However, transfer learning inherently relies on speech data from other languages, resulting in a cross-lingual learning scenario that holds strong potential for developing TTS systems for low-resource languages. Differences in phonetic and prosodic characteristics across languages, particularly in tonal languages, can adversely affect the naturalness of synthesized speech.

Thai is a tonal language in which pitch variation plays a fundamental role in distinguishing lexical meaning. Words with identical orthographic forms can convey entirely different meanings when produced with different tones. Although the Thai writing system explicitly marks only four tone symbols, spoken Thai realizes five distinct lexical tones: mid, low, falling, high, and rising [16]. These tonal contrasts are highly sensitive, and even minor deviations in fundamental frequency (F0) contours or pitch realization can lead to semantic ambiguity, posing significant challenges for accurate speech modeling [17]. Thai has a complex phonological structure character-

ized by a rich inventory of consonants and vowels, as well as phonemic distinctions in vowel length. Moreover, the relationship between written text and spoken realization in Thai is inherently nontransparent [18]. Orthographic forms do not consistently reflect surface acoustic patterns. This linguistic complexity substantially increases the difficulty of data preparation.

As a consequence, Thai speech modeling systems must learn intricate phonological and prosodic patterns directly from limited data, exacerbating data sparsity and reducing model robustness. Despite the large number of Thai speakers, publicly available high-quality speech corpora remain limited compared to high-resource languages [19], such as English and Mandarin. This scarcity of data constrains the performance and generalization ability of data-driven TTS systems, especially those based on deep learning architectures, and highlights the need for effective modeling strategies tailored to low-resource tonal languages.

In this paper, we investigate the impact of cross-lingual transfer learning on the intelligibility and naturalness of Thai speech synthesis in a low-resource setting. Specifically, we fine-tune pre-trained models using speech data from a tonal language and an accent language. This experimental design enables a systematic comparison of the effects of linguistically similar versus typologically distinct pre-training languages on Thai speech synthesis performance. We further examine the influence of the gender composition and the size of the fine-tuning dataset on model performance.

2. RELATED WORK

2.1 Multi-Speaker Text-to-Speech

Multi-speaker TTS systems represent a significant advancement in speech synthesis technology. A key challenge in multi-speaker TTS is disentangling speaker identity from linguistic content. A widely used approach to address this challenge is the use of speaker embeddings, which are continuous vector representations of speaker identity and are provided to the TTS model as conditioning information to control the characteristics of the synthesized speech. Early neural TTS systems, such as Deep Voice 2 [20] and Deep Voice 3 [21], used speaker IDs alongside an embedding matrix serving as a lookup table, with speaker embeddings learned during training. These approaches demonstrated that speaker-specific embeddings can effectively preserve speaker identity and achieve stable performance in multi-speaker TTS.

However, speaker embeddings that rely directly on speaker identifiers exhibit limited generalization to unseen speakers. To address this limitation, [22], [23] proposed using a speaker encoder, a neural network trained to automatically extract speaker embeddings from reference speech signals. Beyond speaker

identity, [11], [14] further disentangled speaking style, or prosody, from speaker identity by style encoders. Global Style Tokens (GST) [24] employ an unsupervised style encoder to learn style embeddings that capture variations in speaking rate, pitch, and expressive characteristics, independent of speaker identity. This line of work suggests that speaking style can be treated as a distinct, controllable aspect separate from speaker identity.

Although prior studies have demonstrated that model adaptation mechanisms and the use of reference speech can effectively enable TTS systems to generalize to previously unseen speakers, in scenarios where the set of speakers is limited and predefined, the use of a speaker embedding lookup table remains a straightforward and effective approach for conditioning the model on speaker identity. This approach reduces architectural complexity and facilitates stable learning of speaker-specific representations without relying on additional reference-based encoding mechanisms.

2.2 Tonal and Accent Language Transfer Learning

Tonal languages use pitch (F0) at the syllabic level to distinguish word meanings, as in Thai, Chinese, and Vietnamese, which employ multiple pitch levels such that a change in tone can completely alter word meaning. In contrast, accent languages do not use pitch as a lexical contrast but instead rely on phonetic variation and prosodic features, such as pitch and duration, to produce natural and intelligible speech [25]. Accurate synthesis of these prosodic features is therefore crucial for maintaining both naturalness and intelligibility in TTS systems. Consequently, several studies have focused on accent modeling and prosody control within TTS systems.

TTS for both tonal and accent languages remains a major challenge in speech synthesis research, as prosodic characteristics directly affect intelligibility and perceived naturalness. As a result, modern TTS systems have increasingly emphasized explicit prosody control, particularly in non-autoregressive models. Which separately predict pitch and duration in a structured manner. Numerous studies have proposed accent adaptation or fine-tuning strategies using target-accent data, enabling models to learn phonetic and prosodic variations effectively, even with limited training data. Studies [26], [27] have demonstrated that selective adaptation of specific modules, including pitch and duration predictors, can improve a model's ability to convey target accent characteristics without disrupting the core linguistic structure of the source language. However, this line of research has been conducted primarily in languages with the same or closely related linguistic typologies.

