Gender and Dialect Classification for the Vietnamese Language

Tran Nguyen[1,2,3], Uyen Nguyen[1,2,3], Thinh Pham[1,2,3], Truc Nguyen[1,2,3], Binh T. Nguyen[1,2,3*]

[1] Vietnam National University Ho Chi Minh City, Vietnam,
[2] University of Science, Ho Chi Minh City, Vietnam,
[3] AISIA Research Lab, Vietnam

Abstract

Gender and dialect detection in voice recordings play a critical role in personalizing user experience and enhancing the accuracy and effectiveness of speech recognition and natural language processing systems, particularly in Vietnamese — a tonal language where variations in pitch or tone can entirely alter a word's meaning, yet exhibits diverse regional variations. Despite the importance of these tasks, there is a notable lack of labeled Vietnamese datasets. This study introduces a novel benchmark dataset, ViSpeech, containing 10,686 files from 449 speakers totaling more than 14 hours of speech. The dataset offers a balanced class distribution, covering both genders and the three main dialects of Vietnamese. Additionally, this paper comprehensively evaluates various CNN-based models on these classification tasks, focusing on the impact of data augmentation and model architecture. Our analysis demonstrates that ResNet models excel in both tasks, with ResNet18 achieving 98.73% accuracy in gender classification on noise-free recordings and 98.14% on recordings with background noise, while ResNet34 in dialect classification achieves accuracies of 81.47% and 74.8%, respectively. Moreover, the results underscore the importance of data augmentation in enhancing model robustness, particularly in noisy conditions. Our findings highlight the potential for further improvements and the practical applicability of the proposed framework in real-world settings.

Keywords: dialect detection, gender detection, mel spectrogram, CNN-based model

1 Introduction

Vietnamese is a tonal language, meaning that the pitch or tone with which a word is pronounced can entirely change its meaning. The tonal system is characterized by using six distinct tones in the Northern dialect, defining the language's complexity. However, the tonal range varies across the country, with some Southern and Central dialects utilizing fewer tones, adding another layer of regional diversity. Even more challenging is the variation in regional dialects, which differ not only in tonal pronunciation but also in the articulation of vowels and consonants. This variability presents a unique challenge in both human communication and audio-based technologies. Additionally, gender plays an essential role in Vietnamese people's tonal and phonetic landscape. Typically, men and women exhibit differences in pitch, speech rate, and intonation. These differences can impact how tones are realized and perceived, further complicating the task of speech recognition. These factors highlight the importance of dialect and gender classification for Vietnamese.

Accurately recognizing both dialect and gender in Vietnamese is crucial to enhancing the performance of various natural language processing applications. For example, speech-to-text systems can account for regional dialects and the speaker's gender to support accurate transcriptions (Bhukya, 2018). Additionally, as voice assistants become more popular, they need to adapt to these variations to deliver personalized and effective responses.

At present, research on dialect and gender speech classification in Vietnamese remains relatively limited, representing a promising area for further exploration. Moreover, there is a shortage of accessible labeled Vietnamese audio datasets, posing a challenge for such research. In this study, we present a speech dataset that includes recordings extracted from YouTube videos, annotated with both gender and dialect labels. Based on this data set, we will propose a framework to provide a robust solution for gender and regional dialect classification in the Vietnamese language.

The contribution of this study is twofold:

*Corresponding author: Binh T. Nguyen (e-mail: ngtbinh@hcmus.edu.vn).

1. The introduction of a novel Vietnamese speech dataset that features both male and female speakers and encompasses three distinct dialects from the regions of North, Central, and South Vietnam.

2. The implementation and evaluation of a proposed method utilizing convolutional neural networks (CNN)-based architectures and mel spectrogram features for the task of Vietnamese dialect and gender classification.

2 Related Work

Several scientific publications have significantly contributed to enhancing the quality of voice recognition systems, employing a diverse range of methodologies and classification approaches.

In 2021, the study titled "Accent and Gender Recognition from English Language Speech and Audio Using Signal Processing and Deep Learning" investigated the classification of speakers' regional origins and genders from the United Kingdom (Jagjeevan et al., 2021). This research utilized Fourier transforms in conjunction with deep convolutional neural networks (CNNs) to analyze the speech data. The findings revealed that gender classification achieved higher accuracy than accent classification, with the latter being more challenging due to the overlapping nature of regional accents, which hindered accurate classification.

