Lao-English Code-Switched Speech Synthesis Via Neural Codec Language Modeling

Yaping Liu[1,2], Linqin Wang[1,2], Shengxiang Gao[1,2,*], Zhengtao Yu[1,2], Ling Dong[1,2], Tian Tian[1,2]
[1]Faculty of Information Engineering and Automation
Kunming University of Science and Technology, Kunming, China
[2]Yunnan Key Laboratory of Artificial Intelligence, Kunming, China
20232204183@stu.kust.edu.cn, 2424172505@qq.com
gaoshengxiang.yn@foxmail.com, ztyu@hotmail.com
46761956@qq.com, 20222104107@stu.kust.edu.cn

Abstract

This paper addresses the challenges of data scarcity and limited speaker resources in Lao-English code-switched speech synthesis. We propose a neural encoder-decoder-based method for mixed-lingual speech synthesis. The method first extracts phoneme-level speech representations and employs a dot-product attention mechanism to map Lao and English phonemes into a shared latent space, thereby enhancing the model's capability to represent cross-lingual phonetic information. In addition, language ID embedding module is extended to explicitly indicate the language of each input token, helping the model distinguish and adapt to language-specific pronunciation characteristics. Experiments are conducted on the open-source English dataset LibriTTS and a proprietary Lao speech corpus. Both subjective evaluations (MOS, AB preference tests) and objective metrics (RMSE) demonstrate that the proposed approach significantly outperforms the baseline VALL-E X model in terms of naturalness and language-switching fluency. Furthermore, ablation studies confirm that both the shared phoneme latent space and the language ID module play critical roles in improving synthesis quality. This approach offers a novel solution for integrating low-resource languages into mixed-lingual speech synthesis.

Keywords: Code-switched speech synthesis , Lao-English language , Neural codec language model

1 Introduction

Code-switched text-to-speech synthesis (Code-switching TTS) aims to enable a system to generate speech in two different languages within a single utterance while maintaining a consistent speaker identity (Zhou et al., 2020; Chai et al., 2021; NAKAYAMA et al., 2021). Existing code-switching TTS approaches can be broadly classified into two categories. The first approach involves recording multilingual speech data from the same speaker and training separate monolingual TTS models for each language (Zhang et al., 2019; Liu and Mak, 2020). During inference, the input text is first segmented by language, and each segment is synthesized using the corresponding monolingual model. The resulting audio segments are then concatenated to produce the final utterance. While this method supports language switching, it requires maintaining multiple models, leading to increased computational cost and system complexity. The second approach leverages multilingual training data to train a unified multilingual TTS model (Yang and He, 2022; Cai et al., 2023). This strategy is generally more efficient in terms of training and deployment. However, it still faces challenges such as poor alignment between speech and text, variability in data quality, and the limited availability of high-quality multilingual corpora. Both methods fundamentally assume the availability of multilingual recordings from a single speaker, which is often impractical due to the high cost of data collection and the scarcity of speakers fluent in multiple languages.

This issue is particularly pronounced in the context of Lao, a low-resource language with a limited speaker population. Due to socio-economic conditions, geographic location, and historical influences,

* Corresponding Author: Shengxiang Gao. gaoshengxiang.yn@foxmail.com

Proceedings of the 24th China National Conference on Computational Linguistics, pages 1067–1077, Jinan, China, August 11–14, 2025.

