ViASR: A Novel Benchmark Dataset and Methods for Vietnamese Automatic Speech Recognition

Son Thanh Huynh[1,3,5], **Khanh Quoc Tran**[2,3,5], **An Tran-Hoai Le**[1,3,5], **An Trong Nguyen**[2,3,5], **Tung Tran Nguyen Doan**[5], **An Phan Thi Thuy**[1,3,5], **Thanh Nguyen Le**[2,3,5], **Nghia Nguyen Hieu**[2,3,5], **Dang T. Huynh**[4,5], **Binh T. Nguyen**[2,3,5]*

[1] *University of Science, Ho Chi Minh City, Vietnam*
[2] *University of Information Technology, Ho Chi Minh City, Vietnam*
[3] *Vietnam National University, Ho Chi Minh City, Vietnam*
[4] *Fulbright University Vietnam, Vietnam*
[5] *AISIA Research Lab, Vietnam*

Abstract

The need for accurate speech recognition systems has increased in recent years due to the growing demand for speech-based interfaces in various applications, such as mobile devices and smart speakers. However, current solutions for speech recognition in Vietnamese are limited in accuracy and practicality. To address these limitations, we proposed a novel framework for Vietnamese automatic speech recognition (ASR) that leverages the strength of the transformer-based approach and our benchmark dataset to improve the speech recognition's accuracy. We introduce ViASR, a novel human-annotated dataset that is available to the scientific community as a benchmark for the task of **Vi**etnamese **A**utomatic **S**peech **R**ecognition. The ViASR dataset contains 4,276 transcripts with over 30 hours of audio collected from major Vietnamese news videos. This paper provides an overview of the Vietnamese ASR task, the process of creating the ViASR dataset, and the techniques for carrying out the baseline experiments. Through the implementation and evaluation of the proposed framework, we demonstrated the feasibility of finetuning the OpenAI Whisper model for Vietnamese speech recognition, which was confirmed by the improved accuracy compared to state-of-the-art models. Our findings highlight the potential for further improvements and the practical application potential of the proposed framework in real-world settings.

1 Introduction

Automatic Speech Recognition (ASR) is a rapidly developing field of research that focuses on developing computer algorithms and systems that can transcribe, recognize, and understand spoken language (Benzeghiba et al., 2007; Besacier et al., 2014; Malik et al., 2021). The increasing availability of speech data and advances in deep learning have led to significant progress in ASR, with state-of-the-art systems demonstrating remarkable performance on benchmark datasets (e.g. Wav2Letter++ (Pratap et al., 2019), Conformer (Gulati et al., 2020)). However, low-resource languages, such as Vietnamese, still present a challenge due to the limited availability of data and resources.

Vietnamese is a tonal language (Metze et al., 2013) with a unique phoneme set and pronunciation, which makes it challenging for existing ASR systems ((Luong and Vu, 2016), (Huy Nguyen, 2019)) to transcribe and recognize speech in Vietnamese accurately. Despite this, the increasing demand for speech-based services in Vietnamese-speaking communities highlights the need for effective ASR systems for this language (Kaur et al., 2021). In the case of low-resource languages like Vietnamese, the challenge is to develop speech recognition systems with limited data and resources (Srivastava et al., 2018; Reitmaier et al., 2022). To address this challenge, we collect a large and diverse dataset from various sources to ensure that the ASR system is robust to multiple accents and speaking styles. We also propose a new Vietnamese Automatic Speech Recognition framework that utilizes recent deep learning approaches to improve performance on low-resource languages. Our proposed framework is designed to provide an accurate and robust solution for Vietnamese speech recognition for data from various sources, including customer-salesperson calls, meetings, etc. The system's output is a corresponding text string, which can be utilized to solve

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

multiple problems in fields such as business, education, and health.

The rest of the paper is organized as follows: Section 2 provides an overview of related work in ASR, Section 3 details the process of creating the ViASR dataset, Section 4 introduces our proposed Vietnamese ASR system, Section 5 presents experiment results and discussions, and finally, the paper concludes with conclusions and potential future works.

2 Related work

Automatic Speech Processing is a field of study focusing on developing computer algorithms and systems that can transcribe, recognize, and understand spoken language (Spille et al., 2018). A typical benchmark dataset for evaluating speech recognition systems is the Common Voice dataset (Ardila et al., 2020), a publicly available corpus of speech data in multiple languages, including Vietnamese. The state-of-the-art method in Automatic Speech Processing is Deep Learning, which has led to significant advancements in the field in recent years. Deep Neural Networks (DNNs) (Sainath et al., 2017; Song and Cai, 2015), Convolutional Neural Networks (CNNs) (Han et al., 2020; LeCun et al., 1995), and Recurrent Neural Networks (RNNs) (Oruh et al., 2022; Siniscalchi et al., 2014) are commonly used for speech recognition tasks.

