EJSIT
European Journal of Science, Innovation and Technology
ISSN: 2786-4936
www.ejsit-journal.com
Volume 6 | Number 4 | 2026

Low-Resourced Text-to-Speech System for Cuoi Cham

Pham Van Dong[1*], Vu Duy Long[2] and Nguyen Duy Hai[3]
[1]Faculty of Information Technology, Hanoi University of Mining and Geology, Vietnam
[2]Faculty of Computer Science, Hanoi National University of Education, Vietnam
[3]Phenikaa University, Hanoi, Vietnam

ABSTRACT

Text-to-Speech (TTS) has become an important research area due to its applications in education, accessibility tools, digital assistants, and language preservation. Although recent end-to-end neural architectures have significantly improved speech synthesis quality, these advances mainly benefit high-resource languages with abundant speech data and linguistic resources. Developing TTS systems for low-resource languages remains challenging because of limited data and insufficient phonological representations. Cuoi Cham (Tho), a minority language spoken in Nghe An province, represents a typical low-resource scenario and remains underrepresented in modern speech technologies. This paper proposes a Cuoi Cham TTS system based on transfer learning using a pre-trained VITS model. To better capture linguistic characteristics, we introduce a linguistically informed phoneme inventory and a phoneme-aware tokenization strategy that explicitly models syllable structure, including onset, nucleus, coda, and tone. Experimental results demonstrate stable and intelligible speech generation under limited data conditions. Objective and subjective evaluations, including Mel Cepstral Distortion (MCD), spectrogram analysis, and Mean Opinion Score (MOS), indicate that combining transfer learning with linguistically informed phoneme representation is a promising approach for low-resource TTS. This work provides an initial framework for Cuoi Cham speech synthesis and contributes to speech technology development for minority languages in Vietnam.

Keywords: Speech synthesis, low-resource language, transfer learning, VITS, Cuoi Cham

INTRODUCTION

In recent years, deep learning has significantly advanced the field of speech processing (Tan et al., 2021), particularly in Text-to-Speech (TTS) systems, where end-to-end architectures have enabled the generation of highly natural and intelligible speech. Despite these achievements, developing high-quality TTS systems still requires a large amount of paired text–audio data, which poses a major challenge for the majority of the world's languages. It is estimated that a vast number of languages remain under-resourced, lacking sufficient digital data and technological support, thereby limiting their accessibility in modern digital environments.

Low-resource languages, especially those with a small number of speakers, face significant barriers in the development of speech technologies such as TTS. Previous studies have shown that for languages with limited datasets, conventional end-to-end approaches often fail to generalize effectively, requiring alternative strategies such as cross-lingual learning or transfer learning. However, existing approaches primarily focus on model adaptation and often rely on generic phoneme representations, which may not adequately capture the linguistic characteristics of target low-resource languages.

* Corresponding Author

Cuoi Cham, a minority language spoken in Nghe An province in central Vietnam, represents a typical low-resource scenario. The language has a limited number of speakers and lacks publicly available large-scale speech datasets, making it difficult to develop data-driven speech synthesis systems. The absence of TTS technology for Cuoi Cham not only restricts access to digital information but also poses challenges for preserving linguistic heritage in the digital era.

To address these challenges, this paper proposes a Text-to-Speech system for Cuoi Cham by leveraging transfer learning and linguistically informed phoneme modeling (Liu et al., 2020). Specifically, we fine-tune a pre-trained VITS model, which integrates acoustic modeling and waveform generation in an end-to-end framework (Kim et al., 2021), using a small dataset of Cuoi Cham speech.

Unlike existing approaches that rely on generic phoneme representations, we introduce a phoneme-aware tokenization strategy tailored to the linguistic structure of Cuoi Cham. This approach explicitly models syllable components, including onset, nucleus, coda, and tone, enabling more accurate alignment between linguistic units and acoustic features. By combining transfer learning with linguistically informed phoneme design, the proposed system provides an initial solution for speech synthesis in low-resource settings.

The main contributions of this work can be summarized as follows:

• We present one of the first Text-to-Speech systems for the Cuoi Cham language, addressing a critical gap in speech technology for underrepresented languages.

• We propose a phoneme-aware tokenization framework that explicitly models syllable structure and supports phonetic representation and alignment in low-resource scenarios.

• We investigate the combination of transfer learning and linguistically informed phoneme design for speech synthesis under limited-data conditions.

