ViMedCSS: A Vietnamese Medical Code-Switching Speech Dataset & Benchmark

Tung X. Nguyen[1,2], Nhu Vo[1,4], Giang-Son Nguyen[1,2], Duy Mai Hoang[3], Chien Dinh Huynh[3], Iñigo Jauregi Unanue[4], Massimo Piccardi[4], Wray Buntine[1], Dung D. Le[1,2]

[1] College of Engineering and Computer Science, VinUniversity, Vietnam
[2] Center for AI Research, VinUniversity, Vietnam
[3] College of Health Sciences, VinUniversity, Vietnam
[4] University of Technology Sydney, Australia
{tung.nx, nhu.vd, son.ng, duy.hm, chien.hd, wray.b, dung.ld}@vinuni.edu.vn
{DiepNhu.Vo, Inigo.JauregiUnanue, Massimo.Piccardi}@uts.edu.au

Abstract

Code-switching (CS), which is when Vietnamese speech uses English words like drug names or procedures, is a common phenomenon in Vietnamese medical communication. This creates challenges for Automatic Speech Recognition (ASR) systems, especially in low-resource languages like Vietnamese. Current most ASR systems struggle to recognize correctly English medical terms within Vietnamese sentences, and no benchmark addresses this challenge. In this paper, we construct a 34-hour **Vi**etnamese **Med**ical **C**ode-**S**witching **S**peech dataset (ViMedCSS) containing 16,576 utterances. Each utterance includes at least one English medical term drawn from a curated bilingual lexicon covering five medical topics. Using this dataset, we evaluate several state-of-the-art ASR models and examine different specific fine-tuning strategies for improving medical term recognition to investigate the best approach to solve in the dataset. Experimental results show that Vietnamese-optimized models perform better on general segments, while multilingual pretraining helps capture English insertions. The combination of both approaches yields the best balance between overall and code-switched accuracy. This work provides the first benchmark for Vietnamese medical code-switching and offers insights into effective domain adaptation for low-resource, multilingual ASR systems.

Keywords: Automatic Speech Recognition, Vietnamese, Medical, Code-switching, Contextual Biasing

1. Introduction

Code-switching (CS) is pervasive in Vietnamese medical communication, where English clinical terms (drug names, procedures, biomarkers) appear within otherwise Vietnamese utterances. Prior work across languages shows that ASR errors peak precisely on the embedded-language portions of an utterance—i.e., at the points where non-matrix terms are inserted—highlighting the need for language-tagged, degree-controlled evaluation to diagnose model behavior on these spans (Lyu et al., 2010; Ugan et al., 2024; Agro et al., 2025). Yet there is no open benchmark centered on Vietnamese medical CS that simultaneously supports a systematic study of practical remedies, from injecting domain term lists during decoding to CS-oriented adaptation on modern encoder–decoder backbones.

Clear and accurate medical communication is a cornerstone of patient safety and clinical effectiveness (Sharkiya, 2023). In Vietnam, where healthcare professionals frequently alternate between Vietnamese and English medical terminology during consultations, lectures, and patient education, this code-switching reflects the globalization of medicine but also introduces a significant risk of misunderstanding (Chen, 2025). Misrecognition of key terms—such as drug names, anatomical structures, or diagnostic procedures—by automated transcription systems can lead to errors in clinical documentation, medication administration, and data reporting. In medical education and research, inaccurate recognition of bilingual terminology diminishes the clarity of lectures and assessment materials, undermining both comprehension and patient-care competence among trainees (Hamad et al., 2025). A reliable system for detecting and transcribing code-switched speech is therefore not merely a technical goal but a public-health necessity. It ensures that digital records accurately capture the clinician's intent, supports high-quality medical training materials, and facilitates inclusive communication with non-specialist audiences and multilingual patients.

Model capacity and pretraining have rapidly advanced Vietnamese ASR, spanning both multilingual architectures (Radford et al., 2023; Pratap et al., 2024) and Vietnamese-optimized variants (Le et al., 2024; Nguyen, 2021; Zhuo et al., 2025). Across these families, a characteristic trade-off emerges in code-switching: models optimized for Vietnamese tend to reduce sentence-level errors in the matrix language, while broadly trained multilingual models better recognize embedded English segments. A benchmark that explicitly separates

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 5657–5665
11–16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

overall accuracy from accuracy on code-switched spans is therefore needed.

