Speech-to-Text: Translate Dialects to Standard Vietnamese Language

Bang Ly
The Parsing Pals
ly000051@umn.edu
Pre-trained model fine-tuning

Huong Giang To
The Parsing Pals
to000032@umn.edu
Dialect data mining

Khanh Chi Le
The Parsing Pals
le000422@umn.edu
LLMs testing & validation

Abstract

Vietnamese exhibits diverse phonetic variations across regions, posing significant challenges for state-of-the-art speech-to-text (STT) systems. While recent advances in pre-trained models have improved Vietnamese STT overall, their performance on dialectal speech remains limited. To address this issue, our project focuses on developing a pipeline that transcribes dialectal Vietnamese speech into standardized Vietnamese text – instead of providing a direct word-for-word transcription, the system generates standard Vietnamese to help bridge communication gaps within the community. Our work demonstrates the feasibility of this approach, showing that fine-tuning existing models on curated dialectal speech data can significantly improve performance in dialectal STT. In addition, we propose future directions for advancing Vietnamese STT, including more robust model adaptation and dialect-aware dataset development.

1 Introduction

Recent advances in speech recognition have significantly improved the performance of Deep Neural Networks (DNNs) for Speech-to-Text (STT) by leveraging large-scale datasets and high-quality benchmarks. While these systems achieve strong performance in high-resource languages such as English, French, and Mandarin, low-resource languages still suffer from high word error rates (WER) due to limited training data.

Vietnamese is one such low-resource language. Despite being spoken by over 100 million people – making Vietnam the 16th most populous country – and ranking among the most spoken languages outside of English in the U.S., Canada, and Australia, Vietnamese STT systems still lag behind in robustness and generalizability.

A major challenge in Vietnamese STT arises from its rich dialectal diversity (Ahlawat et al., 2025). Although commonly divided into three main dialectal regions – Northern, Central, and Southern – linguistic variation exists at a much finer-grained level. Each of Vietnam's 64 provinces has its own accent, and in many cases, the differences in vocabulary, pronunciation, and tonal patterns are substantial enough to hinder mutual intelligibility. These variations create a significant barrier for current STT systems, which are typically trained on standard Vietnamese.

Recent efforts in Vietnamese NLP and speech processing, including the development of national speech corpora (Nga et al., 2021; Tran et al., 2024) and end-to-end STT systems (Truong et al., 2019), have made important strides. However, these approaches largely treat Vietnamese as a monolithic language, overlooking intra-language dialectal variation.

To address this gap, We aim to make speech recognition systems better at understanding the different ways people speak Vietnamese across regions and convert that into easy-to-understand, standardized Vietnamese. Our project investigates the ability of state-of-the-art Vietnamese STT models to handle dialectal speech. We propose a pipeline that transcribes dialectal Vietnamese speech into standardized Vietnamese text – moving beyond direct phonetic transcription to produce semantically coherent, region-neutral output. This approach not only bridges communication gaps within the Vietnamese community but also enhances accessibility for speakers across dialects and diaspora communities worldwide.

Our work may be of interest to researchers in speech recognition and natural language processing, as well as to those developing Vietnamese speech-to-text systems. By evaluating the performance of current state-of-the-art models

on dialectal input and conducting detailed error analysis, our project provides valuable insights into both the limitations and opportunities in this domain. It also demonstrates potential for broader social impact – improving communication across Vietnamese dialects and supporting accessibility initiatives. Furthermore, we aim to highlight the urgent need for more nuanced dialectal datasets to better facilitate the development of inclusive and effective Vietnamese STT systems.

2 Background

Vietnamese Language Vietnamese is a tonal and monosyllabic language, which means each word usually consists of a single syllable, and the tone plays a major role in determining the word's meaning. A Vietnamese syllable has three parts: the beginning sound (initial), the end sound (final), and the tone.

For example, in the word bạn (friend), b is the initial, an is the final, and the tone . is marked on ạ.

Vietnamese has six tones, and each one changes the meaning of a word. For instance, changing the tone of ba can give you different words like ba (three), bá (uncle), bà (grandmother), bả (poison), bã (waste), and bạ (randomly). Because of this, even a small mistake in tone can lead to a completely different meaning, which makes speech recognition especially difficult (Dinh et al., 2024).

Standard Vietnamese Dialectal Vietnamese On top of these tonal challenges, Vietnamese has a high level of dialectal variation. While Standard Vietnamese is the official version used in government, media, and education, people speak different dialects in their daily lives depending on where they live. The primary differences between Vietnamese dialects are phonological, with some variations in vocabulary (Phạm and McLeod, 2016).

The three main dialect regions are Northern, Central, and Southern. However, the differences go beyond just three regions – each of Vietnam's 64 provinces has its own local accent or way of speaking. In many cases, the pronunciation, vocabulary, and tone patterns can be so different that even native speakers have trouble understanding each other (Alves, 2007).