In the context of cross-lingual transfer learning for TTS, several studies have explored knowledge trans-

fer across different languages by leveraging multilingual pre-training followed by fine-tuning on target-language data [15], [28]. These works mainly focus on improving overall speech quality, enhancing synthesis performance in low-resource languages, or addressing speaker and accent adaptation at the application level. Nevertheless, they generally lack an in-depth analysis of how differences in linguistic typology between source and target languages affect the learning and transfer of prosodic features. This limitation becomes particularly evident when transferring knowledge between structurally distinct languages.

3. METHODOLOGY

3.1 Overview of the Proposed Approach

This paper proposes the transfer learning to develop a multi-speaker Thai text-to-speech system under limited data resource conditions. This study evaluates three transfer learning scenarios, categorized according to the linguistic characteristics of the source languages: (1) Thai fine-tuning, which transfers knowledge from a single-speaker Thai corpus to a multi-speaker Thai TTS system within the same tonal language; (2) Mandarin Chinese fine-tuning, which represents transfer from another tonal language; and (3) English fine-tuning, which represents transfer from an accent language. This experimental design enables a systematic analysis of the impact of linguistic similarity on the performance of multi-speaker Thai TTS systems under low-resource conditions. The two training stages of the proposed transfer learning architecture are shown in Figure 1.

First stage, the model was pre-trained on languages with prominent accent or tonal characteristics, resulting in two separate models: one for accent languages and another for tone languages. Each model uses a single speaker to capture pitch and duration variations of the language.

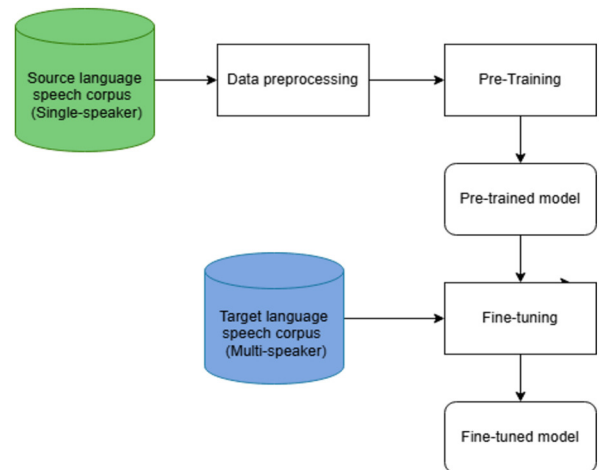


Fig.1: Training stages of the proposed transfer learning architecture.

By learning the relationships between textual input and pitch sequences, the models established a robust foundation for speech synthesis that closely reflects the characteristics of the source language. Training on a single speaker also allowed the model to effectively learn complex pitch contours, dynamic range, and prosody patterns without interference from inter-speaker variability.

In the second stage, the pre-trained models from the first stage were fine-tuned in Thai using multi-speaker data, with speakers distinguished by unique speaker IDs. This setup enabled the model to accurately capture inter-speaker variations while generating synthetic speech that preserves each speaker's unique characteristics. To leverage the speaker embeddings learned during pre-training, speaker weights from the source language were loaded from the pre-trained model's state dictionary and duplicated for each target Thai speaker before fine-tuning. This procedure allowed the model to adapt to new speakers efficiently, without learning pitch and prosody from scratch, while preserving the pitch contours, intonation patterns, and prosody acquired from the source languages. Furthermore, the use of speaker IDs provides a flexible framework for multi-speaker speech synthesis, enabling the system to incorporate new speakers without affecting the existing model while preserving each speaker's individual prosodic characteristics. This capability is a key advantage for high-quality multi-speaker Thai speech synthesis.

3.2 Data Preprocessing

This paper utilizes speech data from three languages: Thai, Chinese, and English, as detailed in Section 4.1. First, the text data undergoes cleaning and normalization, removing special characters and unnecessary spaces. The processed data are stored in a JSON file that includes the filename, duration, text, and speaker information. In the text component, characters in each language are converted into phonemes using a Grapheme-to-Phoneme (G2P) conversion process. For tonal languages, the tone of each word is annotated along with its vowel. In Mandarin phonetics, initial consonants and final consonants are typically combined with vowels to form syllables. In this work, we separated the vowel and the components of the initial and final consonant. For example, the vowel “iang4” could be segmented as “ia4|ng\^”. With respect to speaker information, each speaker is represented by a unique index number, which allows the system to distinguish between different speakers. This index-based labeling helps to maintain speaker consistency across datasets and facilitates subsequent analysis.

3.3 FastPitch Architecture

In this paper, we develop a non-autoregressive end-to-end multi-speaker speech synthesis system using

the FastPitch architecture [29] as the baseline acoustic model. The FastPitch architecture enables fast training and inference while maintaining high speech quality. This architecture is a fully feed-forward transformer-based model that directly predicts mel-spectrograms from phoneme sequences, allowing parallel processing. The model architecture is illustrated in Figure 2.