• We provide objective and subjective evaluation results using Mel Cepstral Distortion (MCD) and Mean Opinion Score (MOS), reporting the observed performance of the proposed system under low-resource conditions.

The remainder of this paper reviews related work, describes the proposed methodology and experimental setup, presents the results and discussion, and concludes with the limitations of the study.

RELATED WORK

Recent advances in Text-to-Speech (TTS) systems (Tan et al., 2021) have been largely driven by deep learning, particularly end-to-end neural architectures. Early neural TTS models such as Tacotron (Wang et al., 2017) and Tacotron 2 (Shen et al., 2018) demonstrated the ability to generate highly natural speech by mapping text directly to mel-spectrograms, followed by waveform synthesis using neural vocoders (van den Oord et al., 2016). However, these models typically require large-scale paired text–audio datasets, often consisting of tens of hours of speech, to achieve high-quality results. This data requirement poses a significant limitation when applying such models to low-resource languages.

To address data scarcity, a growing body of research has focused on low-resource TTS. One common approach is cross-lingual (Azizah et al., 2020; Joshi & Garera, 2023) or multilingual training, where models are pre-trained on high-resource languages and adapted to target low-resource languages. Previous studies have shown that leveraging data from high-resource languages can significantly improve speech synthesis quality in low-resource scenarios. For example, cross-lingual transfer learning (Azizah et al., 2020; Byambadorj et al., 2021; Hwang et al., 2020) approaches have been proposed to train TTS systems using only a small amount of target-language data, sometimes as little as 30 minutes, while still achieving reasonable synthesis quality (Byambadorj et al., 2021). These findings highlight the

effectiveness of combining transfer learning and data augmentation techniques for low-resource speech synthesis.

Another important direction is the use of transfer learning in TTS systems. Transfer learning enables models to inherit acoustic and linguistic knowledge from pre-trained models, thereby reducing the amount of required training data. Several studies have shown that fine-tuning models trained on high-resource datasets can significantly improve the naturalness and intelligibility of synthesized speech in low-resource settings (Azizah et al., 2020). In particular, hierarchical and multilingual (Casanova et al., 2022) transfer learning frameworks have demonstrated promising results by progressively adapting models from high-resource to low-resource languages (Azizah et al., 2020).

In addition to model adaptation, phoneme representation (Liu et al., 2020) plays a crucial role in low-resource TTS. Since end-to-end models rely heavily on input representations, designing appropriate phoneme inventories or phonetic representations has been explored to improve pronunciation accuracy and generalization. Previous studies have shown that phoneme-aware or linguistically informed representations can improve performance, especially in cross-lingual scenarios and tonal languages (Liu et al., 2020).

More recently, advanced architectures such as VITS have been proposed to unify acoustic modeling and waveform generation into a single end-to-end framework, improving both synthesis quality and inference efficiency (Kim et al., 2021). Furthermore, modern variants such as YourTTS demonstrate that high-quality speech synthesis can be achieved in low-resource and zero-shot scenarios, further highlighting the effectiveness of transfer learning-based approaches (Casanova et al., 2022).

Despite these advances, the development of TTS systems for many minority languages remains limited. In particular, there is a lack of research on Cuoi Cham, a low-resource language with unique phonological characteristics and limited available data. Existing approaches do not fully address the challenges of phoneme representation (Liu et al., 2020) and data scarcity specific to this language.

In this work, we build upon prior research in low-resource TTS and transfer learning by proposing a phoneme-aware TTS system tailored to Cuoi Cham. Our approach combines transfer learning with a linguistically informed phoneme design and a rule-based tokenization pipeline, aiming to improve speech quality under limited data conditions.

METHODOLOGY

In this section, we describe the proposed Text-to-Speech (TTS) system for Cuoi Cham, including the phoneme design, tokenization process, model architecture, and training strategy.

System Overview

The proposed system follows an end-to-end TTS framework built upon the VITS architecture, which integrates acoustic modeling and waveform generation into a unified model. Given an input phoneme sequence, the system directly generates the corresponding speech waveform without requiring a separate vocoder. This design simplifies the overall pipeline and improves synthesis efficiency.

The processing pipeline consists of three main stages. First, input text is converted into phoneme sequences based on a predefined phoneme inventory. Next, the phoneme sequence is transformed into a structured token representation using a phoneme-aware tokenization method. Finally, the tokenized sequence is fed into the VITS model to generate the output waveform. This design enables the system to capture both linguistic and acoustic characteristics of the target language.