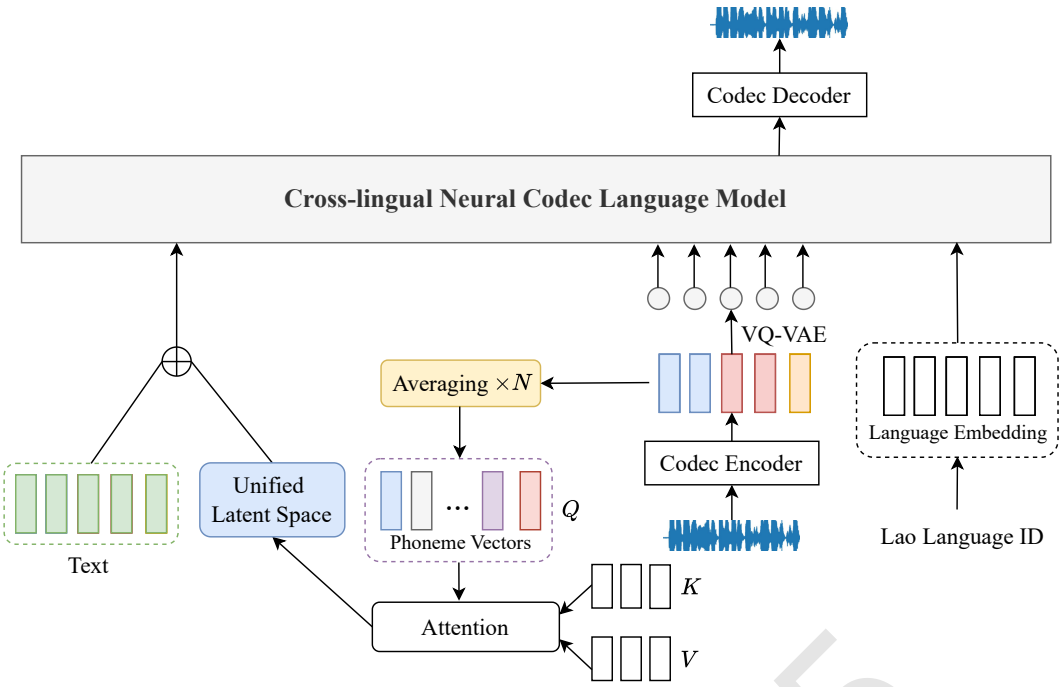

Figure 1: The main architecture of model.

Lao texts frequently contain embedded English words—such as named entities, brand names, or loanwords—making Lao-English code-switching a common phenomenon. However, acquiring bilingual or code-switched recordings from the same speaker is prohibitively expensive for Lao.

To address this challenge, we propose a Lao-English code-switched TTS method based on the large-scale neural codec language model VALL-E X (Zhang et al., 2023). Specifically, we map phoneme-level representations from both Lao and English into a shared latent space, allowing the model to learn unified cross-lingual phonetic representations. In addition, we introduce language ID embeddings to indicate the language of each input token, enabling the model to correctly identify and handle language-specific pronunciation patterns. This approach significantly improves the naturalness of synthesized speech and enhances the fluency of language transitions, providing an effective and practical solution for applying code-switching TTS to low-resource languages.

- We propose a phoneme-level cross-lingual representation learning method that maps Lao and English phonemes into a unified latent space using dot-product attention, enabling the model to capture language-shared pronunciation features and improve synthesis consistency in code-switched speech.

- We introduce a language ID embedding module to explicitly encode language information, allowing the model to dynamically adapt to language-specific acoustic traits and effectively handle intra-sentence language switching.

- Extensive experiments on the LibriTTS and a proprietary Lao speech dataset demonstrate that our model significantly outperforms the baseline VALL-E X in both naturalness and fluency, as validated by MOS, RMSE, and AB preference tests, with further ablation studies confirming the effectiveness of each proposed component.

2 Prior Knowledge

2.1 International Phonetic Alphabet (IPA)

The International Phonetic Alphabet (IPA) is a phonetic transcription system based on the Latin alphabet, widely used in fields such as linguistics, text-to-speech synthesis (TTS), automatic speech recognition (ASR), and language education. IPA is designed to accurately represent the pronunciation characteristics of spoken language, including phonemes, intonation, tone, and syllabic structure. Its core principle is "one sound, one symbol; one symbol, one sound," meaning that each phoneme corresponds to a unique symbol and each symbol denotes a specific phoneme.

IPA symbols are mainly divided into two categories: letters and diacritics. Specifically, IPA comprises 107 letters to represent consonants and vowels, 31 diacritics to modify these basic sounds, and 19 additional symbols to denote suprasegmental features such as length, pitch, stress, and intonation. For example, the English letter ⟨t⟩ may be transcribed in IPA as the simple [t] or as a more detailed variant such as [t^h] depending on the precise phonetic context.

There are two common notations for IPA transcription: broad transcription, enclosed in slashes (/ /), which captures phonemes in a more abstract manner, and narrow transcription, enclosed in square brackets ([ ]), which reflects finer phonetic details. For instance, the phoneme /t/ may be realized as [t^h] or [t] in different linguistic environments, depending on the contextual and language-specific pronunciation rules.

The advantages of IPA lie in its universality, precision, and standardization, which enable a unified and systematic description of phonetic phenomena across languages. This makes IPA particularly valuable in the processing of low-resource languages, offering an efficient and reliable tool for phonetic annotation and representation.

