VietMed: A Dataset and Benchmark for Automatic Speech Recognition of Vietnamese in the Medical Domain

Khai Le-Duc*
University of Toronto, Canada
duckhai.le@mail.utoronto.ca

Abstract

Due to privacy restrictions, there's a shortage of publicly available speech recognition datasets in the medical domain. In this work, we present *VietMed* - a Vietnamese speech recognition dataset in the medical domain comprising 16h of labeled medical speech, 1000h of unlabeled medical speech and 1200h of unlabeled general-domain speech. To our best knowledge, *VietMed* is by far the world's largest public medical speech recognition dataset in 7 aspects: total duration, number of speakers, diseases, recording conditions, speaker roles, unique medical terms and accents. *VietMed* is also by far the largest public Vietnamese speech dataset in terms of total duration. Additionally, we are the first to present a medical ASR dataset covering all ICD-10 disease groups and all accents within a country. Moreover, we release the first public large-scale pre-trained models for Vietnamese ASR, *w2v2-Viet* and *XLSR-53-Viet*, along with the first public large-scale fine-tuned models for medical ASR. Even without any medical data in unsupervised pre-training, our best pre-trained model *XLSR-53-Viet* generalizes very well to the medical domain by outperforming state-of-the-art *XLSR-53*, from 51.8% to 29.6% WER on test set (a relative reduction of more than 40%). All code, data and models are made publicly available here.

Keywords: medical speech recognition, dataset, semi-supervised learning

1. Introduction

Machine learning models require large amounts of training data. However, the scarcity of language resources for Vietnamese and especially for the medical domain has been hindering the advancement of corresponding automatic speech recognition (ASR) systems. Also, the lack of publicly available speech datasets and models in these domains has led to difficulties in reproducing experiments.

Recently, research efforts have been directed towards ASR tasks in the medical field, such as the works (Lüscher et al., 2023; Vieting et al., 2023) focused on the development of hybrid ASR systems to transcribe multilingual telephone speech data from patient-physician conversations. Besides, the works (Edwards et al., 2017; Chiu et al., 2018) tackled difficult acoustic conditions and the absence of domain-specific data. Nevertheless, none of these studies released their own datasets or pre-trained models.

Out of the limited number of public medical speech datasets we identified, to the best of our knowledge, one of them offers a total of 8 hours of English speech data; however, the dataset's quality is low, as indicated by the authors on their webpage[1], where they mentioned issues such as incorrect labels and audio files. The second public English medical speech dataset (Fareez et al., 2022) comprises simulated data, with a predominant focus on respiratory diseases. This situation restricts investigations to a single disease topic, hindering researchers from exploring experiments related to other medical conditions. Also, as pointed out by the authors, this dataset collected speech exclusively from the West England population, which might hurt generalizability to other accents.

Regarding Vietnamese ASR, to the best of our knowledge, there are currently no public large-scale pre-trained models that are peer-reviewed and reproducible[2]. The *XLSR-53* model (Conneau et al., 2021), was unsupervised pre-trained on 56k hours of 53 languages, but it includes only 200 hours of Vietnamese data. Therefore, the constrained performance when fine-tuning the *XLSR-53* model on Vietnamese is conceivable (Le-Duc, 2023).

To handle the concerns above, we present a high-quality dataset for Vietnamese medical speech recognition. To the best of our knowledge, *VietMed* is by far the world's largest public medical speech dataset in terms of total duration, number of speakers, diseases, recording conditions, speaker roles, unique medical terms and accents. Also, *VietMed* is by far the largest public Vietnamese speech dataset in terms of total duration. Moreover, *VietMed* is the first medical ASR dataset covering all ICD-10 disease groups and all accents within a country. We then empirically evaluate baseline models on our dataset. Our key contributions are:

(*)Work done during the bachelor thesis at Lehrstuhl Informatik 6 - Machine Learning and Human Language Technology Group, RWTH Aachen University, Germany

[1] https://www.kaggle.com/datasets/paultimothy-mooney/medical-speech-transcription-and-intent

[2] Several pre-trained models for Vietnamese ASR are available on HuggingFace and GitHub, but none of them have undergone peer review. Their results are self-reported, and we were unable to reproduce them.

LREC-COLING 2024, pages 17365–17370

- We present *VietMed* dataset, which includes 16 hours of labeled medical speech, 1000 hours of unlabeled medical speech and 1200 hours of unlabeled general-domain speech.
- We release the first public large-scale pre-trained models for Vietnamese ASR, which are peer-reviewed and reproducible.
- We release the first public large-scale fine-tuned models for medical ASR.