Recently, transformer-based methods have been proposed for Automatic Speech Processing and have shown promising results (Kim et al., 2022; Karita et al., 2019; Moritz et al., 2020). One such method is the Whisper model (Radford et al., 2023), a transformer-based speech recognition architecture. The novelty of Whisper compared to other transformer models in speech recognition lies in its approach to weakly supervised pretraining. While prior work has focused on self-supervision or self-training techniques, Whisper takes a simple scaling approach by training on a significantly larger dataset of 680,000 hours of labeled audio data. This dataset is several times larger than existing high-quality datasets and includes multilingual and multitask training. The authors demonstrate that models trained with this approach can transfer well to existing datasets without the need for dataset-specific fine-tuning, achieving high-quality results comparable to fully supervised approaches and approaching human accuracy and robustness.

Overall, the state-of-the-art method in this area is Deep Learning, and developing speech recognition systems in low-resource languages like Vietnamese requires the use of transfer learning and unsupervised learning methods (Reitmaier et al., 2022; Yi et al., 2018).

Overall, the state-of-the-art method in this area is Deep Learning, and novel methods in transfer learning and unsupervised learning would be beneficial to developing speech recognition systems in low-resource languages such as Vietnamese (Reitmaier et al., 2022; Yi et al., 2018).

3 Dataset Creation

The ViASR dataset creation process goes through three phases: Dataset Collection, Data Annotation, and Validation of Annotation. These phases are described in detail as follows.

3.1 Data Collection

The data used for this research was sourced from publicly available online platforms, focusing on videos related to financial and monetary topics, primarily from the YouTube platform. The data collection process involved the use of Selenium[1] to automate browsing and access YouTube video pages to obtain their URLs for downloading. Subsequently, the moviepy[2] was employed to convert the downloaded videos into MP3 audio format, facilitating further processing.

To segment the audio data for easier handling, the pydub library[3] was utilized to divide the MP3 audio files into smaller chunks, each approximately 30 seconds in duration. To collect speech data, the YouTube Transcript API3[4] was employed. This API proved to be an efficient and convenient method to access the transcripts of YouTube videos, which was particularly beneficial for speech recognition projects and other endeavors requiring access to such information. Additionally, the API supported transcripts in various languages, making it a versatile tool for diverse applications.

The data collection process specifically targeted finance-related videos using domain-specific keywords such as “vay vốn"$_{loan}$,

[1] https://pypi.org/project/selenium/
[2] https://pypi.org/project/moviepy/
[3] https://pypi.org/project/pydub/
[4] https://pypi.org/project/youtube-transcript-api/

“dịch vụ tín dụng"$_{credit\ service}$, "ngân hàng nhà nước"$_{state-owned\ bank}$, and “thị trường tài chính"$_{financial\ market}$. To ensure the quality of the collected speech data for proper annotation and the development of a reliable speech recognition system, a filtering step was implemented to remove chunks that contain little to no speech (e.g. chunks of mostly silence, background music, or other noises). This control measure was crucial to facilitate efficient annotation and foster the development of a reliable speech recognition system.

The resultant dataset constitutes a new resource for researchers in Vietnamese automatic speech recognition (ViASR), comprising 4,276 transcribed chunks that totaled around 32 hours of audio. Its significance is further amplified by its diversity, demonstrated by its wide coverage of multiple types of voice from all three geographic areas of Vietnam: Northern Vietnam, Central Vietnam, and Southern Vietnam (see Figure 1). This variety provides potential benefits for research and development in the field of Vietnamese automated speech recognition.

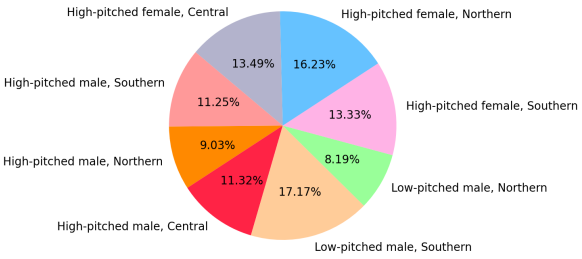

Figure 1: Distribution ratio of types of voice in the ViASR dataset.

In particular, the diversity in the dataset enhances the adaptability and inclusiveness of the ASR system. By encompassing voices from different geographical regions, the dataset reflects a broader spectrum of pronunciation, intonation, and vocabulary variations in the Vietnamese language, and the dataset is aimed to enable the ASR system to gain the ability to learn and adapt to a wider array of vocal characteristics, diverse speaking styles, regional accents, and variations in speech tempo.

3.2 Data Annotation Process

3.2.1 Pilot Annotation Phase

The annotation process involved the participation of four undergraduate students who possessed experience in annotating several datasets in Vietnamese Natural Language Processing. The primary objective of the Pilot Annotation Phase was to familiarize the annotators with the specific task at hand with a small set of samples. An initial set of annotation guidelines, accompanied by illustrative examples, was prepared and provided to the annotators.

