Multi-Dialect Vietnamese: Task, Dataset, Baseline Models and Challenges

Nguyen Van Dinh[1,2]**, Thanh Chi Dang**[1,2]**, Luan Thanh Nguyen**[1,2]**, Kiet Van Nguyen**[1,2]
[1]Faculty of Information Science and Engineering, University of Information Technology,
Ho Chi Minh City, Vietnam
[2]Vietnam National University, Ho Chi Minh City, Vietnam
{20520657, 20520761}@gm.uit.edu.vn
{luannt, kietnv}@uit.edu.vn

Abstract

Vietnamese, a low-resource language, is typically categorized into three primary dialect groups that belong to Northern, Central, and Southern Vietnam. However, each province within these regions exhibits its own distinct pronunciation variations. Despite the existence of various speech recognition datasets, none of them has provided a fine-grained classification of the 63 dialects specific to individual provinces of Vietnam. To address this gap, we introduce **V**ietnamese **M**ulti-**D**ialect (**ViMD**) dataset, a novel comprehensive dataset capturing the rich diversity of 63 provincial dialects spoken across Vietnam. Our dataset comprises 102.56 hours of audio, consisting of approximately 19,000 utterances, and the associated transcripts contain over 1.2 million words. To provide benchmarks and simultaneously demonstrate the challenges of our dataset, we fine-tune state-of-the-art pre-trained models for two downstream tasks: (1) Dialect identification and (2) Speech recognition. The empirical results suggest two implications including the influence of geographical factors on dialects, and the constraints of current approaches in speech recognition tasks involving multi-dialect speech data. Our dataset is available for research purposes[1].

1 Introduction

The Vietnamese language can be divided into three major dialects (regional dialects): Northern, Central, and Southern dialect (Thanh Phu’o’ng, 1982; Hoàng, 2009; Phạm and McLeod, 2016), each associated with distinct region and characterized by unique phonetic characteristic (Pham, 2013; Shimizu, 2013; Hung et al., 2019; Ta et al., 2024). However, even within these regions, the dialects unique to each province (provincial dialect) maintain noticeable differences (Alves, 2007; Shimizu, 2013). Therefore, to obtain a more detailed insight into the Vietnamese dialects, we need to consider a more granular level, particularly provincial dialects, rather than limiting our examination to regional dialects alone. It is worth noting that Vietnamese is a monosyllabic language in which each word is a single syllable and words are separated by spaces. Moreover, in the Vietnamese context, the terms ‘accent’ and ‘dialect’ are used interchangeably (Ta et al., 2024) and consequently, for the sake of consistency, this research adopts the term ‘dialect’.

In recent years, Vietnamese speech-related research has made remarkable strides; however, the issue of multi-dialectal variations within the language has posed a significant challenge (Nga et al., 2021; Hung et al., 2016a; Phung et al., 2024). In an effort to tackle this obstacle, several multi-dialect Vietnamese corpora have been published (Le et al., 2004; Tran et al., 2024; Nguyen et al., 2023; Hung et al., 2016b). However, these datasets exhibit two limitations: (1) All of the mentioned datasets encompass only three to five groups of dialects, (2) Several datasets are not publicly available.

Motivated by these two limitations, we introduce the ViMD speech dataset, a novel resource encompassing 63 provincial dialects, representing all 63 provinces of Vietnam. The audio data is curated from publicly available sources. The transcripts undergo a semi-automatic labeling process, followed by a rigorous manual verification to guarantee the quality of dataset. Furthermore, each record includes additional attributes such as speaker identification codes and gender, allowing the dataset to support various speech-related tasks.

The contributions of our study are as follows:

- We release the first comprehensive multi-dialect Vietnamese speech dataset, offering a fine-grained classification of the 63 dialects, with each dialect being unique to a

[1] https://github.com/nguyen-dv/ViMD_Dataset

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 7476–7498
November 12-16, 2024 ©2024 Association for Computational Linguistics

specific province of Vietnam. The dataset comprises 102.56 hours of audio recordings, nearly 19,000 utterances, and over 1.2 million syllables.

- We conduct experiments on two tasks: dialect identification (DI) and speech recognition (SR) to provide benchmarks and demonstrate the challenging nature of the dataset.
- Based on the experimental results, we present two in-depth analyses, including the impact of geographical factors on dialects and the limitations of the speech recognition approach when dealing with multi-dialect speech data.

2 Basic Phonetic Structure of Vietnamese

Vietnamese, as a monosyllabic tonal language, is structured with three key components: Initial, Final, and Tone. In particular, the Final segment is composed of three elements: Onset, Nucleus, and Coda. These components are detailed in Table 1. The two mandatory elements to construct a syllable, highlighted in bold, are Tone and Nucleus. For instance, the word ‘bạn’ (friend) exemplifies all three key components, where ‘b’ is the Initial, ‘an’ is the Final, and the High-Broken tone is represented by the diacritic below the Nucleus ‘a’. On the other hand, the word ‘u’ (lump) has only the two mandatory elements, with ‘u’ as the Nucleus and the Mid Tone.

There are six tones in Vietnamese: Mid Tone, High-Rising, Low-Falling, Low-Rising, High-Broken, and Low-Broken. Each tone, when combined with a syllable, conveys a distinct meaning. For instance, considering a word with the initial consonant ‘b’ and the nucleus ‘a’: Mid Tone (ba - three), High-Rising (bá - uncle), Low-Falling (bà - grandmother), Low-Rising (bả - poison), High-Broken (bã - waste), and Low-Broken (bạ - randomly). The differences in Pitch Contour between the tones are represented in Table 2 (Phạm and McLeod, 2016).

Tone			
Initial	Final		
	Onset	**Nucleus**	Coda

Table 1: Structure of Vietnamese syllable.

Pronunciation across provinces and regions in Vietnam has its own distinct characteristics. These differences can appear in any component of the syllable. We present more details about these variations in Appendix A, and their impact on the Speech Recognition task in Section 5.6.

Pitch Contour	**Flat**	**UnFlat**	
		Broken	Unbroken
High	No mark	High-broken	High-rising
Low	Low-falling	Low-rising	Low-broken

Table 2: Structure of Vietnamese tones.

3 Related Work

Numerous speech corpora spanning diverse linguistic and dialectical backgrounds have been curated to facilitate dialect classification and speech recognition tasks.

3.1 Global Corpora

QASR (Mubarak et al., 2021), a large-scale Arabic speech and transcription dataset, contains 2,000 hours of 16kHz audio recordings from Aljazeera news channel. It covers 5 Arabic dialects with samples from 19,000 speakers, making it valuable for speech recognition and dialect classification research. The dataset exhibits a gender imbalance (69% male, 6% female), with gender unidentified for speakers having fewer than 20 audio samples.

KeSpeech (Tang et al., 2021) comprises 1,542 hours of recorded speech from 27,237 speakers in 34 cities, covering standard Mandarin and its eight subdialects. This extensive dataset supports a variety of speech processing tasks such as speech recognition, speaker recognition, and sub-dialect identification, promoting the development of multitask learning and conditional learning models. Notably, KeSpeech’s parallel recording of standard Mandarin and specific sub-dialects opens new avenues for applications like dialectal voice conversion.

STT4SG-350 (Plüss et al., 2023) represents a substantial advancement in speech technology resources for Swiss German dialects. A total of 343 hours of recordings spread across seven dialects are included in this dataset, featuring 217,687 unique sentences voiced by 316 speakers. This corpus, annotated with standard German text at the sentence level, addresses the challenges of each dialect and has a large vocabulary size of 42,980.

Thai Dialect Corpus (Suwanbandit et al., 2023b) approximately 840 hours of recordings:

Thai-central with 700 hours of the main Thai dialect; Thai-dialect comprising three Thai dialects including Khummuang, Korat, and Pattani recorded from local people from three corresponding regions: North, Northeast, and South Thailand, with each dialect containing about 40 hours of data. However, this dataset faces a significant gender imbalance, with a ratio of 80% male and 20% female.

Thai-Dialect (Suwanbandit et al., 2023a) contains speech-to-text data for 10 Thai dialects from various regions of Thailand. It includes the standard Thai-central (THA) dialect, along with northern dialects (Khummuang, Nan, Yno), northeastern dialects (Korat, Khmer, Laos), and southern dialects (Krabi, Pattani, Phangnga). All transcriptions adhere to the Thai writing standard.

3.2 Vietnamese Corpora

Although the Vietnamese language has numerous datasets to facilitate ASR task such as VIVOS (Luong and Vu, 2016), speech corpus by (Nguyen et al., 2017), VinBigdata-VLSP2020-100h[2], FPT Open Speech Dataset (FOSD)[3], very few datasets incorporate regional dialectical elements.

VNSpeechCorpus (Le et al., 2004), published in 2004, is considered one of the first datasets on Vietnamese dialects, divided into three region dialects of Vietnam. However, speakers representing these regions were limited to four localities: Hanoi represents the northern dialect; Nghe An, Ha Tinh represents the central dialect; and Ho Chi Minh City represents the southern dialect. The dataset consists of 100 hours of reading-style speech data, recorded in noiseless and office settings. Although the authors did not conduct dialect identification or speech recognition, they designed a vocabulary and a phonetic dictionary. The dataset is not publicly accessible.

VDSPEC (Hung et al., 2016b), published in 2016, comprises a duration of 45.12 hours. The authors used the dialects of Hanoi, Hue and Ho Chi Minh City to represent the Northern, Central, and Southern Vietnamese dialects, respectively. The recordings for each dialectal subset were obtained in a reading style. Furthermore, the authors performed dialect identification on this dataset using LDA and GMM models. Public access to this dataset was not granted.

ViASR (Nguyen et al., 2023), released in 2023, consists of a 32-hour speech collection featuring three distinct regional dialects. The data was collected from openly accessible online sources, focusing on finance-related topics. The authors conducted baseline speech recognition experiments on transformer-based models including Whisper (Radford et al., 2023), Wav2vec 2.0 (Baevski et al., 2020), and MMS (Pratap et al., 2024). This dataset has not been publicly released [4].