2.2 Attention Mechanism

The attention mechanism (Vaswani et al., 2017) is a technique in deep learning inspired by human visual and cognitive systems. It enables neural networks to dynamically focus on the most relevant parts of the input when processing information. Widely adopted in architectures such as the Transformer, attention enhances model performance and generalization by allowing the network to selectively emphasize important features and suppress irrelevant ones. Its primary goal is to extract the most informative content from large-scale input data, especially in sequence modeling tasks such as text, speech, and image sequences.

At its core, the attention mechanism assigns varying weights to different parts of the input based on their relevance to the current task. It achieves this by computing a weighted sum of values, allowing the model to concentrate on the most salient information while ignoring less important content.

The standard attention mechanism operates through a triplet of components: Query (Q), Key (K), and Value (V). The query represents the focus of attention, the keys denote all possible elements in the input, and the values contain the actual content to be aggregated. The process begins by computing the similarity between the query and each key using the scaled dot-product:

$$S(q, k_i) = \frac{q^T k_i}{\sqrt{d_n}} \tag{1}$$

Here, d_n is the dimensionality of the key vectors and serves as a scaling factor to prevent the dot-product values from becoming excessively large or small.

The similarity scores S are then passed through a Softmax function to produce normalized attention weights W, ensuring that the weights are positive and sum to 1:

$$W = \text{Softmax}(S) = \frac{\exp(S(q, k_j))}{\sum_{m=j}^{1} \exp(S(q, k_j))} \tag{2}$$

Finally, the output is computed as the weighted sum of the value vectors:

Proceedings of the 24th China National Conference on Computational Linguistics, pages 1067–1077, Jinan, China, August 11–14, 2025.

$$\text{output} = \text{Attention}(Q, K, V) = \sum w_i v_i \tag{3}$$

A higher attention score indicates a stronger relevance between the query and the corresponding key, and thus a greater contribution from the associated value to the final output. These dynamic weights allow the model to adaptively select contextually important information, enabling it to generate more accurate and context-aware representations.

2.3 Language Identification

Language Identification (LID) (Handoyo et al., 2024) is a fundamental task in natural language processing and speech processing, aimed at determining the language category of a given text or speech input. It plays a critical role in a variety of applications, including text-to-speech synthesis (TTS), automatic speech recognition (ASR), and machine translation systems.

For textual data, LID typically involves analyzing character patterns, lexical features, and language-specific grammatical structures to classify the input language. For speech data, LID relies on acoustic features such as phoneme distributions, speech rhythm, and intonation patterns. In recent years, the use of pretrained models has significantly improved LID performance, especially for supporting low-resource languages.

3 Method

Recent advances in monolingual text-to-speech (TTS) synthesis based on deep learning models (Sotelo et al., 2017; Ping et al., 2018; Kim et al., 2021; Casanova et al., 2022; Ren et al., 2020; Wang et al., 2017) have achieved impressive results in generating natural and fluent speech. However, code-switched TTS remains more challenging than monolingual synthesis, as the system must handle differing phonetic characteristics and pronunciation habits of two languages within a single model. The baseline model used in this study is VALL-E X, a multilingual extension of Microsoft's VALL-E model (Chen et al., 2025; Zhang et al., 2023). Building upon this foundation, we propose an improved approach that enables Lao-English code-switched speech synthesis by integrating phoneme-level representations across languages. Specifically, Lao and English phoneme embeddings are projected into a shared latent space, and language IDs are introduced to guide the model in identifying the language of each input segment. This design aims to generate natural and fluent speech that seamlessly transitions between Lao and English.

3.1 Model Architecture

A straightforward method for building a code-switched TTS system is to use a bilingual speech corpus recorded by a bilingual speaker. However, finding a speaker who is both proficient in multiple languages and consistent in pronunciation across those languages is extremely difficult. Therefore, this work maps the phoneme-level representations of different languages into a shared latent space, minimizing representation differences and enabling a unified code-switched TTS model, as illustrated in Fig. 1.

The overall framework is built on the structure of the VALL-E X model, which includes an autoregressive Transformer decoder and a non-autoregressive Transformer decoder. To enhance the model's ability to process code-switched input, phoneme representations from Lao and English are first unified in a shared latent space. Attention mechanisms are then applied to extract more informative acoustic features from these representations. Additionally, a language ID module is introduced to condition the phoneme embeddings based on the language of the input, helping guide the speech generation process. In practice, language IDs are embedded into dense vectors and combined with the acoustic token embeddings, enabling the model to generate speech aligned with the characteristics of the target language.