In 2022, Chrisina et al. conducted a comprehensive review of contemporary research on automated recognition of geographical origin and gender based on six regional dialects of the United Kingdom (Chrisina et al., 2022). This study assessed the performance of various machine learning classifiers, including Artificial Neural Networks (ANNs), Support Vector Machines (SVMs), Random Forests (RF), and k-nearest neighbors (k-NN). The evaluation, conducted on a dataset of 17,877 voice samples categorized by gender and dialect, showed that ANNs, SVMs, and k-NN outperformed RF in classification tasks, although all models demonstrated reasonable performance.

In the Vietnamese context, a 2016 study titled "Automatic Identification of Vietnamese dialect" (Hung et al., 2016) employed acoustic features like MFCCs and F0 variations combined with Gaussian Mixture Models (GMMs) to improve dialect recognition. Using the VDSPEC corpus, which includes recordings from 150 speakers across Northern, Central, and Southern dialects, the Hanoi voice is chosen for the northern dialect, the Hue voice for the central dialect, and the Ho Chi Minh City voice for the southern dialect, the study achieved a recognition rate of up to 75.1% by varying GMM components. The findings highlight the effectiveness of combining MFCCs, formant frequencies, and F0 in enhancing Vietnamese speech recognition accuracy.

In 2020, Hung introduced a methodology for predicting the gender and regional origin of Vietnamese voices using a deep learning approach based on acoustic features (Hung, 2020). The study involved extracting Mel Spectrogram features from 270 samples corresponding to two genders and three regions from the ZaloAI dataset. These features were then utilized to train and optimize a Convolutional Neural Network (CNN). The evaluation of this method, conducted on a sample of 37 recordings from the VIVOS [1] corpus, achieved an accuracy of 86.48% for gender classification and 51.45% for regional classification.

In 2021, the Viettel Cyberspace Center (Tien and Hai, 2021) introduced an accent corpus for Vietnamese speech and conducted a comparative study of various accent classification methods, including Random Forests, CNNs, and ResNet50 models. The corpus consisted of 3,000 audio files, which were divided into training, development, and test sets. The experimental results indicated that the CNN-based model outperformed other methods, achieving an accuracy of 76.1% on the development set and 73.9% on the test set, underscoring the effectiveness of CNNs in Vietnamese accent recognition tasks.

Most research on Vietnamese speech recognition primarily relies on proprietary corpora, lacking large publicly available datasets. However, recent efforts have resulted in collecting several datasets, summarized in Table 1. These datasets offer valuable resources for gender and dialect recognition but also present challenges related to data quality and dialect balance that need to be addressed for optimal system performance.

3 Dataset

The creation of the Vispeech dataset involves three primary stages: Dataset Collection, Data Annotation, and Annotation Validation. Each of these phases is elaborated in the following subsections.

[1] VIVOS Dataset: http://ailab.hcmus.edu.vn/vivos

Table 1: Recent Datasets from Vietnam.

Dataset	Overview	Label	Properties
VIVOS	· 15 hours · 12,420 utterances from 50 Vietnamese speakers	· Transcript · Gender	The dataset exclusively includes speakers from Southern Vietnam.
FOSD (Chung, 2020)	· 30 hours · 25,921 utterances	· Transcript · Timestamp · Gender	The presence of some unclean data files may impact the quality of text-to-speech (TTS) and speech-to-text (STT) engines.
ViASR (Binh et al., 2023)	· 32 hours · 4,276 transcribed chunks	· Transcript	The dataset is up to request.
Vietnam-Celeb (Pham et al., 2023)	· 187 hours · 87,000 utterances from 1,000 Vietnamese speakers	· Transcript · Gender · Region	The dataset includes a skewed dialect representation, with fewer Central dialect speakers.

3.1 Data Collection

The dataset was sourced from YouTube. The data collection process involved manually selecting videos featuring speakers with identifiable dialects. Google API was utilized to download the selected content. Subsequently, the downloaded videos were converted into MP3 format using the PyDub[2] library. Given that most of these files are in 2-channel audio format, the "libsora" library was used to convert them into single-channel audio, ensuring consistency and ease of processing.