We introduce ViMedCSS,[1] a Vietnamese medical code-switching speech dataset in which every utterance contains at least one code-switched medical term drawn from a bilingual lexicon. The corpus comprises 34.57 hours and 16,576 utterances across five topics, and includes a held-out hard split of rare/unseen terms to test generalization beyond the training vocabulary. We establish zero-shot baselines with state-of-the-art multilingual and Vietnamese ASR systems, then systematically compare fine-tuning on a Whisper-based Vietnamese backbone across complementary approaches to code switching—most notably contextual biasing during decoding versus language-identity–guided adaptation—together with parameter-efficient adapters, post-decoding normalization, and their hybrids. Our evaluation separates overall accuracy from performance on code-switched spans metrics, yielding practical guidance on which strategies most effectively handle medical code switching in Vietnamese ASR.

2. Related Work

Vietnamese ASR has benefited from both large multilingual pretraining and targeted Vietnamese adaptation. Whisper provides strong multilingual zero-shot performance and serves as a widely used encoder–decoder baseline (Radford et al., 2023), while PhoWhisper adapts the same architecture to Vietnamese via fine-tuning on an 844 h corpus covering diverse speakers and styles (Le et al., 2024). MMS scales wav2vec 2.0 (Baevski et al., 2020) to over one thousand languages with competitive CTC baselines (Pratap et al., 2024). On the monolingual side, wav2vec2-base-vi leverages large-scale unlabeled YouTube audio and is fine-tuned on VLSP labels (Nguyen, 2021), and VietASR employs a Zipformer encoder with ASR-biased self-supervision, pre-trained on roughly 70k h and fine-tuned on 50 h of labeled Vietnamese speech (Zhuo et al., 2025). Public Vietnamese resources such as VIVOS and multilingual corpora like FLEURS further support training and evaluation (Luong and Vu, 2016; Conneau et al., 2023). Together these systems span key design axes—multilingual vs. Vietnamese-only training, encoder–decoder vs. CTC decoding, and compact vs. large capacity—and constitute the primary baselines against which we study Vietnamese medical code switching.

Privacy constraints limit open medical speech corpora. For Vietnamese, VietMed provides a mix of labeled and large unlabeled medical audio with ASR baselines and recipes (Le-Duc, 2024). For multilingual clinical communication that includes Vietnamese, MultiMed-ST offers a large many-to-many medical speech–translation corpus with analyses that also examine code-switching phenomena (Le-Duc et al., 2025). On the text side, MedEV introduces a sizeable Vietnamese–English medical parallel corpus and benchmarks multiple Machine Translation (MT) systems, showing clear gains from domain-specific fine-tuning (Vo et al., 2024). Together, these efforts advance Vietnamese medical ASR and MT, but none target systematic evaluation of code-switched medical terminology within ASR or the role of contextual biasing, which motivates our benchmark.

A recent review synthesizes datasets, metrics, and modeling patterns for end-to-end CS ASR, emphasizing language-split reporting and CS-degree slicing (Agro et al., 2025). Canonical testbeds include SEAME, with time-aligned language boundary tags (Lyu et al., 2010), and the ASRU 2019 Mandarin–English challenge (Shi et al., 2020), while DECM contributes a German–English evaluation set with word-level tags and explicit low/mid/high CS bins (Ugan et al., 2024). CS-FLEURS broadens with a benchmark spanning 52 languages and over one hundred code-switched pairs in general domain (Yan et al., 2025). On the modeling side, Whisper-based adaptations that leverage language identity (LID) have proven effective: attention-guided, parameter-efficient finetuning selects and steers LID-sensitive heads (Aditya et al., 2024), and complementary work refines Whisper via encoder improvements and language-aware decoding (Zhao et al., 2025).

Beyond architectural adaptation, contextual biasing targets rare, domain terms at inference. Neural–symbolic approaches such as TCPGen integrate a prefix-trie of bias words into end-to-end decoders and reduce errors on long-tail entities (Sun et al., 2023) To handle large catalogs, ranking/selection methods forward only the top-k most relevant items to the decoder (Hou et al., 2025). Dynamic vocabulary further injects bias entries as single tokens on the fly, avoiding heavy external Language Models (LMs) or rescoring (Sudo et al., 2024, 2025).