LSVSC (Tran et al., 2024), published in 2024, comprises 100.5 hours of audio in a spontaneous style. The dataset includes five dialects: Northern, Central, Southern, Central Highlands, and minority ethnic group dialects. Nonetheless, the dataset shows a significant disproportion in dialects, with the Northern dialect accounting for 88.1% of the dataset. To assess speech recognition performance on this dataset, the authors employed LAS (Chan et al., 2016) and the Speech-Transformer Model (Dong et al., 2018) for experiments. Notably, this dataset is publicly available.

In summary, all the aforementioned Vietnamese corpora only categorize dialects into regional clusters. Consequently, further subdividing these regional clusters into smaller provincial dialects is an extremely challenging task. Moreover, some of these datasets are modest in size, are not publicly available, or exhibit imbalances across dialects, hindering the development of research on Vietnamese dialects. This motivates us to construct a dataset that addresses these shortcomings, spanning 102.56 hours, with relatively balanced representation across dialects in various aspects. Of paramount importance, it encompasses all 63 provincial dialects of Vietnamese, and from these 63 dialects, they can be organized into 3 regional dialects or any other dialect groupings based on the objectives of the research. Table 3 presents a comparison of the multi-dialect Vietnamese speech datasets.

4 ViMD Dataset

4.1 Data Collection

We describe the process of building the ViMD according to Figure 1. The process consists of five phases: Video collection, Audio extraction, Human-Annotated transcription, Quality control and Data splitting.

Video Collection. We gather videos featuring interviews with local residents from official Television and Broadcasting Station as the data for the

[2] https://vlsp.org.vn/resources

[3] https://data.mendeley.com/datasets/k9sxg2twv4/4

[4] Access status last verified on 21st April 2024.

Dataset	Style	Duration	Availability Status	Number of Dialects	DI	SR
VNSpeechCorpus (Le et al., 2004)	Reading	100h	Restricted	3		
VDSPEC (Hung et al., 2016b)	Reading	45.12h	Restricted	3	✓	
ViASR (Nguyen et al., 2023)	Spontaneous	32h	Restricted	3		✓
LSVSC (Tran et al., 2024)	Spontaneous	100.5h	Public	5		✓
ViMD (our)	Spontaneous	102.56h	Public	3/63	✓	✓

Table 3: Comparison of the multi-dialect Vietnamese speech datasets. Compared aspects include Style, Duration, Availability Status, Number of Dialects, and whether Dialect Identification (DI) or Speech Recognition (SR) was conducted.

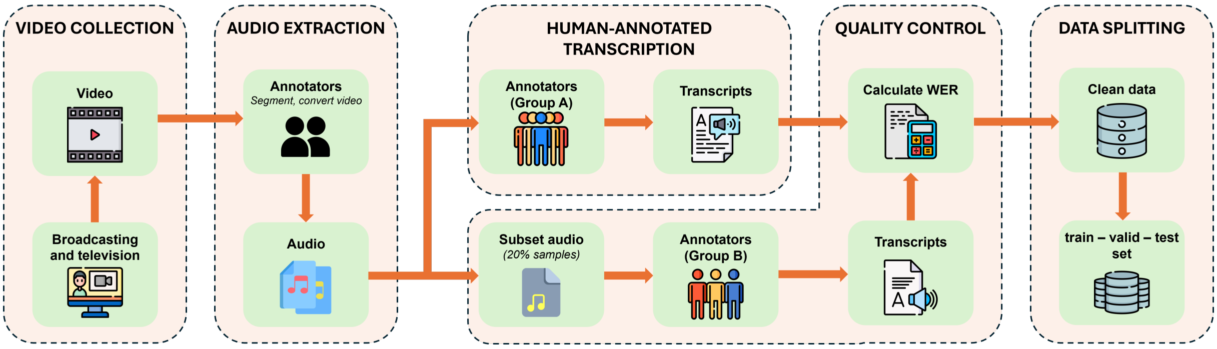


Figure 1: Data Collection Pipeline for the ViMD Dataset.

respective provincial dialect.

Audio Extraction. 10 student annotators with backgrounds in information technology segment videos into short clips containing local speakers' voices with the help of the Open Source Data Labeling Platform – Label Studio (Tkachenko et al., 2020-2022). Three mandatory requirements to ensure the dataset's quality are: (1) Each segment includes only one person's voice, (2) The audio for each person is limited to 180 seconds, (3) Filtering out news presenter's audio, (4) No segment exceeds 30 seconds. Any video segments that did not meet these three requirements were eliminated. The remaining video segments are converted to audio with the .wav format, preserving their original sample rates without any standardization.

Human-Annotated Transcription. This phase is conducted by our annotation group A, consisting of 10 individuals with diverse linguistic backgrounds, representing the Northern, Central, and Southern regions of Vietnam. We use an semi-automated labeling procedure, starting with audio transcription generated by API of AssemblyAI [5]. Subsequent to the initial transcription, transcripts from each provincial dialect are reviewed by an individual annotator, who corrects any inaccuracies to produce refined transcripts.

Quality Control. To verify the quality of Group A's work, two members of Group B independently transcribe 20% of the samples. We then use transcripts of Group B as the ground truth and calculate the Word Error Rate (WER) by comparing them to transcripts of Group A. If the WER is below 8%, we consider the transcript valid. Otherwise, Group A must re-transcribe the entire samples of that dialect. This threshold is considered as perfect, high quality data, as outlined in (Galvez et al., 2021).

Dataset Splitting. We split the data into train, validation, and test sets in an 8:1:1 ratio based on duration, with the additional considerations of gender proportion and speaker exclusivity across sets.

Finally, we stored the metadata information for the audio files in JSON format, with 8 attributes described in Table 4. Further details about dataset are described in Appendix B. In addition, each annotator was compensated with 1.92 USD and 2.4 USD per audio hour for Audio Extraction and Human-Annotated Transcription phase, respectively.

4.2 Dataset Statistics

Overall, our dataset offers a comprehensive representation of Vietnamese dialects, including 63 provincial dialects. It consists of 102.56 hours of audio recordings, with nearly 19,000 records obtained from 12,955 speakers. The accompanying transcripts consist of over 1.2 million words, with a distinct vocabulary of 5,155 unique words. The train, validation, and test sets were split in an 8:1:1

[5] https://www.assemblyai.com

Key	Data type	Description
set	String	The set of audio, which can take values from {‘train‘, ‘valid‘, ‘test‘}.
filename	String	Filename of the audio.
text	String	Transcript of the audio.
length	Float	Length of the audio in seconds.
province	String	The provincial dialect of the sample.
region	String	The regional dialect of the sample, which can be ‘North‘, ‘Central’, or ‘South‘.
speakerID	String	The speaker Identification.
gender	Int	Gender of the speaker, where 0 represents female and 1 represents male.

Table 4: Detailed Attribute Descriptions for Audio Samples in ViMD Dataset.

	Per Provincial Dialect				Data Set			Total
	Min.	Max.	Mean	Std.	Train	Valid.	Test	
Duration	89.11m	117.98m	97.68m	4.18m	81.43h	10.26h	10.87h	102.56h
#record	263	363	301	21	15,023	1,900	2,026	18,949
#speaker	88	309	206	47	10,291	1,320	1,344	12,955
#word	17,038	24,557	19,669	1,174	981,391	125,305	132,471	1,239,167
#unique-word	1,120	1,639	1,405	103	4,813	2,660	2,773	5,155

Table 5: ViMD dataset statistics on duration, number of records, speakers, words and unique words.

ratio based on duration, resulting in 81.43 hours, 10.26 hours, and 10.87 hours, respectively. This ratio extended to the number of records, speakers, and words as well. Notwithstanding such a ratio, the unique word counts in the validation set (2,660 unique words) and test set (2,723 unique words) does not differ significantly from from the training set (4,813 unique words), thus preserving the vocabulary diversity. Across the 63 provincial dialects, with the exception of the number of speakers exhibiting imbalance, the remaining attributes – duration, number of records, words, and unique words – are well-balanced. The relevant statistics are provided in Table 5.

Figure 2 displays the distribution of audio duration, which shows a higher frequency in the range of 10 to 30 seconds. Mean and standard deviation are 19.5 and 6.2 seconds, respectively.

Figure 3a reveals that in our dataset, the duration, records, speakers, and syllables for males are three times greater than those for females. Meanwhile, Figure 3b illustrates a considerable overlap in unique words between males and females, with 3,171 overlaps out of 4,600 unique male words and 3,726 unique female words. These findings suggest that while there are significant differences in duration and other attributes between males and females, there is a notable diversity of word diversity between the two genders.

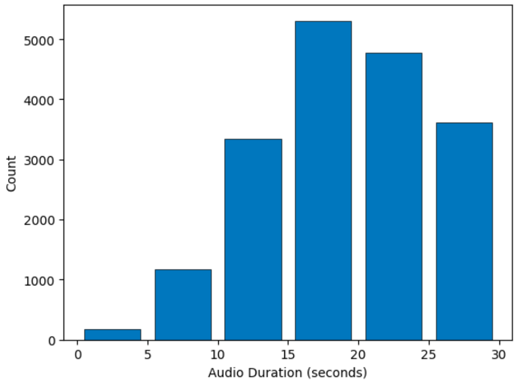


Figure 2: Distribution of Audio Duration.

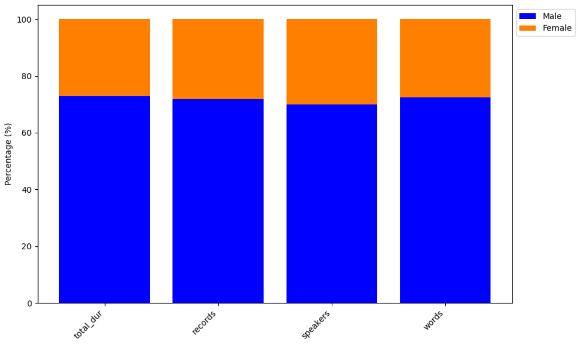
5 Experiments and Results

We conduct experiments on our dataset through two tasks encompassing **Dialect identification** and **Speech recognition**. The process is presented through the following sections: experimental design, baselines evaluation metrics, data preprocessing, and results.