To extract samples containing human speech, a filtering step was implemented to remove segments containing minimal or no speech, such as those primarily consisting of silence, background music, laughter, or other noises. By incorporating the Voice Activity Detection (VAD) model (Tan et al., 2020), the focus was placed solely on the human speech signal, effectively removing non-speech elements. This allowed the extraction of relevant speech segments and divided the MP3 audio files into smaller chunks. If the speech is too short, it may be possible to recognize the dialects. Therefore, we retained audio segments of approximately no less than 1.5 seconds in length to provide a sufficient duration for capturing distinct pronunciation patterns, intonations, and other dialect-related features.

The dataset consists of two sections: one with clean speech, free from background noise, and another with ambient noise. A clean speech dataset is essential because it ensures that the features extracted from the speech data are more representative of the actual speech content, as noise can distort the signal. It also facilitates easier error analysis and supports data augmentation techniques, enriching the dataset and improving model generalization to more complex, noisy environments. On the other hand, a noisy dataset is essential to test the model's robustness to recordings in real-world settings. To ensure the quality of the clean dataset, an additional step was taken where human reviewers meticulously verified the audio files, retaining only those free of noise and extraneous sounds.

The overall data collection process is depicted in Figure 1, which is then followed by the annotation and validation process.

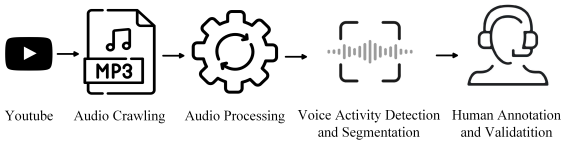

Figure 1: Workflow of the creation of ViSpeech dataset.

3.2 Data Annotation Process

The annotation process engaged four undergraduate students, all of whom had prior experience working with various datasets in Vietnamese Natural Language Processing. Prior to commencing their work on the assigned data samples, the annotators were instructed to strictly adhere to the provided guidelines. The guidelines were designed to assist annotators in accurately identifying and labeling audio samples, with particular emphasis on assigning correct speakers' gender, dialect, and identity. While identifying gender was relatively straightforward, differentiating between dialects, especially Northern and Central dialects, proved more challenging. The guidelines included several key instructions. First, in dialect annotation, the focus was on the speaker's intonation and pronunciation rather than their place of origin, as speakers might adopt different dialects over time. Second,

[2]https://pypi.org/project/pydub/

due to the presence of mixed dialects, only samples with high confidence in their dialect label were kept. Finally, the annotators were also required to verify that the VAD model accurately segmented the speech, ensure that there was no background noise for the clean dataset creation, and confirm that each sample strictly contained only a single speaker.

3.3 Validation of Annotations

The annotated data was subjected to a validation process to ensure its reliability and quality. Annotators engaged in self-validation by reviewing their work after every 300 samples, carefully documenting and correcting any errors. This method was implemented to maintain a high standard of annotation accuracy. Additionally, a cross-validation stage was conducted, where each annotator reviewed the work of a different annotator. The primary objective of this validation process was to preserve the integrity of the annotated data, making it suitable for academic and professional research.

3.4 Dataset Analysis

3.4.1 Overview

The dataset [3] comprises 10,686 mp3 files, totaling slightly over 14 hours of speech data from 449 speakers representing both genders across the three primary Vietnamese dialects: Northern, Central, and Southern. It is divided into three subsets: a training set with clean recordings and two test sets—one with clean recordings and the other with ambient noise. Notably, the speakers in the training set are independent of those in the test sets. The dataset is designed to provide a diverse and comprehensive resource for audio classification research. Table 2 presents key statistics for the subsets.

Table 2: Overview statistics of the ViSpeech dataset.

	Train set	Test set	
		Clean	Noisy
Audio samples	8,166	1,500	1,020
Max length (s)	14.0	13.0	14.3
Avg. length (s)	4.8	4.6	4.9
Min length (s)	1.6	1.8	2.7
Unique speakers	310	84	66

The duration of each audio in the dataset ranges from 1.5 to 15 seconds. The distribution by gender and dialect is detailed in table 3. It is noteworthy that the two test sets are perfectly balanced across classes concerning both the number of files and speakers. Similarly, the distribution within the training set is also nearly uniform, ensuring minimal bias.