Given the transferability of medical terms across languages at some degree, our aim is to contribute to future research in medical ASR for other languages. All code, data and models are published online[3,4].

2. Data

VietMed data comprises of 3 sets, namely *VietMed-L* for labeled medical speech, *VietMed-U* for unlabeled medical speech, and *Viet-U* for unlabeled general domain speech. We then split *VietMed-L* into 3 subsets, train (*VietMed-Train*), dev (*VietMed-Dev*) and test (*VietMed-Test*) with duration being 5 hours, 5 hours, and 6 hours respectively, avoiding speaker overlap between the train, dev and test sets.

2.1. Metadata

Audio name	Rec.	Role	Accent
VietMed_001	Tel.	Doctor	North
Speaker ID	ICD-10	Gender	Hours
VietMed_001_a	J00-J99	Male	0.06

Table 1: Example of Metadata_labeled.xlsx. *Rec.* stands for *Recording condition*, in this example is *Tel.* (*Telephone*). Details of ICD-10 codes are shown in Table 7 of the Appendix. The speaker role is defined by common roles of speakers in conversations, which typically are: doctor, patient, host, broadcaster, etc.

We saved all the metadata information to files named Metadata_labeled.xlsx and Medical_terms.txt. As shown in Table 1, we designed metadata in a way that can support multiple tasks apart from ASR, for example: speaker recognition, keyword recognition, or accent recognition.

2.2. Data Collection

We first legally crawled audio data from YouTube under Fair Use Policies[5,6] (Details of Fair Use and Consent are in the Appendix). We manually removed non-speech elements like music, noise, long silences, and any parts that could reveal speaker identities. Specifically, we removed speaker names, locations where they live, organizations where they work, personal contacts (phone numbers, emails, etc.), personal identifier (date of birth, bank account, id number, etc.), etc. We converted MP3 audio files to 8kHz wav format, creating 10-30 second segments for *VietMed-U* and *Viet-U*, and <10 second segments for *VietMed-L*. Also, we encoded segment names, retaining only ICD-10 code tags to enhance privacy. Finally, we shuffled all segments of *VietMed-U* and *Viet-U*, making about 500k meaningless segments. The purpose here is to prevent immoral users from concatenating segments into meaningful conversations to learn more about speakers.

2.3. Annotation Process

Manual annotation of medical spontaneous speech is challenging for humans (Edwards et al., 2017). Annotators may produce varying transcripts. Also, applying the fully automated approach (Chen et al., 2021) requires large-scale ASR models, which are unavailable in the medical domain and suffer from low quality due to limited human supervision. We therefore implemented a computer-assisted workflow for medical annotation, outlined as follows:

1. We initially gathered transcripts generated by YouTube.
2. A native Vietnamese with a Biomedical Engineering degree corrected the automatically generated transcripts manually. This reduced annotation time by 70% and improved transcript quality, as it could address issues like stuttering words and speaking rate variations common in real-world conversations.
3. Another native Vietnamese independently annotated using the same approach.
4. The resulting two computer-assisted annotation versions were merged and compared. Segments with large differences were excluded.
5. Finally, we divided the merged transcripts into 3 small validation subsets, where three other Vietnamese with medical backgrounds assessed quality through manual annotation without assistance by automatic transcription. We then merged the computer-assisted and non-computer-assisted versions as in step 4.

Detailed concerns about the noisy speech in our dataset are shown in the Appendix.

[3] https://github.com/leduckhai/MultiMed
[4] https://github.com/rwth-i6/returnn-experiments
[5] https://support.google.com/youtube/answer/9783148
[6] https://www.copyright.gov/fair-use/

	Labeled	Unlabeled	
	Medical		General
Length [hours]	16	966	1204
#Speakers	61	2352	202
#Record. cond.	8	9	1
#Med. terms	978	-	-
#Accents	6	6	2
#Roles	6	6	2

Table 2: Statistics of *VietMed-L*, *VietMed-U*, *Viet-U*, retrieved from file "Metadata" in the dataset.