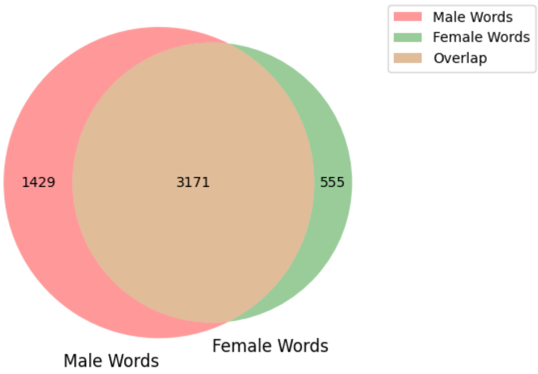
5.1 Experimental Design

Dialect Identification. We use only audio data to fine-tune pre-trained models, proceeding through five experiments.

- In the initial task, we categorize dialects into three labels including Northern, Central, and Southern, based on the ‘region’ attribute spec-



(a) Gender-wise Distribution of ViMD. The blue portion at the bottom representing males and the orange portion on top representing females.



(b) Gender unique word counts and overlap.

Figure 3: (a) Gender-wise Distribution of ViMD, and (b) Gender unique word counts and overlap.

ified in Table 2. This experiment, employing the entire 102.56 hours of data, is referred to as **[DI_VN_3]**.

- We divide our dataset into three sub-datasets, comprising 25 provinces in the Northern region (40.59 hours), 19 provinces in the Central region (31.47 hours), and 19 provinces in the Southern region (30.5 hours). Subsequently, we perform provincial dialect classification on these three sub-datasets, called **[DI_North]**, **[DI_Central]**, and **[DI_South]**, respectively.
- The last experiment in our DI task is denoted as **[DI_VN_63]**, in which we undertake the comprehensive classification of 63 labels corresponding to the 63 provincial dialects, utilizing the full 102.56 hours of data.

Speech Recognition. For the SR task, we utilize both audio data and accompanying transcripts. We carry out four experiments, with two stages in each experiment including direct inference from the models, and inference after fine-tuning on our data.

- Similar to the DI task, we employ three sub-datasets representing the Northern, Central, and Southern region of Vietnam. The experiments are respectively named **[SR_North]**, **[SR_Central]**, and **[SR_South]**.
- For the most comprehensive experiment of SR task, we leverage the entirety of the dataset to perform speech recognition across all dialects of Vietnam. This task is named **[SR_VN_63]**.

5.2 Baseline Models

To evaluate this challenging dataset, we conduct experiments on both tasks using the state-of-the-art transformer-based pre-trained models. These models achieve impressive results in speech-related tasks, particularly in the field of automatic speech recognition.

Wav2vec 2.0 (Baevski et al., 2020) is a state-of-the-art self-supervised Learning model. It uses CNN, transformer, and quantization modules. During training, raw audio is mapped to quantized speech representations used in a contrastive task. Cosequently, the model is fine-tuned on labeled data for transcription. We use wav2vec2-base-vi[6], wav2vec2-large-vi[7] for DI task and wav2vec2-base-vietnamese[8], wav2vec2-base-vietnamese-160h[9], wav2vec2-base-vietnamese-250h[10], wav2vec2-base-vi-vlsp2020[11] for SR task.

XLSR (Ruder et al., 2019) and **XLS-R** (Babu et al., 2021) are multilingual extensions of wav2vec 2.0 for cross-lingual speech recognition. XLSR uses a shared quantizer for pretraining on diverse languages. XLS-R is an enhanced version with larger models and more language coverage. We apply XLSR and XLS-R for only DI task, including wav2vec2-xls-r-300m[12] and wav2vec2-large-xlsr-53[13].

Whisper (Radford et al., 2023) is OpenAI's advanced multilingual ASR system using self-supervised learning on 680,000 hours of speech

[6] huggingface.co/nguyenvulebinh/wav2vec2-base-vi

[7] huggingface.co/nguyenvulebinh/wav2vec2-large-vi

[8] huggingface.co/dragonSwing/wav2vec2-base-vietnamese

[9] huggingface.co/khanhld/wav2vec2-base-vietnamese-160h

[10] huggingface.co/nguyenvulebinh/wav2vec2-base-vietnamese-250h

[11] huggingface.co/nguyenvulebinh/wav2vec2-base-vi-vlsp2020

[12] huggingface.co/facebook/wav2vec2-xls-r-300m

[13] huggingface.co/facebook/wav2vec2-large-xlsr-53

	#Params	**Vietnam** [DI_VN_3]	**North** [DI_North]	**Central** [DI_Central]	**South** [DI_South]	**Vietnam** [DI_VN_63]
Model/Num. of labels		**3**	**25**	**19**	**19**	**63**
wav2vec2-base-vi	95M	0.9102	0.4322	0.5863	**0.3560**	0.3522
wav2vec2-large-vi	317M	**0.9147**	0.4229	0.5981	0.3528	0.3570
wav2vec2-xls-r-300m	300M	0.8901	0.3216	0.3282	0.2787	0.3728
wav2vec2-large-xlsr-53	300M	0.8736	0.2830	0.1990	0.2440	0.3255
$\text{whisper}_{\text{base}}$	74M	0.8559	0.4336	0.5854	0.3383	0.3976
$\text{PhoWhisper}_{\text{base}}$	74M	0.8697	**0.4470**	**0.6251**	0.3257	**0.4107**

Table 6: Dialect identification experimental results with F1-macro metric.

data. It employs transformer models with attention, utilizing multi-task learning. The encoder processes audio spectrograms, and the decoder generates transcripts from the audio features. PhoWhisper (Le et al., 2024) is a fine-tuned version of Whisper, trained on an 844-hour Vietnamese speech dataset. We employ $\text{whisper}_{\text{base}}$ and $\text{PhoWhisper}_{\text{base}}$ for both DI and SR tasks.

We provide detailed information about the computational resources and hyperparameter settings in Appendix C.

5.3 Evaluation Metrics

F1-macro (Fujino et al., 2008) is utilized for the tasks of dialect identification, while Word Error Rate (WER) (Levenshtein et al., 1966) is employed for speech recognition tasks. Detailed information regarding these two metrics will be provided in Appendix D.

5.4 Data Pre-processing

All audio files are resampled to a 16kHz sampling rate and converted to mono channel. For the dialect identification task, files with a duration under 10 seconds are kept intact, whereas longer files are segmented into segments not exceeding 10 seconds. This segmentation approach stems from established practices in related research (Lu et al., 2020; Umapathy et al., 2007). As for the speech recognition task, the text data undergoes several common pre-processing techniques, such as converting to lowercase and removing punctuation marks.

5.5 Dialect Identification Experimental Results

The results of the DI experiment are shown in Table 6. Our analysis focuses on two aspects: first, we compare the performance across the different models used. Second, we look at how the models perform when tested on experiments with different audio data and label categories.

In the [DI_VN_3], the wav2vec 2.0 family achieves the highest F1-macro scores, with 91.02% for the base model and 91.47% for the large model. However, in the most challenging task - [DI_VN_63], their performance is relatively poor, with scores of only 35.22% for the base model and 35.28% for the large model. Interestingly, the whisper model group exhibits the opposite trend, with the lowest F1-macro scores in the [DI_VN_3] (85.59% for whisper-base and 86.97% for phowhisper-base) but the highest scores in the [DI_VN_63] (39.76% for whisper-base and 41.07% for phowhisper-base). The XLSR and XLS-R models' F1-macro scores are average across both tasks, with the exception of wav2vec2-large-xlsr-53, which has the lowest score of 32.55% in the [DI_VN_63]. Regarding categorizing provincial dialects within each regional dialect, the wav2vec and whisper groups outperform the XLSR and XLS-R groups. PhoWhisper performs the best with 44.70% for the [DI_North] and 62.51% for the [DI_Central], while wav2vec2-base-vi achieves the best 35.60% for the [DI_South].

Out of the five experiments we conduct, the [DI_VN_3] is the simplest with the fewest labels, and the models achieve very high F1-macro scores, all above 85%. For the identification of provincial dialects within each regional dialect, the provinces in the Central region [DI_Central] seem to have the most distinct characteristics, with the highest F1-macro score of 62.31%, followed by the North [DI_North] at 44.70%, and the model performs the worst for the Southern region [DI_South] with only 35.60%. When using all 63 provincial dialects for the [DI_VN_63], the F1 score is 41.07%.

These experiments established a benchmark for this dataset. Simultaneously, the results also demonstrate the challenging nature of the DI task on this dataset.

Model	**#Params**	**North**	**Central**	**South**	**Vietnam**
		[SR_North]	[SR_Central]	[SR_South]	[SR_VN_63]
w/o Fine-tuned					
wav2vec2-base-vietnamese	95M	0.2032	0.2728	0.2248	0.2307
wav2vec2-base-vietnamese-160h	95M	0.2750	0.3812	0.3093	0.3174
wav2vec2-base-vietnamese-250h	95M	0.1498	0.2097	0.1724	0.1747
wav2vec2-base-vi-vlsp2020	95M	**0.1364**	**0.1926**	**0.1481**	**0.1568**
whisper$_{\text{base}}$	74M	0.2637	0.3991	0.2946	0.3138
PhoWhisper$_{\text{base}}$	74M	0.1496	0.2415	0.1787	0.1861
Fine-tuned					
wav2vec2-base-vietnamese	95M	0.1464	0.2027	0.1772	0.1580
wav2vec2-base-vietnamese-160h	95M	0.1670	0.2456	0.2051	0.1749
wav2vec2-base-vietnamese-250h	95M	0.1229	**0.1715**	0.1526	0.1356
wav2vec2-base-vi-vlsp2020	95M	**0.1217**	0.1719	0.1508	**0.1224**
whisper$_{\text{base}}$	74M	0.2005	0.2789	0.2089	0.1993
PhoWhisper$_{\text{base}}$	74M	0.1320	0.1826	**0.1354**	0.1630

Table 7: Speech recognition experimental results with WER metric.