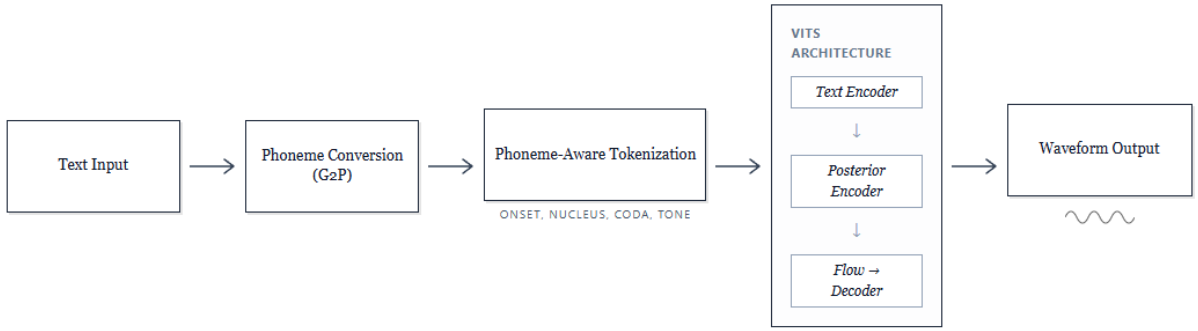

Figure 1: Overview of the proposed TTS system pipeline, including phoneme conversion, phoneme-aware tokenization, and waveform generation using the VITS model.

Phoneme Inventory Design

Because Cuoi Cham does not have a standardized computational phoneme inventory, we derive the inventory from the phoneme annotations available in the Pangloss Collection (Michailovsky et al., 2014). The set includes consonants, vowels, and tone markers and distinguishes onset and coda positions within syllables.

To improve phonetic modeling, we explicitly distinguish between onset and coda consonants by assigning separate representations to final consonants using a suffix-based notation. This distinction allows the model to differentiate phonemes that share the same symbol but differ in phonological roles, thereby improving pronunciation accuracy. The phoneme inventory is designed to maintain compatibility with the annotated dataset while preserving the phonological characteristics of the language.

Manner of articulation \ Position of articulation			Bilabial	Coronal			Glottal
				Alveolar	Palatal	Velar	
Plosive	*obstruent, voiceless*	*unaspirated*	p	t	c	k	ʔ
		aspirated	pʰ	tʰ		kʰ	
	obstruent, voiced		b	d			
	sonorant, nasal		m	n	ɲ	ŋ	
Fricative	*obstruent, voiceless*			s			h
	obstruent, voiced		v	z		ɣ	
	sonorant, lateral			l			
trill				r			

(a) Consonant Onsets Table

Manner of Articulation \ Position of Articulation	Bilabial	Tongue	
		Apical	Velar
plosive, voiceless	p	t	k
plosive, nasal	m	n	ŋ
fricative, latteral		l	
semi vowels	u̯	i̯	

(c) Coda Consonants Table

Mouth Opening \ Line		Front	Back unrounded	Back rounded
Monophthong	high	i	ɯ	u
	close-mid	e	ɤ/ɤ̆	o
	open-mid	ɛ/ɛ̆	ʌ/ʌ̆	ɔ/ɔ̆
	low		ɑ/ɑ̆	ɒ/ɒ̆
Diphthong		ie	ɯɤ	uo

(b) Vowel System Table

Figure 2: Phoneme inventory of Cuoi Cham, compiled from the Pangloss annotations (Michailovsky et al., 2014), including (a) consonant onsets, (b) the vowel system, and (c) coda consonants.

Phoneme-Aware Tokenization

To convert phoneme sequences into a representation suitable for neural modeling, we propose a rule-based tokenization approach that explicitly captures the internal structure of syllables. Each syllable is decomposed into its constituent components, including onset, nucleus, coda, and tone.

During tokenization, tone markers are first extracted and normalized into a consistent format. Final consonants are then identified and annotated with a dedicated suffix to distinguish them from onset consonants. In cases where vowel sequences are ambiguous, a greedy

matching strategy is applied to ensure consistent segmentation. As a result, each syllable is represented as a sequence of structured tokens that reflect its phonological composition.