3. Vietnamese Medical Code-Switching Speech dataset - ViMedCSS

3.1. Construction

As seen from Figure 1, we start the dataset construction pipeline from the *Meddict*[2] dictionary, an

[1] https://huggingface.co/datasets/tensorxt/ViMedCSS

[2] https://meddict-vinuni.com/

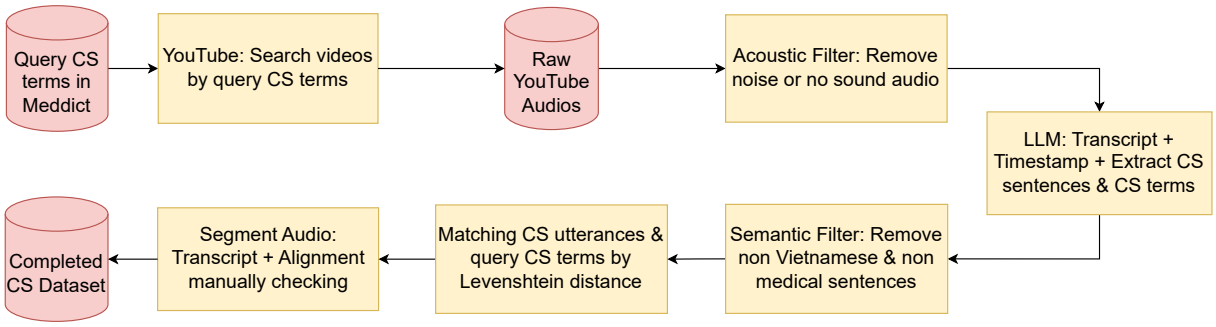

Figure 1: Dataset construction pipeline for Vietnamese medical code-switching.

English–Vietnamese medical lexicon created at VinUniversity with 64,232 entries. Meddict offers curated translations of specialized clinical terminology and is protected under Intellectual Property Rights Certificate No. 3365/2024/QTG. It is released for academic and healthcare use under the institution's license. From this bilingual source, we select entries whose Vietnamese usage retains an English or foreign-root surface form; these define our set of code-switched (CS) medical terms. In total, we extract 3,203 CS terms from the dictionary.

Using these terms as queries, we retrieve Vietnamese medical videos from public platforms. Candidates must have Vietnamese titles and belong to the medical domain; both conditions are automatically checked with a large language model (LLM). We crawl more than 13,000 YouTube videos and discard items with music-only or non-speech audio before transcription. Each remaining audio track is transcribed with Gemini 2.5 Pro to produce time-aligned text and to automatically flag candidate CS sentences and terms. In aggregate, over 700 hours of audio are processed in this step.

We provide the LLM with the following instruction to obtain sentence-level timestamps and CS spans in a machine-readable format:

> **Transcription Task:** You are an advanced transcription assistant for Vietnamese medical audio. Do the following:
>
> 1) Transcribe the Vietnamese speech.
> 2) Segment into sentences of approximately 5–15 seconds.
> 3) For each segment, output `start_time`, `end_time`, and `text`.
> 4) Detect segments that contain any non-Vietnamese terms (e.g., English, Chinese, technical product/brand names).
> 5) Return the complete set of segments, and separately the subset of code-switch segments. For each code-switch segment, list the non-Vietnamese terms that appear.
>
> All output must be a single valid JSON object with keys:
>
> - `"segments"`: an array of all segment objects {`start_time`, `end_time`, `text`}.
> - `"code_switch_segments"`: an array of code-switch segment objects {`start_time`, `end_time`, `text`, `cs_term`: [...]}.
>
> Return only JSON; do not include any additional prose.

We then apply LLM-assisted *semantic filtering* to remove utterances that are non-Vietnamese or off-domain, and we *normalize* surface forms to a canonical dictionary (orthography, hyphenation, casing, common variants) to ensure consistent term identity across transcripts. Because the CS spans returned by the LLM may not exactly match queried dictionary entries, we further align terms by computing the Levenshtein distance between each dictionary item and each sentence, assigning the closest canonical entry to the detected span. After this filtering and normalization pass, a little over 34 hours of audio remain as valid Vietnamese medical CS data.

Finally, we segment the raw audio into 3–29 s utterances and perform manual alignment checks for quality control. The resulting corpus is domain-focused and guarantees at least one CS medical term per utterance, enabling evaluation along both contextual-biasing and code-switching dimensions.

3.2. Sampling and Quality Verification

To assess annotation reliability, we sampled 500 utterances (approximately one hour) from the 34.57-hour corpus, stratified across the five topics to preserve domain diversity. Two trained annotators independently reviewed each utterance following the project guidelines, assigning labels for (i) transcription errors on Vietnamese words and (ii) errors on code-switched terms.