Before annotating assigned samples from the collected data, the annotators were instructed to adhere to the provided annotation guidelines properly. These guidelines were meticulously drafted to enable annotators to identify and label transcripts effectively. Noteworthy guidelines included converting numerical representations (e.g., price) into standard formats (e.g., one hundred thousand Vietnam Dongs to 100,000 Vietnam Dongs), placing appropriate punctuation, and differentiating speakers in multi-speaker audio. Regarding numerical representations, we chose to annotate numbers (e.g. "100") instead of pronunciation (e.g. "một trăm") because our study initially focused on improving the speech understanding capacity of our system, especially when the input audio is finance-related.

The Label Studio (Tkachenko et al., 2020-2022) tool was employed for efficient annotation. Each annotator would check, modify, and supplement the transcript while listening to the corresponding audio. The annotated data was then exported to the JSON format and subsequently converted to the CSV format for further use.

3.2.2 Main Annotation Phase

In the primary annotation process, each annotator was assigned 1069 samples to annotate, followed by a cross-checking process to ensure the consistency and accuracy of annotations among annotators. The total number of annotated audio files is about 4,276 (equal to 32 hours of audio).

The dataset was split into two sets to facilitate model training and evaluation: a training set comprising 3,420 audio samples and a test set containing 856 audio samples. To enhance accessibility and encourage further research, the dataset was duly registered with Hugging Face[5], thus amplifying its potential for reuse and knowledge dissemination.

[5] https://huggingface.co

3.3 Validation of Annotations

The annotated data underwent a validation process to ensure its reliability and quality. Annotators performed self-validation by reviewing their work after every 500 samples, documenting and revising errors. This process aimed to maintain a high standard of annotation quality.

Moreover, a cross-validation stage was implemented at the end. During this step, each annotator reviewed and validated the work of another annotator, resolving any conflicts in collaboration with fellow researchers. The objective of the validation process was to uphold the integrity of the annotated data, rendering it appropriate for further academic and professional research endeavors.

3.4 Dataset Analysis

3.4.1 Overview

The dataset has been partitioned into two distinct subsets, the training set and the test set, aimed at creating a diverse and comprehensive data repository to facilitate research in Vietnamese Automatic Speech Recognition. Table 1 presents essential statistics regarding the ViASR subsets.

Table 1: Overview statistics of the ViASR dataset. Note that vocabulary size and comment length are computed at the word level.

	Training set	**Test set**
Audio samples	3,420.0	856.0
Max audio length (s)	28.0	27.9
Avg. audio length (s)	25.3	26.2
Min audio length (s)	5.6	8.1
Max transcript length	150.0	148.0
Avg. transcript length	87.5	88.7
Min transcript length	1.0	1.0

The training set comprises a total of 3,420 audio samples, the test set consists of 856 audio samples, and all samples from each set are restricted to a duration of less than 30 seconds.

3.4.2 Characteristics of ViASR Dataset

The Vietnamese Automatic Speech Recognition (ViASR) dataset presents various technical and linguistic challenges that can impact the quality and accuracy of speech recognition and transcription. These challenges are crucial to acknowledge, as they can influence the overall performance of the ASR system:

1. **Interference Audio:** 1,388 (32.46%) audio samples within the dataset contain additional noise, such as background music, ambient noise, or reverberation resulting from unfiltered microphone pickup. These extraneous elements can obscure speech clarity, making it challenging to discern and accurately transcribe the underlying content.

2. **Regional Accent Speakers:** The presence of speakers with regional accents introduces potential errors in the transcription process due to variations in pronunciation, intonation, and word usage across different regions. These accent-induced discrepancies pose a considerable hurdle to achieving precise and consistent transcriptions.

3. **Multiple Speakers in One Audio:** Certain audio samples comprise conversations involving two or more speakers, necessitating the processing of overlapping speech and accurate segmentation between speakers. Dealing with such complexities demands sophisticated signal processing and advanced speech recognition techniques.

Despite these challenges, the Vietnamese ASR dataset has a significant advantage in its diversity, as discussed in subsection 3.1, and may serve as a cornerstone for developing and refining practical Vietnamese ASR technology.

4 Our Proposed System

Figure 2 depicts the proposed workflow of our Vietnamese ASR system. This system consists of two main components: Speech Preprocessing and the Speech Recognition model. The Speech Preprocessing component involves cleaning and preparing the speech data for feature extraction. Finally, the Speech Recognition model utilizes the extracted features to perform speech recognition. The proposed system has been demonstrated to improve accuracy compared to existing solutions. It has potential practical applications in real-world settings, such as speech-based interfaces for mobile devices or smart speakers.