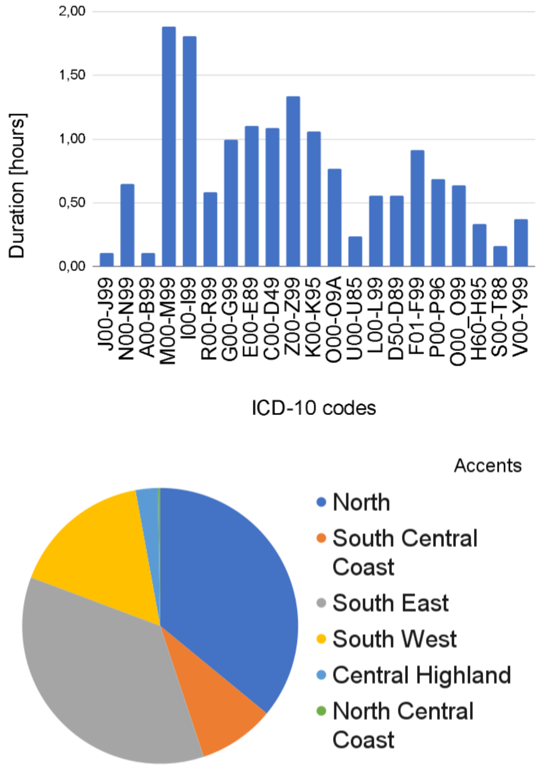

Figure 1: Distribution of ICD-10 codes and accents in *VietMed-L*.

2.4. Data Statistics

2.4.1. Labeled Medical Data *VietMed-L*

In Table 2, *VietMed-L* contains 16 hours of annotated audio, surpassing other private medical ASR datasets (Qorib and Adriani, 2018; Chung et al., 2021). Also, *VietMed-L* has a much higher number of speakers and unique medical terms. Unlike most datasets that only use simulated scenarios (Lüscher et al., 2023; Fareez et al., 2022), *VietMed-L* captures real-life situations across 8 recording conditions, including telephone (e.g. telemedicine), lectures (e.g. in university hospitals), news (e.g. in medical centers), audiobooks (e.g. medical textbooks), where 85% of the content is spontaneous speech. Additionally, we include speech from various roles such as lecturers, hosts, broadcasters, beyond just doctors and patients. Furthermore, we ensure diversity by gathering 6 accents representing all regions.

In Figure 1, rather than primarily focusing on the respiratory disease group (J00-J99) as in (Fareez et al., 2022), *VietMed-L* has data from 22/22 disease groups as per World Health Organization (WHO)'s ICD-10 code[7], supporting the dataset's generalizability. Also, the accents closely match the real accent distribution[8] (see Table B.2 in the Appendix), and the male/female ratio (54.7%-45.3%) is quite balanced.

2.4.2. Unlabeled Medical Data *VietMed-U*

In Table 2, we collected *VietMed-U* in a manner similar to *VietMed-L*, assuring a comparable generalizability as in Figure 1. Distribution of ICD-10 codes and accents is in Figure 2 and Figure 3 of the Appendix.

2.4.3. Unlabeled General Domain Data *Viet-U*

In real world, audiobooks are typically recorded using major Northern and Southern accents. In Table 3, statistics of *Viet-U* is shown.

Northern Male	Southern Male
213h	183h
Northern Female	Southern Female
518h	290h

Table 3: Genders and accents in *Viet-U*.

2.5. Extra Text Data *ExtraText*

In Table 4, besides *VietMed-Train* for language model (LM), we used extra text data *ExtraText* to gain lower PPLs. Sources are: VIVOS[9] (Luong and Vu, 2016), BABEL[10], CommonVoice[11] (Ardila et al., 2020), FOSD[12] (Tran, 2020), VNTC-Health[13], VLSP 2020[14], ViHealthBERT-FAQ (Minh et al., 2022) and PhoNER-Covid19 (Truong et al., 2021).

2.6. Lexicon

We used the BABEL project's seed lexicon and augmented it with either *VietMed-Train* or *VietMed-*

[7] https://www.icd10data.com/ICD10CM/Codes
[8] https://www.gso.gov.vn/en/population/
[9] http://ailab.hcmus.edu.vn/vivos
[10] https://www.iarpa.gov/research-programs/babel
[11] https://commonvoice.mozilla.org/
[12] https://www.kaggle.com/datasets/thinh127/fpt-open-speech-dataset-fosd-vietnamese
[13] https://github.com/duyvuleo/VNTC
[14] https://vlsp.org.vn/

Trained lexicon		LM		*VietMed-Dev*		*VietMed-Test*	
#words	#vocab	#words	Size [MB]	OOV	PPL	OOV	PPL
VietMed-Train (70k)	5295	*VietMed-Train* (70k)	1	0.76%	149	0.66%	210
		VietMed-Train + *ExtraText* (8.5M)	**98**		**66**		**84**
VietMed-Train + *ExtraText* (8.5M)	33904		103	-	69	-	87

Table 4: Results of 4-gram LMs for 2 lexica.