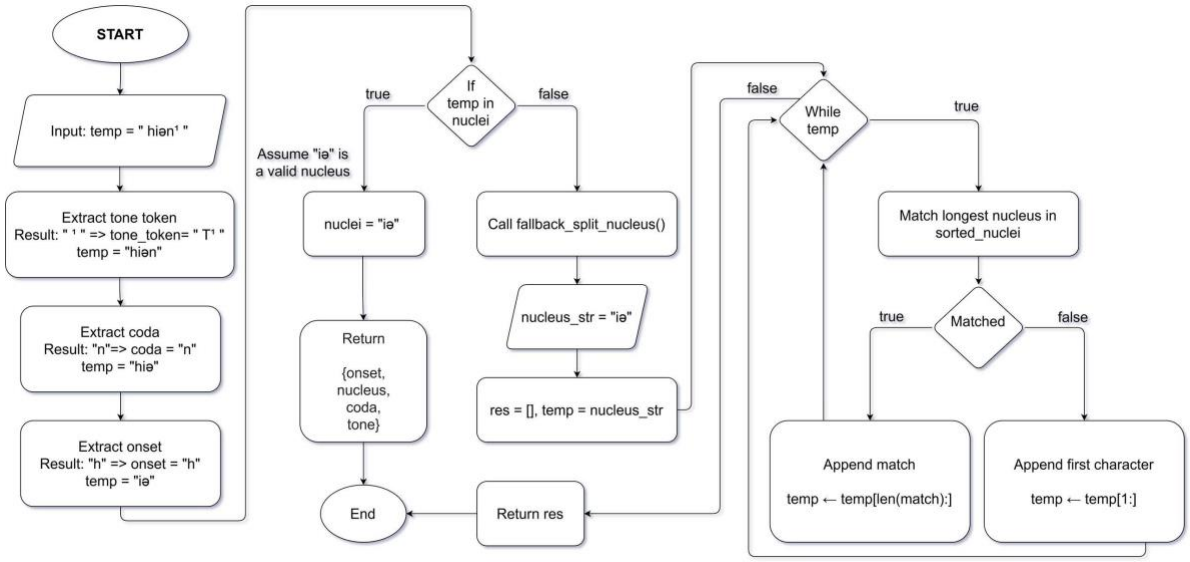

Figure 3: Flowchart of the proposed phoneme-aware tokenization algorithm.

This phoneme-aware representation enables the model to learn more precise alignments between linguistic units and acoustic features, which is particularly important in low-resource scenarios where data is limited.

Syllable representation = [onset, nucleus, coda, tone]

Data Preprocessing

Given the limited and potentially noisy nature of the dataset, a preprocessing pipeline is applied to improve data quality prior to training. Audio samples with excessive background noise or distortion are removed to ensure consistency. The remaining audio data is resampled to 22.05 kHz and normalized.

To maintain speaker consistency, only recordings from a single speaker are selected. In addition, phoneme annotations are normalized by removing ambiguous or inconsistent representations. These preprocessing steps aim to reduce noise in both the acoustic and textual domains, thereby facilitating more stable model training.

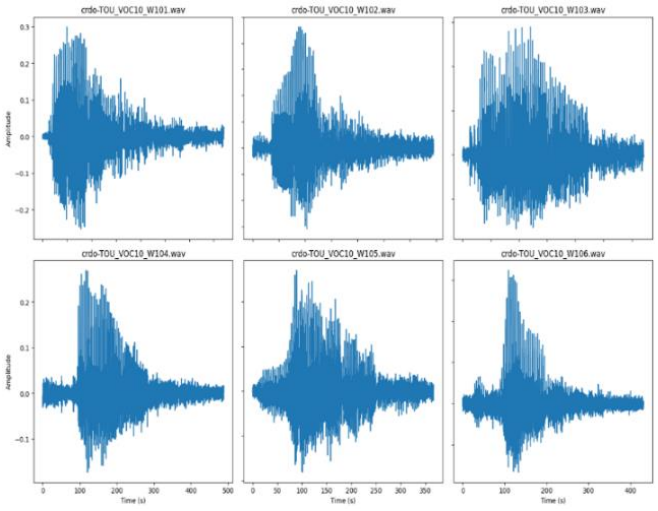

Figure 4: Visualization of waveform samples in the dataset after preprocessing.