4.1 Speech Preprocessing

To process the audio data, we use the librosa[6] library to first convert the audio recordings into a

[6] https://librosa.org/

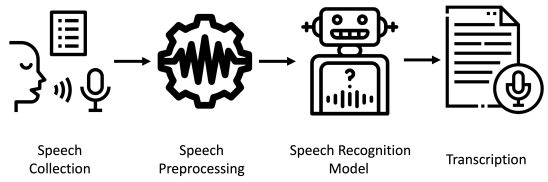

Figure 2: We employ a straightforward workflow for our Vietnamese ASR system that can be adapted for a variety of applications

format the machine can understand and process. This involves converting the audio recordings into 1D arrays, where each value in the array represents the amplitude at the sampling time. The sampling rate used in our pipeline is 16000Hz. Raw audio data can be difficult to work with, so we convert the 1D arrays into Mel-spectrograms, which are more manageable for the machine. To further enhance the training data for the model, we apply two data augmentation techniques, including time masking and frequency masking[7], which artificially increase the size of the training dataset by generating new samples from the existing ones.

4.2 Baseline Model for Vietnamese ASR using Pretrained Models

The primary objective of this study is to develop a highly accurate, efficient, and robust Automatic Speech Recognition system for Vietnamese audio recordings with diverse potential applications in various domains. To achieve this goal, our study thoroughly evaluates different ASR models and algorithms to determine the optimal model for our pipeline. We conduct experiments with various architectures and assess their performance using standard metrics to identify the best-performing model for our ASR system.

In the field of speech recognition, the Transformers architecture has undergone various developments and implementations, such as Wave2Vec (Baevski et al., 2020), XLS-R (Babu et al., 2022), and XLSR-Wave2Vec (Conneau et al., 2021). For our ASR system, we adopt Whisper, the latest model for ASR tasks published by OpenAI (Radford et al., 2023). Our baseline ASR model for Vietnamese involves evaluating the performance of various pre-trained ASR models, with a specific focus on those provided by OpenAI's Whisper (Radford et al., 2023), Wav2Vec2 (Baevski et al., 2020), and MMS (Pratap et al., 2023) transformers architectures.

Whisper, a state-of-the-art ASR system developed by OpenAI (Radford et al., 2023), represents a significant advancement in the field. It incorporates weakly-supervised learning on a massive labeled audio dataset of 680,000 hours to train the ASR model. This allowed Whisper to achieve competitive accuracy and robustness on diverse datasets.

Wave2vec, an innovative self-supervised learning method developed by Facebook AI Research (FAIR) (Baevski et al., 2020), aims to learn contextualized representations directly from raw waveform data without relying on phonetic or linguistic annotations. This departure from traditional ASR methods, which often relied on handcrafted features or linguistic knowledge, allows Wave2vec to excel in capturing discriminative acoustic features and context, making it suitable for downstream ASR tasks.

MMS-300m stands for "Multilingual Masked Sequence-to-Sequence 300 million," a pre-training method introduced by OpenAI (Pratap et al., 2023). Designed to learn multilingual representations from a massive dataset consisting of 300 million sentences across various languages, MMS-300m's multilingual nature empowers it to capture universal linguistic features shared across different languages. Consequently, fine-tuning MMS-300m on downstream tasks, such as ASR for specific languages like Vietnamese, facilitates effective cross-lingual knowledge transfer. This attribute proves especially valuable for low-resource languages, where monolingual pre-training might not suffice.

5 Experimental and Results

5.1 Evaluation Metrics

This section provides an overview of the evaluation metrics employed to assess the performance of ASR models. Evaluating the accuracy and efficacy of ASR systems is crucial to gauge their performance and facilitate comparative analysis. In this regard, we adopt well-established metrics, namely Word Error Rate (WER) and Character Error Rate (CER), to quantitatively evaluate the quality of our ASR system.

Both WER and CER provide valuable insights into the performance of ASR systems, albeit focusing on distinct aspects of accuracy. WER pri-

[7] https://www.mathworks.com/help/audio/ug/time-frequency-masking-for-harmonic-percussive-source-separation.html

marily addresses word-level errors and serves as a primary metric in ASR evaluation, particularly in scenarios where precise word recognition is critical. Conversely, CER operates at the character level, providing a more nuanced analysis, which proves particularly useful for languages featuring intricate pronunciation or applications necessitating precise character-level transcription.

5.2 Experimental Configures

We operate all experiments on a computer with Intel(R) Core(TM) i9 7900X with 128GB of RAM, and an Nvidia GeForce RTX 2080Ti GPU with 11 GB VRAM. To replicate the experiments rigorously, one can utilize the available Hugging Face's Docker image[8], which includes all the necessary libraries for model training.