Studies have shown that the Northern, Central, and Southern dialects differ in five main ways (Phạm and McLeod, 2016): how many consonants they use and where they appear in words, how they use certain vowels and diphthongs, whether they include semivowels, and even the number of tones (for example, some dialects use 5 tones instead of 6).

These dialectal differences make it harder for current speech-to-text systems – typically trained on Standard Vietnamese – to recognize or transcribe regional speech accurately. As a result, Vietnamese is a particularly challenging language for building effective speech technology, both because of its tonal structure and because of how much it varies across regions.

3 Related Work

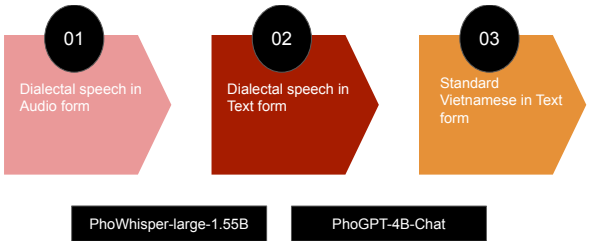
In recent years, Vietnamese speech-related research has made remarkable strides (Nguyen et al., 2017). One of the first efforts involved building a Vietnamese speech corpus and recognition system for the customer service domain (Nguyen et al., 2017). In parallel, another ASR system was developed using a large dataset that included various accents from Northern, Central, and Southern Vietnam (Truong et al., 2019).

Building on these efforts, PhoWhisper (Le et al., 2024) – a fine-tuned version of OpenAI's Whisper model – was trained on a large-scale Vietnamese ASR dataset with diverse regional accents. It has achieved state-of-the-art performance on Vietnamese ASR benchmarks.

Despite these advances, dialectal variation remains a major challenge in Vietnamese ASR systems (Nga et al., 2021; Phung et al., 2024). To address this, several multi-dialect corpora have been developed in recent years (Tran et al., 2024; Nguyen et al., 2023a). However, these datasets often have two key limitations: (1) they cover only three to five major dialect groups, and (2) many of them are not publicly accessible.

The ViMD speech dataset (Dinh et al., 2024) was recently introduced to overcome these limitations. It is the first resource to include speech data from all 63 provinces of Vietnam, offering much finer-grained dialectal coverage.

To our knowledge, no existing work has proposed an end-to-end system for transcribing dialectal Vietnamese speech into standardized Vietnamese text. Our project addresses this gap by developing



Hình 1: End-to-End Dialect-to-Standard Vietnamese Transcription Pipeline

such a system and evaluating its effectiveness on dialectal inputs. We aim to explore how fine-tuning ASR models with dialect-specific data can improve transcription quality and bridge communication gaps within the Vietnamese-speaking community.

4 Methodology

4.1 End-to-End Dialect-to-Standard Vietnamese Transcription Pipeline

We propose a two-stage pipeline for transcribing dialectal Vietnamese speech into standardized Vietnamese text (Figure 1). The first stage involves converting speech audio into a dialectal transcription using a pre-trained ASR model. The second stage translates this dialectal text into standard Vietnamese using a large language model.

Specifically, we use PhoWhisper-large-1.55B (Le et al., 2024) for automatic speech recognition (ASR) of dialectal audio. The resulting transcription is then passed to PhoGPT-4B-Chat (Nguyen et al., 2023b), which generates the equivalent standardized Vietnamese text.

To our knowledge, this is the first end-to-end system that translates dialectal Vietnamese speech into standardized text using a combination of fine-tuned ASR and large language models. Unlike prior work that only transcribes speech as-is, we focus on generating region-neutral, intelligible Vietnamese, addressing an overlooked yet critical need in Vietnamese NLP

4.2 Experiment Design

The goal of our experiments is to evaluate the performance of this pipeline under various conditions and investigate the potential of fine-tuning for improving dialectal transcription. Our design includes:

- Evaluating the baseline pipeline without any fine-tuning.
- Fine-tuning PhoWhisper-large-1.55B on dialectal audio from the Multi-Dialect Vietnamese dataset (Dinh et al., 2024).
- Comparing our fine-tuned pipeline with existing state-of-the-art systems, including OpenAI Whisper and GPT-4.

Our hypothesis is that fine-tuning PhoWhisper on dialect-specific data will significantly improve its transcription accuracy and semantic fidelity over the pre-trained model, especially for phonetically challenging dialects such as Nghệ An. We further hypothesize that combining this improved transcription with a strong Vietnamese language model like PhoGPT will yield standardized outputs that better preserve meaning across dialects.

4.3 Dataset and Dialect Selection

We use the Multi-Dialect Vietnamese (ViMD) dataset (Dinh et al., 2024), which is composed of 102.56 hours of data, representing 63 dialects in Vietnam. It includes audio, transcription, and province-level dialect labels. This allows us to both fine-tune and evaluate our models with dialect-specific granularity.

