TSPC: A TWO-STAGE PHONEME-CENTRIC ARCHITECTURE FOR CODE-SWITCHING VIETNAMESE-ENGLISH SPEECH RECOGNITION

Minh N. H. Nguyen⋆*, Anh Nguyen Tran*⋆*, Dung Truong Dinh, Nam Van Vo*

Vietnam - Korea University of Information and Communication Technology,
The University of Danang, Vietnam
⋆ Equal Contribution

ABSTRACT

Code-switching (CS) presents a significant challenge for general Auto-Speech Recognition (ASR) systems. Existing methods often fail to capture the subtle phonological shifts inherent in CS scenarios. The challenge is particularly difficult for language pairs like Vietnamese and English, where both distinct phonological features and the ambiguity arising from similar sound recognition are present. In this paper, we propose a novel architecture for Vietnamese-English CS ASR, a Two-Stage Phoneme-Centric model (TSPC). The TSPC employs a phoneme-centric approach, built upon an extended Vietnamese phoneme set as an intermediate representation to facilitate mixed-lingual modeling. Experimental results demonstrate that TSPC consistently outperforms existing baselines, including PhoWhisper-base, in Vietnamese-English CS ASR, achieving a significantly lower word error rate of 19.9% with reduced training resources. Furthermore, the phonetic-based two-stage architecture enables phoneme adaptation and language conversion to enhance ASR performance in complex CS Vietnamese-English ASR scenarios.

Index Terms— code-switching speech recognition, phoneme recognition, multilingual speech recognition.

1. INTRODUCTION

Automatic Speech Recognition (ASR) has gained significant attention in recent years and proliferated as an emerging interaction method between humans and machines. Despite the continuous development of multilingual ASR approaches in the globalization era, code-switching remains extremely challenging, where speakers randomly shift between languages within a conversation. Existing multilingual ASR approaches are still limited due to the inability to capture the unique pronunciation nuances of each language. To investigate this limitation, we conducted a comparative analysis of multilingual and monolingual models' performance in recognizing the mix of Vietnamese and English languages, the illustration is presented in Table 1.

Label	**PhoWhisper-Large**[1]	**Whisper-Large** [2]	**mms-1b-all** [3]
kiểm tra ***camera*** tòa nhà một lần nữa	kiểm tra **cả mẹ ra** toà nhà một lần nữa	Kiểm tra **cả mẹ ra** tòa nhà một lần nữa	kicm tra cc me gia toa nha mt lan nga
khi mình đi dự ***concert***	khi mình đi giữ **con sót**	khi mình đi giữ **con sót**	khi minh di d **con sot**

Table 1. The results shown by monolingual and multilingual models, where the English words in **bold** are incorrect prediction.

Recent End-to-End ASR approaches often overlook inherent phonological structures, leading to significant performance degradation in the mix of language scenarios. Hence, several approaches have been proposed to enhance the traditional ASR architectures, such as the work in Deep context [4], which introduced a bias-encoder for embedding contextual phrases into dimensional representations. However, the approach encounters limitations when processing a large number of context phrases since the model struggles to differentiate highly correlated phrases within extensive corpora. Other approaches integrate a separate Language Identification (LID) module [5, 6] into the main architecture to identify both primary and secondary languages, while the effectiveness of the LID component relies heavily on a large volume of code-switched data, which poses a significant challenge for low-resource languages. Additionally, [7] proposed AdaCS, employing a bias attention module to improve the generalization for unseen language mixes without retraining the model, especially in Vietnamese scenarios.

In most of the ASR approaches, linguistic factors such as tone, which plays a crucial role in ASR, are often overlooked, especially for tonal languages like Vietnamese, since Vietnamese relies on pitch variations to express distinct word meanings, a characteristic absent in non-tonal languages like English. Hence, tonal dependency, coupled with substantial phonological overlap between

Vietnamese and English (e.g., "*lít*" vs. "*list*"), poses significant challenges for ASR systems in CS scenarios. As a result, tone-insensitive ASR models frequently face recognition errors and fail to differentiate phonetically similar lexical items. Addressing tone and language-specific phonological features is crucial to enhance ASR performance in multilingual and code-switched environments. In this work, we propose the TSPC model, a novel framework for Vietnamese-English CS ASR. In particular, the proposed model adopts a phoneme-centric approach, leveraging extended Vietnamese phonemes as a linguistic bridge for mixed-lingual modeling by comprising two stages: a Speech-to-Phone (S2P) module that represents acoustic input to phoneme sequences, and a Phone-to-Text (P2T) module that translates phoneme sequences into text.