Hyperparameter	Whisper	Wav2Vec
Number of Epochs	5	30
Learning Rate	5×10^{-5}	10^{-4}
Batch Size	8	24
Max grad norm	1.0	1.0
Seed	42	42
Max grad norm	1.0	1.0
Optimizer	AdamW	Adam
β_1	0.900	0.900
β_2	0.980	0.999
ϵ	10^{-6}	10^{-8}
Weight Decay	0.1	0.1
Weight Init	Gaussian Fan-In	Gaussian Fan-In
Learning Rate Schedule	Linear Decay	Linear DeCay

Table 2: The training hyperparameters of Whisper models and various wav2vec2 model including wave2vec$_{base}$, wav2vec2$_{large}$, wav2vec2$_{large-lv60-self}$, MMS-300m, and wav2vec2$_{large-xlsr-300m}$..

In our research, the choice of setting the seed to 42 ensures that the weights are initialized in a principled manner to facilitate result reproducibility through training. Besides, The differences in Experimental Configurations between the two model structures, Whisper and Wav2vec2, lie in the Batch Size, Learning Rate, Optimizer, and, notably, the Number of Epochs. The Wav2vec2 structure is more difficult to converge than Whisper, requiring a significantly larger number of epochs for proper tuning than Whisper. One can find more details in Table 2.

[8] https://hub.docker.com/r/huggingface/transformers-pytorch-gpu

5.3 Performance Comparison of ASR Models

Table 3 presents evaluation results of baseline models from the experiments. Among the evaluated ASR baseline models, the **finetuned-whisper-medium** model demonstrates the highest performance with the lowest Word Error Rate (WER) of 0.1527 and Character Error Rate (CER) of 0.0966. This model's exceptional accuracy at both the word and character levels underscores the effectiveness of fine-tuning in optimizing ASR recognition. Fine-tuning the models, indicated by the **finetuned-whisper** variants, leads to substantial improvements compared to the **openai-whisper** baseline models, highlighting the crucial role of domain-specific fine-tuning in speech recognition tasks.

Table 3: Evaluation results of several baseline models on the ViASR datasets.

Model	**WER**	**CER**
openai-whisper$_{tiny}$	0.6773	0.4804
openai-whisper$_{small}$	0.3506	0.2299
openai-whisper$_{medium}$	0.2669	0.1869
openai-whisper$_{large}$	0.2815	0.2005
openai-whisper$_{large-v2}$	0.2560	0.1733
finetuned-whisper$_{tiny}$	0.2896	0.1813
finetuned-whisper$_{small}$	0.1807	0.1134
finetuned-whisper$_{medium}$	0.1527	0.0966
finetuned-whisper$_{large}$	0.1637	0.1005
finetuned-whisper$_{large-v2}$	0.1559	0.0982
wave2vec$_{base}$	0.4129	0.1926
wav2vec2$_{large}$	0.2493	0.1304
wav2vec2$_{large-lv60-self}$	0.3075	0.1479
wave2vec$_{large-xlsr-300m}$	0.3296	0.1533
MMS-300m	0.2778	0.1345

The ASR model **wav2vec2-large-960h** emerges as a close contender with a competitive WER of 0.2493 and CER of 0.1304. The adoption of the wav2vec2 architecture exhibits a clear advantage over traditional ASR methods (e.g., **wave2vec-base** and **wave2vec-large-xlsr-300m**), suggesting the efficacy of self-supervised pretraining in enhancing ASR accuracy. The consistent outperformance of **wav2vec2** models in both WER and CER metrics signifies the potential of this architecture in the field of speech recognition.

Model size proves to be a significant factor affecting ASR performance. Larger ASR models consistently outperform their smaller coun-

terparts within both the **openai-whisper** and **finetuned-whisper** variants. This relationship is evident in the progressively decreasing WER and CER values as the model size increases. For instance, **openai-whisper-large_v2** achieves the lowest WER of 0.2560, while **openai-whisper-tiny** records the highest WER of 0.6773. These findings highlight the positive correlation between model size and recognition accuracy.

While the ASR model **MMS-300m** demonstrates competitive performance compared to the openai-whisper baseline models, it falls short when compared to the top-performing **finetuned-whisper** and **wav2vec2** models. This discrepancy suggests that fine-tuned models outperform multilingual pre-trained models for ASR tasks, emphasizing the importance of customizing models to specific languages and domains.

In conclusion, this analysis reveals valuable insights into the performance of various ASR baseline models. The results indicate that fine-tuning, self-supervised pretraining, and model size are critical factors in optimizing ASR accuracy. Understanding the trade-offs involved in selecting ASR models can help practitioners and researchers make informed decisions based on their specific requirements and resource constraints. Future research could delve into domain adaptation techniques and transfer learning across related languages or low-resource settings to enhance ASR performance further and make speech recognition technology more versatile and robust.

5.4 Error Analysis and Discussion

This section provides a comprehensive error analysis of our ASR system's performance, focusing on three typical types of errors: Hallucination, Punctuation Errors, and Math Operations Errors. To facilitate a rigorous evaluation, we selected the most optimal baseline model, **finetuned-whisper-medium**, based on Word Error Rate (WER) scores. Table 4 presents an error case study using the ViASR development set to gain valuable insights into the system's limitations and areas for improvement.