Given the limited time frame, we focus on four representative dialects to test the model's generalizability:

Level 1 (Easy) Standard Vietnamese, to verify model performance in a non-dialectal setting. The specific dialect used for this Ha Noi dialect.

Level 2 (Medium - Lexical) Huế dialect (Central Vietnam), selected for its distinct vocabulary.

Level 3 (Medium - Tonal) Hồ Chí Minh City dialect (Southern Vietnam), chosen for its tonal shifts.

Level 4 (Hard) Nghệ An/Hà Tĩnh dialect, widely regarded as the most difficult due to heavy tone merging and extensive vocabulary divergence from the standard.

4.4 Model Choices

PhoWhisper-large (1.55B) This ASR model, developed by VinAI, is pre-trained on a large-scale Vietnamese speech dataset and has achieved state-of-the-art results on Vietnamese ASR benchmarks such as VLSP Task-1, VLSP Task-2 (Nguyen et al., 2020), and VIVO (Luong and Vu, 2016). We evaluate both its original (pre-trained) version and

a fine-tuned version trained on the ViMD dataset. This serves as the ASR backbone of our pipeline.

PhoGPT-4B-Chat We use this large language model for converting dialectal transcriptions to standardized Vietnamese text. PhoGPT-4B-Chat outperforms other Vietnamese-specific LLMs (e.g., Vistral-7B-Chat (Chien Van Nguyen, 2023), SeaLLM-7B-v2 (Xuan-Phi Nguyen*, 2023), Sailor-7B-Chat (Dou et al., 2024), VBD-LLaMA2-7B-50b-Chat (LR-AI-Labs, 2023)) as well as closed models such as GPT-4 and Gemini Pro on the ViTruthfulQA benchmark (Nguyen et al., 2023c). Due to the absence of a supervised dataset for dialect-to-standard Vietnamese text rewriting and time constraints, we do not fine-tune this model.

OpenAI Whisper-large + GPT-4o To establish a performance benchmark, we evaluate a comparative pipeline using OpenAI's Whisper-large for ASR and GPT-4 for text normalization. These models represent the global state-of-the-art in speech and language processing.

5 Experiments and Results

5.1 Evaluation Metrics

To evaluate the performance of our pipeline, we assess each step separately to better understand where improvements are needed.

For the first step – transcribing dialectal speech into dialectal text – we use Word Error Rate (WER) to measure transcription accuracy. WER is a standard metric in speech recognition that calculates the proportion of insertions, deletions, and substitutions required to align the model output with the reference transcription. We also use BERTScore, which measures semantic similarity between the model output and the reference using contextual embeddings. This allows us to evaluate whether the meaning is preserved, even when the exact words differ. Using both WER and BERTScore provides a more comprehensive assessment of transcription quality by capturing both literal accuracy and semantic fidelity.

For the second step – converting dialectal text into standardized Vietnamese – we peform manual evaluation. As all team members are native Vietnamese speakers, we assess the outputs based on fluency, meaning preservation, and alignment with the reference standardized text.

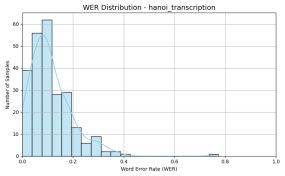
By evaluating each step separately, we can identify which component of the pipeline requires further improvement and gain more insights for future improvements.

5.2 Testing Pre-trained PhoWhisper-large-1.55B without finetune

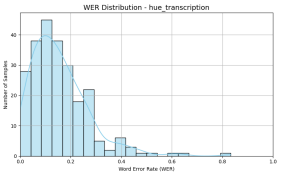
We began by evaluating how the pre-trained PhoWhisper-large-1.55B model performs on dialectal speech before applying any fine-tuning. Our hypothesis was that the model would struggle moderately with Level 2 and Level 3 dialects, and significantly with Level 4.

Bảng 1: Initial model WER Rate & BERT Score

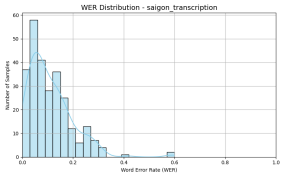
	Ha Noi	Hue	Saigon	Nghe An
WER	0.1141	0.1567	0.1105	0.1285
BERT	0.9176	0.9057	0.9215	0.9124



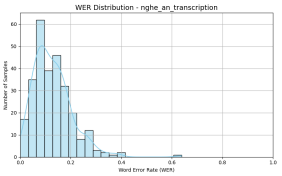
(a) WER Hanoi



(b) WER Hue



(c) WER Sai Gon



(d) WER Nghe An

Hình 2: WER for each province

The results of this preliminary evaluation are shown in Figure 2 and Table 1. We observed that PhoWhisper performs reasonably well in transcribing dialectal audio, with only around 10–15% word error rate between the output and reference transcriptions. Surprisingly, the Nghệ An dialect (Level 4), which we expected to be the most challenging, was not the worst-performing, and the Hà Nội dialect (Level 1) was not the most accurate.