2. PHONEME-CENTRIC MODEL

2.1. Extended Vietnamese Phonemes with English Conversion

In code-switched scenarios, phoneme recognition is the fundamental linguistic representation that bridges the acoustic signal and text sequences [8, 9]. The phoneme-level representation provides an intermediate representation to enhance the generalization of ASR models across languages and dialects. The problem becomes increasingly critical in many languages in the globalization era. Representation at the phonetic level enables the identification of linguistic intersections, revealing cross-linguistic similarities in phonetic notation and pronunciation, especially in naturalistic contexts where languages mix, such as Vietnamese and English. Despite existing differences in phonological systems across languages, comparative analysis in [10] shows that both Vietnamese and English utilize a range of similar consonants and vowels, such as ([p], [b], [m], [n], [i]). The overlapping issue that appears in code-switched scenarios can cause recognition errors for acoustic models that struggle to differentiate between highly similar phonemes. Moreover, Vietnamese is a highly tonal language, characterized by short, simple syllables and six distinct lexical tones [11] that determine the meaning of the words. Particularly, tone acts as a primary component and significantly influences Vietnamese speakers when speaking non-tonal languages like English, e.g.: English diphthong *"eɪ"* is replaced with the Vietnamese syllable *"ây"*.

We perform a comprehensive comparative validation to examine the overlapping phonemes in the context of Vietnamese-English CS based on sound similarities extracted based on the pre-trained Vietnamese ASR model, as shown in Fig. 1. From this study, we obtained

English (example)	Prefix			Postfix		
	IPA	vi-syllable	phone	IPA	vi-syllable	phone
zoo	z	d	z	uː	u	u - 0
play	pl	p, l	p, l	eɪ	ây	ə - 0 iz
go	g	g	ɣ	əʊ	âu	ə - 0 uz
come	k	c	k	ʌm	âm	ə - 0 mz
young	j	gi	z	ʌŋ	ăng	a - 0 ŋz
sing	s	s	s	ɪŋ	ing	i - 0 ŋz
bee	b	b	b	iː	i	i - 0
pet	p	p	p	et	ét	ɛ - 4 tz
core	k	c	k	ɔː	o	ɔ - 0
foot	f	ph	f	ʊt	út	u - 4 tz
tea	t	t	t	iː	i	i - 0
think	θ	th	t^h	ɪŋk	in	i - 0 nz
view	v	v	v	juː	iu	i - 0 uz
ship	ʃ	s	s	ɪp	íp	i - 4 pz
lamp	l	l	l	æmp	am	aː - 0 mz
tour	t	t	t	ʊər	ua	uə - 0

Table 2. Comparative analysis of English words, an English word is separated to prefix and postfix parts, where Vietnamese syllable (vi-syllable) is mapped with IPA, and converted to Vietnamese phoneme (phone).

preliminary results for converting the pronunciation of English words into Vietnamese syllables, as illustrated in the Table 2. Subsequently, we utilized existing phonetic vocabularies and conversion modules to convert represented Vietnamese syllables into precise phonetic representations. For instance, the English word "*a*" is mapped with the Vietnamese syllable "*ây*", which was then converted to the phonetic form "ə - *0 iz*". The phoneme conversion methodology demonstrates a robust method based on the intricate phonetic interactions and adaptations observed in cross-linguistic contexts.

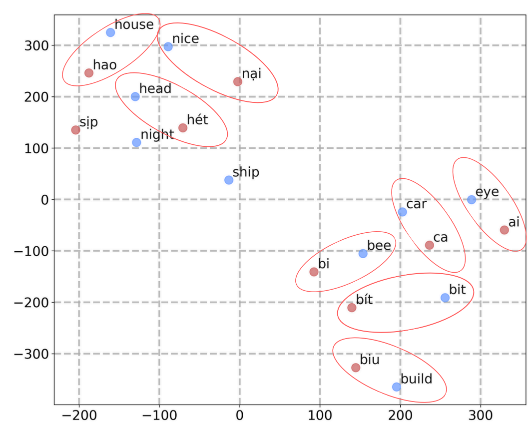

Fig. 1. T-SNE visualization of English words in blue and Vietnamese syllables in red.