Hallucination: The prevalent errors observed in the ASR system pertain to hallucination, whereby the model generates words or phrases not present in the ground-truth transcript. For instance, the original utterance "Lên 100 tỷ... mình có...Anh ta làm..." (English: To reach 100 billion... I have... He does...) was erroneously transcribed as "cần một trăm tí anh có bằng anh ạ" (English: In need of one hundred billion, then you will have a degree). The prominence of this error type raises concerns regarding the model's tendency to overgenerate and the challenges associated with maintaining fidelity to the source speech.

Punctuation Errors: Another significant issue encountered in the ASR system's output is related to punctuation errors. Throughout the transcription, the model exhibited inaccuracies in placing appropriate punctuation marks while omitting necessary capitalizations. For instance, the ground truth sentence "nhiệt tình và nhiều năng lượng, hi vọng có thể cảm nhận được cái năng lượng tích cực đó từ cái video này của Thảo. Cảm ơn chúc mọi người nhiều sức khỏe." (English: enthusiastic and full of energy, I hope you can feel the positive energy from this video of Thảo. Thank you, and I wish everyone good health.) was transcribed with errors: "nhiệt tình và nhiều năng lượng hi vọng có thể cảm nhận được cái năng lượng tích cực đó từ cái video này của thảo. cảm ơn chúc mời người nhiều sức khỏe." These deficiencies adversely impact the intelligibility and readability of the transcribed content.

Math Operations Errors: Furthermore, the ASR system demonstrated limitations in accurately transcribing mathematical expressions and equations. A case in point is the sentence "3620 x 8.52 là 30842.4 x 9.120" (English: 3620 times 8.52 is 30842.4 times 9.120). The ASR system's errors in recognizing mathematical operations and numbers undermine the reliability and practicality of the transcriptions, particularly in technical contexts.

The identified errors in the ASR system underscore the complexities and challenges inherent in current ASR technology. The presence of hallucination errors signals a pressing need to fine-tune the model to achieve a better balance between generalization and over-generation (Serai et al., 2022; Ji et al., 2023). Further research and model refinement should prioritize preserving the integrity of the source content during transcription to minimize such errors.

Addressing punctuation errors represents a critical aspect in enhancing the natural flow and coherence of the transcribed text (Liu et al., 2018).

Ground-truth transcript	finetuned-whisper$_{\text{medium}}$ prediction	Error Type
Lên 100 tỷ... mình có...Anh ta làm... (**English:** To reach 100 billion... I have... He does...)	cần một trăm tí anh có bằng anh ạ (**English:** In need of one hundred billion, then you will have a degree)	**Hallucination**
nhiệt tình và nhiều năng lượng, hi vọng có thể cảm nhận được cái năng lượng tích cực đó từ cái video này của Thảo. Cảm ơn chúc mọi người nhiều sức khỏe. (**English:** enthusiastic and full of energy, I hope you can feel the positive energy from this video of Thảo. Thank you, and I wish everyone good health.)	nhiệt tình và nhiều năng lượng hi vọng có thể cảm nhận được cái năng lượng tích cực đó từ cái video này của thảo. cảm ơn chúc mời người nhiều sức khỏe. (**English:** Enthusiastic and full of energy I hope you can feel the positive energy from this video of Thảo. thank you and I wish you all good health.)	**Punctuation Errors**
3620 x 8.52 là 30842.4 x 9.120	phẩy năm hai, 8,52 nhá, thì bằng thôi 3620 nhân 8,52 thì sẽ ra là đây 308 42,4 được chưa nào rồi nhá đó xong rồi, xong xoay tiêu tiêu tụ a rồi, có tiêu tiêu tụ khác thì chúng ta có rồi đúng không, đây 9120 vậy thì suy ra là tổng tiêu tiêu	**Math Operations Errors**

Table 4: The proposed user interface of our automatic speech recognition system.

Investigating advanced techniques for punctuation prediction and incorporating context-aware language models can significantly contribute to ameliorating this aspect of the ASR system's performance (Nozaki et al., 2022; Nguyen et al., 2019).

The ASR system's struggle with mathematical expressions necessitates specialized fine-tuning of mathematical content and exploring domain-specific language models (Ruan et al., 2020; Lu et al., 2019). Such endeavors are fundamental to improving the system's accuracy and reliability, especially when dealing with technical discourse.

In conclusion, while the **`finetuned-whisper-medium`** model demonstrated promising performance, our thorough error analysis reveals areas that demand careful attention to elevate the overall efficacy and utility of the ASR system. Advancements in ASR technology hinge on embracing these insights and pursuing dedicated research efforts to propel the field toward greater accuracy and adaptability across diverse contexts and applications.