One possible explanation lies in the composition of the ViMD dataset. Much of the audio is sourced from news anchors, where pronunciation is often more standardized, and dialectal variation – particularly in vocabulary and tonal shifts – is less pronounced. As a result, the dataset contains

relatively few challenging dialectal cases, which may limit the model's exposure to distinctive phonetic or lexical differences. This could explain why PhoWhisper performs similarly across dialects and maintains relatively low error rates overall.

From this point, we considered two potential directions: (1) curating a new dataset with more representative and dialect-rich audio samples, particularly emphasizing tonal variation and region-specific vocabulary; or (2) fine-tuning PhoWhisper on the existing ViMD dataset to explore whether model performance could still be improved. While PhoWhisper's baseline error rate was relatively low, an 11% WER is still significant in applications requiring precise transcription accuracy. After consulting with our mentor, we chose the second option due to the limited project timeline and the potential for meaningful improvements through fine-tuning.

5.3 Fine-tuning PhoWhisper-large-1.55B

Fine-tuning on all 4 dialects

We began our fine-tuning experiments by using 830 samples for training, 104 for validation, and 104 for testing, drawn from all four dialects. The model was fine-tuned using the AdamW optimizer with a learning rate of 1e-5 over five epochs.

Bảng 2: Post fine-tuning WER Rate & BERT Score

	Ha Noi	**Hue**	**Saigon**	**Nghe An**
WER	0.5785	0.2134	0.3522	0.4424
BERT	0.8505	0.9088	0.8797	0.8543

However, the fine-tuned model exhibited significantly worse performance compared to the pre-trained version. While the original model achieved word error rates (WER) between 10% and 15%, the fine-tuned model's WER ranged from 21% to 57%. Notably, the Hà Nội dialect, which previously performed best, now yielded the highest WER at 57%, indicating substantial degradation.

Upon further analysis, we observed that while the training loss consistently decreased, the validation loss began increasing after the first epoch. This divergence suggests that the model was overfitting to the training data and failing to generalize to unseen dialectal speech.

Our hypothesis is that PhoWhisper requires more extensive fine-tuning before any conclusions can be drawn about its performance relative to the original model. However, given the time constraints of this project, we decided to narrow our scope and focus on a single dialect to allow for more thorough fine-tuning. We selected the Nghệ An dialect (Level 4), as it is widely regarded as the most challenging. We believe that if performance can be improved on this dialect, similar improvements are likely achievable for other, less complex dialects.

Fine-tuning on Nghe An dialect

To fine-tune the model on a single dialect (Nghệ An), we experimented with multiple training setups by varying the optimizer type, batch size, learning rate, and train-validation-test data split. In total, we tested 24 unique configurations by combining different values for each of these parameters.

The hyperparameter values we used are as follows:

- Optimizer: Adam, AdamW, SGD
- Batch size: 3, 12
- Learning rate: 1e-5, 1e-6
- Train-validation-test split: 80-10-10, 90-5-5

These choices were based on a combination of insights from prior research and practical considerations related to our dataset and computational constraints. Adam and AdamW are widely used in fine-tuning transformer-based models due to their adaptive learning rate mechanisms and training stability. SGD was included as a baseline to compare against more advanced optimizers.

The batch sizes of 3 and 12 were selected to balance training stability with GPU memory limitations, allowing us to observe how smaller versus moderately larger update steps affect learning. Learning rates of 1e-5 and 1e-6 are commonly used in fine-tuning pre-trained models, as they offer a good trade-off between learning speed and model stability, while reducing the risk of catastrophic forgetting.

We also varied the data split ratios to explore the trade-off between training data quantity and evaluation reliability. The 80-10-10 split provides balanced coverage across training, validation, and test sets, while the 90-5-5 split maximizes training data at the cost of less validation feedback. This variation helps us assess generalization and potential overfitting.

Bảng 3: WER and BERTScore under different training configurations

Split	Batch	Optimizer	LR	WER	BERT
80-10-10	3	AdamW	1e-5	0.1521	0.9389
			1e-6	0.2442	0.9117
		Adam	1e-5	0.1515	0.9425
			1e-6	0.2753	0.9043
		SGD	1e-5	0.2165	0.9155
			1e-6	0.2188	0.9147
	12	AdamW	1e-5	0.1522	0.9366
			1e-6	0.2176	0.9206
		Adam	1e-5	0.1458	0.9382
			1e-6	0.2165	0.9200
		SGD	1e-5	0.2182	0.9147
			1e-6	0.2188	0.9147
90-5-5	3	AdamW	1e-5	0.1755	0.9350
			1e-6	0.2194	0.9202
		Adam	1e-5	0.1665	0.9361
			1e-6	0.2194	0.9233
		SGD	1e-5	0.2452	0.9129
			1e-6	0.2452	0.9122
	12	AdamW	1e-5	0.1716	0.9368
			1e-6	0.2194	0.9226
		Adam	1e-5	0.1729	0.9354
			1e-6	0.2245	0.9212
		SGD	1e-5	0.2452	0.9122
			1e-6	0.2465	0.9121

To obtain a comprehensive view of performance changes across models, we first calculated the WER and BERTScore for the pre-trained PhoWhisper model, as shown in Table 4.