1) Feed-Forward Transformer (FFT) Encoder: The feed-forward Transformer encoder [5] processes a sequence of phonemes through encoder blocks, each comprising a multi-head self-attention mechanism and a position-wise feed-forward network. The encoder is designed to learn and model long-range contextual dependencies among phonetic units in the input text. The self-attention mechanism enables the model to attend to all phoneme positions simultaneously, facilitating effective contextual representation learning. At the same time, the position-wise feed-forward network applies non-linear transformations to the representations at each position.

2) Duration Predictor: The duration predictor is a multi-layer convolutional network that estimates the duration (number of frames) for each phoneme. This component addresses the alignment problem between textual and acoustic representations, enabling fully parallel synthesis.

3) Pitch Predictor: The pitch predictor estimates the fundamental frequency (F0) at the frame level, which is particularly crucial for modeling tonal characteristics in Thai speech. Convolutional layers effectively capture pitch contour patterns.

4) FFT Decoder: The FFT decoder converts the upsampled encoder outputs—based on the predicted durations—into mel-spectrograms through M decoder blocks that share the same architecture as the encoder.

5) Speaker Embedding Layer: Speaker embedding lookup models each speaker as a learnable continuous vector indexed by a speaker ID (s) defined as:

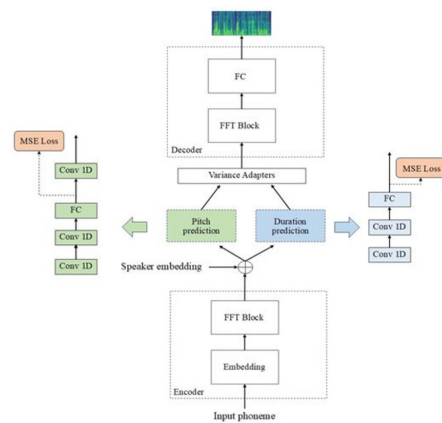


Fig.2: The architecture of the FastPitch multi-speaker speech synthesis [29].

$$e_s = \text{Embedding}(s) \quad (1)$$

The text encoder generates a sequence of linguistic representations at each time step (t). Given an input text sequence $x = (x_1, x_2, \dots, x_N)$, where N is the length of the input text sequence, the encoding process can generally be formulated as follows.

$$h_t = \text{Encoder}(x_t) \quad (2)$$

The speaker embedding is then integrated with the output of the text encoder at every time step through a speaker conditioning mechanism, which can be generally formulated as

$$\tilde{h}_t = h_t + e_s \quad (3)$$

The corresponding embedding is injected into the text encoder output at each time step through speaker conditioning, which enables multi-speaker speech synthesis by disentangling speaker-dependent acoustic variations from linguistic content. The model is trained end-to-end using a composite loss function defined as:

$$L_{total} = L_{mel} + \lambda_{dur}L_{dur} + \lambda_{pitch}L_{pitch} \quad (4)$$

where L_{mel} denotes the mean squared error (MSE) between the predicted and ground-truth mel-spectrograms, L_{dur} represents the MSE loss for the duration prediction, L_{pitch} represents the MSE loss for pitch prediction, and λ_{dur} and λ_{pitch} are weighting coefficients that balance the contribution of each loss term.

For waveform generation, we used the HiFi-GAN [30] vocoder to convert mel-spectrograms into audio files. The vocoder is pre-trained and not modified during the experiment.

Table 1: Summary of Pre-training Datasets.

Dataset	Language	Speaker	Duration	Sentence	Tone System
Nun-TTS	Thai	1 Female	5 h 25 m	2,710	5 tones
Call-To-Arm	Mandarin	1 Female	3 h 55 m	1,815	4 tones + neutral
LJSpeech	English	1 Female	23 h	13,100	Stress-accent

4. EXPERIMENTAL SETUP

4.1 Dataset

The dataset is split into two parts: pre-training data and fine-tuning data.

1) Pre-train dataset: A single-speaker speech corpus from three languages with distinct linguistic characteristics is employed, as shown in Table 1.

Thai corpus: The dataset used public data from “Nun-TTS” in “AiForThai” [31]. It comprises 2,710 Thai sentences. The speech data includes 1 female speech with a sample rate of 44.1 kHz at 16 bits/sample. The total time is 5 hours and 25 minutes. The corpus consists of phonetically balanced

sentences and covers all five Thai tones: mid, low, falling, high, and rising.

Mandarin Chinese corpus: The dataset used readings of the short story “Call-to-Arms”, which is in the public domain of “LibriVox” [32]. The recordings were divided into 1,815 sentences, totaling 3 hours and 55 minutes, and were recorded by a single female speaker. The sampling rate is 44,100 Hz. Chinese is a tonal language with four tones (level, rising, falling-rising, and falling) in addition to a neutral tone, making Chinese prosodically similar to Thai.