Inter-annotator agreement, measured with Cohen's kappa (Cohen, 1960), was $\kappa = 0.65$, indicating substantial consistency. This suggests that the guidelines were clear and that the sampled set is representative of the broader corpus.

The main source of discrepancy arose from imperfect segment boundaries: timestamps were occasionally misaligned, making end points ambigu-

ous. We therefore added an automatic boundary-refinement step to the pipeline (timestamp smoothing and alignment correction) before downstream processing, which reduced these mismatches in subsequent audits.

3.3. Statistics

Table 1 illustrates the linguistic phenomenon targeted by the corpus: each utterance contains at least one code-switched medical term (boldface), ranging from single to multiple insertions within fluent Vietnamese contexts.

Examples for CS Utterances
Tiếp theo số ba đó là cái dạng **peptide** mà nó ức chế dẫn truyền thần kinh.
Các loại **protit** có nguồn gốc động vật có giá trị dinh dưỡng cao, còn các **protit** thực vật có giá trị dinh dưỡng thấp.
Các nghiên cứu cho thấy **quercetin** có thể hoạt động bằng cách ngăn chặn hoạt động của các hóa chất gây viêm trong cơ thể như **prostan** và **leukotriene**.

Table 1: Representative utterances with increasing numbers of code-switched medical terms (1, 2, 3).

The utterances are grouped into five medical topics—Medical Sciences, Pathology & Pathogens, Treatments, Nutrition, and Diagnostics—using automatic assignment with Gemini 2.5 Pro followed by manual checks. Figure 2 visualizes the segment-duration distribution, Table 2 reports hours and utterance counts per topic.

Topics	Duration	# Utterances
Medical Sciences	16.33 h	7,477
Pathology & Pathogens	10.27 h	4,937
Treatments	3.82 h	1,976
Nutrition	2.14 h	1,155
Diagnostics	2.02 h	1,031

Table 2: Per-topic distribution by total duration and number of utterances.

Overall, the dataset contains 16,576 utterances (34.6 hours). Segment lengths range from 3–29 s and follow a unimodal, mildly right-tailed distribution (Fig. 2); the mean is slightly above the median and most segments are under about 12 s, which helps reduce padding and improves batch efficiency. The topic mix is intentionally skewed to mirror real usage (Table 2): Medical Sciences contributes roughly half of the hours, Pathology & Pathogens forms a substantial second share, and Treatments, Nutrition, and Diagnostics make up the remainder, yielding both broad scientific coverage and focused procedural or lifestyle content.

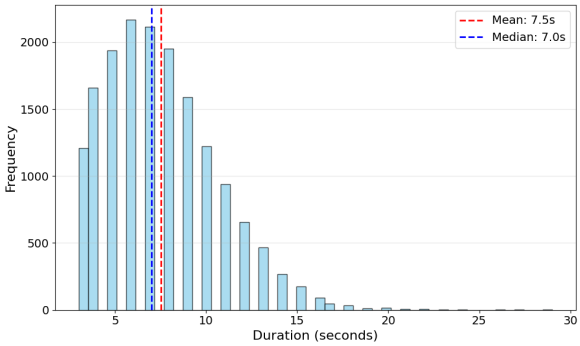

Figure 2: Histogram of utterance durations.

We collected 889 distinct code-switched medical terms from the dataset. The distribution is long-tailed: 160 terms appear exactly once (18.0%), 435 appear at most five times (48.9%), and 207 appear at least twenty times (23.3%). This skew mirrors domain practice—many specialized items are rare—allowing evaluation on both frequent and infrequent terminology, with rare/unseen items further isolated in the Hard split.

4. Experiments

4.1. Setup

4.1.1. Split

We partition the corpus into four *mutually exclusive* sets: *Train*, *Valid*, *Test*, and a dedicated *Hard* split. The Hard split contains only code-switched medical terms that occur once or twice in the entire collection; all occurrences of those terms are removed from Train/Valid/Test to prevent leakage. The remaining pool is divided 8:1:1 into Train/Valid/Test while preserving speaker and topic diversity (Table 3).

Split	Duration	# Utterances	CS terms
Train	24.31 h	11,833	610
Dev	3.56 h	1,714	523
Test	3.38 h	1,615	509
Hard	1.38 h	658	338

Table 3: Dataset splits (rows are sets; columns are statistics). The Hard set is disjoint from Train/Dev/Test.