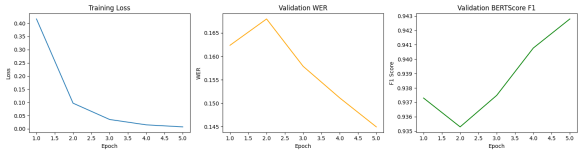
Bảng 4: Pre-trained model's WER and BERTScore by Train-Validation-Test Split

Train-Validation-Test Split	WER Score	BERTScore
80-10-10	0.2188	0.9147
90-5-5	0.2465	0.9121

These baseline scores provide a reference point to compare the fine-tuned models and highlight performance improvements or regressions. Table 5 summarizes the relative performance changes across the 24 fine-tuning configurations.

Bảng 5: Percentage change in WER and BERTScore relative to baseline

Split	Batch	Optimizer	LR	WER % Change	BERT % Change
80-10-10	3	AdamW	1e-5	-30.48	2.65
			1e-6	11.61	-0.33
		Adam	1e-5	-30.76	**3.04**
			1e-6	25.82	-1.14
		SGD	1e-5	-1.05	0.09
			1e-6	0.00	0.00
	12	AdamW	1e-5	-30.39	2.39
			1e-6	-0.55	0.65
		Adam	1e-5	**-33.37**	2.57
			1e-6	-1.05	0.58
		SGD	1e-5	-0.27	0.00
			1e-6	0.00	0.00
90-5-5	3	AdamW	1e-5	-19.79	2.22
			1e-6	0.27	0.60
		Adam	1e-5	-23.91	2.34
			1e-6	0.27	0.94
		SGD	1e-5	12.06	-0.20
			1e-6	12.06	-0.27
	12	AdamW	1e-5	-21.56	2.42
			1e-6	0.27	0.87
		Adam	1e-5	-20.99	2.26
			1e-6	2.61	0.71
		SGD	1e-5	12.06	-0.27
			1e-6	12.66	-0.28



Hình 3: Training loss, validation WER, and validation BERTScore F1 over epochs for the PhoWhisper model fine-tuned with Adam optimizer (learning rate = 1e-5), batch size = 3, and 80-10-10 train-validation-test split.

Hình 4: Training loss, validation WER, and validation BERTScore F1 over epochs for the PhoWhisper model fine-tuned with Adam optimizer (learning rate = 1e-5), batch size = 12, and 80-10-10 train-validation-test split.

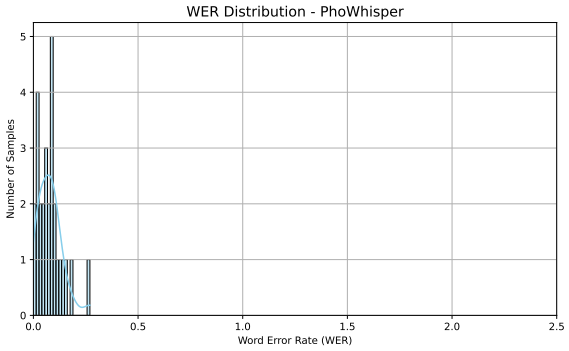
From Table 5, the best-performing model in terms of semantic similarity (BERTScore) was fine-tuned using the Adam optimizer, a batch size of 3, a learning rate of 1e-5, and a train-validation-test split of 80-10-10, achieving a BERTScore of 94%, representing a >3% improvement over the baseline. Meanwhile, the best WER improvement came from the same configuration but with a batch size of 12, resulting in a WER of 14%, a 33% reduction compared to the pre-trained model.

Figure 3 and Figure 4 show training and validation loss trends for the best-performing models.

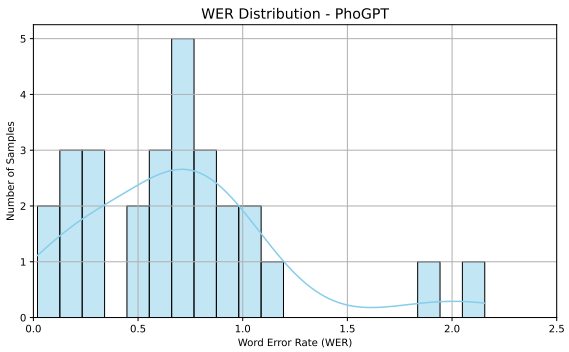
Based on this exhaustive fine-tuning process, we conclude that the best-performing setup is PhoWhisper-large (1.55B) fine-tuned using the Adam optimizer with a learning rate of 1e-5. Increasing the batch size introduces a trade-off between maximizing BERTScore and minimizing WER, suggesting different optimization paths depending on which metric is prioritized. In general, models trained with a 80-10-10 split consistently outperformed those trained with a 90-5-5 split, likely due to more reliable validation and test evaluation.