3.2 Cross-Lingual Latent Space for Code-Switched TTS

To enable cross-lingual sharing of phoneme-level representations and improve the naturalness and fluency of synthesized speech, we propose a latent space construction method based on phoneme-aligned speech representations using dot-product attention.

Proceedings of the 24th China National Conference on Computational Linguistics, pages 1067–1077, Jinan, China, August 11–14, 2025.

First, frame-level speech features are extracted from the audio to form phoneme-level vectors, which serve as the queries (Q) in the attention module. For each phoneme-level vector, its time boundaries—start frame s_i and end frame e_i —are located within the corresponding audio. All frame-level features c_j within this range are averaged to compute a temporary phoneme representation r_c , capturing the acoustic profile of the phoneme across frames. Since the same phoneme appears across multiple utterances, the final phoneme-level vector is obtained by averaging all its temporary representations:

$$r_c = \frac{1}{e_i - s_i + 1} \sum_{j=s_i}^{e_i} c_j \tag{4}$$

Here, c_j denotes the frame-level feature at index j, and s_i, e_i are the phoneme's start and end frame indices.

The outputs of a pretrained neural encoder are used as the keys (K) and values (V) in the attention mechanism. Let $P = \{r_{c1}, r_{c2}, \ldots, r_{cn}\}$ denote the phoneme-level vectors. For each query vector r_{ci}, we compute attention scores using scaled dot-product attention:

$$S = \frac{r_{ci} \cdot e_i^T}{\sqrt{m}} \tag{5}$$

The attention scores are normalized using the Softmax function to obtain attention weights w_j , which are then used to compute a weighted sum over the value vectors V, producing the final phoneme-level latent representation H:

$$H = \text{softmax}(S)V = w_i V \tag{6}$$

The resulting latent space H is fused with the phoneme sequence and serves as part of the input to the cross-lingual neural codec language model. The training strategy follows the autoregressive and non-autoregressive schemes proposed in (Zhang et al., 2023).

3.3 Integration of Lao Language ID

As the baseline model already supports multilingual speech synthesis using language IDs and is trained on English data, our work focuses on introducing Lao-specific language identifiers. For monolingual inputs, the corresponding language ID is retrieved from a predefined language dictionary and converted into a tensor, which is then added to the acoustic token embeddings. For multilingual sequences, each language is assigned a distinct ID, and all IDs are similarly embedded and added to the acoustic embeddings.

These language IDs are projected into a high-dimensional embedding space via an embedding layer. The resulting language embeddings encode phonetic characteristics of each language and guide the model to distinguish among them, enabling generation of speech with accurate linguistic and acoustic traits. This mechanism allows the model to dynamically switch between languages, facilitating natural and seamless code-switched speech synthesis.

4 Experiments and Analysis

4.1 Datasets

The proposed model is trained on both the open-source English dataset LibriTTS (Zen et al., 2019) and a proprietary Lao speech dataset.

LibriTTS is a multi-speaker English corpus containing approximately 585 hours of read speech. The audio files are sourced from LibriVox, and the corresponding texts are derived from Project Gutenberg. The audio is sampled at 24kHz and segmented at sentence-level pauses. The dataset includes both raw and normalized transcripts, with utterances containing obvious background noise removed. LibriTTS is further divided into multiple subsets, as detailed in Table 1.

Table 1: Detailed Statistics of the LibriTTS Dataset.

Subset	Duration (hours)	Male Speakers	Female Speakers	Total Speakers
dev-clean	8.97	20	20	40
test-clean	8.56	19	20	39
dev-other	6.43	16	17	33
test-other	6.49	17	16	33
train-clean-100	53.78	123	124	247
train-clean-360	191.29	430	474	904
train-other-500	310.08	560	600	1160
Total	585.80	1185	1271	2456

The proprietary Lao dataset consists of audiobook recordings produced by native Lao speakers using personal computers or mobile devices. The recordings are mono-channel, sampled at 24kHz, and amount to a total of 15,965 audio clips from 380 speakers, with a total duration of approximately 100 hours. All recordings were manually reviewed and corrected by university students in Laos collaborating with our research team. This dataset is a parallel corpus, with each audio clip aligned with a corresponding text. The corpus is split into training, validation, and test sets at a ratio of 8:1:1. Each clip in the training and validation sets lasts between 8 to 20 seconds. The test set consists of manually segmented speech clips, each approximately 3 seconds in duration. Due to the manual segmentation process, there may be slight variations in length across test samples, typically ranging from 0.1 to 0.3 seconds.