Table 3: Distribution of the ViSpeech dataset by Gender and Dialect Categories.

		Northern	Central	Southern
Number of samples				
Male	Training set	1304	1228	1374
	Clean test set	250	250	250
	Noisy test set	170	170	170
Female	Training set	1509	1244	1506
	Clean test set	250	250	250
	Noisy test set	170	170	170
Number of unique speakers				
Male	Training set	52	52	51
	Clean test set	14	14	14
	Noisy test set	11	11	11
Female	Training set	51	53	51
	Clean test set	14	14	14
	Noisy test set	11	11	11

3.4.2 Characteristics of Dataset

The dataset is meticulously curated to ensure high quality and diversity, which enables robust model training and accurate analysis.

Diversity: The dataset comprises a wide range of pitch variations, including both high-pitched and low-pitched voices within each gender, to ensure comprehensive coverage. Additionally, it incorporates voices with various qualities, such as breathy, creaky, and nasal tones, enabling the model to manage diverse vocal characteristics across different genders and dialects effectively. The dialect diversity spans Southern dialects from regions like the Cuu Long Delta and Southeast, Northern dialects, and Central dialects, which are notably diverse, with each province exhibiting its own variations. These Central dialects include those from areas such as Thanh Hoa-Nghe Tinh, Quang Nam, Quang Ngai, Hue, Phu Yen-Binh Dinh, and Đak Lak. A major challenge in collecting Central dialect data is the scarcity of clean, high-quality videos available on YouTube. Most videos with high-quality audio come from the entertainment industry, where many individuals from the Central region, particularly artists who are prominent speakers in the dataset,

[3] https://github.com/TranNguyenNB/ViSpeech

often adopt Southern and Northern dialects. This switch is frequently due to the prevalent use of local expressions and strong regional accents in Central dialects, which can sometimes hinder effective communication. Consequently, we were able to find only a limited number of sources for Central dialects, with the Hue dialect being particularly prevalent, making it the most dominant Central dialect in the dataset.

Noise Level: Noise can obscure or distort phonetic features crucial for distinguishing dialects, often involving subtle variations in pronunciation, intonation, and stress patterns. To ensure high-quality audio, files are manually selected to be noise-free. While low-level white noise may still be present, it is maintained at a level that does not interfere with phonetic clarity. Non-verbal sounds like laughter, coughing, and filler words ("uhm", "ah") are minimized to maintain the audio's clarity. A noise-free dataset also allows for enrichment through data augmentation, introducing variations that simulate different recording conditions or speech patterns.

One Speaker in One Audio: Each audio sample is restricted to a single speaker to prevent the presence of multiple dialects or genders within a single audio file. This approach guarantees precise labeling and minimizes confusion during feature extraction, training, and inference phases.

4 Methodology

In this section, we will present our approaches to the main problem. Figure 2 illustrates the proposed workflow for both gender and dialect classification pipelines. The pipelines encompass several stages, including data loading and preprocessing. After preprocessing, the data is transformed into features, which are subsequently used by the models to perform classification.

4.1 Training and Testing Datasets

The training set was further divided into training and validation subsets with an 85:15 file ratio using the `train_test_split` function from the "scikit-learn" library, implementing a stratified approach to preserve balanced class distributions across both subsets. The evaluation will be performed on both test datasets. Notably, the training and test sets consist of distinct speakers, thereby preventing data leakage, ensuring an unbiased evaluation, and improving the accuracy of the model's generalization assessment.