5.4 Evaluation of the full pipeline ffter fine-tuning

Using the best-performing fine-tuned model from our previous experiments, we tested the complete pipeline by passing its dialectal transcription output to PhoGPT-4B-Chat. To prompt PhoGPT for dialect-to-standard Vietnamese translation, we used the instruction "Dịch câu này sang tiếng Việt phổ thông"which translates to "Translate this sentence into standard Vietnamese."



(a) PhoWhisper WER



(b) PhoGPT WER

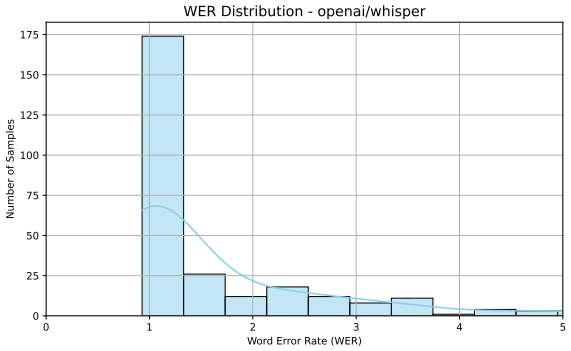
Hình 5: WER distribution before and after translating to standard Vietnamese

On the test set using an 80-10-10 split, the average WER for the dialectal transcription produced by PhoWhisper was 0.0796. After passing the transcription to PhoGPT for standardization, the WER increased significantly to 0.7022. While some increase in WER is expected – since PhoGPT outputs standard Vietnamese rather than a word-for-word match – the magnitude of this increase was unexpectedly high.

Upon manual inspection, we found that PhoGPT did not consistently translate dialectal words into their standard equivalents. Instead, it often avoided unfamiliar dialect-specific terms by omitting them entirely or rephrasing the sentence structure without altering its core meaning. This behavior suggests that PhoGPT prioritizes grammatical fluency and sentence-level coherence over word-level fidelity.

5.5 OpenAI-Whisper + GPT4 pipeline baseline

We evaluated the performance of OpenAI Whisper on the full Nghệ An dialect subset, consisting of 280 audio recordings. Despite explicitly specifying the language as Vietnamese, the model produced incoherent outputs that included a mix of Vietnamese and unrelated foreign words. The resulting transcriptions were largely unintelligible and inconsistent, rendering them unsuitable for downstream processing.



Hình 6: WER of OpenAI Whisper on Nghệ An dialect speech.

As shown in Figure 6, the Word Error Rate (WER) for this experiment was 1.741, indicating severe transcription failure. Given the extremely poor performance, we decided not to proceed with the second step of the pipeline, which would have involved feeding the transcriptions into GPT-4o for standardization.

Thus, our findings suggest that OpenAI Whisper

is not robust to dialectal Vietnamese speech, particularly in challenging cases such as Nghệ An. This highlights the importance of using models that are specifically trained or fine-tuned on Vietnamese data when working on dialect-sensitive STT tasks.

The transcription outputs generated by OpenAI Whisper for the Nghệ An dialect can be found in our project's GitHub repository.

6 Error Analysis

6.1 PhoWhisper-large-1.55b

To understand further how the best fine-tuned PhoWhisper models performs, we will conduct a qualitative and quantitative error analysis on the models output. The 2 models's output that we will analyze are:

Bảng 6: Best-performing fine-tuned models by data split

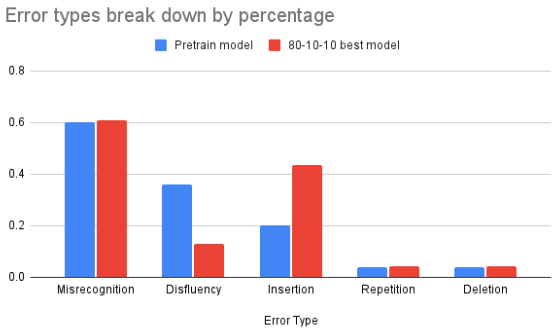
Split	Optimizer	Learning Rate	Batch Size
80-10-10	Adam	1e-5	12
90-5-5	Adam	1e-5	3

The most common errors are:

- **Misrecognition:** Some words were misrecognized and substituted with similar-sounding but incorrect words.
- **Disfluencies:** Irrelevant words were inserted. Common insertions included filler sounds such as "à", "thì", "ấy", and "ờ".
- **Insertion:** Irrelevant words were inserted.
- **Repetition:** Unnecessarily repeated words or phrases, resulting in unnatural output. Examples: "là là là", "đã đã đã"
- **Deletion:** Key words or phrases were omitted entirely. In several cases, the transcribed text was noticeably shorter than the reference text.