4.2 Evaluation Metrics

To evaluate the performance of the code-switched speech synthesis system, both objective and subjective evaluation metrics are employed. The objective metric used is Root Mean Square Error (RMSE), consistent with the metric introduced in Chapter 3.

For subjective evaluation, A/B preference tests and Mean Opinion Score (MOS) are adopted, and an additional metric—Comparative Mean Opinion Score (CMOS)—is introduced. CMOS is a relative subjective evaluation method in which listeners are asked to compare two speech samples and assign a score based on preference. Unlike MOS, which rates a single sample independently, CMOS uses a 7-point scale ranging from -3 to +3: +3 means A is much better than B, +2 means A is better, +1 means A is slightly better, 0 indicates no preference, and negative values indicate B is better to corresponding degrees.

Subjective evaluations were conducted with 10 Lao-speaking and 10 English-speaking participants. Each test included five audio samples per model, and the evaluation was performed under a 95% confidence level. Due to the lack of real recordings of Lao speakers speaking English and English speakers speaking Lao, objective evaluation could only be performed on synthesized speech from speakers matching the language (i.e., Lao for Lao and English for English).

4.3 Experimental Results and Analysis

To verify the effectiveness of the proposed method, we conducted comparative and ablation experiments across different speakers and languages. Both the baseline model and the proposed model were evaluated using multiple objective and subjective metrics.

4.3.1 Comparative Experiments

To demonstrate the effectiveness of the proposed method in Lao-English code-switched speech synthesis, a series of comparative experiments were conducted. As shown in Table 2, we compare the performance

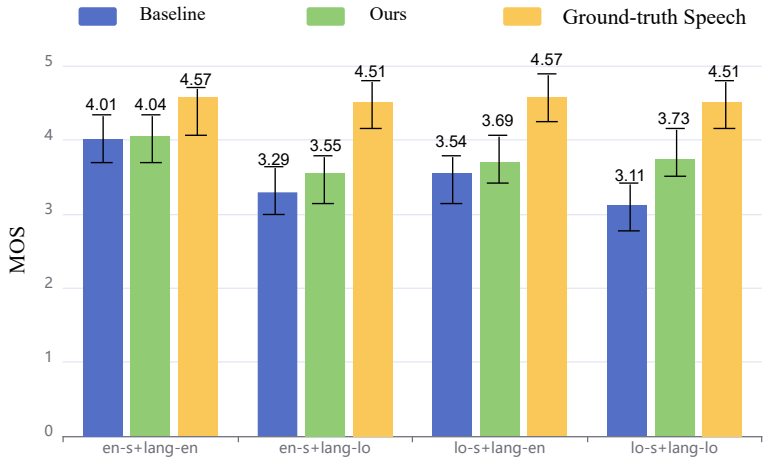

Figure 2: MOS Evaluation Results of Different Models on Lao-English Code-Switching Speech Synthesis.

Table 2: CMOS and RMSE Evaluation Results Across Different Languages.

Model	Text	Speaker	CMOS	RMSE
VALL-E-X	English	en-speaker	-0.11	55.43
		lo-speaker	-0.97	-
	Lao	en-speaker	-0.78	-
		lo-speaker	-1.21	56.04
Ours	English	en-speaker	-0.09	52.25
		lo-speaker	-0.27	-
	Lao	en-speaker	-0.30	-
		lo-speaker	-0.10	54.27

of the baseline and improved models in synthesizing both Lao and English speech. In the table, "en-speaker" and "lo-speaker" refer to speakers whose native languages are English and Lao, respectively.

As observed from Table 2, since the baseline model is a multilingual model pretrained on English data, both the baseline and the improved models perform better when synthesizing English speech using English speakers, with RMSE scores of 55.43 and 52.25, respectively.

For cross-lingual synthesis, the improved model outperformed the baseline by +0.7 CMOS points when generating English speech using Lao speakers, and by +0.48 CMOS points when generating Lao speech using English speakers. These results indicate that the proposed method significantly improves synthesis quality in cross-lingual settings, especially for low-resource scenarios.