6 Conclusion and Future Works

This paper introduced ViASR, a novel benchmark dataset developed specifically for Vietnamese Automatic Speech Recognition. The dataset includes a vast collection of 4,276 transcripts with over 30 hours of audio that have been carefully collected for the ASR task. In addition, we conducted comprehensive evaluations of state-of-the-art ASR models, including variants of Whisper, Wav2Vec2, and MMS, to assess their performance. When compared to other ASR baseline models, the **`fintuned-whisper-medium`** model performed best, with an outstanding Word Error Rate (WER) of 0.1527 and Character Error Rate (CER) of 0.0966. Moreover, we carried out a thorough error analysis to gain valuable insights into the challenges and limitations faced by the ASR models. This investigation provided valuable insights, highlighting opportunities for improvement and inspiring future research efforts.

As discussed in Section 5.4, while this framework has achieved promising results, there is still room for further improvement. Some potential avenues for future work include training on a larger, more diverse dataset to improve accuracy and stability (Ardila et al., 2020); incorporating different feature extraction approaches, such as Mel-frequency cepstral coefficients (MFCCs) (Logan et al., 2000; Sigurdsson et al., 2006), to improve performance; Integrating the proposed framework with additional natural language processing (NLP) approaches, such as language modeling and named entity recognition (Sudoh et al., 2006), to create more powerful speech-based applications.

7 Limitations and Ethics

The proposed benchmark dataset and methods acknowledge certain limitations and ethical considerations. Firstly, the dataset may not fully encompass the extensive linguistic variability in the Vietnamese language, potentially impacting the gener-

alization ability of Automatic Speech Recognition (ASR) models. Consequently, the performance of ASR systems on this benchmark may not entirely reflect their real-world capabilities due to the limited data representation. Moreover, the involvement of human annotation introduces the possibility of errors, which could influence benchmark evaluations and ASR model performances.

The study addresses these concerns diligently, acknowledging the challenges in representing Vietnamese linguistic diversity and accounting for annotation errors. Ethical principles underpin the research, ensuring informed consent and data privacy. The emphasis on responsible AI contributes positively to ASR technology advancement while safeguarding societal interests. Overall, this benchmark dataset and approach strike a balance between acknowledging limitations and upholding ethical standards, providing valuable insights for ASR development.

References

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

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. 2022. XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale. In *Proc. Interspeech 2022*, pages 2278–2282.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. Wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

Mohamed Benzeghiba, Renato De Mori, Olivier Deroo, Stephane Dupont, Teodora Erbes, Denis Jouvet, Luciano Fissore, Pietro Laface, Alfred Mertins, Christophe Ris, et al. 2007. Automatic speech recognition and speech variability: A review. *Speech communication*, 49(10-11):763–786.

Laurent Besacier, Etienne Barnard, Alexey Karpov, and Tanja Schultz. 2014. Automatic speech recognition for under-resourced languages: A survey. *Speech communication*, 56:85–100.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2021. Unsupervised Cross-Lingual Representation Learning for Speech Recognition. In *Proc. Interspeech 2021*, pages 2426–2430.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented Transformer for Speech Recognition. In *Proc. Interspeech 2020*, pages 5036–5040.

Wei Han, Zhengdong Zhang, Yu Zhang, Jiahui Yu, Chung-Cheng Chiu, James Qin, Anmol Gulati, Ruoming Pang, and Yonghui Wu. 2020. ContextNet: Improving Convolutional Neural Networks for Automatic Speech Recognition with Global Context. In *Proc. Interspeech 2020*, pages 3610–3614.

Van Huy Nguyen. 2019. An end-to-end model for vietnamese speech recognition. In *2019 IEEE-RIVF International Conference on Computing and Communication Technologies (RIVF)*, pages 1–6.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Shigeki Karita, Nanxin Chen, Tomoki Hayashi, Takaaki Hori, Hirofumi Inaguma, Ziyan Jiang, Masao Someki, Nelson Enrique Yalta Soplin, Ryuichi Yamamoto, Xiaofei Wang, et al. 2019. A comparative study on transformer vs rnn in speech applications. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 449–456. IEEE.

Jaspreet Kaur, Amitoj Singh, and Virender Kadyan. 2021. Automatic speech recognition system for tonal languages: State-of-the-art survey. *Archives of Computational Methods in Engineering*, 28:1039–1068.

Sehoon Kim, Amir Gholami, Albert Eaton Shaw, Nicholas Lee, Karttikeya Mangalam, Jitendra Malik, Michael W. Mahoney, and Kurt Keutzer. 2022. Squeezeformer: An efficient transformer for automatic speech recognition. In *Advances in Neural Information Processing Systems*.

Yann LeCun, Yoshua Bengio, et al. 1995. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995.

Xin Liu, Yi Liu, and Xiao Song. 2018. Investigating for punctuation prediction in chinese speech transcriptions. In *2018 International Conference on Asian Language Processing (IALP)*, pages 74–78. IEEE.