English corpus: Data have been collected from “LJSpeech” [33], consisting of 13,100 sentences, over a total time of 23 hours and 55 minutes, uttered by a woman. For training, the audio sample rate is converted to 44,100 Hz and a bit depth of 16 bits per sample. English employs a stress-accent system rather than lexical tones, making it a suitable contrastive case for evaluating the role of tonal similarity in cross-lingual transfer learning.

2) Fine-tune dataset: We collected Thai speech data from three native speakers (two males and one female). The speakers will read 200 sentences and divide them into two sets: a set of 100 sentences (15 minutes) and a set of 200 sentences (30 minutes). All recordings were made in a quiet recording room with a sampling rate of 44.1 kHz and a 16-bit resolution. Each recording session was manually inspected for audio quality.

4.2 Model Construction

In this experiment, Chinese and Thai are selected as source domains for tonal languages due to their well-defined pitch patterns, which facilitate the model’s learning of the relationship between pitch and pronunciation. English is chosen as the source domain for accent languages to provide a contrast in prosodic and phonetic characteristics. A total of 6 models is initially pre-trained on data from three languages consisting of Thai, Mandarin, and English.

Table 2: Fine-tuning Configurations for Thai Target Speech.

Model	Prosodic Typology	Pre-train Languages	Sentence	Duration (min)
Th-200	Tonal (Thai)	Thai	200	30
Th-100		Thai	100	15
Ch-200	Tonal (Thai)	Chinese	200	30
Ch-100		Chinese	100	15
En-200	Tonal (Thai)	English	200	30
En-100		English	100	15

Subsequently, the models are fine-tuned in the target domain using two different data configurations, 200 sentences per speaker and 100 sentences per speaker, in a separate experimental setup, to evaluate the effectiveness of different data sizes for speaker-specific adaptation. In the fine-tuning process, we used data from 2 male and 1 female speakers to investigate potential effects of speaker gender, as shown in Table 2.

4.3 Training Configuration

The main hyperparameters remain similar to the standard FastPitch [29]. An additional module for adapter training is included, which allows the model to learn adapter weights instead of training the entire network. The encoder and decoder consist of 1-D convolution layers with a kernel size of 6 and 384/1,536 input/output channels. The duration and pitch predictors have an input size of 384, a filter size of 256, and 2 layers. The optimizer used is LAMB with a learning rate of 1e-2. For spectral analysis, a sampling rate of 44.1 kHz is used, with an FFT size of 2,048, a window size of 2,048, a hop size of 512, and 80 Mel channels. The pre-training models were trained for 2,000 epochs, and the fine-tuned models were trained for 1,000 epochs.

5. RESULT AND DISCUSSION

To evaluate the performance, we assessed the proposed multi-speaker Thai speech synthesis system, focusing on the impact of source languages with tonal and accent characteristics on fine-tuning. Evaluation was conducted using both objective metrics and subjective listening tests. A total of 8 types of speech were compared, including synthesized speech from 4.2, original speech recordings, and a multi-speaker model trained from scratch on 200 sentences to examine the effect of fine-tuning versus training from scratch.

For testing, a set of 10 sentences per speaker, not included in the training data, was used. Speech synthesis was conducted separately for each speaker, resulting in 30 synthesized sentences per model. Across all models and the original speech, the total number of speech samples used for evaluation was 240.

5.1 Objective Evaluation Results

In the objective evaluation, two primary aspects were assessed: spectral accuracy and pitch accuracy. Spectral accuracy was evaluated using Mel-Cepstral Distortion (MCD), a standard metric that quantifies the spectral difference between synthesized and original speech based on Mel-frequency cepstral coefficients (MFCCs).

Table 3: Mel-Cepstral Distortion comparison pre-trained models and speakers (dB).

Model	Male		Female	AVG
	Speaker1	Speaker2	Speaker3	
Th-200	5.63	6.64	8.59	6.95
Ch-200	5.68	6.78	8.50	6.99
En-200	5.69	7.01	8.31	7.00
Th-100	5.85	7.13	8.70	7.23
Ch-100	6.37	7.45	8.93	7.58
En-100	6.22	7.29	8.80	7.44
From Scratch	6.06	7.05	8.53	7.21

Table 4: F0 RMSE comparison of pre-trained models and speakers (Hz).

Model	Male		Female	AVG
	Speaker1	Speaker2	Speaker3	
Th-200	16.33	14.04	29.68	20.01
Ch-200	16.99	14.50	30.88	20.79
En-200	17.63	14.13	33.18	21.65
Th-100	15.05	12.99	37.51	21.85
Ch-100	16.81	14.30	40.20	23.77
En-100	16.88	14.79	40.24	23.97
From Scratch	19.53	14.09	37.94	23.85

Pitch accuracy, particularly important in Thai as a tonal language, was assessed using the fundamental frequency Root Mean Square Error (F0 RMSE), which measures the deviation of the pitch contours between synthesized and original speech signals. During F0 extraction, different pitch ranges were specified according to speaker gender to ensure reliable pitch tracking and reduce estimation errors. Specifically, the F0 range was set to 50–250 Hz for male speakers and 120–500 Hz for female speakers. The results of these objective evaluations for all models are summarized in Tables 3 and 4, which report the MCD and F0 RMSE values, respectively.