Train + *ExtraText*. Using the toolkit Sequitur Grapheme-To-Phoneme[15] (Bisani and Ney, 2008) - the conversion tool on these pronunciation lexica, the seed lexicon was extended, creating the lexica for training.

3. Experimental Setups

For language modelling and initial Gaussian Mixture - Hidden Markov Model (GM-HMM), we followed the same setups and hyperparameters as in (Lüscher et al., 2023). The acoustic model labels were generalized triphone states obtained by classification and regression trees with 4501 labels. For unsupervised wav2vec 2.0 training (Baevski et al., 2020) and fine-tuning, we used the same vanilla setups and hyperparameters in (Le-Duc, 2023). All models had 118M parameters including 7 CNN layers and 8 Transformer layers. The last CNN layer had a stride halved for the 8kHz data. We then chose the pre-training epoch to fine-tune with Framewise Cross-Entropy (fCE) loss that led to the best WERs on dev. The SpecAugment (Park et al., 2019) was used during 33 fine-tuning epochs.

We used RETURNN (Zeyer et al., 2018) for supervised training and Fairseq (Ott et al., 2019) for unsupervised wav2vec 2.0 training. Decoding was performed with RASR (Rybach et al., 2011). Fairseq models were converted to RETURNN models with our PyTorch-to-RETURNN toolkit[16].

4. Experimental Results

4.1. Language Model

In Table 4, augmenting the seed lexicon with only *VietMed-Train* to train *VietMed-Train+ExtraText* for LM yields the best PPLs.

4.2. GM-HMM Alignments

In Table 5, understanding that WER isn't always a precise metric for alignment quality assessment, we found that WER of SAT was quite similar to SAT+VTLN. Therefore, we chose SAT alignments as input for hybrid wav2vec 2.0 training to bypass some steps in GM-HMM process.

WER [%] on *VietMed-Dev*				
Mono	Tri	SAT	VTLN	SAT+VTLN
71.7	61.3	52.6	61.3	52.2

Table 5: Word-Error-Rates (WERs) [%] of GMM-HMM on *VietMed-Dev*. Steps go from Monophone, Triphone to Speaker Adaptive Training + Vocal Tract Length Normalization.

Pre-trained model	WER [%]	
	dev	test
None	Non-converged	
XLSR-53	45.2	51.8
w2v2-Viet	45.3	49.5
XLSR-53-Viet	**26.8**	**29.6**

Table 6: WERs of wav2vec 2.0 baselines on *VietMed-Dev* and *VietMed-Test*. *w2v2-Viet* was pre-trained from scratch on *Viet-U*. *XLSR-53-Viet* was pre-trained with *XLSR-53* as initialization on *Viet-U*. All models have the same architecture and hyperparameters.

4.3. Hybrid wav2vec 2.0 Baselines

As shown in Table 6, training from scratch did not converge, possibly due to the limited 5-hour fine-tuning data. *XLSR-53* is a state-of-the-art model pre-trained on 56k hours of 53 languages. Fine-tuning *XLSR-53* on *VietMed-Train* helped reduce WER from 52.6% to 45.2% on *VietMed-Dev*. Our *w2v2-Viet* model was competitive to *XLSR-53* despite using 46 times less data for pre-training. We obtained further improvements by applying our *XLSR-53-Viet* model, which reduced WERs to 26.8% and 29.6% on dev and test set respectively, equivalent to relative WERR of 41.8% compared to the *XLSR-53* model. In both our models, we didn't adapt the in-domain data *VietMed-U* during the unsupervised pre-training, although we believed doing so could further enhance WERs and we leave it for future work.

[15] https://github.com/sequitur-g2p/sequitur-g2p
[16] https://github.com/rwth-i6/pytorch-to-returnn

5. Conclusion

In this work, we present *VietMed* - a medical speech recognition dataset for Vietnamese. We introduce a high-quality annotation approach for medical ASR dataset that saves 70% of time. Also, we outline our work on creating a LM with acceptable PPL and a compact size. Finally, our best pre-trained model *XLSR-53-Viet* outperforms the vanilla state-of-the-art *XLSR-53* by reducing WERs from 51.8% to 29.6% WER on test set (a relative reduction of more than 40%) without using any medical data in unsupervised pre-training.