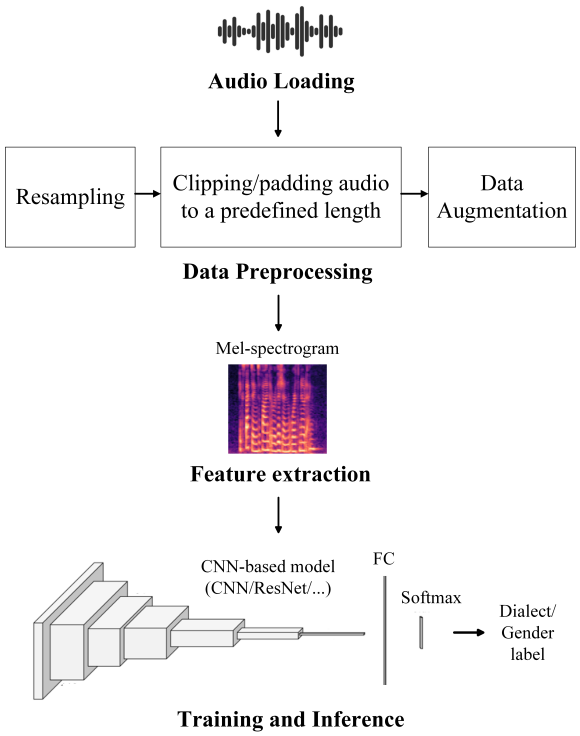

Figure 2: Workflow of the proposed framework for both gender and dialect pipelines.

4.2 Data Preprocessing and Feature Extraction

Most YouTube audio is recorded at 44,100 Hz, a standard rate for high-quality audio but resource-intensive. The audio was resampled to a lower sampling rate for both tasks to reduce computational demands. Next, each audio clip was clipped or padded to a predefined length by repeating the initial segment to ensure uniform duration and preserve natural sound characteristics across clips. Data augmentation techniques were employed to enhance model robustness, including Gaussian noise injection, reverberation, speed perturbation, and background noise injection, such as adding instrumental music, street sounds, rain, and footsteps. Finally, the audio was converted into Mel-spectrograms. The configurations for speech preprocessing and feature extraction are presented in Table 4.

4.3 Models

In this study, we explore CNN-based architectures on both gender and dialect classification tasks.

ResNet34 and ResNet18: ResNet models are known for their use of residual connections, which effectively mitigate the vanishing gradient problem, allowing deeper networks to be trained success-

Table 4: Configuration for Speech Processing and Feature Extraction.

Hyperparameters	Gender Model	Dialect Model
Sampling Rate (Hz)	16000	22050
Audio Length (s)	4	5
Number of Mel Bands	40	64
Window Length (ms)	25	25
Step Size (ms)	10	10

fully. ResNet18, with its shallower architecture, offers faster processing and is well-suited for real-time applications, while ResNet34, being deeper, can capture more complex audio patterns. The residual connections in these models enable efficient training, crucial for distinguishing subtle differences in gender and dialect features.

DenseNet121: DenseNet121's unique architecture connects each layer to every other layer, ensuring maximum feature reuse and efficiency. This dense connectivity reduces the number of parameters, making the model more efficient and capable of learning rich and detailed features. This characteristic is particularly valuable in tasks requiring high precision, such as dialect differentiation, where subtle acoustic variations are critical.

MobileNet: MobileNet is a lightweight CNN model that uses depthwise separable convolutions to significantly reduce the number of parameters and computational costs while maintaining strong performance. Despite its compact architecture, it effectively extracts relevant features for audio classification tasks, making it a strong candidate for scenarios where balancing model size and performance is crucial.

Each model backbone, initialized with a pre-trained model from the Hugging Face Hub, is followed by a fully connected layer that projects the extracted features to the target classes, with a subsequent Softmax layer to output class probabilities. With a learning rate of 1e-4 and batch size of 32, the Adam optimizer is utilized to update model parameters, guided by cross-entropy loss. The random seed is set to 42 for reproducibility.

5 Results and Discussion

We conducted all experiments on a computer Intel(R) Xeon(R) Gold 5320 CPU @ 2.20GHz with 32GB of RAM and an Nvidia A30 Tensor Core GPU with 24GB VRAM. The results are detailed as follows.