5.2 Subjective Evaluation

In the subjective test, we used the Mean Opinion Score (MOS) to assess the naturalness of each model's synthesized speech. Rate the naturalness of speech employing a score of 1-5, where 1 represents unnaturalness and 5 represents naturalness. The evaluation was designed to cover multiple speakers' accents to reflect each model's performance in a multi-speaker, tonal-language setting, such as Thai.

In addition, a paired t-test was used to examine differences in MOS across the models, based on the listeners' MOS ratings.

Table 5: Mean Opinion Score results.

Model	Male		Female	Total Average
	Speaker1	Speaker2	Speaker3	
Original	4.94 ± 0.37	4.91 ± 0.33	4.95 ± 0.35	4.93 ± 0.32
From Scratch	3.01 ± 1.09	3.35 ± 1.01	2.61 ± 1.03	2.99 ± 1.08
Th-200	3.75 ± 0.89	3.94 ± 0.91	3.26 ± 1.22	3.65 ± 1.06
Ch-200	3.82 ± 0.81	3.74 ± 0.92	3.08 ± 1.07	3.55 ± 0.99
En-200	3.52 ± 1.15	3.52 ± 1.07	3.28 ± 0.99	3.44 ± 1.08
Th-100	3.28 ± 1.09	3.41 ± 1.10	3.12 ± 1.06	3.27 ± 1.09
Ch-100	2.91 ± 1.18	3.51 ± 1.05	2.71 ± 1.08	3.04 ± 1.13
En-100	3.02 ± 1.05	3.36 ± 1.02	2.80 ± 1.05	3.06 ± 1.07

A total of 17 listeners (male and female, aged 18–30) with normal hearing participated in the test. All were native Thai speakers, enabling accurate evaluation of pronunciation, tonal characteristics, and naturalness. Before the evaluation, participants received clear instructions on the procedure and the MOS scale to ensure consistent scoring.

Each participant evaluated the same set of speech samples, which included synthesized speech from all compared models, totaling 240 samples. The evalu-

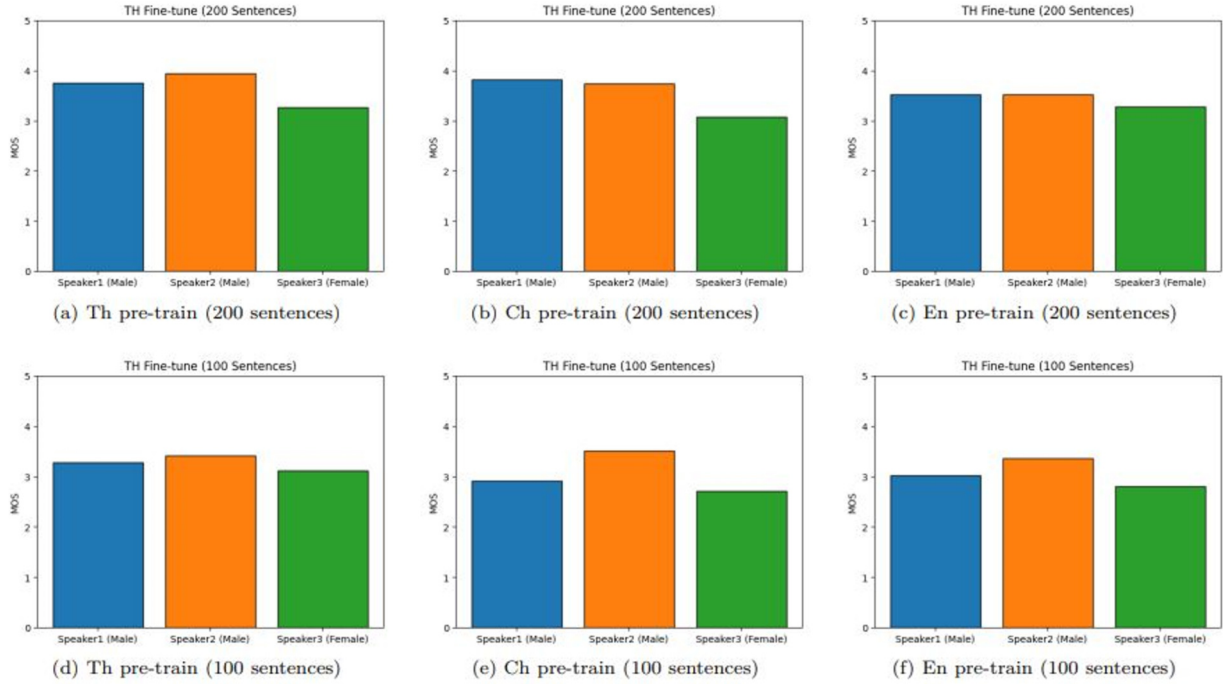


Fig.3: MOS comparison across languages and speakers using 200 sentences (top row) and 100 sentences (bottom row). Each bar represents a different speaker.

ation was conducted under identical listening conditions, with participants using headphones during the test. Speech samples were presented in random order to minimize potential order effects and listener bias. This experimental design helped ensure consistency across listeners and contributed to more reliable MOS averages. The results are summarized in Table 5, highlighting the differences in MOS performance among the evaluated models.