The notable improvement in synthesizing English speech from Lao speakers can be attributed to the incorporation of a shared phoneme-level latent space, which enhances the model's ability to model Lao phoneme pronunciation despite the relatively small amount of Lao training data. This further confirms that aligning phoneme representations across languages helps the model better capture the characteristics of each language, thus improving the naturalness and fluency of code-switched speech synthesis.

In addition to the previously mentioned evaluations, MOS (Mean Opinion Score) was also used to

Table 3: CMOS and RMSE evaluation results for different VITS training strategies.

Model	Text	Speaker	CMOS ↓	RMSE ↓
VITS (Separate)	English	en-speaker	-0.42	58.10
		lo-speaker	-1.30	–
	Lao	en-speaker	-0.82	–
		lo-speaker	-1.35	58.70
VITS (Joint)	English	en-speaker	-0.72	59.30
		lo-speaker	-1.45	–
	Lao	en-speaker	-0.88	–
		lo-speaker	-1.50	59.80

assess the performance of different input text representations. The final MOS scores were obtained by averaging listener ratings, and the results are presented in Fig. 2.

In the figure, "en-s" and "lo-s" denote speakers whose native languages are English and Lao, respectively, while "lang-en" and "lang-lo" refer to English and Lao input texts. The evaluation results on Lao-English code-switched speech synthesis show that the baseline model generally underperforms compared to the proposed model. Except for the case where the English speaker synthesized English speech, the MOS scores of the baseline model across other speaker-text combinations were relatively low—specifically, 3.29, 3.54, and 3.11.

This performance degradation is attributed to the phonemic mismatches between Lao and English. Without a unified latent space for Lao-English phoneme representations, the model struggles to accurately learn phoneme-level pronunciation patterns across languages. This limitation becomes particularly apparent in the synthesis of Lao phonemes during language switching, leading to lower naturalness and fluency in the generated speech.

By incorporating a shared phoneme-level latent space and embedding Lao language IDs, the proposed model facilitates more effective learning of cross-lingual pronunciation knowledge. As a result, the model demonstrates improved audio quality in both monolingual and code-switched speech synthesis tasks.

4.3.2 Comparison of Modeling Strategies in TTS Architecture

To explore mixed-language speech modeling strategies under the condition of lacking real code-switched audio data, we designed two typical modeling approaches based on the mainstream VITS framework. These approaches were trained and evaluated on the same monolingual Lao and English corpora. VITS (Separate): Two independent VITS models were trained on English and Lao data separately. During synthesis, the mixed-language text was segmented by language, and the corresponding speech segments were generated by each model before being concatenated into the complete audio; VITS (Joint): The monolingual Lao and English corpora were merged to directly train a unified VITS model.

We evaluated these methods using the same CMOS and RMSE metrics as the main model. The experimental results are shown in Table 3. The experimental results show that the VITS (Separate) model tends to produce timbre discontinuities and prosodic inconsistencies during language switching, which negatively affect the overall naturalness of synthesized speech. Although the VITS (Joint) model can handle mixed-language input, it lacks an explicit language distinction mechanism, often leading to phoneme confusion and unstable control during code-switching. These comparisons further validate the effectiveness and necessity of the proposed shared latent space and language ID embedding in improving the naturalness and cross-lingual consistency of mixed-lingual speech synthesis.

Proceedings of the 24th China National Conference on Computational Linguistics, pages 1067–1077, Jinan, China, August 11–14, 2025.

Table 4: Evaluation Results of Ablation Studies by Removing Different Modules (where "w/o" indicates "without").

Method	MOS	CMOS (vs loGT)
w/o PC	3.59 (±0.10)	-1.26
w/o LID	3.63 (±0.06)	-0.54
Our Model	3.69 (±0.14)	-0.31

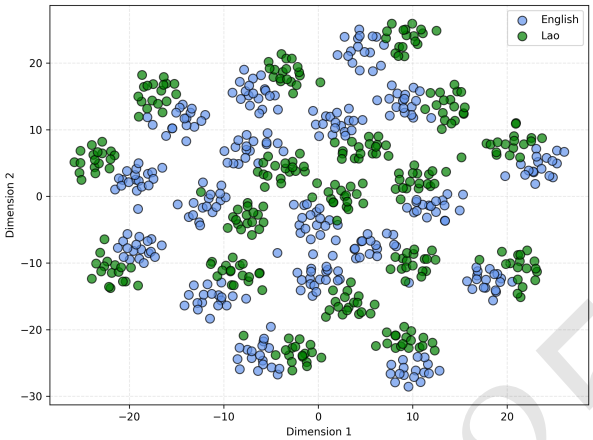

Figure 3: t-SNE of Phoneme-Level Latent Space.