2.2. Two-stage Model Development

The proposed two-stage ASR model utilizes phonemic representation by combining two foundation mod-

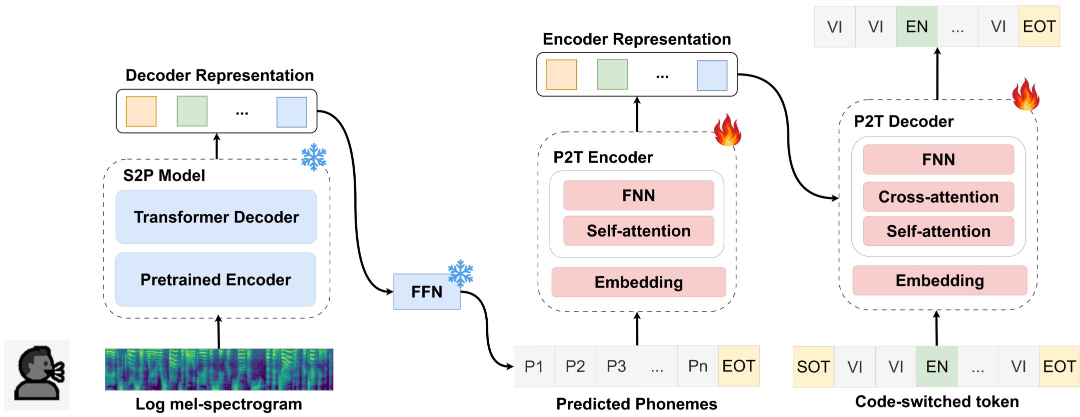

Fig. 2. Two-stage phoneme-centric model (TSPC), where predicted phoneme sequence represents the input for P2T model.

els, such as Speech-to-Phone (S2P) and Phone-to-Text (P2T) that are independently pre-trained, then integrated and fine-tuned for code-switched scenarios.

Speech-to-Phone: For the S2P model, we adopt the well-established Sequence-to-Sequence (Seq2Seq) paradigm [12] for speech-to-text tasks due to its efficient architecture for learning complex acoustic features and mapping variable-length input speech to a sequence of phonemes. Specifically, the S2P model employs a pre-trained Encoder built on large-scale Vietnamese datasets that can capture rich acoustic feature extraction capabilities. Following the acoustic Encoder, different variants of Decoders are validated for generating phoneme sequences. Throughout our comprehensive experiments in our prior work in [13], we can figure out the most suitable Decoder for phoneme recognition as shown in the in Table 3.

Encoder - Decoder	**Phoneme Error Rate** (%) ↓			
	LSVSC	*Vietbud_500*	*CmV*	*VLSP 2020*
PhoWhisper - GRU	1070	2230	3280	1970
PhoWhisper - LSTM	720	1290	1410	1403
Wav2vecVN - Transformer	46.20	37.73	17.66	49.09
PhoWhisper - Transformer	**8.43**	**5.71**	**10.13**	**15.4**

Table 3. Vietnamese phoneme recognition results of architectural combination.

Phone-to-Text: Phone-to-Text (P2T) conversion is framed as a sequence-to-sequence (Seq2Seq) translation problem, drawing inspiration from Machine Translation (MT) [14]. In this work, the phoneme sequence serves as the *"source language"* and the text as the *"target language"*. Seq2Seq architecture enables an inherent design for the P2T module to map a phoneme input sequence to a text sequence. In this work, we choose T5 model [15] for the P2T model, which frames all NLP problems as text2text generation tasks. As discussed in [16], the model demonstrates its effectiveness in handling phonetic transformations, successfully applying T5 to Grapheme-to-Phoneme (G2P) conversion. Similarly, we also leverage T5 model design for the inverse problem of phoneme-to-text translation.

Two-stage TSPC model: The integration of two separate models in a unified architecture is not straightforward, as the incorrect phone sequences generated by S2P model can introduce noises in the input of P2T model, which reduces the performance of the translation task. To resolve this issue, the TSPC model carefully combines S2P and P2T modules and performs joint fine-tuning as shown in Fig 2.1. To integrate the pre-trained S2P model with P2T model for robust text generation, we freeze S2P parameters during the tuning phase to ensure consistent phonetic sequence production. The phoneme sequence from the S2P model plays a vital role in association with the P2T model, in which the parameters are updated to adapt to predicted phonemes.