5.3 Impact of Pre-training Language Speech Data

The experimental results show that the language used during pre-training significantly impacts synthesized speech quality in both objective and subjective evaluations, particularly when analyzed within the framework of prosodic typology. In terms of objective performance, models pre-trained on tonal languages consistently outperform those pre-trained on accent-based languages across all experimental conditions, indicating a superior ability to preserve speech spectral structure. As reported in Table 3, with 200 pre-training utterances, Th-200 achieves the lowest average MCD of 6.95 dB, followed by Ch-200 at 6.99 dB, whereas En-200 yields a noticeably higher MCD of 7.00 dB. These results indicate that fine-tuning from a model pre-trained on a tonal language with prosodic characteristics similar to the target language is more effective in reducing spectral distortion than fine-tuning from a model pre-trained on an accent language.

A similar trend becomes even more pronounced

when considering the F0 RMSE, as shown in Table 4. F0 RMSE directly reflects the accuracy of pitch contour and tonal pattern modeling. Th-200 achieves the lowest F0 RMSE of 20.01 Hz, while Ch-200 and En-200 exhibit higher values of 20.79 Hz and 21.65 Hz, respectively. This finding suggests that tonal pre-training enables more effective phonemic-level control of F0. In contrast, accent-based pre-training inherently emphasizes pitch variation for stress and rhythmic purposes rather than tonal encoding.

The results of the subjective evaluation further support these objective findings. As summarized in Table 5, models employing tonal pre-training consistently achieve higher MOS ratings than the accent-based model. In the case of 200 pre-training utterances, Th-200 obtains the highest average MOS of 3.65 ± 1.06 , followed by Ch-200 with 3.55 ± 0.99 , while En-200 records a lower score of 3.44 ± 1.08 .

To verify whether the observed differences in MOS scores are statistically significant, a paired t-test was conducted. The results show that the proposed model (Th-100) significantly outperforms the other models (Ch-100 and En-100) at the $p < 0.05$ level. However, the comparison between Ch-200 and Th-200 is statistically significant only at the $p < 0.10$ level, suggesting that increasing the number of training utterances reduces the performance gap and makes the two models more comparable.

This finding can be attributed to the phonetic similarities between Thai and Mandarin. Both languages are tonal languages in which F0 contours play an important role in distinguishing lexical meaning. In

addition, Thai and Mandarin share several similar phonemes, and in some cases, identical phoneme representations can be used. These similarities facilitate more effective knowledge transfer, particularly regarding prosodic patterns and phoneme representation. However, the two languages still differ in their tonal systems. In contrast, English is an accent-based language, which primarily uses pitch variation to express stress and rhythm rather than to distinguish lexical meaning. Furthermore, the phonemic system of English differs considerably from that of Thai. As a result, knowledge transfer from English tends to be less effective compared with that from Mandarin [34].

In summary, these findings indicate that knowledge transfer from source languages with prosodic typologies closely aligned with that of the target language leads to superior performance. Such alignment enhances spectral quality, improves F0 control accuracy, and results in higher perceived speech naturalness.

5.4 Impact of fine-tune Data Size

The experimental results indicate that pre-training with 200 utterances (approximately 30 minutes of speech) yields the best overall performance in this study. Models trained on this data consistently achieve lower objective errors and higher perceived naturalness than those pre-trained on only 100 utterances. In addition, the differences in MOS scores between models of the same language trained with 100 and 200 utterances are statistically significant at the $p < 0.05$ level. This indicates that increasing the amount of training data can significantly improve the perceived quality of synthesized speech.

These findings suggest that approximately 30 minutes of pre-training speech constitutes an effective and sufficient data scale for learning both spectral and prosodic representations under the low-resource conditions considered. Increasing the amount of pre-training data yields more robust acoustic representations, which, in turn, improve the effectiveness of subsequent fine-tuning on limited target-speaker data.

5.5 Impact of Speaker Gender in Pre-training Data

In the reduced training data setting, the model trained with 200 utterances consistently achieved higher MOS values than the model trained with 100 utterances. When examining the speaker-dependent results shown in Figure 3, performance varies across speakers. In particular, Speaker-3 (female) exhibits a smaller degradation in MOS compared to the two male speakers when the synthesized language differs from the pre-training language and the amount of pre-training data is reduced. Since the pre-training dataset consists exclusively of female speech, these results suggest that speaker gender plays an important role in preserving the naturalness of synthesized

speech, especially under limited training data conditions.