Beth Logan et al. 2000. Mel frequency cepstral coefficients for music modeling. In *Ismir*, 1, page 11. Plymouth, MA.

Y. Lu, Mark J.F. Gales, Kate M. Knill, P. Manakul, L. Wang, and Y. Wang. 2019. Impact of ASR Performance on Spoken Grammatical Error Detection. In *Proc. Interspeech 2019*, pages 1876–1880.

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

Mishaim Malik, Muhammad Kamran Malik, Khawar Mehmood, and Imran Makhdoom. 2021. Automatic speech recognition: a survey. *Multimedia Tools and Applications*, 80:9411–9457.

Florian Metze, Zaid AW Sheikh, Alex Waibel, Jonas Gehring, Kevin Kilgour, Quoc Bao Nguyen, et al. 2013. Models of tone for tonal and non-tonal languages. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 261–266. IEEE.

Niko Moritz, Takaaki Hori, and Jonathan Le. 2020. Streaming automatic speech recognition with the transformer model. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6074–6078. IEEE.

Binh Nguyen, Vu Bao Hung Nguyen, Hien Nguyen, Pham Ngoc Phuong, The-Loc Nguyen, Quoc Truong Do, and Luong Chi Mai. 2019. Fast and accurate capitalization and punctuation for automatic speech recognition using transformer and chunk merging. In *2019 22nd conference of the oriental COCOSDA international committee for the co-ordination and standardisation of speech databases and assessment techniques (O-COCOSDA)*, pages 1–5. IEEE.

Jumon Nozaki, Tatsuya Kawahara, Kenkichi Ishizuka, and Taiichi Hashimoto. 2022. End-to-end Speech-to-Punctuated-Text Recognition. In *Proc. Interspeech 2022*, pages 1811–1815.

Jane Oruh, Serestina Viriri, and Adekanmi Adegun. 2022. Long short-term memory recurrent neural network for automatic speech recognition. *IEEE Access*, 10:30069–30079.

Vineel Pratap, Awni Hannun, Qiantong Xu, Jeff Cai, Jacob Kahn, Gabriel Synnaeve, Vitaliy Liptchinsky, and Ronan Collobert. 2019. Wav2letter++: A fast open-source speech recognition system. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6460–6464.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, et al. 2023. Scaling speech technology to 1,000+ languages. *arXiv preprint arXiv:2305.13516*.

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*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Thomas Reitmaier, Electra Wallington, Dani Kalarikalayil Raju, Ondrej Klejch, Jennifer Pearson, Matt Jones, Peter Bell, and Simon Robinson. 2022. Opportunities and challenges of automatic speech recognition systems for low-resource language speakers. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–17.

Weitong Ruan, Yaroslav Nechaev, Luoxin Chen, Chengwei Su, and Imre Kiss. 2020. Towards an ASR Error Robust Spoken Language Understanding System. In *Proc. Interspeech 2020*, pages 901–905.

Tara N Sainath, Ron J Weiss, Kevin W Wilson, Bo Li, Arun Narayanan, Ehsan Variani, Michiel Bacchiani, Izhak Shafran, Andrew Senior, Kean Chin, et al. 2017. Multichannel signal processing with deep neural networks for automatic speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(5):965–979.

Prashant Serai, Vishal Sunder, and Eric Fosler-Lussier. 2022. Hallucination of speech recognition errors with sequence to sequence learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:890–900.

Sigurdur Sigurdsson, Kaare Brandt Petersen, and Tue Lehn-Schiøler. 2006. Mel frequency cepstral coefficients: An evaluation of robustness of mp3 encoded music. In *ISMIR*, pages 286–289.

Sabato Marco Siniscalchi, Torbjørn Svendsen, and Chin-Hui Lee. 2014. An artificial neural network approach to automatic speech processing. *Neurocomputing*, 140:326–338.

William Song and Jim Cai. 2015. End-to-end deep neural network for automatic speech recognition. *Standford CS224D Reports*, pages 1–8.

Constantin Spille, Birger Kollmeier, and Bernd T Meyer. 2018. Comparing human and automatic speech recognition in simple and complex acoustic scenes. *Computer Speech & Language*, 52:123–140.

Brij Mohan Lal Srivastava, Sunayana Sitaram, Rupesh Kumar Mehta, Krishna Doss Mohan, Pallavi Matani, Sandeepkumar Satpal, Kalika Bali, Radhakrishnan Srikanth, and Niranjan Nayak. 2018. Interspeech 2018 low resource automatic speech recognition challenge for indian languages. In *SLTU*, pages 11–14.

Katsuhito Sudoh, Hajime Tsukada, and Hideki Isozaki. 2006. Incorporating speech recognition confidence into discriminative named entity recognition of speech data. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 617–624.

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.

Jiangyan Yi, Jianhua Tao, Zhengqi Wen, and Ye Bai. 2018. Language-adversarial transfer learning for low-resource speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(3):621–630.