4.1.2. Metrics

Following prior work on Vietnamese code-switching ASR, we report WER and CER together with CS-WER and N-WER to disentangle accuracy on code-switched spans from the rest of the transcript (Chu et al., 2025). Concretely, *CS-WER* is the word error

Models	Test set				Hard set			
	WER ↓	CER ↓	CS-WER ↓	N-WER ↓	WER ↓	CER ↓	CS-WER ↓	N-WER ↓
MMS	58.30	32.11	68.44	59.80	64.19	35.51	73.25	65.85
wav2vec2-base-vi	44.74	25.48	69.05	46.56	58.21	30.72	73.08	61.11
Whisper-Small	50.03	34.34	61.25	51.35	69.20	44.72	66.75	71.81
PhoWhisper-Small	36.31	22.64	62.55	37.59	46.27	28.11	66.01	48.13
Whisper-Large-v3	34.47	24.61	**46.69**	35.26	39.35	27.46	**50.10**	40.37
PhoWhisper-Large	31.24	**19.25**	55.05	32.36	37.37	**23.02**	57.37	38.67
VietASR	**27.56**	20.38	58.38	**28.25**	**34.43**	25.28	60.71	**35.34**

Table 4: Zero-shot baselines with different models.

Methods	Test Set				Hard Set			
	WER ↓	CER ↓	CS-WER ↓	N-WER ↓	WER ↓	CER ↓	CS-WER ↓	N-WER ↓
Frozen	36.31	22.64	62.55	37.59	46.27	28.11	66.01	48.13
DV	36.29	22.63	62.46	37.57	46.07	27.95	66.01	47.94
RS	35.60	23.68	60.07	36.54	43.89	29.64	61.52	45.10
AdaCS	39.40	28.62	32.91	34.49	50.29	37.23	67.43	48.12
LoRA	27.13	18.01	30.26	27.90	37.27	24.55	60.71	38.69
LoRA + AdaCS	30.85	19.52	30.14	27.90	40.93	26.90	60.92	38.81
AG	**23.67**	**14.73**	**19.50**	**24.40**	**33.73**	**21.48**	**57.29**	**34.80**
AG + AdaCS	25.82	15.52	20.86	25.19	35.00	22.21	**57.29**	34.88

Table 5: Fine-tuning results on PhoWhisper-small across methods.

rate computed only on tokens inside code-switched regions, *N-WER* is the word error rate restricted to tokens that do not require normalization (i.e., outside CS spans), and *WER* is computed over the full, normalized output sequence. We compute all metrics on the mixed *Test* set and report the *Hard* set separately to diagnose generalization to rare/unseen medical terms.

4.1.3. Models & Methods

To situate our benchmark, we report zero-shot results from representative systems along the above axes: MMS (multilingual CTC) and its Vietnamese-only counterpart wav2vec2-base-vi (self-supervised pretraining plus VLSP fine-tuning) (Pratap et al., 2024; Nguyen, 2021); Whisper (Small, Large-v3; multilingual encoder–decoder) and the Vietnamese-adapted PhoWhisper (Small/Large) (Radford et al., 2023; Le et al., 2024); and VietASR (Zipformer with ASR-biased self-supervision) (Zhuo et al., 2025). This set balances multilingual and monolingual training and decoder paradigms while avoiding architectural redundancy; fuller background appears in Section 2.

We probe adaptation strategies on a common backbone and choose *PhoWhisper-small* as the base: Whisper-style models are the prevailing substrate for recent CS-ASR, and PhoWhisper offers a Vietnamese-optimized instantiation of practical size. We group methods into four families. (i) *Contextual biasing in the decoder*: *Dynamic Vocabulary* (DV) extends the output inventory at inference so that each entry in a bias list is represented as a single token, enabling phrase-level biasing without external LMs (Sudo et al., 2024, 2025); and *Rank & Selection* (RS) ranks a large bias list with an auxiliary scorer and forwards only the top-k items to the decoder for scalable contextualization (Hou et al., 2025). (ii) *Post-processing with contextualization*: *AdaCS* adds a bias-attention normalization module to identify and normalize code-switched phrases given an external list (Chu et al., 2025). (iii) *Parameter-efficient adapters*: *LoRA* inserts low-rank adapter matrices into transformer blocks to fine-tune a small parameter subset (Hu et al., 2022). (iv) *LID-guided adaptation*: *Attention Guide* (AG) selects attention heads indicative of language identity and guides them during adaptation to handle switches; prior work reports strong results on SEAME using this approach (Aditya et al., 2024; Agro et al., 2025). Because AdaCS operates after decoding, we also evaluate hybrids (*LoRA+AdaCS*, *AG+AdaCS*). For all contextual-biasing methods (DV, RS, AdaCS), the bias list is built from the code-switched medical terms present in the corresponding split and used when decoding that split (train/test/hard), ensuring split-consistent contextualization; for AG, we employ bilingual Vietnamese–English prompts to reflect the intended CS setting.