6. CONCLUSION AND FUTURE WORK

Our research examines the efficiency of transfer learning for developing a multi-speaker Thai text-to-speech system under low-resource conditions, using the non-autoregressive FastPitch architecture. We compare three transfer scenarios: Thai fine-tune Thai, Mandarin Chinese fine-tune Thai, and English fine-tune Thai. The work systematically analyzes how linguistic similarity between the source and target languages influences speech synthesis quality in both objective and subjective evaluations.

The experimental results demonstrate that fine-tuning from Thai speech yields the best performance across all metrics, confirming the importance of in-language pre-training when sufficient data are available. Among cross-lingual settings, Mandarin Chinese transfer consistently outperforms English transfer, as evidenced by lower RMSE F0, reduced MCD, and higher MOS scores. This finding highlights the critical role of tonal and prosodic compatibility in cross-lingual TTS transfer, as both Thai and Mandarin are tonal languages. While English relies primarily on stress and intonation patterns, the results suggest that pre-trained models learned from tonal languages can be more effectively applied to Thai without significantly degrading tonal pronunciation.

Furthermore, the amount of fine-tuning data was shown to have a substantial impact on synthesis quality. Models fine-tuned with 200 utterances per speaker consistently outperform those trained on only 100, indicating that even modest increases in target-language data can significantly improve adaptation performance. Speaker-dependent analysis also revealed that female speakers tend to exhibit higher F0 RMSE and MCD values. However, they benefit more from pre-training on female speech, suggesting that speaker gender and pitch range should be considered when selecting source data for transfer learning.

However, this study was conducted with a limited number of speakers. In future work, a more diverse set of speakers will be incorporated to provide a more comprehensive evaluation and further enhance the model's ability to generate diverse, natural-sounding speech. In addition, improvements to the model architecture will be explored, including more effective speaker representations and conditioning mechanisms, to better capture speaker variability and enhance the overall quality and controllability of synthesized speech.

ACKNOWLEDGEMENT

This research was partially supported by the Department of Computer Science, College of Computing, Khon Kaen University, Khon Kaen, Thailand.

AUTHOR CONTRIBUTIONS

Conceptualization, K.C., P.S., and P.J.; methodology, K.C.; software, K.C.; investigation, and validation, K.C., P.S., and P.J.; data curation, K.C. and P.J.; writing-original draft preparation, K.C.; writing-review and editing, K.C., P.S., and P.J.; supervision, P.S. and P.J. All authors have read and agreed to the published version of the manuscript.

References

- [1] H. Zen, A. Senior and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, Vancouver, BC, Canada, pp. 7962-7966, 2013.
- [2] Y. Wang *et al.*, "Tacotron: Towards end-to-end speech synthesis," *arXiv preprint arXiv:1703.10135*, 2017.
- [3] J. Shen *et al.*, "Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, AB, Canada, pp. 4779-4783, 2018.
- [4] M. Lee, J. Lee and J. -H. Chang, "Non-Autoregressive Fully Parallel Deep Convolutional Neural Speech Synthesis," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 1150-1159, 2022.
- [5] Y. Ren *et al.*, "Fastspeech: Fast, robust and controllable text to speech," in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, no. 285, pp. 3171-3180, 2019.
- [6] Y. Ren *et al.*, "Fastspeech 2: Fast and high-quality end-to-end text to speech," *arXiv preprint arXiv:2006.04558*, 2020.
- [7] M. Chen *et al.*, "Multispeech: Multi-speaker text to speech with transformer," *arXiv preprint arXiv:2006.04664*, 2020.
- [8] T. Li, X. Wang, Q. Xie, Z. Wang and L. Xie, "Cross-Speaker Emotion Disentangling and Transfer for End-to-End Speech Synthesis," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 1448-1460, 2022.
- [9] M. Adibian and H. Zeinali, "End-to-End Multi-Speaker FastSpeech2 With Hierarchical Decoder," in *IEEE Access*, vol. 13, pp. 127805-127814, 2025.
- [10] E. Casanova *et al.*, "Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone," in *Proceedings of the 39th International Conference on Machine Learning (PMLR)*, vol. 162, pp. 2709-2720, 2022.
- [11] K. Azizah and W. Jatmiko, "Transfer Learning, Style Control, and Speaker Reconstruction Loss for Zero-Shot Multilingual Multi-Speaker Text-to-Speech on Low-Resource Languages," in *IEEE Access*, vol. 10, pp. 5895-5911, 2022.
- [12] M. A. Hedderich, L. Lange, H. Adel, J. Strötgen and D. Klakow, "A survey on recent approaches for natural language processing in low-resource scenarios," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2545-2568, 2021.
- [13] A. R. Gladston and K. V. Pradeep, "Exploring Solutions for Text-to-Speech Synthesis of Low-Resource Languages," *2023 4th International Conference on Signal Processing and Communication (ICSPC)*, Coimbatore, India, pp. 168-172, 2023.
- [14] K. Azizah, M. Adriani and W. Jatmiko, "Hierarchical Transfer Learning for Multilingual, Multi-Speaker, and Style Transfer DNN-Based TTS on Low-Resource Languages," in *IEEE Access*, vol. 8, pp. 179798-179812, 2020.
- [15] A. Debnath, S. S. Patil, G. Nadiger and R. A. Ganesan, "Low-Resource End-to-end Sanskrit TTS using Tacotron2, WaveGlow and Transfer Learning," *2020 IEEE 17th India Council International Conference (INDICON)*, New Delhi, India, pp. 1-5, 2020.
- [16] S. Chomphan and T. Kobayashi, "Implementation and evaluation of an HMM-based Thai speech synthesis system," *Interspeech 2007*, Antwerp, Belgium, pp. 2849-2852, 2007.
- [17] T. Daengsi and P. Pornpongtechavanich, "Quality of Experience: Comparison of Synthesized Speech Naturalness Between Apple's Siri and Google Translate Referring to Thai Language," *2021 International Conference on Computer Communication and Informatics (ICCCI)*, Coimbatore, India, pp. 1-4, 2021.
- [18] P. Janyoi and A. Thangthai, "Investigation of an Input Sequence on Thai Neural Sequence-to-Sequence Speech Synthesis," *2021 24th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, Singapore, Singapore, pp. 218-223, 2021.
- [19] A. Suwanbandit *et al.*, "Thai-Dialect: Low Resource Thai Dialectal Speech to Text Corpora," *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Taipei, Taiwan, pp. 1-8, 2023.
- [20] S. Arik *et al.*, "Deep voice 2: Multi-speaker neural text-to-speech," *arXiv preprint arXiv:1705.08947*, 2017.
- [21] W. Ping *et al.*, "Deep voice 3: Scaling text-to-speech with convolutional sequence learning," *arXiv preprint arXiv:1710.07654*, 2017.
- [22] Y. Jia *et al.*, "Transfer learning from speaker verification to multispeaker text-to-speech syn-

