PHOWHISPER: AUTOMATIC SPEECH RECOGNITION FOR VIETNAMESE

Thanh-Thien Le, Linh The Nguyen, Dat Quoc Nguyen
VinAI Research, Vietnam
{v.thienlt3, v.linhnt140, v.datnq9}@vinai.io

ABSTRACT

We introduce **PhoWhisper** in five versions for Vietnamese automatic speech recognition. PhoWhisper's robustness is achieved through fine-tuning the Whisper model on an 844-hour dataset that encompasses diverse Vietnamese accents. Our experimental study demonstrates state-of-the-art performances of PhoWhisper on benchmark Vietnamese ASR datasets. We have open-sourced PhoWhisper at: `https://github.com/VinAIResearch/PhoWhisper`.

1 INTRODUCTION

Automatic speech recognition (ASR) technology, also referred to as speech-to-text, has experienced significant advancements (Baevski et al., 2020; Barrault et al., 2023; Pratap et al., 2023), expanding its applicability across a wide range of applications. The state-of-the-art ASR model, Whisper (Radford et al., 2023), has become extremely popular, being widely used in both academia and industry.

In this paper, we present an empirical study exploring Whisper for Vietnamese. Specifically, we further fine-tune the multilingual Whisper model on a large-scale ASR dataset that includes a diverse array of Vietnamese accents from different regions in Vietnam. This results in a fine-tuned model that we name PhoWhisper. Our empirical results demonstrate state-of-the-art performances of PhoWhisper, outperforming the previous best baselines on the Vietnamese Common Voice, VIVOS, VLSP 2020 Task-1 and VLSP 2020 Task-2 test sets.

We publicly release PhoWhisper, which can be used with `transformers` (Wolf et al., 2019) and `openai-whisper` (Radford et al., 2023). We hope that PhoWhisper can serve as a strong baseline for future Vietnamese ASR research and applications.

2 PHOWHISPER

Our PhoWhisper has five versions, including PhoWhisper$_{\text{tiny}}$, PhoWhisper$_{\text{base}}$, PhoWhisper$_{\text{small}}$, PhoWhisper$_{\text{medium}}$ and PhoWhisper$_{\text{large}}$, using the same architectures of the multilingual models Whisper$_{\text{tiny}}$, Whisper$_{\text{base}}$, Whisper$_{\text{small}}$, Whisper$_{\text{medium}}$ and Whisper$_{\text{large-v2}}$, respectively.

Table 1: Data statistics.

Dataset	**Training size** (hours)	**Validation size** (hours)	**Test size** (hours)	**#syllables in training set** (min – max | average)
CMV–Vi 14	3.04	0.41	1.35	1 – 14 | 7.55
VIVOS	13.94	0.98	0.75	2 – 30 | 13.25
VLSP 2020 Task-1	240.91	2.53	7.50	1 – 349 | 17.52
VLSP 2020 Task-2	–	–	6.01	
Our private data	585.90	–	–	11 – 24 | 16.90
Total	843.79	3.92	15.61	–

We fine-tune our models on a large-scale ASR training set consisting of 844 hours of audio collected from four different resources, including **CMV–Vi**, the Vietnamese part of the Common Voice 14

(Ardila et al., 2020), **VIVOS** (Luong & Vu, 2016), VLSP 2020 ASR challenge,[1] and our private data, as shown in Table 1. Our "private data" is instrumental in providing the much-needed diversity of accents from 26K people spanning 63 provinces and municipalities, offering a profound understanding of the diverse ways in which Vietnamese is spoken. Finally, to enhance the robustness of our models against natural noises, we incorporate environmental sounds sourced from Piczak (2015) and leverage the `audiomentations` library to add noise to half of the training set. That is, we randomly split the training set into two equal parts, A and B. We then augment part A with noise and combine the noise-augmented part A with the original part B to create the final training set of 844 hours of audio.

For fine-tuning, we use `transformers` (Wolf et al., 2019), initializing PhoWhisper models from the corresponding multilingual Whisper models. We employ 8 A100 GPUs (40GB memory each) with a per-device batch size fixed at 4 and the number of gradient accumulation steps at 2 for all model versions, resulting in a global batch size of 64. The peak learning rates are set at 3.75e-5, 2.5e-5, 1.25e-5, 6.25e-6, and 5e-6 for $\text{PhoWhisper}_{\text{tiny}}$, $\text{PhoWhisper}_{\text{base}}$, $\text{PhoWhisper}_{\text{small}}$, $\text{PhoWhisper}_{\text{medium}}$, and $\text{PhoWhisper}_{\text{large}}$, respectively. We perform a total of 48,000 updating steps, which is approximately equivalent to 5 epochs.