Model Architecture

The backbone of the proposed system is the VITS architecture, which combines variational inference, normalizing flows, and adversarial training into a unified framework. Unlike traditional TTS systems that rely on separate acoustic models and vocoders, VITS directly generates waveform signals from input representations.

The model consists of multiple components, including a text encoder that processes phoneme sequences, a posterior encoder that encodes acoustic features, and a decoder that generates the waveform. A stochastic duration predictor is also incorporated to model the temporal alignment between phoneme sequences and speech signals. This architecture allows the system to jointly model the linguistic, acoustic, and temporal aspects of speech.

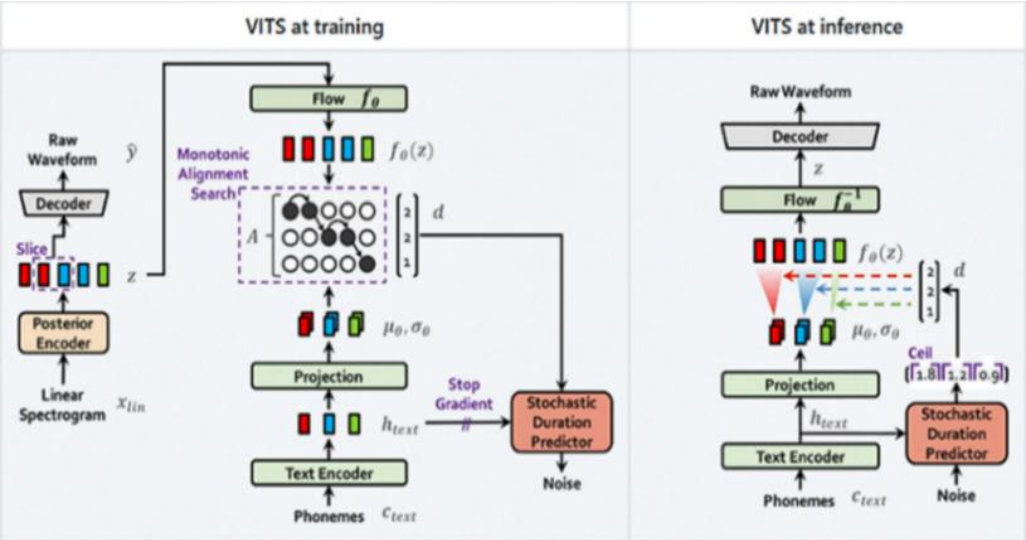

Figure 5: Architecture of the VITS-based text-to-speech model used in this study (adapted from (Kim et al., 2021)).

Transfer Learning Strategy

To address the scarcity of training data, we adopt a transfer learning approach by fine-tuning a pre-trained VITS model trained on a high-resource language. The pre-trained model provides a strong initialization by capturing general acoustic patterns and speech characteristics.

During adaptation, the phoneme embedding layer is modified to accommodate the Cuoi Cham phoneme inventory. The model is then fine-tuned using the target dataset, allowing it to adapt to the phonological and acoustic properties of the language. This strategy is intended to reduce the amount of target-language data required for training and to support stable convergence.

Training Configuration

The model is trained using the preprocessed dataset with a phoneme-based input representation. Optimization is performed using standard gradient-based methods, with training progress monitored through loss functions defined on both acoustic reconstruction and duration prediction.

The training process aims to achieve stable alignment between phoneme sequences and generated speech while maintaining high-quality waveform synthesis. Transfer learning and structured phoneme representation (Liu et al., 2020) are used to improve robustness under low-resource conditions.

EXPERIMENTS

Dataset

The dataset used in this study was collected from the Pangloss Collection (Michailovsky et al., 2014) and consists of approximately 4,382 raw audio samples with corresponding phoneme annotations. After preprocessing and speaker filtering, the dataset was reduced to approximately 2,180 high-quality utterances from a single speaker to ensure consistency in pronunciation and acoustic characteristics.

All audio files were resampled from the original sampling rate of 44.1 kHz to 22.05 kHz to reduce computational complexity while preserving sufficient audio quality for speech synthesis.

The dataset was further processed to remove noisy samples, clipping artifacts, and structurally inconsistent phoneme annotations. These preprocessing steps ensured that the final dataset is clean and suitable for training a neural TTS model in a low-resource setting.

Experimental Setup

The proposed system is implemented using the VITS architecture (Kim et al., 2021) and trained with transfer learning (Azizah et al., 2020; Byambadorj et al., 2021). A VITS model pre-trained on a high-resource English dataset is used to initialize the model parameters, after which the model is fine-tuned on the Cuoi Cham dataset.