- thesis,” in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 4485-4495, 2018.
- [23] O. Nazir, A. Malik, S. Singh and A.-S. K. Pathan, “Multi speaker text-to-speech synthesis using generalized end-to-end loss function,” *Multimedia Tools and Applications*, vol. 83, pp. 64205-64222, 2024.
- [24] Y. Wang *et al.*, “Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis,” in *Proceedings of the 35th International Conference on Machine Learning (PMLR)*, Stockholm, Sweden, 2018.
- [25] L. Loots and T. Niesler, “Automatic conversion between pronunciations of different english accents,” *Speech Communication*, vol. 53, no. 1, pp. 75–84, 2011.
- [26] R. Liu, B. Sisman, G. Gao and H. Li, “Controllable accented text-to-speech synthesis,” *arXiv preprint arXiv:2209.10804*, 2022.
- [27] X. Zhou, M. Zhang, Y. Zhou, Z. Wu and H. Li, “Accented Text-to-Speech Synthesis With Limited Data,” in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 1699-1711, 2024.
- [28] Z. Cai, Y. Yang and M. Li, “Cross-lingual multi-speaker speech synthesis with limited bilingual training data,” *Computer Speech & Language*, vol. 77, p. 101427, 2023.
- [29] A. Łańcucki, “Fastpitch: Parallel Text-to-Speech with Pitch Prediction,” *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, ON, Canada, pp. 6588-6592, 2021.
- [30] J. Kong, J. Kim and J. Bae, “Hifi-gan: Generative adversarial networks for efficient and high-fidelity speech synthesis,” in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, no. 1428, pp. 17022-17033, 2020.
- [31] C. Wutiwawatchai, S. Saychum and A. Rugchatjaroen, “An intensive design of a Thai speech synthesis corpus,” *International Symposium on Natural Language Processing (SNLP 2007)*, 2007.
- [32] J. Li, “Call to Arms,” LibriVox, 2012. [Online]. Available: <https://librivox.org/call-to-arms-by-xun-lu/>.
- [33] K. Ito and L. Johnson, “The LJ Speech Dataset,” 2017. [Online]. Available: <https://keithito.com/LJ-Speech-Dataset/>.



Kittikan Charoenrattana received the B.Sc., and M.Sc. degrees in Computer Science from Khon Kaen University, Thailand. Currently, he is studying in Ph.D. degree in computer science, College of Computing, Khon Kaen University, Thailand. His research interests include speech synthesis and natural language processing.



Pusadee Seresangtakul received the B.Sc. degree in physics from Khon Kaen University, Thailand, the M.Sc. degree in computer science from Chulalongkorn University, Thailand, and the Ph.D. degree in interdisciplinary intelligent systems engineering from the University of the Ryukyus, Japan. She is currently an Assistant Professor with Khon Kaen University. Her research interests include natural language and speech processing, machine learning and intelligent systems, and artificial intelligence.



Pongsathon Janyoi received the B.Sc., M.Sc., and Ph.D. degrees in computer science from Khon Kaen University, Khon Kaen, Thailand, in 2010, 2015, and 2020, respectively. He is currently a Lecturer with the College of Computing, Khon Kaen University. His research interests include speech synthesis, automatic speech recognition, and natural language processing.