5.6 Speech Recognition Experimental Results

Model Performance. The outcomes of the SR task are summarized in Table 7. We analyze the discrepancy between before and after fine-tuning, based on two main criteria: performance across different models and different experiments.

All direct inference results yield lower performance compared to fine-tuning, except for the case of wav2vec2-base-vi-vlsp2020 in the [SR_South] experiment where fine-tuning leads to a 0.27% higher WER. The model that shows the best improvement across all tasks after fine-tuning is wav2vec2-base-vietnamese-160h, although this improvement is still not sufficient to outperform other models. Prior to fine-tuning, wav2vec2-base-vi-vlsp2020 demonstrates superiority by achieving the best results across all experiments. However, after fine-tuning, wav2vec2-base-vietnamese-250h performs comparably, sometimes outperforming and sometimes underperforming wav2vec2-base-vi-vlsp2020, with the highest gap being only 1.32% across all experiments. phowhisper-base shows significant improvement in [SR_South], outperforming other models and achieving a result of 13.54% in this experiment.

When fine-tuning on the entire dataset in the [SR_VN_63] experiment, most models show the best improvement, which can be influenced by the large data quantity. The best result on [SR_VN_63] is 12.24% with the wav2vec2-base-vi-vlsp2020 model. When comparing the three experiments with smaller data sizes, [SR_North], [SR_Central], and [SR_South], both before and after fine-tuning, [SR_Central] always has the highest WER while [SR_North] has the lowest WER across all experiments. However, an encouraging observation is that [SR_Central] shows the most significant improvement, while [SR_North] exhibits the least improvement after fine-tuning. This suggests that although the results for [SR_Central] are not yet optimal, our dataset has a considerably positive impact on the current models in the Central region of Vietnam. The best results obtained after fine-tuning for [SR_North], [SR_Central], and [SR_South] are 12.17%, 17.15%, and 13.54%, respectively.

In general, our dataset helped improve the performance of models in the SR task. Concurrently, it has also poses challenges for models in dealing with certain specific dialects.

The Effect of Dialectal Variations on SR Performance. We select the best-performing model based on WER to analyze the results from four dialectal ASR experiments. We assert that using dialect-specific datasets significantly reduces spelling errors. Vietnam's linguistic diversity, with each province having at least one unique dialect, is a key factor in recognition errors. Our analysis reveals that provinces in the Northern, Central, and Southern regions often display similar dialect-specific spelling mistakes. These errors are compiled in Table 13, along with model predictions before and after fine-tuning on our dataset.

The Northern region of Vietnam typically has the clearest pronunciation. However, some local pronunciation mistakes are still prevalent, particularly the confusion between the syllables 'n' and 'l', as seen in 'linh bình' for 'ninh bình' and 'việc nàm' for 'việc làm'. Model performance significantly improved after fine-tuning, as evidenced by a notable reduction in the WER metric.

In the Central region, the accent exhibits the most distinct variations, resulting in a significantly higher WER in our experiments compared to other regions. Certain vowels undergo notable changes; for example, 'a' becomes 'o' or 'a' morphs into 'e', turning 'thắt chặt' into 'thất chẹt', 'lắng nghe' into 'lấn nghe', 'bán' into 'bón', and 'năm' into 'nem'. In provinces like Hue (74) and Quang Tri (75), question tones tend to be pronounced with a heavier inflection, such as 'phát triện' instead of 'phát triển'.

In the Southern region, words beginning with the letters 'v' and 'd' are commonly mispronounced as beginning with 'd', for example, 'dội dả' instead of 'vội vã'; and 'tr' is often misheard as 'ch', like 'nổi chội' instead of 'nổi trội'. Additionally, final consonants 'n' and 'ng' are frequently confused as 'ng', 'c' and 't' are both misinterpreted as 'c', such as 'bạng' for 'bạn', 'các' for 'cát'. There are also other errors such as unclear pronunciation of complex words or failure to distinguish between different but relatively similar tonal marks.

6 Dicussion

Our analysis of the results yields two notable findings encompassing (1) Geographical influences on dialects and (2) Multi-Dialect data challenges for speech recognition approach. The following are our conjectures based on experimental results. The cause could also stem from the training data for pre-trained models or other factors.

6.1 Geographical Influences on Dialects

The detailed confusion matrix for the DI task and the table of province codes are presented in the Appendix E; a map of Vietnam is also included[14]. In the [DI_VN_63] experiment, 12 provinces achieved F1-macro scores of 0.6 or above. What is truly remarkable is that 10 (coded as 17 - Thai Binh, 18 - Nam Dinh, 35 - Ninh Binh, 37 - Nghe An, 73 - Quang Binh, 76 - Quang Ngai, 77 - Binh Dinh, 78 - Phu Yen, 86 - Binh Thuan, 59 - Ho Chi Minh) out of these 12 provinces are coastal regions, while the remaining 2 provinces (coded as 28 - Hoa Binh and 98 - Bac Giang) are only one province away from the sea. This observation underscores the potential influence of coastal factors on the unique features of local speech patterns.

A noteworthy finding in the Central region is that although the highest DI scores ([DI_Central]) demonstrate the highly distinctive nature of provincial dialects within the region, the SR result ([SR_Central]) for the Central provinces are the poorest. The Central region's narrow and latitudinally elongated shape, unlike the Northern and Southern regions, could be a potential cause for this characteristic.

6.2 Multi-Dialect data challenges for speech recognition approach

We select the wav2vec2-base-vi-vlsp2020 model as it exhibits the best performance on the [SR_VN_63] experiment. We calculate the WER for each regional dialect in [SR_VN_63] and compare it with the corresponding WER on [SR_North], [SR_Central], and [SR_South]. The improvements over training on a entire dataset are 1.86%, 3.07%, and 2.34% for the Northern, Central, and Southern dialects, respectively. Details on the WER differences for each provincial dialect are listed in Appendix E. These differences demonstrate that despite training the model on a combined dataset containing various dialects with a duration approximately 2-3 times larger than the individual datasets, the performance gain over training on separate dialects is relatively small. The findings indicate a need for more effective methods of cross-dialect knowledge transfer, rather than merely aggregating the dialects and training on the combined dataset as a separate dataset.

7 Conclusion

We introduce ViMD, a novel dataset covering all 63 provincial dialects in Vietnam. We carry out experiments on the two tasks of dialect identification and speech recognition, employing various state-of-the-art models to establish baseline benchmarks. The results facilitate a more in-depth investigation of dialects, including the impact of geographical factors on dialectal variations, and pose challenges for speech recognition models in tackling the multi-dialect aspect of the Vietnamese language. In addition, we hope that our process will help expand both the scale and quality of the other datasets. We also expect that this dataset will facilitate future research aimed at enhancing the performance of DI and SR tasks, as well as other related speech tasks, especially for the Vietnamese language.

[14] https://bandovn.vn/vi/page/mau-ban-do-hanh-chinh-nuoc-cong-hoa-xa-hoi-chu-nghia-viet-nam-181

Limitations

Although the majority of residents within a given province tend to speak the local dialect, a minority who have previously resided in other regions retain their original dialectal forms, even giving rise to "hybrid dialects". The audio duration of the dataset is quite modest at only 102.56 hours. The transcripts contained within the dataset show a few inaccuracies resulting from regional pronunciation patterns as well as vocabulary highly specific to certain areas. Additionally, there is a gender disparity, with the number of male speakers being three times that of female speakers. While we have not yet categorized the audio content by topic, the majority of the recordings are from television news programs, which typically address a broad spectrum of issues from daily life.

In our experiments, we primarily employed base version of pre-trained models, employing only a few large version of pre-trained models owing to constraints in computational resources, thus limiting our ability to conduct a comprehensive assessment of state-of-the-art model capabilities. Subsequent investigations will potentially include the use of larger versions, notably Wav2vec2-BERT (Barrault et al., 2023) and MMS-1B (Pratap et al., 2024).

Ethics Statement

The entirety of the data utilized in this research study is from publicly available sources, ensuring no infringement of privacy rights. All data has been published by the 63 Television and Broadcasting Stations corresponding to all 63 provinces of Vietnam, guaranteeing the verification of all included content. Our study is aimed at furthering research efforts into the regional dialects found throughout Vietnam, and it is not intended to target any specific individuals or organizations.

Acknowledgement

This research was supported by The VNU-HCM University of Information Technology's Scientific Research Support Fund.

References

M. J. Alves. 2007. A look at north-central vietnamese. In *Proceedings of the 12th Annual Meeting of the Southeast Asian Linguistics Society 2002 (SEALS XII), Canberra, Australia*.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2021. Xls-r: Self-supervised cross-lingual speech representation learning at scale. *Preprint*, arXiv:2111.09296.

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.

Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, et al. 2023. Seamless: Multilingual expressive and streaming speech translation. *arXiv preprint arXiv:2312.05187*.

William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals. 2016. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4960–4964. IEEE.

Linhao Dong, Shuang Xu, and Bo Xu. 2018. Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5884–5888. IEEE.

Akinori Fujino, Hideki Isozaki, and Jun Suzuki. 2008. Multi-label text categorization with model combination based on f1-score maximization. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-II*.

Daniel Galvez, Greg Diamos, Juan Ciro, Juan Felipe Cerón, Keith Achorn, Anjali Gopi, David Kanter, Maximilian Lam, Mark Mazumder, and Vijay Janapa Reddi. 2021. The people's speech: A large-scale diverse english speech recognition dataset for commercial usage. *arXiv preprint arXiv:2111.09344*.

Thị Châu Hoàng. 2009. *Phương ngữ học tiếng Việt*. Đại học Quốc gia Hà Nội.

Pham Ngoc Hung, Nguyen Thu Ha, Trinh Van Loan, Vu Xuan Thang, and Nguyen Dinh Chien. 2019. Vietnamese dialect identification on embedded system. *UTEHY Journal of Science and Technology*, 24:82–87.