The system employs a phoneme-aware tokenization strategy as described in Section 3, allowing the model to capture phonological structures more effectively. During training, optimization is performed with monitoring of both training and validation losses to ensure stable convergence. The overall goal is to learn a mapping from phoneme sequences to raw waveform signals under limited data conditions.

Evaluation Metrics

The performance of the proposed system is evaluated using objective and subjective measures. Mel Cepstral Distortion (MCD) (Kubichek, 1993) is used as the primary objective metric, while Mean Opinion Score (MOS) is adopted for subjective evaluation of perceptual speech quality and naturalness. Spectrogram analysis is additionally conducted for qualitative assessment.

RESULTS AND DISCUSSION

Training Behavior

The training curves show an overall decrease in the reported loss components, indicating that the model learns the relationship between phoneme sequences and acoustic features. However, the validation behavior is not uniformly monotonic, so the curves should not be interpreted as conclusive evidence that overfitting is absent.

The observed gap between training and validation losses remains limited for part of the training process, but some divergence is visible. Additional experiments with early stopping, repeated runs, and a held-out test protocol are needed to assess generalization more rigorously.

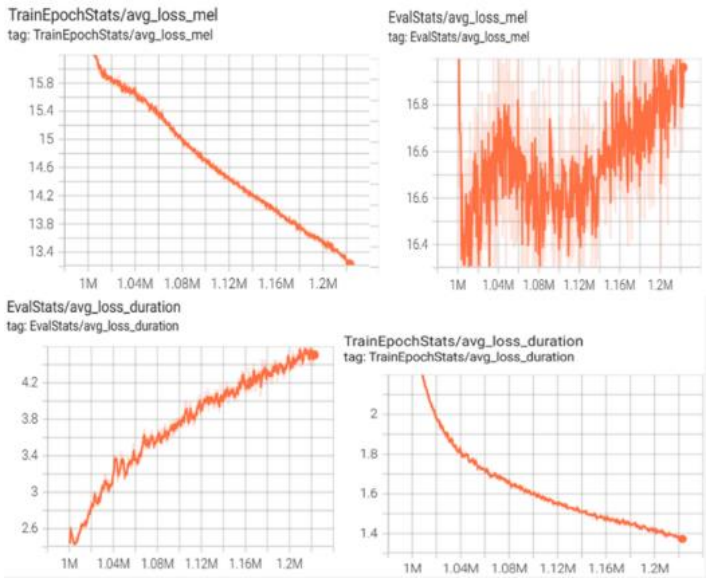

Figure 6: Training and validation loss curves; the validation trends should be interpreted cautiously because some divergence is visible.

Spectrogram Analysis

The comparison between ground truth and synthesized spectrograms reveals that the proposed system is capable of capturing the main acoustic characteristics of Cuoi Cham speech. The generated spectrograms exhibit clear harmonic structures and consistent temporal alignment with the original signals.

Although the overall structure is well preserved, some discrepancies can be observed in high-frequency regions and during transitions between certain phonemes. These differences are likely caused by the limited size and diversity of the training data, which restricts the model's ability to capture all phonetic variations.

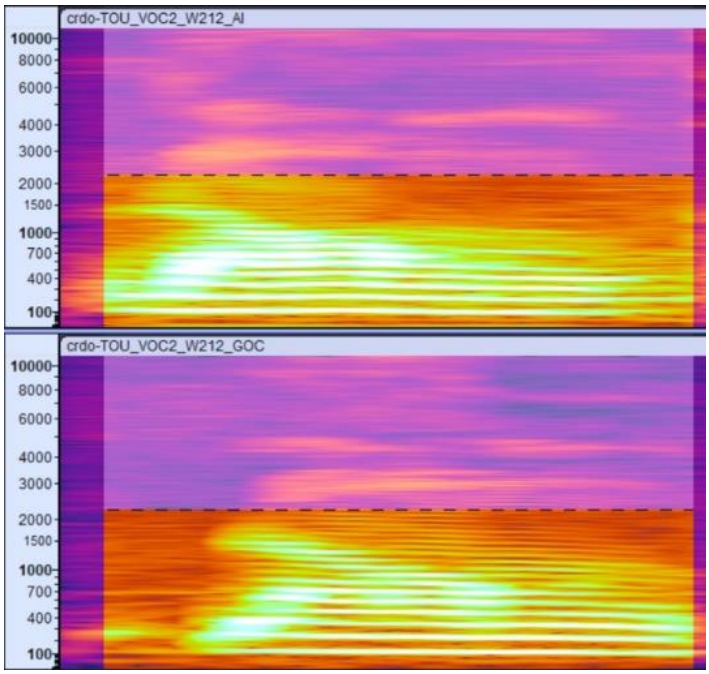

Figure 7: Comparison of generated and ground-truth spectrograms for synthesized speech.