5.1 Performance comparision

Gender classification: Table 5 shows the performance of various CNN-based models in gender classification, emphasizing the relative simplicity of the task for these models. Even the baseline CNN model achieves around 98% accuracy on both test sets, indicating that this task is effectively handled by CNN architectures. ResNet18 outperforms the clean test set with 98.73% accuracy, while ResNet34 leads on the noisy test set at 98.63%. However, The marginal gap of about 0.5% in accuracy between the two models on each test set is negligible, making ResNet18 a more appealing choice due to its lower resource demands. Additionally, data augmentation on ResNet18 slightly reduces its clean test set accuracy but substantially enhances its performance on the noisy test set, underscoring the value of augmentation in noisy environments. The DenseNet121 model, although generally considered more powerful than ResNet models due to its dense connectivity pattern, does not result in a significant performance increase compared to the ResNet models in this context and is also more resource-intensive. Given that even simpler models like CNN perform well, we explored using resource-efficient models like MobileNet_v2. The performance of MobileNet_v2 was found to be comparable to that of ResNet18, making it a suitable choice for deployment on mobile devices.

Table 5: Evaluation results of gender classification using CNN-based models on the ViSpeech dataset.

Gender Model	Accuracy (%)	
	Clean test set	Noisy test set
$\text{ResNet18}_{\text{w/o augment}}$	98.80	97.06
ResNet18	98.73	98.14
ResNet34	98.20	98.63
MobileNet_v2	98.07	98.33
DenseNet121	98.27	97.65
Shallow CNN	98.07	97.45

Dialect classification: Table 6 presents the performance of the CNN-based models in the task of dialect classification. The results clearly demonstrate that data augmentation not only enhances generalization on clean data but also significantly boosts the model's resilience to noise. Specifically, ResNet34 shows an approximately 3% increase in accuracy on the clean test set and a substantial 8.62% improvement on the noisy test set with data

augmentation. The baseline CNN model, even with augmentation, has the lowest accuracy on both test sets, with less than 6% gaps compared to ResNet34 without augmentation, highlighting its limitations in this task. DenseNet121, while more complex and resource-intensive than the ResNet models, does not perform better. ResNet18 and ResNet34 share the same accuracy on the clean test set (81.47%). However, on the noisy test set, ResNet18's accuracy drops to 73.14%, while ResNet34 slightly outperforms with 74.8%, though the difference is minimal. The notable decrease of approximately 7% in dialect classification performance on the noisy test set highlights the need for further analysis and refinement.

Table 6: Evaluation results of dialect classification using CNN-based models on the ViSpeech dataset.

Dialect Model	Accuracy (%)	
	Clean test set	Noisy test set
$\text{ResNet34}_{\text{w/o augment}}$	78.20	66.18
ResNet34	81.47	74.80
ResNet18	81.47	73.14
DenseNet121	81.00	73.24
Shallow CNN	72.53	63.92

5.2 Error Analysis and Discussion

This section provides an error analysis of the performance of the gender and dialect classification models. The most effective baseline models were chosen for this analysis: ResNet18 for gender classification and ResNet34 for dialect classification.

Gender classification error: Misclassification was observed in both gender categories. Female voices were incorrectly classified as male, often due to their low-pitch, deep, and husky vocal characteristics. Conversely, some male voices were misclassified as female, likely because of their high-pitched, light, and clear tones. However, it is noteworthy that for each speaker involved in these misclassification cases, most of their other audio samples were correctly classified, with only a few instances being misclassified. This indicates that while the model generally performs well, it may encounter challenges with edge cases where vocal characteristics overlap between genders.

Dialect classification error: An analysis of the performance on the clean test set, where linguistic features are expected to be unaffected by noise, reveals that out of the 84 speakers, 19 were classified correctly with no errors. Although some utterances were misclassified for the remaining speakers, no speaker was entirely misclassified. The overall error rate across speakers was determined to be 18.7%.

Distinguishing Vietnamese dialects presents challenges due to several factors, including the similarities shared across different dialects. There has been discussion regarding the number of dialects within Vietnam. Various studies have identified between one and nine distinct dialects of Vietnamese spoken throughout the country. However, the most widely accepted classification divides Vietnamese dialects into three primary categories: northern, central, and southern (Phạm and McLeod, 2016). Despite this division, dialects in certain regions may exhibit greater similarity to another dialect group (Pham, 2005) (Thi, 2004). For instance, in terms of tonal characteristics, the dialects of the south-central regions exhibit similarities with those of the southern regions and are often classified as part of the Southern dialect group. In the misclassification cases by the ResNet34 model, among the three instances of Central dialect misclassification with error rates exceeding 50%, two involved speakers from South Central Vietnam—one from Binh Dinh and the other from Quang Ngai. Specifically, the speaker from Binh Dinh had 10 out of 13 cases misclassified as the Southern dialect, while the speaker from Quang Ngai had 15 out of 18 cases similarly misclassified.