80-10-10 split best model:

The make up of each error type is:

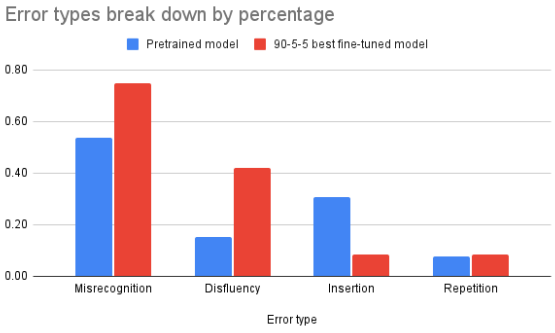


Hình 7: Error breakdown for 80-10-10 split best model and pretrain model

Compared to the pretrain model, the fine-tuned model doesn't make as many mistakes in Disfluency. However, the fine-tuned model makes more misrecognition and significantly more insertion mistakes. Between the output of the pretrain model and the finetuned model, many outputs from the same audio has the same mistakes. For example, a misrecognized mistake that is made frequently is:

- "những công việc ở sát thành phố"(original text)
- "những công việc ở pháp thành phố"(both pretrain and fine-tuned model have this error)

90-5-5 split best model:



Hình 8: Error breakdown for 90-5-5 split best model and pretrain model

The fine-tuned model is better in avoiding insertion errors compared to pretrain model. Specifically, the fine-tuned models makes insertion errors 8% of the time and the pretrain model makes insertion errors 31% of the time. However, the fine-tune model performs significant worse in making more misrecognition and disfluency mistakes in

comparison to the pretrain model.

Overall, based on thorough error analysis of the outputs from PhoWhisper-large (1.55B) fine-tuned model on Adam optimizer, learning rate 1e-5. It is uncertain that the finetuned model is absolutely better than the pretrained model due to inconsistent improvements.

7 Discussion

7.1 Replicability

Our results are moderately replicable. All code, model configurations, and evaluation scripts are available in our GitHub repository, and we have documented all hyperparameter choices and experimental setups. However, due to the large size of the PhoWhisper model and the GPU requirements for training and evaluation, reproducing our experiments may be challenging for researchers without access to substantial computational resources.

7.2 Dataset Limitations and Community Contribution

While the Multi-Dialect Vietnamese (ViMD) dataset marks a significant advancement by covering all 63 provinces, its dialectal representation remains shallow. Most audio samples are derived from formal sources, such as news anchors, where regional variation is minimal and standardized speech is prioritized. As a result, critical aspects of dialectal identity – such as unique vocabulary, idiomatic expressions, regional syntax, and tonal shifts – are underrepresented.

This limitation affects both model training and evaluation, as models fine-tuned on such data cannot fully learn the nuances that differentiate dialects in real-world usage. As Vietnamese STT research advances, we believe the next step is to develop smaller but deeper datasets focused on individual dialects. These datasets should prioritize natural, spontaneous speech and rich regional variation, rather than aiming solely for broad provincial coverage. We encourage researchers and local communities to contribute to the creation of such resources to promote better dialect modeling. While we did not create a new dataset, our experiments highlight the limitations of ViMD and suggest new directions for dataset design. We hope future work will build on our evaluation criteria and benchmarks to improve dialect modeling.

7.3 Model Limitations and Future Directions

Our pipeline reveals several key limitations. First, although fine-tuning on ViMD improved performance, PhoWhisper still exhibits difficulty in generalizing to highly informal or phonologically distant dialectal input. This highlights the need for multi-style or conversational training data and potentially more dialect-aware architectures.

Second, we found that the performance of the language model (PhoGPT) is highly sensitive to the prompt. While our prompt—"Dịch câu này sang tiếng Việt phổ thông"– generally worked, it sometimes led to unexpected behaviors, such as omitting unknown words or paraphrasing rather than translating dialect-specific terms. Future work may benefit from instruction tuning, few-shot learning, or prompt engineering using dialect–standard pairs to guide generation more reliably.

Additionally, our two-stage pipeline is fragile: errors in the ASR step propagate directly to the standardization step. A more robust future direction could involve joint modeling, where transcription and standardization are learned together within a unified architecture.

Evaluation also presents challenges. While WER and BERTScore provide useful signals, they may not fully reflect the quality of dialect-to-standard text transformation. These tasks require a balance between fidelity and fluency, which cannot always be captured by existing metrics. Human evaluation, task-specific metrics, or semantic preservation scoring may offer better insights.

7.4 Ethical Considerations

While standardizing dialectal speech can increase accessibility and interoperability, it also raises ethical concerns. Excessive focus on standardization may contribute to the marginalization or erasure of regional dialects, especially if the tools are widely deployed in education, media, or government systems. It is important to preserve linguistic diversity by promoting inclusive STT systems that also support dialectal speech recognition or offer bidirectional translation between dialects and standard Vietnamese. Ensuring transparency, consent, and community involvement in dataset collection is also critical to addressing representation bias and promoting equitable language technology.