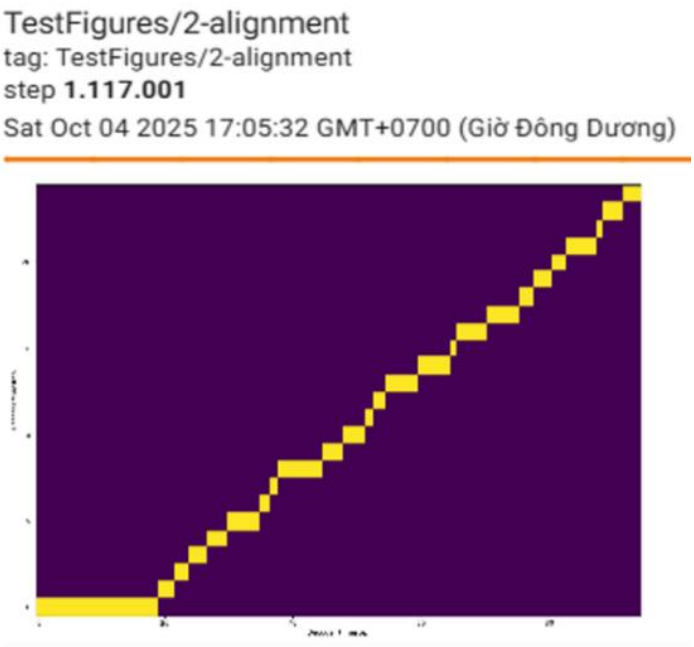

Figure 8: Alignment between phoneme tokens and acoustic frames learned by the model.

Objective Evaluation

To quantitatively evaluate the quality of synthesized speech, we use Mel Cepstral Distortion (MCD) (Kubichek, 1993) with Dynamic Time Warping (DTW) alignment. MCD measures the spectral distance between synthesized speech and ground-truth recordings, where lower values indicate higher similarity.

The proposed system achieves an average MCD of 5.5053 dB on the test set, with a standard deviation of 0.9992 dB. The minimum and maximum MCD values are 4.2205 dB and 9.4901 dB, respectively.

Table 1: Detailed MCD statistics

Metric	Value (dB)
Mean MCD (↓)	5.5053
Standard Deviation	0.9992
Minimum MCD	4.2205
Maximum MCD	9.4901

The results indicate that the synthesized speech is acoustically close to the ground truth within the evaluated test set. The relatively low standard deviation suggests reasonably consistent performance across different utterances.

However, the relatively high maximum value indicates that certain samples remain challenging, likely due to limited phonetic coverage and variability in the dataset.

Because no explicit baseline or ablation model was trained, the MCD result should be interpreted as an initial within-system evaluation rather than evidence that the proposed method outperforms other low-resource TTS systems.

Subjective Evaluation (MOS)

In addition to objective evaluation, a subjective listening test was conducted using Mean Opinion Score (MOS), which is widely used for assessing the perceptual quality and naturalness of synthesized speech in Text-to-Speech systems. MOS evaluates speech quality

based on human judgments using a five-point rating scale ranging from 1 (poor) to 5 (excellent).

A total of 14 participants with diverse demographic backgrounds participated in the evaluation process. The listening test consisted of 15 natural speech recordings and 15 synthesized speech samples generated by the proposed system. Each participant listened to all utterances and assigned a score according to the perceived overall speech quality. To reduce variability, all participants received the same evaluation instructions and rating criteria. Previous studies indicate that experimental conditions and evaluator instructions may significantly influence MOS outcomes.

For each speech sample, the MOS value was calculated as the arithmetic mean of listener ratings:

$$\boldsymbol{MOS} = \frac{\sum_{i=1}^{N} R_i}{N}$$

where R_i denotes the rating score assigned by participant i, and N represents the total number of evaluators.