Pham Ngoc Hung, Trinh Van Loan, and Nguyen Hong Quang. 2016a. Statistical analysis of vietnamese dialect corpus and dialect identification experiments. *International Journal of Scientific Engineering and Applied Science//(IJSEAS)–Volume-2, Issue-8*.

Pham Ngoc Hung, Trinh Van Loan, and Nguyen Hong Quang. 2016b. Statistical analysis of vietnamese dialect corpus and dialect identification experiments. *International Journal of Scientific Engineering and Applied Science//(IJSEAS)–Volume-2, Issue-8*.

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*.

Viet Bac Le, Do Dat Tran, Eric Castelli, Laurent Besacier, and Jean-François Serignat. 2004. Spoken and written language resources for vietnamese. In *LREC*, volume 4, pages 599–602.

Vladimir I Levenshtein et al. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union.

Haoye Lu, Haolong Zhang, and Amit Nayak. 2020. A deep neural network for audio classification with a classifier attention mechanism. *arXiv preprint arXiv:2006.09815*.

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.

Hamdy Mubarak, Amir Hussein, Shammur Absar Chowdhury, and Ahmed Ali. 2021. QASR: QCRI aljazeera speech resource a large scale annotated Arabic speech corpus. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2274–2285, Online. Association for Computational Linguistics.

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. 2023. 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.

Quoc Bao Nguyen, Ba Quyen Dam, Minh Hung Le, et al. 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. IEEE.

T. B. Pham. 2013. Phát âm nhầm /l/-/n/ trong tiếng việt: Tình huống điều chỉnh và cần thiết. *Ngôn ngữ*, 10:25–32.

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.

Michel Plüss, Jan Deriu, Yanick Schraner, Claudio Paonessa, Julia Hartmann, Larissa Schmidt, Christian Scheller, Manuela Hürlimann, Tanja Samardžić, Manfred Vogel, and Mark Cieliebak. 2023. STT4SG-350: A speech corpus for all Swiss German dialect regions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1763–1772, Toronto, Canada. Association for Computational Linguistics.

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.

Sebastian Ruder, Anders Søgaard, and Ivan Vulić. 2019. Unsupervised cross-lingual representation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 31–38, Florence, Italy. Association for Computational Linguistics.

M. Shimizu. 2013. Vị trí của tiếng quảng nam trong quá trình biến đổi âm cuối gốc lưỡi. In *Hội thảo ngôn ngữ học toàn quốc lần thứ II năm 2013, Hà Nội, Việt Nam*.

Artit Suwanbandit, Jaturong Chitiyaphol, Sutthinan Chuenchom, Kanyarat Kwiecien, Husen Sawal, Ruslan Uthai, Orathai Sangpetch, and Ekapol Chuangsuwanich. 2023a. Thai-dialect: Low resource thai dialectal speech to text corpora. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE.

Artit Suwanbandit, Burin Naowarat, Orathai Sangpetch, and Ekapol Chuangsuwanich. 2023b. Thai Dialect Corpus and Transfer-based Curriculum Learning Investigation for Dialect Automatic Speech Recognition. In *Proc. INTERSPEECH 2023*, pages 4069–4073.

Bao Thang Ta, Nhat Minh Le, et al. 2024. Transfer learning methods for low-resource speech accent recognition: A case study on vietnamese language. *Engineering Applications of Artificial Intelligence*, 132:107895.

Zhiyuan Tang, Dong Wang, Yanguang Xu, Jianwei Sun, Xiaoning Lei, Shuaijiang Zhao, Cheng Wen, Xingjun Tan, Chuandong Xie, Shuran Zhou, Rui Yan, Chenjia Lv, Yang Han, Wei Zou, and Xiangang Li. 2021. Ke-speech: An open source speech dataset of mandarin and its eight subdialects. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

VU Thanh Phu'o'ng. 1982. Phonetic properties of vietnamese tones across dialects. *n Bradley. D.(Edt.), Tonation (see Bradley, 1982)*.

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020-2022. Label Studio: Data labeling software. Open source software available from https://github.com/heartexlabs/label-studio.

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.

Karthikeyan Umapathy, Sridhar Krishnan, and Raveendra K Rao. 2007. Audio signal feature extraction and classification using local discriminant bases. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(4):1236–1246.

A Linguistic Variations Across Dialects

The diversity across the three regions of Vietnam is reflected in the pronunciation of syllable elements (Phạm and McLeod, 2016):

- Initial consonant: The Northern dialect possesses the smallest number of initial consonants at 20, followed by the Southern dialect with 21, and the Central dialect has the highest count with 23 initial consonants.
- Tone: the Northern dialect has the most tones, with 6, while the Central and Southern dialects have 5 tones each.
- Final consonant: The Northern dialect features 10 final consonants, the Central dialect also has 10 final consonants, and the Southern dialect has 8.
- Vowel: The pronunciation of vowels adheres to specific word contexts.

Differences in pronunciation can lead to distortions in meaning. In the Northern region, some areas interchange the pronunciation of initial consonant 'l' and 'n' (Pham, 2013). For instance, the word 'lầm' (mistake) might be pronounced as 'nầm' (breast of a mammal). Quang Nam province, representing the Central region, showcases a phonetic shift where the vowel 'a' is pronounced as 'o' (Shimizu, 2013), seen in the pronunciation of 'tám' (eight) as 'tóm' (catch). In the Southern region, there's a tendency to pronounce the final consonants 'n' and 'ng' similarly (Tran et al., 2024), as evidenced by the identical pronunciation of 'lươn' (eel) and 'lương' (salary). The differences in pronunciation are not only at the regional level, but they also exist among different provinces within the same region. For example, within the Central dialect, provinces in the North-Central region pronounce the letter 'gi' as [z], whereas some provinces in the South-Central region (represented by Quang Nam) pronounce it as [j]. Furthermore, within a province, there are intra-provincial differences in pronunciation, illustrated by variations in the pronunciation of 'bật lửa' (lighter) across districts in Nghe An province (Alves, 2007).

Besides a word having multiple pronunciations, different regions also have distinct words expressing the same meaning. Table 8 illustrates some of the varying words across regions (Hung et al., 2019; Ta et al., 2024).

Northern	Central	Southern	Meaning
bố, thầy	bọ	ba, tía	father
u, mẹ	mế, bầm, mạ	má, me	mother
chúng tui	bầy tui	tụi tui	we
mày	mi	mầy	you
gì	chi	gì	what
đâu thế	mô rứa	đâu vậy	where
thế nào	răng	sao	how

Table 8: Variations in Vietnamese Words with the Same Semantics Across Regions.

B Dataset

B.1 Data Construction

In the audio transcription task described in Section 4.1, we followed specific guidelines to ensure our transcripts were clear and consistent: (1) Numerals were converted to their word form, (2) Units of measurement were phonetically transcribed into Vietnamese, and (3) Local vernacular terms were preserved without modification. The distinction between a sentence written in the common language and our prescribed transcript format will be illustrated in Table 9. In the table, **red** highlights Guideline (1), **blue** for Guideline (2), and **green** for Guideline (3).

Although there is a clear process for transcribing and quality control, some errors still exist in the dataset. Common errors in transcripts typically stem from the use of local vocabulary, which is often spoken rather than written, leading to inaccuracies. Additionally, mistakes frequently occur with proper names of villages, towns, districts, or individuals due to annotators' unfamiliarity with these locations. To address these issues, we propose researching and providing a list of villages, towns, cities and districts for annotators. Furthermore, referring to online sources or Vietnamese dictionaries can help ensure accurate transcription of dialectal terms, allowing for better alignment between spoken and written language. However, the error rate is limited to under 8% word error rate.

Table 4 presents the 8 attributes associated with each audio sample in our dataset. Below is an example of a sample. The filename follows the syntax {province code}_{Sequence Number of Audio}, and similarly, the speaker identification code adheres to the syntax spk_{province code}_{Sequence Number of Speaker}.

Commonly written	Chúng tôi đã **làm** được trên **50 ha** lúa.
Transcript	Chúng tôi đã **mần** được trên **năm chục héc ta** lúa.
English	We have **cultivated** over **50 hectares** of rice paddies.

Table 9: Divergences Between Common Writing and Transcript Format.

```
{
    "set": "train",
    "filename": "19_0001.wav",
    "text": "Vật dụng để phòng chống cháy nổ ở khu vực này vẫn còn thiếu rất là nhiều.",
    "speakerID": "spk_19_0001",
    "gender": 1,
    "length": 5.244
}
```

Table 10 presents a comprehensive list of all 63 provinces in Vietnam, including the province name in Vietnamese, the region to which the province belongs, the province code and several other attributes. The province codes are assigned based on the vehicle registration plate designations for automobiles and motorcycles in Vietnam, as regulated by the Vietnamese Government[15]. For provinces with multiple codes, we have selected a representative code.

B.2 Dataset Additional Statistics

Statistics by Provincial Dialects and Gender. Figure 6 presents a stacked visualization of the duration across 63 provincial dialects in Vietnam, with the blue area at the bottom reflecting the duration for males, and the area above representing the duration for females. Accompanying these bars are two lines: one with blue markers depicting the number of male speakers, and another with orange markers indicating the female speaker count. While the total duration appears relatively uniform across the provincial dialects, the figure highlights a significant disparity in duration and speaker count between the two genders. Table 10 illustrates the number of words and the number of unique words across 63 provincial dialects, with the blue line representing the word count and the orange line indicating the count of unique words.

Statistics by Regional Dialects. The statistical representation in Figure 4 highlights the comparison among the three regions concerning total duration (total_dur), number of records (records), duration of male speakers (male_dur), duration of female speakers (female_dur), number of speakers (speakers), and number of words (words). The distribution among the three regions appears to be relatively balanced, with only a slight predominance in the North, attributed to the larger number of provinces (25) compared to the Central (19) and Southern (19) regions. Figure 5 shows the vocabulary overlap among regions. The intersection among all three regions is small, with only 2506 words out of 5167. Adjacent regions like Northern and Central, or Central and Southern, have more overlap than the distant Northern and Southern regions. The Northern and Central dialects have a quite similar number of unique words, 697 and 662 respectively, while the Southern dialect has fewer unique words, with 504.