Methods	Test Set				Hard Set			
	WER ↓	CER ↓	CS-WER ↓	N-WER ↓	WER ↓	CER ↓	CS-WER ↓	N-WER ↓
Whisper-Small	50.03	34.34	61.25	51.35	69.20	44.72	66.75	71.81
PhoWhisper-Small	36.31	22.64	62.55	37.59	46.27	28.11	66.01	48.13
Whisper-Small (LoRA)	39.96	26.41	33.24	41.35	46.31	31.91	60.92	47.78
PhoWhisper-Small (LoRA)	27.13	18.01	30.26	27.90	37.27	24.55	60.71	38.69
Whisper-Small (AG)	24.54	15.05	20.86	25.23	35.26	22.07	56.84	36.36
PhoWhisper-Small (AG)	**23.67**	**14.73**	**19.50**	**24.40**	**33.73**	**21.48**	**57.29**	**34.80**

Table 6: Effect of fine-tuning approaches by LoRA & Attention Guide on multilingual (Whisper-Small) models vs. monolingual (PhoWhisper-Small).

4.2. Zero-shot results

As shown in Table 4, there is a stable split between sentence-level accuracy and code-switched spans across both test and hard sets. Within each model pair, Vietnamese-optimized systems reduce utterance errors (WER/CER/N-WER) relative to their multilingual counterparts: wav2vec2-base-vi improves over MMS, PhoWhisper-Small over Whisper-Small, and PhoWhisper-Large over Whisper-Large-v3. Overall, VietASR is strongest on sentence-level metrics (best WER and N-WER on both splits), and PhoWhisper-Large attains the lowest CER among the large-capacity models. These outcomes are consistent with extensive Vietnamese-only pretraining and targeted fine-tuning that better capture matrix-language phonotactics and style.

On the code-switched regions, the pattern reverses at the high-capacity end: Whisper-Large-v3 delivers the lowest CS-WER on both splits, outperforming VietASR and the PhoWhisper variants (e.g., on the test set it leads by a clear margin). At smaller scale the gap narrows and can be comparable, but PhoWhisper-Small still retains its advantage on overall WER/CER/N-WER. Taken together, the results reinforce a common trade-off also observed on external CS benchmarks: monolingual or Vietnamese-adapted models dominate sentence-level accuracy, whereas broad multilingual exposure improves recognition of embedded English "islands" within Vietnamese utterances (Ugan et al., 2024; Agro et al., 2025).

4.3. Fine-tuning results

Results of finetuning across methods on PhoWhisper-small (Table 5) show a consistent pattern. Decoder-side contextual methods (DV, RS) provide only modest changes relative to the frozen model, while AdaCS sharply reduces CS-WER but can raise overall WER & CER in isolation—typical of precision–recall trade-offs when bias spans are sparse or noisy. In contrast, adapter-based fine-tuning improves all metrics, with AG yielding the strongest overall and hard-set scores and LoRA a solid second. Combining adapters with post-processing further stabilizes performance (LoRA+AdaCS, AG+AdaCS), preserving low CS-WER while recovering sentence-level accuracy. Taken together, Vietnamese medical CS benefits more from parameter-efficient adaptation—especially LID-guided AG—than from decoder-only contextualization, while contextual normalization remains a useful complement for difficult terms.

To compare monolingual and multilingual initializations under the same adaptations, Table 6 show experiments across two strongest adaptation strategies-LoRA and Attention Guide (AG) on Whisper-small and PhoWhisper-small. Both gain markedly, but PhoWhisper ends up stronger overall and on most CS metrics. With LoRA, PhoWhisper cuts CS-WER by roughly half—more than the reduction observed for Whisper—and also achieves lower WER/CER/N-WER on the test split while keeping an edge on the hard split. AG pushes performance further: PhoWhisper attains the lowest test errors, with CS-WER slightly below Whisper's and sentence-level metrics clearly in its favor; on the hard split, CS-WER is comparable across models, but PhoWhisper retains better WER and N-WER. In short, after fine-tuning, the Vietnamese-optimized backbone learns code-switched medical terms more effectively and delivers consistently stronger accuracy than its multilingual counterpart.