3. EXPERIMENTS

3.1. Dataset Preparation

The P2T dataset is constructed by generating a mix of English-Vietnamese transcriptions, mapping English words to Vietnamese syllables to form initial localized code-switched text. To account for pronunciation diversity, each English word is matched with various syllable variants, as illustrated in Fig. 3, while Vietnamese transcriptions are concurrently gathered from an independent source. Both the mapped English-Vietnamese text and these collected Vietnamese transcriptions are then transformed into phonemes. Supporting S2T and S2P model development, a speech dataset is curated: for S2T, the code-switched text is synthesized into audio via a high-quality Text-to-Speech (TTS) service and

Model	#params	CS	Vi	Vi			
			avg	*LSVSC*	*Vietbud_500*	*CMV*	*VLSP 2020*
Whisper-base	71.8M	59.45	74.83	52.01	92.9	92.9	61.5
Wav2vec2VN	94.4M	38.06	21.7	13.7	16.4	32.4	24.3
PhoWhisper-base	71.8M	27.9	**14.05**	**9.4**	14.3	**16.08**	**16.42**
TSPC - Zero	90.6M	25.35	18.13	14.9	11.59	20.93	25.13
TSPC	90.6M	**19.9**	16.47	13.64	**10.63**	18.3	23.33

Table 4. WER results of methods on Code-switching (CS) and Vietnamese (Vi) test sets. Best results in **bold**.

Reference:
Hôm qua tớ vừa xem cái **video** này hay lắm.
(*Yesterday, I watched this video, it was really cool.*)

Input:
Hôm qua tớ vừa xem cái **vi deo** này hay lắm.
(*"vi deo": Vietnamese conversion of "video."*)

Variants (word replacements):
video → vi đêu (*variation with different pronunciation*)
video → vi đê ô (*different speaking of "video"*)

Fig. 3. Example of variants in Phone-to-Text dataset

integrated with an existing Open Vietnamese Dataset; for S2P, audio transcriptions within the S2T dataset are converted to phonetic form.

In our study, the corpus for training CS ASR system consists of existing Vietnamese speech datasets, including VLSP 2020 [17] (92.03 hours), VietBud500 [18] (43.55 hours), Common Voice [19] (20.3 hours), LSVSC [20] (49.17 hours), and VSV [21] (31.51 hours). In addition, we incorporated 7.32 hours of Vietnamese-English CS speech, which includes both the Capleaf [22] and synthetic CS data. For CS evaluation, a subset of 1.18 hours of CS speech was used to assess the system performance.

3.2. Training Strategy

In the pre-training phase, the S2P model leverages the frozen PhoWhisper-base Encoder for acoustic feature extraction, and the Transformer Decoder was pre-trained for 15 epochs. While the P2T model was pre-trained on phoneme-text pairs for 40 epochs. Both models shared an identical architecture of 6 Encoder-Decoder blocks with a model dimension of 512, and hyperparameters were uniform: a batch size of 16 and a learning rate of $1e-4$ with a linear warm-up scheduler. During the joint fine-tuning, the S2P model was frozen to maintain its phoneme generation capabilities. The fine-tuning stage was conducted with a batch size of 8 for 10 epochs.

3.3. Experimental Results

Code-switching speech recognition: Table 4 reveals high error rates of the existing Whisper-base model [2] (59.45%) and Wav2vec2VN [23] (38.06%) in CS speech, demonstrating their inability to handle mixed language scenarios. Although pre-trained PhoWhisper-base model achieves a superior 27.9% error rate, TSPC models yield the most compelling results by using significantly fewer training resources. TSPC models consistently surpass PhoWhisper-base: TSPC - Zero achieves 25.35%, and TSPC improves to 19.9%. Note that TSPC - Zero is independent operate two stage S2P and P2T models without joint fine-tuning.

Vietnamese speech recognition: For Vietnamese utterances evaluation in Table 4, Whisper-base without Vietnamese speech fine-tuning obtained a poor performance with a high average WER of 74.83%. In contrast, Wav2vec2VN demonstrates efficiency with 21.7% WER after 250 hours of Vietnamese speech training. The proposed TSPC models achieved good results with significantly fewer training data in low-resource training scenarios. The best TSPC variant achieves an average WER of 16.47%, approximately 2.5% error rate higher than PhoWhisper-base, which was fine-tuned on extensive large Vietnamese Speech data.

4. CONCLUSION

In this study, we proposed a two-stage phoneme-centric architecture designed for Vietnamese-English CS speech recognition. The extensive experiments demonstrate the efficiency of Vietnamese phoneme representation by enhancing the model's ability to handle the mix of Vietnamese-English languages. The proposed method emphasizes the importance of converted phonemes in advancing the speech recognition performance in CS and future multilingual systems.