3 Empirical Results

We compare our models with the previous state-of-the-art "wav2vec2"-based baselines from Nguyen (2021). These baselines are obtained by first pre-training Wav2Vec2.0 "base" and "large" models (Baevski et al., 2020) on 13K hours of unlabeled Vietnamese YouTube audio and then fine-tuning them using 240+ hours of labeled training data from the VLSP 2020 ASR challenge.

Table 2: Results on Vietnamese ASR benchmarks. "#paras" denotes the number of parameters.

Model	**#paras**	**Word Error Rate**			
		CMV–Vi	VIVOS	VLSP Task-1	VLSP Task-2
wav2vec2-base-vietnamese-250h	95M	102.04	10.83	21.02	50.35
wav2vec2-base-vi-vlsp2020	95M	103.71	9.90	16.82	44.91
wav2vec2-large-vi-vlsp2020	317M	101.41	8.61	15.18	36.75
$\text{PhoWhisper}_{\text{tiny}}$	39M	19.05	10.41	20.74	49.85
$\text{PhoWhisper}_{\text{base}}$	74M	16.19	8.46	19.70	43.01
$\text{PhoWhisper}_{\text{small}}$	244M	11.08	6.33	15.93	32.96
$\text{PhoWhisper}_{\text{medium}}$	769M	<u>8.27</u>	<u>4.97</u>	<u>14.12</u>	<u>26.85</u>
$\text{PhoWhisper}_{\text{large}}$	1.55B	**8.14**	**4.67**	**13.75**	**26.68**

Table 2 presents our Word Error Rate (WER) results obtained for PhoWhisper and the baselines. Our $\text{PhoWhisper}_{\text{small}}$, $\text{PhoWhisper}_{\text{medium}}$, and $\text{PhoWhisper}_{\text{large}}$ outperform all the "wav2vec2"-based baselines. Meanwhile, the remaining $\text{PhoWhisper}_{\text{tiny}}$ and $\text{PhoWhisper}_{\text{base}}$ are competitive with "wav2vec2-base-vi-vlsp2020" and perform better than "wav2vec2-base-vietnamese-250h". Here, $\text{PhoWhisper}_{\text{large}}$ establishes a new state-of-the-art WER score on each benchmark dataset.

4 Conclusions

In this paper, we have presented an empirical study exploring Whisper-based models, specifically PhoWhisper, for Vietnamese ASR. Our experimental results showcase PhoWhisper's state-of-the-art performance. We hope that our study and the public release of PhoWhisper will pave the way for further advancements and collaborations in this evolving field.

URM Statement

The authors acknowledge that at least one key author of this work meets the URM criteria of ICLR 2024 Tiny Papers Track.

[1] `https://vlsp.org.vn/vlsp2020/eval/asr`

REFERENCES

R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber. Common Voice: A Massively-Multilingual Speech Corpus. In *Proceedings of the 12th Conference on Language Resources and Evaluation*, pp. 4211–4215, 2020.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In *Proceedings of the 34th Conference on Neural Information Processing Systems*, pp. 12449–12460, 2020.

Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, et al. SeamlessM4T-Massively Multilingual & Multimodal Machine Translation. *arXiv preprint arXiv:2308.11596*, 2023.

Hieu-Thi Luong and Hai-Quan Vu. A non-expert Kaldi recipe for Vietnamese Speech Recognition System. In *Proceedings of the 3rd International Workshop on Worldwide Language Service Infrastructure and 2nd Workshop on Open Infrastructures and Analysis Frameworks for Human Language Technologies*, pp. 51–55, 2016.

Thai Binh Nguyen. Vietnamese end-to-end speech recognition using wav2vec 2.0, 2021. URL `https://github.com/vietai/ASR`.

Karol J. Piczak. ESC: Dataset for Environmental Sound Classification. In *Proceedings of the 23rd Annual ACM Conference on Multimedia*, pp. 1015–1018, 2015.

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

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 28492–28518, 2023.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.