Lexical Statistics by Gender We have listed the 6 most frequently used words for each gender, excluding one word that was a proper name. Here they are: For males: (1) 'cồn' (nồng độ cồn - Alcohol Concentration), (2) 'loạt' (đồng loạt – simultaneously), (3) 'cọc' (tiền cọc – deposit), (4) 'cưỡng' (cưỡng chế - force), (5) 'container' (xe container - container truck). For females: (1) 'hò' (hẹn hò, hát hò – date, sing) , (2) 'dance', (3) 'piano', (4) 'lứt' (gạo lứt - brown rice), (5) inox (đồ dùng inox - stainless steel utensils). None of these words are inherently gendered in Vietnamese grammar. Instead, they seem to reflect different topics or areas of interest that may be more common among males or females in the context of our dataset.

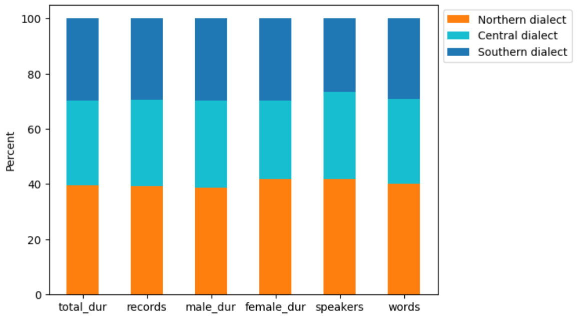


Figure 4: Comparison of Duration and Number of Speakers Between Genders.

[15] https://congbao.chinhphu.vn/noi-dung-van-ban-so-58-2020-tt-bca-31631

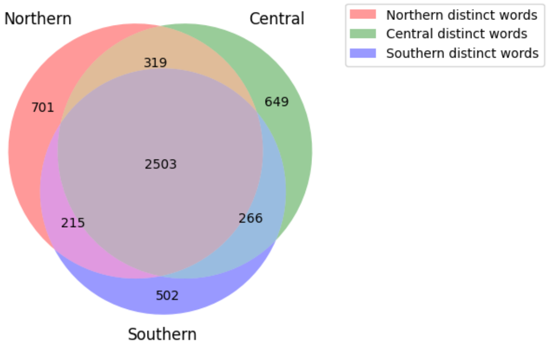


Figure 5: Words and Unique Words Count Across Provinces.

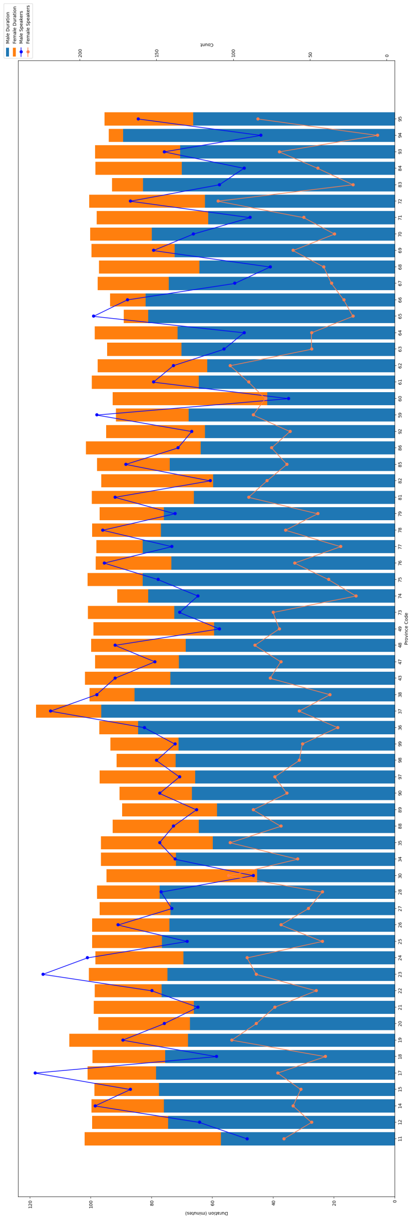


Figure 6: Comparison of Duration and Number of Speakers Between Genders.

No.	Province Name	Region	Province Code	Duration	Speakers	Records	Words	Unique Words
1	Cao Bằng	North	11	101.97	158	357	19,621	1,355
2	Lạng Sơn	North	12	99.48	171	304	20,304	1,317
3	Quảng Ninh	North	14	99.77	251	316	20,282	1,423
4	Hải Phòng	North	15	98.72	223	306	20,714	1,545
5	Thái Bình	North	17	100.96	300	329	21,488	1,534
6	Nam Định	North	18	99.38	151	300	20,403	1,143
7	Phú Thọ	North	19	106.99	273	313	22,407	1,483
8	Thái Nguyên	North	20	97.45	230	314	20,780	1,438
9	Yên Bái	North	21	98.98	196	289	20,187	1,376
10	Tuyên Quang	North	22	98.59	199	286	20,595	1,124
11	Hà Giang	North	23	100.56	309	325	20,401	1,330
12	Lào Cai	North	24	98.38	286	313	19,754	1,377
13	Lai Châu	North	25	99.46	172	295	20,007	1,233
14	Sơn La	North	26	99.53	244	308	19,702	1,399
15	Điện Biên	North	27	97.07	191	283	19,025	1,316
16	Hòa Bình	North	28	97.89	187	281	19,406	1,322
17	Hà Nội	North	30	94.82	190	316	19,930	1,531
18	Hải Dương	North	34	96.59	196	290	19,730	1,451
19	Ninh Bình	North	35	96.57	250	268	20,143	1,441
20	Vĩnh Phúc	North	88	92.71	208	266	18,543	1,280
21	Hưng Yên	North	89	89.64	211	277	17,940	1,473
22	Hà Nam	North	90	90.48	213	272	18,229	1,290
23	Bắc Kạn	North	97	97.02	208	291	19,169	1,341
24	Bắc Giang	North	98	91.44	207	272	17,998	1,315
25	Bắc Ninh	North	99	93.53	193	277	18,956	1,355
26	Thanh Hóa	Central	36	97.19	188	303	20,068	1,399
27	Nghệ An	Central	37	117.98	275	363	24,549	1,410
28	Hà Tĩnh	Central	38	100.35	226	305	20,472	1,487
29	Đà Nẵng	Central	43	101.82	253	337	21,460	1,430
30	Đắk Lắk	Central	47	98.54	220	306	18,935	1,456
31	Đắk Nông	Central	48	99.79	263	304	20,296	1,498
32	Lâm Đồng	Central	49	99.05	179	303	19,416	1,408
33	Quảng Bình	Central	73	100.94	208	345	20,857	1,638
34	Quảng Trị	Central	74	91.19	143	292	17,893	1,365
35	Thừa Thiên Huế	Central	75	101.04	187	303	1,9870	1,403
36	Quảng Ngãi	Central	76	98.29	242	330	20,432	1,627
37	Bình Định	Central	77	98.13	169	307	20,754	1,428
38	Phú Yên	Central	78	99.54	249	305	19,841	1,525
39	Khánh Hòa	Central	79	97	182	281	19,522	1,352
40	Gia Lai	Central	81	99.66	267	314	19,696	1,361
41	Kon Tum	Central	82	96.52	192	305	17,670	1,439
42	Ninh Thuận	Central	85	97.94	235	314	19,789	1,426
43	Bình Thuận	Central	86	101.58	211	325	20,455	1,575
44	Quảng Nam	Central	92	94.94	190	291	19,526	1,527
45	Hồ Chí Minh	South	59	91.68	276	318	19,823	1,349
46	Đồng Nai	South	60	92.78	142	275	18,738	1,422
47	Bình Dương	South	61	99.66	242	314	20,431	1,407
48	Long An	South	62	97.68	241	308	19,576	1,505

Continued on next page

Table 10 – continued from previous page

No.	Province Name	Region	Province Code	Duration	Speakers	Records	Words	Unique Words
49	Tiền Giang	South	63	94.57	154	289	18,629	1,467
50	Vĩnh Long	South	64	98.67	142	284	19,820	1,397
51	Cần Thơ	South	65	89.11	213	263	16,970	1,175
52	Đồng Tháp	South	66	93.64	196	273	19,281	1,409
53	An Giang	South	67	97.65	135	285	18,929	1,409
54	Kiên Giang	South	68	97.21	117	278	18,521	1,505
55	Cà Mau	South	69	99.74	213	302	19,097	1,500
56	Tây Ninh	South	70	100.15	160	302	20,052	1,316
57	Bến Tre	South	71	97.98	143	289	18,627	1,378
58	Bà Rịa - Vũng Tàu	South	72	100.45	277	319	19,935	1,461
59	Sóc Trăng	South	83	92.95	131	273	17,389	1,326
60	Trà Vinh	South	84	98.39	138	293	19,050	1,383
61	Bình Phước	South	93	98.57	215	319	19,319	1,559
62	Bạc Liêu	South	94	93.99	88	278	17,851	1,257
63	Hậu Giang	South	95	95.38	246	309	19,756	1,419

Table 10: List of 63 Provinces of Vietnam with Language Data Statistics.

C Experimental Settings

The pretrained models were originally trained for the SR task. Therefore, when fine-tuning them for the DI task, we add two linear layers on top of the pretrained models and the cross-entropy loss function during the training.

The models tasked with Dialect Identification are configured with the hyperparameters listed in Table 11, while the models responsible for Speech Recognition utilize the hyperparameters detailed in Table 12. All experimental training are carried out on an NVIDIA GeForce RTX 4090 (24GB).

Hyperparameter	**wav2vec 2.0**		**XLSR**	**Whisper**
	Base	**Large**	**XLS-R**	
Epochs	15	15	15	15
Learning rate	3e-5	6e-5	6e-5	3e-5
Batch size	64	64	64	64
Optimizer	AdamW	AdamW	AdamW	AdamW
Weight Decay	0	0	0	0
Warmup Ratio	0.1	0.1	0.1	0.1

Table 11: Dialect Identification experimental configurations.