Topics	Frozen	LoRA	AG
Medical Sciences	68.17	41.28	**34.28**
Pathology & Pathogens	63.86	37.07	**26.92**
Treatments	73.74	34.30	**27.44**
Nutrition	53.82	16.92	**13.85**
Diagnostics	54.07	52.01	**44.40**

Table 7: CS-WER by different methods based on PhoWhisper-Small divided by topics.

Moreover, Table 7 shows consistent gains across topics after adaptation. Treatments is the most error-prone category in the frozen model but no longer the worst once fine-tuned, while Nutrition

starts easiest and remains so. Medical Sciences, the largest and most diverse topic, continues to be comparatively challenging even after adaptation, and Diagnostics also trails the middle group. Across all topics, AG yields the lowest CS-WER and LoRA is a close second, indicating that adapter-based methods substantially narrow cross-topic gaps and shift the error peak away from Treatments, though broad scientific content still stresses the model.

5. Conclusion

In this paper, we introduced **ViMedCSS**, the first publicly available benchmark dataset for Vietnamese medical code-switching (CS) speech, containing 34.6 hours and 16,576 utterances. This resource addresses a critical gap in ASR development, as we demonstrated that standard models struggle to recognize English medical terms embedded in Vietnamese. Our zero-shot experiments revealed a clear performance trade-off: multilingual models like Whisper-Large-v3 excel at recognizing English CS terms, whereas Vietnamese-optimized models like VietASR are superior for the surrounding Vietnamese text, resulting in lower overall word error rates.

To resolve this, we investigated several fine-tuning strategies, finding that parameter-efficient adaptation offers the most effective solution. Notably, applying the Attention Guide (AG) adaptation method to a Vietnamese-specialized model (PhoWhisper-Small) yielded the best performance, significantly reducing errors on both CS terms and general speech. This work not only provides a valuable dataset for the community but also identifies a clear and effective fine-tuning approach for building robust, domain-specific ASR models in low-resource and code-switching contexts. Future work can leverage this benchmark to explore further enhancements in contextual biasing and model architecture.

6. Acknowledgments

This research was supported by the VinUniversity Cross-College Research Grant (Grant ID: VUNI.2324.CC06).

7. Ethics Statement

The data were collected from a publicly available source, YouTube. The content extracted from this source is used for research purposes only, and it does not contain any private information about patients.

8. Bibliographical References

Bobbi Aditya, Mahdin Rohmatillah, Liang-Hsuan Tai, and Jen-Tzung Chien. 2024. Attention-Guided Adaptation for Code-Switching Speech Recognition. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10256–10260.

Maha Tufail Agro, Atharva Kulkarni, Karima Kadaoui, Zeerak Talat, and Hanan Aldarmaki. 2025. Code-switching in end-to-end automatic speech recognition: A systematic literature review.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc.

Stacy S Chen. 2025. A "code-switching" model for healthcare communication. *Healthcare Management Forum*, 38(4):391–394.

The Chuong Chu, Vu Tuan Dat Pham, Trung Kien Dao, Ngoc Hoang Nguyen, and Steven Truong. 2025. AdaCS: Adaptive Normalization for Enhanced Code-Switching ASR. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2023. FLEURS: FEW-Shot Learning Evaluation of Universal Representations of Speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 798–805.

Abdullah Ashraf Hamad, Doaa B Mustaffa, Asmaa Zakria Alnajjar, Raghad Amro, Mohammad Ghassab Deameh, Bassant Amin, and Ibraheem M Alkhawaldeh. 2025. Decolonizing medical education: a systematic review of educational language barriers in countries using foreign languages for instruction. *BMC Medical Education*, 25(1):701.

Haoxiang Hou, Xun Gong, Wangyou Zhang, Wei Wang, and Yanmin Qian. 2025. Ranking and Selection of Bias Words for Contextual Bias Speech Recognition. In *Interspeech 2025*, pages 5183–5187.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

Thanh-Thien Le, Linh The Nguyen, and Dat Quoc Nguyen. 2024. PhoWhisper: Automatic Speech Recognition for Vietnamese. In *Proceedings of the ICLR 2024 Tiny Papers track*.