References

Harsh Ahlawat, Naveen Aggarwal, and Deepti Gupta. 2025. Automatic speech recognition: A survey of deep learning techniques and approaches. *International Journal of Cognitive Computing in Engineering*, 6:201–237.

Mark J Alves. 2007. A look at north-central vietnamese. In *SEALS XII: Papers from the 12th meeting of the Southeast Asian Linguistics Society (2002)*, pages 1–7.

Quan Nguyen Huy Nguyen Björn Plüster Nam Pham Huu Nguyen Patrick Schramowski Thien Nguyen Chien Van Nguyen, Thuat Nguyen. 2023. Vistral-7b-chat - towards a state-of-the-art large language model for vietnamese.

Nguyen Van Dinh, Thanh Chi Dang, Luan Thanh Nguyen, and Kiet Van Nguyen. 2024. Multi-dialect Vietnamese: Task, dataset, baseline models and challenges. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7476–7498, Miami, Florida, USA. Association for Computational Linguistics.

Longxu Dou, Qian Liu, Guangtao Zeng, Jia Guo, Jiahui Zhou, Xin Mao, Ziqi Jin, Wei Lu, and Min Lin. 2024. Sailor: Open language models for south-East Asia. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*.

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*.

LR-AI-Labs. 2023. Vbd-llama2-7b-50b-chat: A conversationally-tuned llama2 for vietnamese. https://huggingface.co/LR-AI-Labs/vbd-llama2-7B-50b-chat. Accessed: 2025-05-08.

Hieu-Thi Luong and Hai-Quan Vu. 2016. A non-expert Kaldi recipe for Vietnamese speech recognition system. In *Proceedings of the Third International Workshop on Worldwide Language Service Infrastructure and Second Workshop on Open Infrastructures and Analysis Frameworks for Human Language Technologies (WLSI/OIAF4HLT2016)*, pages 51–55, Osaka, Japan. The COLING 2016 Organizing Committee.

Cao Hong Nga, Chung-Ting Li, Yung-Hui Li, and Jia-Ching Wang. 2021. A survey of vietnamese automatic speech recognition. In *2021 9th International Conference on Orange Technology (ICOT)*, pages 1–4. IEEE.

Binh Nguyen, Son Huynh, Quoc Khanh Tran, An Le Tran-Hoai, Trong An Nguyen, Nguyen Tung Doan Tran, Thuy An Phan Thi, Hieu Nghia Nguyen, Dang Huynh, et al. 2023a. Viasr: A novel benchmark dataset and methods for vietnamese automatic speech recognition. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 387–397.

Dat Quoc Nguyen, Linh The Nguyen, Chi Tran, Dung Ngoc Nguyen, Dinh Phung, and Hung Bui. 2023b. PhoGPT: Generative Pre-training for Vietnamese. *arXiv preprint*, arXiv:2311.02945.

Minh Thuan Nguyen, Khanh Tung Tran, Nhu Van Nguyen, and Xuan-Son Vu. 2023c. ViGPTQA - state-of-the-art LLMs for Vietnamese question answering: System overview, core models training, and evaluations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 754–764, Singapore. Association for Computational Linguistics.

Quoc Bao Nguyen, Van Hai Do, Ba Quyen Dam, and Minh Hung Le. 2017. Development of a vietnamese speech recognition system for viettel call center. In *2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*, pages 1–5.

Thi Thu Trang Nguyen, Hoang Ky Nguyen, Quang Minh Pham, and Duy Manh Vu. 2020. Vietnamese text-to-speech shared task VLSP 2020: Remaining problems with state-of-the-art techniques. In *Proceedings of the 7th International Workshop on Vietnamese Language and Speech Processing*, pages 35–39, Hanoi, Vietnam. Association for Computational Lingustics.

Trung-Nghia Phung, Duc-Binh Nguyen, and Ngoc-Phuong Pham. 2024. A review on speech recognition for under-resourced languages: A case study of vietnamese. *International Journal of Knowledge and Systems Science (IJKSS)*, 15(1):1–16.

Ben Phạm and Sharynne McLeod. 2016. Consonants, vowels and tones across vietnamese dialects. *International Journal of Speech-Language Pathology*, 18(2):122–134. PMID: 27172848.

Linh Thi Thuc Tran, Han-Gyu Kim, Hoang Minh La, and Su Van Pham. 2024. Automatic speech recognition of vietnamese for a new large-scale corpus. *Electronics*, 13(5):977.

Do Quoc Truong, Pham Ngoc Phuong, Tran Hoang Tung, and Luong Chi Mai. 2019. Development of high-performance and large-scale vietnamese automatic speech recognition systems. *Journal of Computer Science and Cybernetics*, 34(4):335–348.

Xin Li* Mahani Aljunied* Zhiqiang Hu Chenhui Shen$^{Y}ewKenChia^{X}ingxuanLiJianyuWangQingyuTanLiyingCh$
$PhiNguyen*, Wenxuan Zhang*$.2023. $Seallms-largelanguagemodelsforsoutheastasia$.