4.3.3 t-SNE Analysis of the Shared Latent Space

To verify the cross-lingual modeling capability of the shared phoneme latent space, we conduct a t-SNE visualization of phoneme embeddings from different languages, as shown in Fig. 3. It can be observed that Lao and English phonemes form a mixed yet structured distribution in the 2D space, which further confirms the effectiveness of our proposed method in achieving unified modeling for code-switched TTS.

4.3.4 Ablation Study

To validate the effectiveness of the proposed modules, we conducted ablation experiments to analyze the contribution of each component to the overall model performance. For the MOS evaluation, the assessment was based on synthesized audio from code-switched inputs. For the CMOS evaluation, due to the absence of ground-truth Lao-English code-switched speech, we compared synthesized Lao speech with reference Lao utterances.

The ablated components include two key modules introduced in this work: the Phoneme-level shared latent space (PC) and the Language Identification module (LID). Each was removed independently to observe its impact on performance. The results of the ablation study are shown in Table 4.

4.3.5 Ablation Study Analysis

As shown in Table 4, the performance of models using only the Phoneme-level latent space (PC) or only the Lao language ID is inferior in terms of naturalness and fluency compared to the proposed full model. When the PC module is removed, the MOS score drops from 3.69 to 3.59, a decrease of 0.10. The CMOS score reaches -1.2, which is lower than -1, indicating that the synthesized audio is perceptibly worse than the reference audio. In contrast, the CMOS score of the proposed model is -0.31, suggesting that the unified phoneme space significantly contributes to improved speech quality.

This improvement can be attributed to the phoneme-level latent space's ability to better model the phonetic similarity between Lao and English, enabling shared representations and smoother transitions

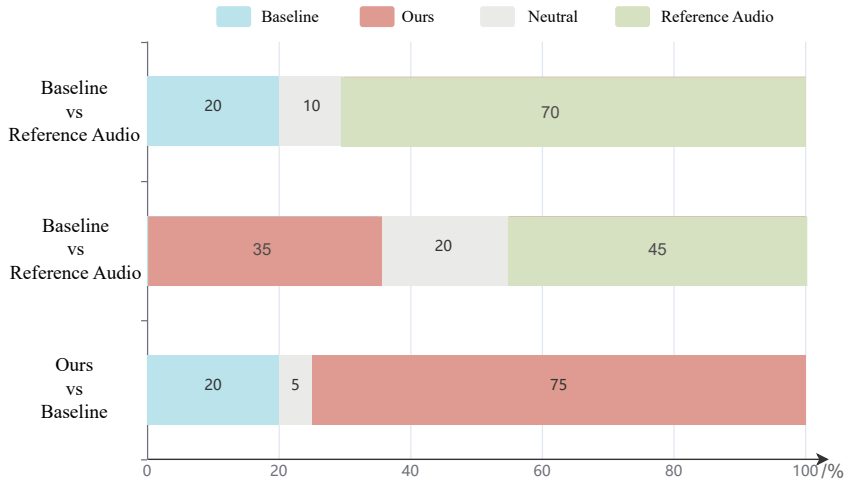

Figure 4: AB Test Results of Code-Switching Speech.

between languages.

When the Lao language ID is removed, the MOS score drops to 3.63, 0.06 points lower than the full model, and the CMOS score is -0.54. These results show that language IDs help the model distinguish phonemes across languages and apply correct pronunciation rules during code-switched synthesis.

Overall, although both modules contribute to performance, the results indicate that removing the PC module causes a more significant degradation than removing the LID module, demonstrating that the phoneme-level latent space has a greater impact on synthesis quality.

4.3.6 AB Test Analysis

An A/B preference test was also conducted, involving 10 Lao international students who were fluent in both Lao and English. As shown in Fig. 4, participants were asked to compare audio samples synthesized by the baseline model, the proposed model, and real reference recordings.

The results indicate that the proposed model outperformed the baseline model in Lao-English code-switched speech synthesis. The audio generated by the baseline model was generally less preferred by participants, while the proposed model's audio received significantly more favorable ratings. Specifically, in the A/B test comparing the proposed model with the baseline, 75% of participants preferred the speech generated by the proposed method, while only 5% preferred the baseline output.