Table 2: Subjective evaluation results

System	MOS
Natural Speech	4.56
Proposed TTS	4.18

Effect of Transfer Learning

Transfer learning provides an initialization based on acoustic knowledge learned from a high-resource language. The observed training behavior is consistent with stable adaptation; however, a from-scratch baseline is required to quantify the specific contribution of transfer learning.

Effect of Phoneme Representation

The proposed phoneme-aware tokenization explicitly models syllable structure, including onset, nucleus, coda, and tone information. The alignment results suggest that this representation is suitable for Cuoi Cham, although an ablation without phoneme-aware tokenization is required to quantify its effect.

Limitations and Discussion

Despite the promising results, several limitations remain. The relatively small dataset limits phonetic diversity and may affect model generalization. Certain phoneme combinations and tonal variations remain insufficiently represented.

In addition, the subjective MOS evaluation involved only 14 participants. The current report does not provide listener-level variability or 95% confidence intervals, so the MOS comparison should be interpreted cautiously. Future evaluations should report confidence intervals and use a larger, more diverse listener group.

Overall, the results indicate that combining transfer learning with linguistically informed phoneme design is a promising initial strategy for low-resource TTS. Baseline and ablation experiments are needed before the individual contribution of each component can be established.

REFERENCES

Azizah, K., Adriani, M., & Jatmiko, W. (2020). Hierarchical transfer learning for multilingual, multi-speaker, and style transfer DNN-based TTS on low-resource languages. *IEEE Access, 8,* 179798–179812. https://doi.org/10.1109/ACCESS.2020.3027619.

Byambadorj, Z., Nishimura, R., Ayush, A., Ohta, K., & Kitaoka, N. (2021). Text-to-speech system for low-resource language using cross-lingual transfer learning and data augmentation. *EURASIP Journal on Audio, Speech, and Music Processing, 2021*(1), 42. https://doi.org/10.1186/s13636-021-00225-4.

Casanova, E., Weber, J., Shulby, C. D., Junior, A. C., Gölge, E., & Ponti, M. A. (2022). YourTTS: Towards zero-shot multi-speaker TTS and zero-shot voice conversion for everyone. In *Proceedings of the 39th International Conference on Machine Learning* (pp. 2709–2720). PMLR. https://proceedings.mlr.press/v162/casanova22a.html.

Hwang, M.-J., Yamamoto, R., Song, E., & Kim, J.-M. (2020). TTS-by-TTS: TTS-driven data augmentation for fast and high-quality speech synthesis. arXiv:2010.13421. https://doi.org/10.48550/arXiv.2010.13421.

Joshi, R., & Garera, N. (2023). Rapid speaker adaptation in low resource text to speech systems using synthetic data and transfer learning. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation* (pp. 267–273).

Kim, J., Kong, J., & Son, J. (2021). Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *Proceedings of the 38th International Conference on Machine Learning* (Vol. 139, pp. 5530–5540).

Kubichek, R. (1993). Mel-cepstral distance measure for objective speech quality assessment. In *Proceedings of the IEEE Pacific Rim Conference on Communications, Computers and Signal Processing* (Vol. 1, pp. 125–128). https://doi.org/10.1109/PACRIM.1993.407206.

Liu, R., Wen, X., Lu, C., & Chen, X. (2020). Tone learning in low-resource bilingual TTS. In *Interspeech 2020* (pp. 2952–2956). ISCA. https://doi.org/10.21437/Interspeech.2020-2180.

Michailovsky, B., Mazaudon, M., Michaud, A., Guillaume, S., François, A., & Adamou, E. (2014). Documenting and researching endangered languages: The Pangloss Collection. *Language Documentation & Conservation, 8,* 119–135.

Shen, J., et al. (2018). Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions. arXiv:1712.05884. https://doi.org/10.48550/arXiv.1712.05884.

Tan, X., Qin, T., Soong, F., & Liu, T.-Y. (2021). A survey on neural speech synthesis. arXiv:2106.15561. https://doi.org/10.48550/arXiv.2106.15561.

van den Oord, A., et al. (2016). WaveNet: A generative model for raw audio. arXiv:1609.03499. https://doi.org/10.48550/arXiv.1609.03499.

Wang, Y., et al. (2017). Tacotron: Towards end-to-end speech synthesis. arXiv:1703.10135. https://doi.org/10.48550/arXiv.1703.10135.