6. Acknowledgements

This work was partially supported by the project HYKIST funded by the German Federal Ministry of Health on the basis of a decision of the German Federal Parliament (Bundestag) under funding ID ZMVI1-2520DAT04A.
We thank Minh-Nghia Phan, Peter Vieting, Robin Schmitt, Moritz Gunz, Julian Dierkes for their precious assistance in experimental setups.
We also appreciate Christoph Lüscher, Ralf Schlüter, Hermann Ney for their valuable feedback on this work.

7. Bibliographical 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.

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.

Maximilian Bisani and Hermann Ney. 2008. Joint-sequence models for grapheme-to-phoneme conversion. *Speech communication*, 50(5):434–451.

Guoguo Chen, Shuzhou Chai, Guanbo Wang, Jiayu Du, Wei Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, Mingjie Jin, Sanjeev Khudanpur, Shinji Watanabe, Shuaijiang Zhao, Wei Zou, Xiangang Li, Xuchen Yao, Yongqing Wang, Zhao You, and Zhiyong Yan. 2021. Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio. Proc. Interspeech 2021, pages 4376–4380.

Chung-Cheng Chiu, Anshuman Tripathi, Katherine Chou, Chris Co, Navdeep Jaitly, Diana Jaunzeikare, Anjuli Kannan, Patrick Nguyen, Hasim Sak, Ananth Sankar, Justin Tansuwan, Nathan Wan, Yonghui Wu, and Xuedong Zhang. 2018. Speech Recognition for Medical Conversations. In *Proc. Interspeech 2018*, pages 2972–2976.

Sheng-Luen Chung, Yi-Shiuan Li, and Hsien-Wei Ting. 2021. Data centric approach to chinese medical speech recognition. In *Proceedings of the 33rd Conference on Computational Linguistics and Speech Processing (ROCLING 2021)*, pages 72–80.

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.

Erik Edwards, Wael Salloum, Greg P Finley, James Fone, Greg Cardiff, Mark Miller, and David Suendermann-Oeft. 2017. Medical speech recognition: reaching parity with humans. In *Speech and Computer: 19th International Conference, SPECOM 2017, Proc. 19*, pages 512–524.

Faiha Fareez, Tishya Parikh, Christopher Wavell, Saba Shahab, Meghan Chevalier, Scott Good, Isabella De Blasi, Rafik Rhouma, Christopher McMahon, Jean-Paul Lam, et al. 2022. A dataset of simulated patient-physician medical interviews with a focus on respiratory cases. *Scientific Data*, 9(1):313.

Khai Le-Duc. 2023. Unsupervised pre-training for vietnamese automatic speech recognition in the hykist project. *arXiv preprint arXiv:2309.15869*. Bachelor thesis at FH Aachen University of Applied Sciences.

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.

Christoph Lüscher, Mohammad Zeineldeen, Zijian Yang, Peter Vieting, Khai Le-Duc, Weiyue Wang, Ralf Schlüter, and Hermann Ney. 2023. Development of hybrid asr systems for low resource medical domain conversational telephone speech. In *ITG Speech Communication*.

Nguyen Minh, Vu Hoang Tran, Vu Hoang, Huy Duc Ta, Trung Huu Bui, and Steven Quoc Hung Truong. 2022. Vihealthbert: Pre-trained language models for vietnamese in health text mining. In *Proceedings of the Language Resources and Evaluation Conference*, pages 328–337.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53.

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

Muhammad Reza Qorib and Mirna Adriani. 2018. Building medisco: Indonesian speech corpus for medical domain. In *2018 International Conference on Asian Language Processing (IALP)*, pages 133–138.

David Rybach, Stefan Hahn, Patrick Lehnen, David Nolden, Martin Sundermeyer, Zoltán Tüske, Simon Wiesler, Ralf Schlüter, and Hermann Ney. 2011. RASR - the RWTH Aachen University open source speech recognition toolkit. In *2011 IEEE Workshop on Automatic Speech Recognition & Understanding*.

Duc Chung Tran. 2020. FPT open speech dataset (FOSD) - vietnamese.

Thinh Hung Truong, Mai Hoang Dao, and Dat Quoc Nguyen. 2021. COVID-19 Named Entity Recognition for Vietnamese. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Peter Vieting, Christoph Lüscher, Julian Dierkes, Ralf Schlüter, and Hermann Ney. 2023. Efficient utilization of large pre-trained models for low resource asr. In *2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*.

Albert Zeyer, Tamer Alkhouli, and Hermann Ney. 2018. RETURNN as a generic flexible neural toolkit with application to translation and speech recognition. In *Annual Meeting of the Assoc. for Computational Linguistics*.