Another challenge arises from speakers exhibiting a mix of dialects due to migration and prolonged exposure to different linguistic environments. For instance, an individual born in the northern region of Vietnam who later relocates to the southern region may adapt to the consonant pronunciation of the local dialect while retaining the tonal features of their original northern dialect.

A further complication in distinguishing dialects stems from the dominance of the Hue dialect in the dataset, which possesses unique tonal patterns and vocabulary that differ significantly from other Vietnamese dialects. Due to this distinctiveness, models can struggle to accurately categorize other Central dialects that deviate from the Hue dialect, sometimes leading to their misclassification as either Northern or Southern dialects and vice versa.

6 Conclusion

The paper has presented ViSpeech, a novel benchmark dataset tailored for Vietnamese gender and

dialect speech detection. The dataset comprises 10,686 files from 449 speakers and more than 14 hours of meticulously curated audio, ensuring a balanced representation across different classes and encompassing diverse Vietnamese dialects. While its primary focus is on gender and dialect detection, ViSpeech is versatile and can also be utilized for various other applications, including speech recognition with annotated speaker labels, signal processing, and broader speech processing tasks.

In addition, we have evaluated various CNN-based models to assess their performance, with the ResNet models demonstrating strong performance across both dialect and gender classification tasks. The analysis highlights the significant impact of data augmentation and model architecture on accuracy. Data augmentation for dialect classification proves crucial in enhancing generalization on clean data and significantly improving resilience to noise, as evidenced by ResNet34's performance gains, achieving 81.4% accuracy on the clean test set and 74.8% on the noisy test set. While ResNet18 matches ResNet34 in accuracy on the clean test set, ResNet34 outperforms in noisy environments. In gender classification, the task's relative simplicity is evident, with even the baseline CNN model achieving approximately 98% accuracy on both test sets. ResNet18, with 98.73% accuracy on the clean test set and 98.14% on the noisy test set, is a suitable choice for balancing resource efficiency with accuracy. However, ResNet34 exhibited slightly superior performance in noisy conditions with 98.63% accuracy. Additionally, an error analysis was conducted to identify challenges and limitations faced by the models, offering valuable insights for future research and potential areas for improvement.

7 Limitations and Future Works

While this framework has achieved promising results, there remains room for improvement. The significant drop in dialect classification performance on the noisy test set indicates the need for further analysis and refinement. Enhancements could include incorporating additional data augmentation techniques, such as SpecAugment (S. et al., 2019), pitch shifting (Galic and Grozdić, 2023), and introducing more diverse background noise (Nicolas et al., 2007) (Pervaiz et al., 2020) to boost the model's robustness to diverse real-world speaking environments. Training on a larger, more diverse dataset representing a wider range of accents within each dialect and exploring different feature extraction methods, like Mel-frequency cepstral coefficients (MFCCs) (Silvestre and Ferreira, 2023), Wavenet Features (Tri-Nhan et al., 2020), and experimenting with state-of-the-art models, such as Wav2Vec (Baevski et al., 2020), could also advance dialect classification. For gender classification, considering the good performance of ResNet variants and mel-spectrograms, it can be beneficial to explore more compact, resource-efficient models or other robust feature extraction methods that can maintain strong performance. This approach would facilitate deployment in resource-constrained environments and real-time applications.

Regarding the dataset, the Central dialect class is predominantly represented by the Hue accent despite the rich diversity of dialects across various provinces in the Central region, highlighting the need for more comprehensive data collection. The limited representation of dialects in the dataset may affect the model's ability to perform accurately in real-world scenarios, where a wider variety of dialects might be encountered. Additionally, the involvement of human annotation introduces the possibility of errors. The dataset also has significant potential for improvement; expanding its size and including transcriptions could enhance its utility for various speech research areas, including text-to-speech and speech-to-text tasks.