This demonstrates that the integration of a phoneme-level shared latent space and Lao language ID effectively improves both the naturalness and fluency of Lao-English code-switched speech synthesis.

5 Conclusion

This work addresses the challenges of phoneme set mismatches and the difficulty of constructing bilingual corpora from a single speaker in Lao-English code-switched speech synthesis. We propose a strategy that integrates a shared phoneme-level latent space with a Lao-specific language ID module. By projecting phoneme representations from both Lao and English into a unified latent space and embedding language identifiers, the model can effectively distinguish and adapt to different phonetic characteristics.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (62376111, U24A20334, U23A20388, U21B2027 and 62366027), Science and Technology Planning Projects of Yunnan Province (202502AD080014, 202401BC070021, 202303AP140008 and 202302AD080003).

References

Zexin Cai, Yaogen Yang, and Ming Li. 2023. Cross-lingual multi-speaker speech synthesis with limited bilingual training data. *Comput. Speech Lang.*, 77(C), January.

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 *Proceedings of ICML Conference*, pages 2709–2720.

Mengxin Chai, Shaotong Guo, Cheng Gong, Longbiao Wang, Jianwu Dang, and Ju Zhang. 2021. Learning language and speaker information for code-switch speech synthesis with limited data. In *Proceedings of ASRU Conference*, pages 602–609.

Sanyuan Chen, Chengyi Wang, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei. 2025. Neural codec language models are zero-shot text to speech synthesizers. *IEEE Transactions on Audio, Speech and Language Processing*, 33:705–718.

Ahmad Handoyo, Chung Tran, Dessipuji Lestari, and Sakriani Sakti. 2024. Indonesian-english code-switching speech synthesizer utilizing multilingual sten-tts and bert lid. In *Proceedings of O-COCOSDA Conference*, pages 1–6, 10.

Jaehyeon Kim, Jungil Kong, and Juhee Son. 2021. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *Proceedings of ICML Conference*, pages 5530–5540.

Zhaoyu Liu and Brian Mak. 2020. Multi-lingual multi-speaker text-to-speech synthesis for voice cloning with online speaker enrollment. In *Proceedings of Interspeech Conference*, pages 2932–2936, 10.

Sahoko NAKAYAMA, Andros TJANDRA, Sakriani SAKTI, and Satoshi NAKAMURA. 2021. Code-switching asr and tts using semisupervised learning with machine speech chain. *IEICE Transactions on Information and Systems*, E104.D(10):1661–1677.

Wei Ping, Kainan Peng, Andrew Gibiansky, Sercan O. Arik, Ajay Kannan, Sharan Narang, Jonathan Raiman, and John Miller. 2018. Deep voice 3: 2000-speaker neural text-to-speech. In *Proceedings of ICLR Conference*.

Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2020. Fastspeech 2: Fast and high-quality end-to-end text to speech. *arXiv preprint arXiv:2006.04558*.

Jose M. R. Sotelo, Soroush Mehri, Kundan Kumar, João Felipe Santos, Kyle Kastner, Aaron C. Courville, and Yoshua Bengio. 2017. Char2wav: End-to-end speech synthesis. In *Proceedings of ICLR Conference*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Proceedings of NIPS Conference*, volume 30. Curran Associates, Inc.

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*.

Jingzhou Yang and Lei He. 2022. Cross-lingual text-to-speech using multi-task learning and speaker classifier joint training. *arXiv preprint arXiv:2201.08124*.

Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. 2019. Libritts: A corpus derived from librispeech for text-to-speech. *arXiv preprint arXiv:1904.02882*.

Yu Zhang, Ron J Weiss, Heiga Zen, Yonghui Wu, Zhifeng Chen, RJ Skerry-Ryan, Ye Jia, Andrew Rosenberg, and Bhuvana Ramabhadran. 2019. Learning to speak fluently in a foreign language: Multilingual speech synthesis and cross-language voice cloning. *arXiv preprint arXiv:1907.04448*.

Ziqiang Zhang, Long Zhou, Chengyi Wang, Sanyuan Chen, Yu Wu, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, et al. 2023. Speak foreign languages with your own voice: Cross-lingual neural codec language modeling. *arXiv preprint arXiv:2303.03926*.

Xuehao Zhou, Xiaohai Tian, Grandee Lee, Rohan Kumar Das, and Haizhou Li. 2020. End-to-end code-switching tts with cross-lingual language model. In *Proceedings of ICASSP Conference*, pages 7614–7618.