Khai Le-Duc. 2024. VietMed: A Dataset and Benchmark for Automatic Speech Recognition of Vietnamese in the Medical Domain. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17365–17370, Torino, Italia. ELRA and ICCL.

Khai Le-Duc, Phuc Phan, Tan-Hanh Pham, Bach Phan Tat, Minh-Huong Ngo, Thanh Nguyen-Tang, and Truong-Son Hy. 2025. MultiMed: Multilingual Medical Speech Recognition via Attention Encoder Decoder. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 1113–1150, Vienna, Austria. Association for Computational Linguistics.

Hieu-Thi Luong and Hai-Quan Vu. 2016. Vivos: Vietnamese Speech Corpus for ASR.

Dau-Cheng Lyu, Tien-Ping Tan, Eng Siong Chng, and Haizhou Li. 2010. SEAME: a Mandarin-English code-switching speech corpus in south-east asia. In *Interspeech 2010*, pages 1986–1989.

Thai Binh Nguyen. 2021. Vietnamese end-to-end speech recognition using wav2vec 2.0.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, et al. 2024. Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*, 25(97):1–52.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Samer H Sharkiya. 2023. Quality communication can improve patient-centred health outcomes among older patients: a rapid review. *BMC Health Services Research*, 23(1):886.

Xian Shi, Qiangze Feng, and Lei Xie. 2020. The ASRU 2019 Mandarin-English Code-Switching Speech Recognition Challenge: Open Datasets, Tracks, Methods and Results.

Zheshu Song, Jianheng Zhuo, Yifan Yang, Ziyang Ma, Shixiong Zhang, and Xie Chen. 2024. LoRA-Whisper: Parameter-Efficient and Extensible Multilingual ASR. In *Interspeech 2024*, pages 3934–3938.

Yui Sudo, Yusuke Fujita, Atsushi Kojima, Tomoya Mizumoto, and Lianbo Liu. 2025. OWSM-Biasing: Contextualizing Open Whisper-Style Speech Models for Automatic Speech Recognition with Dynamic Vocabulary. In *Interspeech 2025*, pages 5188–5192.

Yui Sudo, Yosuke Fukumoto, Muhammad Shakeel, Yifan Peng, and Shinji Watanabe. 2024. Contextualized Automatic Speech Recognition With Dynamic Vocabulary. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 78–85.

Guangzhi Sun, Xianrui Zheng, Chao Zhang, and Philip C. Woodland. 2023. Can Contextual Biasing Remain Effective with Whisper and GPT-2? In *Interspeech 2023*, pages 1289–1293.

Enes Yavuz Ugan, Ngoc-Quan Pham, and Alexander Waibel. 2024. DECM: Evaluating Bilingual ASR Performance on a Code-switching/mixing Benchmark. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4468–4475, Torino, Italia. ELRA and ICCL.

Nhu Vo, Dat Quoc Nguyen, Dung D. Le, Massimo Piccardi, and Wray Buntine. 2024. Improving Vietnamese-English medical machine translation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8955–8962, Torino, Italia. ELRA and ICCL.

Brian Yan, Injy Hamed, Shuichiro Shimizu, Vasista Sai Lodagala, William Chen, Olga Iakovenko, Bashar Talafha, Amir Hussein, Alexander Polok, Kalvin Chang, Dominik Klement, Sara Althubaiti, Puyuan Peng, Matthew Wiesner, Thamar Solorio, Ahmed Ali, Sanjeev Khudanpur, and Shinji Watanabe. 2025. CS-FLEURS: A Massively Multilingual and Code-Switched Speech Dataset. In *Interspeech 2025*, pages 743–747.

Zengwei Yao, Liyong Guo, Xiaoyu Yang, Wei Kang, Fangjun Kuang, Yifan Yang, Zengrui Jin, Long

Lin, and Daniel Povey. 2024. Zipformer: A faster and better encoder for automatic speech recognition. In *ICLR*.

Jiahui Zhao, Hao Shi, Chenrui Cui, Tianrui Wang, Hexin Liu, Zhaoheng Ni, Lingxuan Ye, and Longbiao Wang. 2025. Adapting Whisper for Code-Switching through Encoding Refining and Language-Aware Decoding. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Jianheng Zhuo, Yifan Yang, Yiwen Shao, Yong Xu, Dong Yu, Kai Yu, and Xie Chen. 2025. VietASR: Achieving Industry-level Vietnamese ASR with 50-hour labeled data and Large-Scale Speech Pretraining. In *Interspeech 2025*, pages 1163–1167.