8 Acknowledgments

We express our gratitude to the University of Science, Vietnam National University Ho Chi Minh City, and AISIA Research Lab for their support, guidance, and resources, which were vital to the success of this study. Tran Nguyen acknowledges the support from the AISIA Extensive Research Assistant Program 2023 (Batch 1) during this work.

References

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Preprint*, arXiv:2006.11477.

Sreedhar Bhukya. 2018. Effect of gender on improving speech recognition system. *International Journal of Computer Applications*, 179:22–30.

Nguyen; Binh, Huynh; Son, Quoc Khanh Tran, Tran-Hoai; An Le, Nguyen; Trong An, Tran; Nguyen Tung

Doan, Thi; Thuy An Phan, Nguyen; Le Thanh, Nguyen; Hieu Nghia, and Huynh; Dang. 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, Hong Kong, China. Association for Computational Linguistics.

Jayne; Chrisina, Chang; Victor, Bailey; Jozeene, and Xu; Qianwen. 2022. Automatic accent and gender recognition of regional uk speakers.

Tran Duc Chung. 2020. Fpt open speech dataset (fosd) - vietnamese. *Mendeley Data*, 04.

Jovan Galic and Đorđe Grozdić. 2023. Exploring the impact of data augmentation techniques on automatic speech recognition system development: A comparative study. *Advances in Electrical and Computer Engineering*, 23:3–12.

Bui Hung. 2020. Vietnamese voice classification based on deep learning approach. international journal of machine learning and networked collaborative engineering. 04:171–180.

Pham Hung, Loan Trinh Van, and Nguyen Quang. 2016. Automatic identification of vietnamese dialects. *Journal of Computer Science and Cybernetics*, 32:19–30.

Shergill; Jagjeevan, Pravin; Chandresh, and Ojha; Varun. 2021. Accent and gender recognition from english language speech and audio using signal processing and deep learning.

Morales; Nicolas, Gu; Liang, and Gao; Yuqing. 2007. Adding noise to improve noise robustness in speech recognition. volume 2, pages 930–933.

Ayesha Pervaiz, Fawad Hussain, Humayun Israr, Muhammad Ali Tahir, Fawad Riasat Raja, Naveed Khan Baloch, Farruh Ishmanov, and Yousaf Bin Zikria. 2020. Incorporating noise robustness in speech command recognition by noise augmentation of training data. *Sensors (Basel, Switzerland)*, 20.

Hoa Pham. 2005. Vietnamese tonal system in nghi loc. *Toronto Working Papers in Linguistics*, 24.

Viet Thanh Pham, Xuan Thai Hoa Nguyen, Vu Hoang, and Thi Thu Trang Nguyen. 2023. Vietnam-Celeb: a large-scale dataset for Vietnamese speaker recognition. In *Proc. INTERSPEECH 2023*, pages 1918–1922.

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.

Park; Daniel S., Chan; William, Zhang; Yu, Chiu; Chung-Cheng, Zoph; Barret, Cubuk; Ekin D., and Le; Quoc V. 2019. Specaugment: A simple data augmentation method for automatic speech recognition. In *Interspeech 2019*. ISCA.

Carvalho; Silvestre and Gomes; Elsa Ferreira. 2023. Automatic classification of bird sounds: Using mfcc and mel spectrogram features with deep learning. *Vietnam Journal of Computer Science*, 10(01):39–54.

Zheng-Hua Tan, Achintya kr. Sarkar, and Najim Dehak. 2020. rvad: An unsupervised segment-based robust voice activity detection method. *Computer Speech and Language*, 59:1–21.

Chau; Hoang Thi. 2004. *Phương Ngữ Học Tiếng Việt*. Nhà Xuất Bản Đại Học Quốc Gia Hà Nội.

Duong; Quang Tien and Do; Van Hai. 2021. Development of accent recognition systems for vietnamese speech. In *2021 24th Conference of the Oriental COCOSDA International Committee for the Coordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, pages 174–179.

Do; Tri-Nhan, Nguyen; Minh-Tri, Nguyen; Hai-Dang, Tran; Minh-Triet, and Cao; Xuan-Nam. 2020. Hcmus at mediaeval 2020: Emotion classification using wavenet feature with specaugment and efficientnet. In *MediaEval*, volume 2882 of *CEUR Workshop Proceedings*. CEUR-WS.org.