5. ETHICAL STATEMENT

The TSPC architecture was trained using real Vietnamese pronunciation styles of English words due to the ability to handle Vietnamese accents and speech styles for ensuring the nature of sound and promoting high precision. The constructed dataset is built on speeches from different regions of Vietnam, genders, and ages, which can mitigate the biases in the CS ASR system. However, we advocate that the training data still limited and to be largely extended in future, particularly in auditory and text modalities, hindering its ability to generalize to a wide range of code-switched scenarios.

6. REFERENCES

[1] Thanh-Thien Le and Linh The Nguyen et. al., "Phowhisper: Automatic speech recognition for vietnamese," 2024.

[2] Alec Radford and Jong Wook et. al. Kim, "Robust speech recognition via large-scale weak supervision," in *International conference on machine learning*. PMLR, 2023, pp. 28492–28518.

[3] Vineel Pratap and Andros Tjandra et. al., "Scaling speech technology to 1,000+ languages," *arXiv*, 2023.

[4] Golan Pundak, Tara N Sainath, and et. al. Prabhavalkar, "Deep context: end-to-end contextual speech recognition," in *2018 IEEE spoken language technology workshop (SLT)*. IEEE, 2018, pp. 418–425.

[5] Zhiping Zeng, Yerbolat Khassanov, and et. al. Pham, Van Tung, "On the end-to-end solution to mandarin-english code-switching speech recognition," *arXiv preprint arXiv:1811.00241*, 2018.

[6] Hukai Huang and Jiayan Lin et. al., "Enhancing code-switching speech recognition with lid-based collaborative mixture of experts model," 2024.

[7] The Chuong Chu and et. al. Pham, Vu Tuan Dat, "Adacs: Adaptive normalization for enhanced code-switching asr," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[8] Megan Merz and Olga Scrivner, "Discourse on asr measurement: Introducing the arpoca assessment tool," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, 2022, pp. 366–372.

[9] Morris Swadesh, "The phonemic principle," *Language*, vol. 10, no. 2, pp. 117–129, 1934.

[10] Ngo Phuong Anh, "L1 influence on vietnamese accented english," *Voices*, pp. 108–125, 2011.

[11] Tuong Tu Huu and Viet Thanh et. al Pham, "Mispronunciation detection and diagnosis model for tonal language, applied to vietnamese," in *Proc. INTERSPEECH*, 2023, vol. 2023, pp. 1014–1018.

[12] Ilya Sutskever and Oriol et. al. Vinyals, "Sequence to sequence learning with neural networks," *Advances in neural information processing systems*, vol. 27, 2014.

[13] Nguyen Minh and et. al. Anh, "Whisper based cross-lingual phoneme recognition between vietnamese and english," *arXiv preprint arXiv:2508.19270*, 2025.

[14] Felix Stahlberg, "Neural machine translation: A review," *Journal of Artificial Intelligence Research*, vol. 69, pp. 343–418, 2020.

[15] Colin Raffel and Noam Shazeer et. al., "Exploring the limits of transfer learning with a unified text-to-text transformer," 2023.

[16] Markéta Řezáčková, Jan Švec, and Daniel Tihelka, "T5g2p: Using text-to-text transfer transformer for grapheme-to-phoneme conversion," in *Proc. Interspeech 2021*, 2021, pp. 6–10.

[17] Association for Vietnamese Language and Speech Processing, "Automatic speech recognition for vietnamese," `https://vlsp.org.vn/vlsp2020/eval/asr`, 2020, VLSP 2020 ASR evaluation campaign.

[18] Anh Pham and Khanh Linh Tran et. al., "Bud500: A comprehensive vietnamese asr dataset," 2024.

[19] Rosana Ardila and Megan et.al. Branson, "Common voice: A massively-multilingual speech corpus," *arXiv preprint arXiv:1912.06670*, 2019.

[20] Linh Thi Thuc Tran and et. al. Kim, Han-Gyu, "Automatic speech recognition of vietnamese for a new large-scale corpus," *Electronics*, vol. 13, no. 5, pp. 977, 2024.

[21] Pham Quang Nhut, Duong Pham Hoang Anh, and Nguyen Vinh Tiep, "Vietspeech: Vietnamese social voice dataset," 2024.

[22] Capleaf, "vivoice: Vietnamese multi-speaker speech synthesis dataset," `https://huggingface.co/datasets/capleaf/viVoice`, 2024, Hugging Face dataset (gated access; CC-BY-NC-SA 4.0).

[23] Thai Binh Nguyen, "Vietnamese end-to-end speech recognition using wav2vec 2.0," 09 2021.