Hyperparameter	**wav2vec 2.0**	**Whisper**
Epochs	15	10
Learning rate	1e-4	1e-5
Batch size	8	8
Optimizer	AdamW	AdamW
Weight Decay	0.005	0.005
Warmup Ratio	0.1	0.1

Table 12: Speech Recognition experimental configurations.

D Evaluation Metrics

D.1 F1-macro

The F1-score (Fujino et al., 2008) is a metric used in statistical machine learning to evaluate the performance of a classification model. The F1-score is calculated based on precision and recall, and is the harmonic mean of these two measures. The formula for the F1 score is given by:

$$F_1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

To calculate the macro F1-score, the F1 score is computed for each class individually, and then the average of these F1-scores is taken. The formula for the macro F1-score is as follows:

$$F_{1\text{ macro}} = \frac{1}{n}\sum_{i=1}^{n} F_{1i}$$

where n is the number of classes and F_{1i} is the F1 score for the i-th class. The macro F1-score is not affected by imbalances in class distribution, as each class is treated equally when averaging.

In our dialect identification tasks, we choose to use the macro F1-score as the evaluation metric to ensure that the performance of the classification model across each provincial dialect is computed fairly, without being influenced by the disparity in sample sizes among the provinces. This is particularly important in this study, where each province represents a distinct dialect that needs to be treated with equivalent fairness.

D.2 WER

Word Error Rate (WER) (Levenshtein et al., 1966) is a crucial metric used to evaluate the performance of Speech Recognition systems. It is measured based on the accuracy of the transcription generated by the system compared to the reference transcription, by considering three types of errors: substitution errors (S), deletion errors (D), and insertion errors (I), relative to the total number of words in the reference transcript (N). The formula for the WWER is as follows:

$$WER = \frac{S + D + I}{N}$$

E Experimental Results

E.1 Dialect Identification

The presented results originate from the models exhibiting the highest performance for each respective task, as illustrated in Figure 7, Figure 8, Figure 9, Figure 10 and Figure 11. All confusion matrices are normalized with respect to the true conditions.

The results of the **[DI_VN_3]** experiment demonstrate the model's remarkable dialect identification performance across all three regions, correctly identifying 95%, 88%, and 91% of the samples for the Northern, Central, and Southern dialects, respectively. However, in the **[DI_North]** task, the accuracy rates for predicting provincial dialects were uneven, ranging from a high of 95% (Ninh Binh - 35) for the top-performing province to a low of 14% (Bac Giang - 14) for the least accurate one. The largest confusion occurred for label 12 (Lang Son), which was predicted as label 20

(Thai Nguyen) with an error rate of up to 23%. The **[DI_Central]** task demonstrates relatively promising recognition rates for many provinces, with 12 out of the 19 provinces achieving accurate predictions for over 60% of their samples. However, notable confusion persists among certain geographically close provinces. As an example, the province of Dak Nong (48) is frequently misclassified as Dak Lak (47), and Ha Tinh (38) is often predicted as Quang Binh (73). Against all expectations, Da Nang City (43) is predicted as Binh Thuan (86) with the highest confusion rate of 43%, despite the substantial geographical distance separating the two provinces, making this finding quite inexplicable. In the **[DI_South]** dialect identification task, substantial confusion was observed among provincial dialects, potentially attributable to the high degree of similarity between them. In particular, the provincial dialects of Ca Mau (69) and Bac Lieu (94) have very low correct identification rates of only 7% and 9%, respectively. The confusion matrix also indicates that the two provincial dialects of Binh Phuoc (93) and Ba Ria - Vung Tau (72) are the most accurately classified, with rates of 75% and 66%, respectively.

The outcomes of the **[DI_VN_63]** experiment are depicted in Figure 11. We have incorporated red dashed lines to facilitate the tracking of provincial dialects across regional dialect boundaries. Overall, the confusion between provincial dialects is primarily concentrated within the three regional dialect clusters. While some confusion persists between geographically proximate regional dialects, such as Northern - Central and Central - Southern, the most geographically distant pair, Northern - Southern, exhibits the least confusion. Among Vietnam's five municipalities of Vietnam, four cities – Ha Noi (30), Hai Phong (15), Da Nang (43), and Can Tho (65) - have very low prediction accuracy rates, ranging from 29% to 36%, which can be explained by the influx of residents from other provinces; however, surprisingly, Ho Chi Minh (59) has a very high prediction accuracy rate of 63%. The neighboring provinces of Ha Tinh (38) and Quang Binh (73) exhibit a high rate of mutual misprediction, with 42% of samples of label 38 being predicted as 73, and in 34% of cases, label 73 is predicted as 38. The three most distinctive provinces are Binh Dinh (77), Ninh Binh (35), and Binh Thuan (86), with prediction accuracy rates of 92%, 91%, and 88%, respectively.

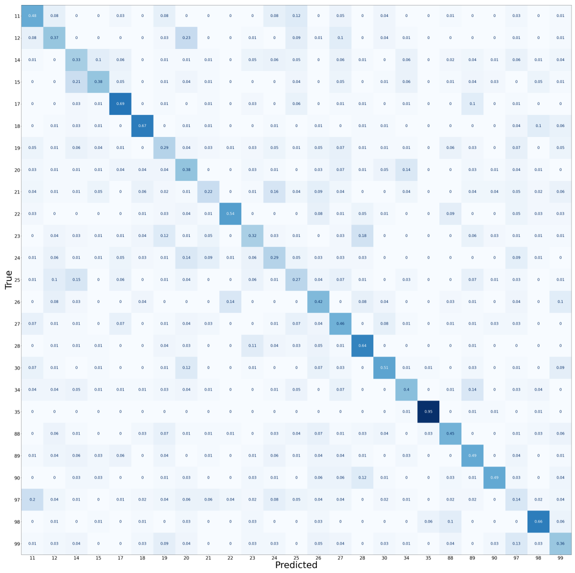


Figure 7: Confusion matrix of [DI_North].

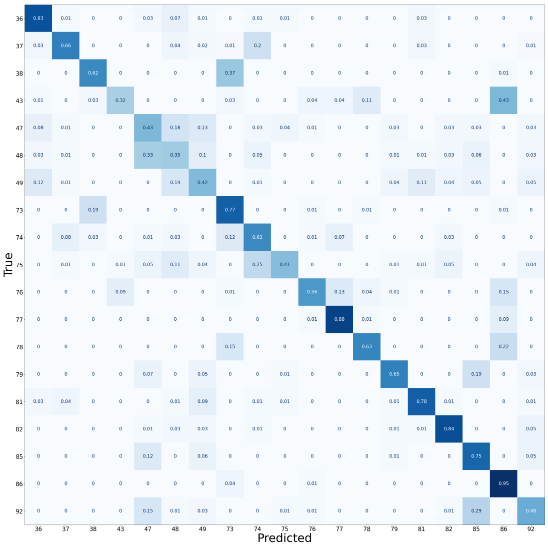


Figure 8: Confusion matrix of [DI_Central].

E.2 Model Improvement with Training on Entire Data

We analyze the results from four dialectal ASR experiments. For each experiment, we select the best-performing model based on WER. The wav2vec2-base-vi-vlsp2020 model performs best in the [SR_North] and [SR_VN_63] experiments, the wav2vec2-base-vietnamese-250h model outperforms others in the [SR_Central] experiment, and the PhoWhisperbase model achieves the top performance in the [SR_South] experiment. The error analysis focuses on the improvement in errors when trained on a combined dataset containing all dialects.

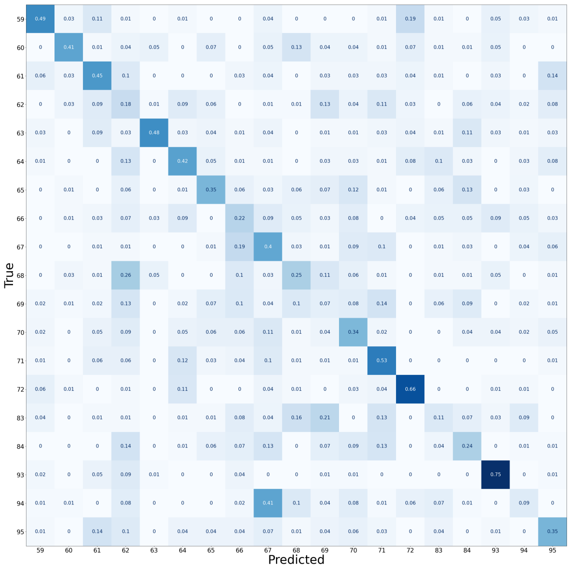


Figure 9: Confusion matrix of [DI_South].

The examples are presented in Table 14. The red text in the 'Regional Data' column indicates errors when trained on a specific regional dialect. For the 'Entire Data' column, red text represents errors that were not resolved, orange text indicates errors that were partially resolved but not entirely, and green text denotes errors that were completely resolved. The models trained on the entire Vietnamese dataset perform better than those trained only on specific regional dialects. For instance, the confusion between 'd' and 'gi' in the Northern dialect was resolved. In the Central Dialect, although the spelling was not perfectly accurate, the model's prediction more closely mirrored the original phonetics than the model from the [SR_Central] experiment. However, this improvement was still quite limited, as exemplified by the sample from the Southern dialect. Figure 12 presents the Word Error Rate (WER) discrepancy when the model is fine-tuned on the entire dataset and fine-tuned on three sub-datasets. Red indicates instances where fine-tuning on the entire dataset performs worse, while blue indicates instances where fine-tuning on the entire dataset performs better.

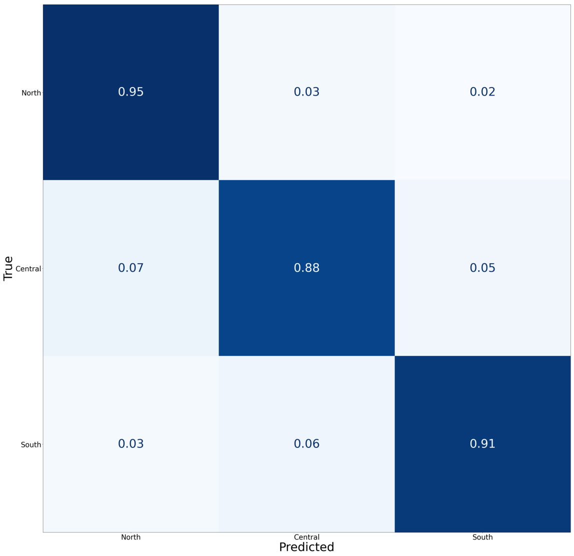


Figure 10: Confusion matrix of [DI_VN_3].

Task	Province code	Reference Transcript	Without Fine-tuned	Fine-tuned	English
[SR_North]	24	năm ngày nay là tất cả các nhà máy đang ngừng hoạt động ở lào cai rồi ninh bình thanh hóa hưng yên ngừng hết	năm ngày nay là tất cả các nhà máy đang ngừng hoạt động ở lào cai rồi linh bình thanh hóa hưng yên ngừng hết	năm ngày nay là tất cả các nhà máy đang ngừng hoạt động ở lào cai rồi ninh bình thanh hóa hưng yên ngừng hết	For the past five days. all factories in Lào Cai, Ninh Bình, Thanh Hóa, and Hưng Yên have been shut down.
	25	người lao động thôn được đào tạo nghề đã tìm được việc làm mới	người lao động thôn được đào tạo nghề thì đã tìm được việt nàm mới	người lao động thôn được đào tạo nghề đã tìm được việc làm mới	The villager workers who were trained in vocational skills have found new jobs.
[SR_Central]	75	các đối tượng là thương binh bệnh binh	két đối tượng là thương binh bệnh binh	các đối tượng là thương binh bệnh binh	The individuals are war invalids and sick soldiers.
	76	tất cả là trâu bò phải bán để cho cháu đi chữa bệnh	tất cả là trâu bò phải bón để cho cháu đi chữa bệnh	tất cả là trâu bò phải bán để cho cháu đi chữa bệnh	All the buffaloes and cows had to be sold to pay for the child's medical treatment.
	77	đã hai năm rồi nhưng mà cũng không có thấy công ty	đã hai nem rồi nhưng mà cũng không có thấy công ty	đã hai năm rồi nhưng mà cũng không có thấy công ty	It has been two years, but there is still no sign of the company.
[SR_South]	67	các anh công an làm sai định danh	các anh công an làm xai định danh	các anh công an làm sai định danh	The police officers made a mistake in the identification.
	94	chúng tôi là có hai mươi mô hình nổi trội	chúng tôi là có hai mươi mô hình nổi chội	chúng tôi là có hai mươi mô hình nổi trội	We have twenty outstanding models.

Table 13: Errors of the best-performing models in [SR_North], [SR_Central], and [SR_South] Experiments.

Reference Transcript	Regional Data		Entire Data	English
	Experiment	Transcript	Transcript	
độ **dày** của lá là **dày** hơn	[SR_North]	độ **giày** của lá là **giày** hơn	độ **dày** của lá là **dày** hơn	The thickness of the leaf is thicker
nghề này là **lúc rảnh rỗi** là mình làm	[SR_Central]	nghề này là **rút rẻn rổi** là mình làm	nghề này là **lút rảng rỗi** là mình làm	During free time, I do this job
chế biến xong rồi sẽ chia **lên khay**	[SR_South]	chế biến xong rồi sẽ chia **lơn khai**	chế biến xong rồi sẽ chia **lên khai**	After cooking, I'll put it into serving trays

Table 14: Speech recognition performance improvement of experiment [SR_VN_63] over remaining experiments.

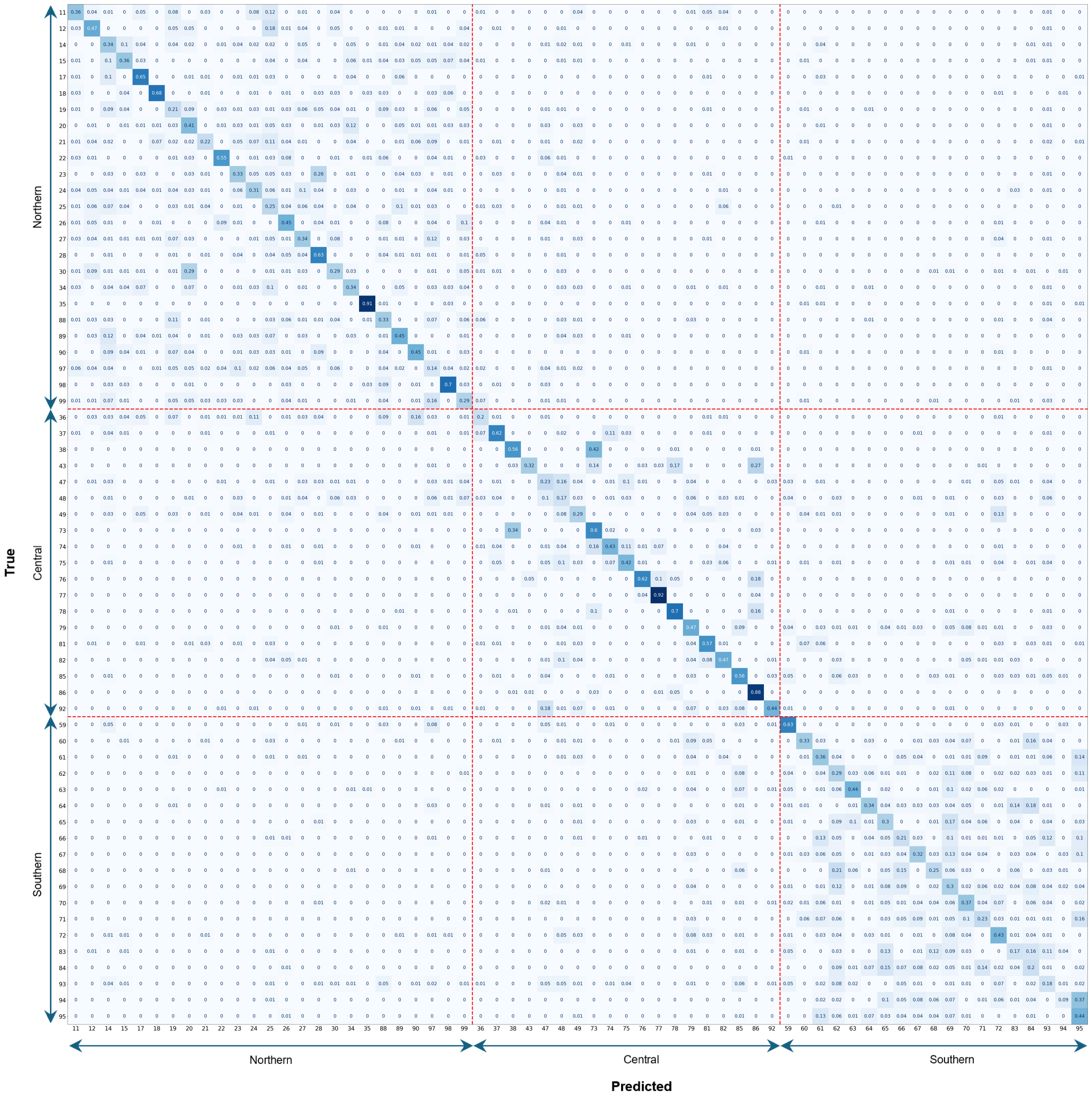


Figure 11: Confusion matrix of [DI_VN_63].

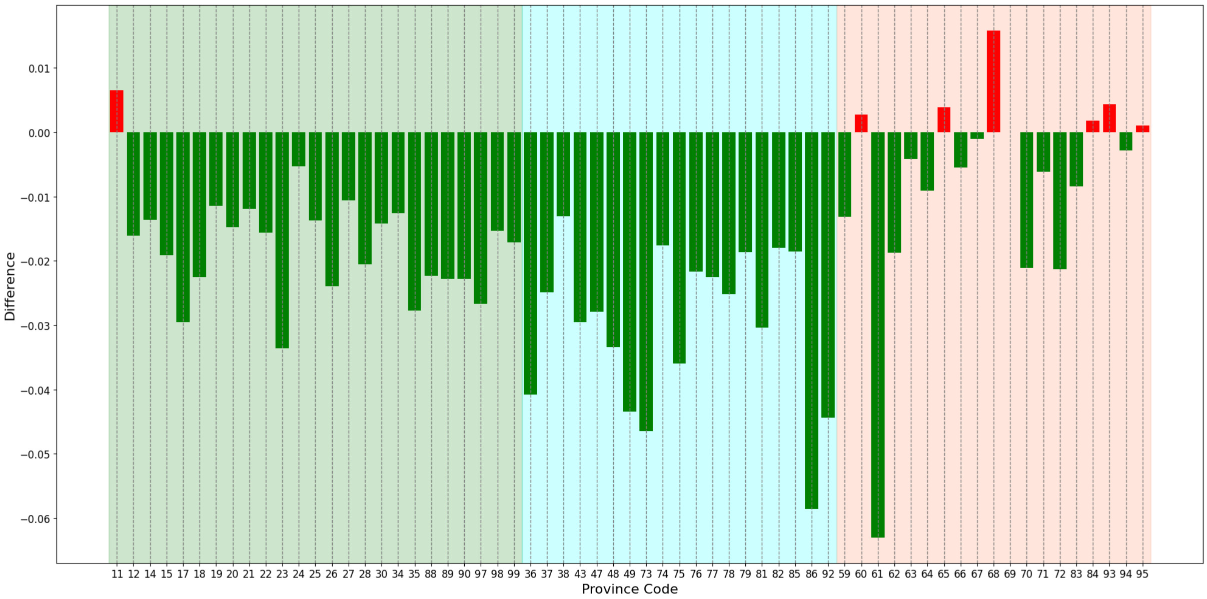


Figure 12: WER discrepancy when fine-tuning the model on the entire dataset versus three sub-datasets.