A. Ethical Statements

A.1. Fair Use

We strictly followed the criteria of Fair Use by The U.S. Copyright Office[17], which also applies to YouTube platform. Section 107 of the Copyright Act provides the statutory framework for determining whether something is a fair use and identifies certain types of uses—such as criticism, comment, news reporting, teaching, scholarship, and research—as examples of activities that may qualify as fair use. Section 107 calls for consideration of the following four factors in evaluating a question of fair use:

- (1) **Purpose and character of the use, including whether the use is of a commercial nature or is for nonprofit educational purposes:** Courts look at how the party claiming fair use is using the copyrighted work, and are more likely to find that nonprofit educational and noncommercial uses are fair. Additionally, “transformative” uses are more likely to be considered fair. Transformative uses are those that add something new, with a further purpose or different character, and do not substitute for the original use of the work.

- (2) **Nature of the copyrighted work:** This factor analyzes the degree to which the work that was used relates to copyright’s purpose of encouraging creative expression. Thus, using a more creative or imaginative work (such as a novel, movie, or song) is less likely to support a claim of a fair use than using a factual work (such as a technical article or news item). In addition, use of an unpublished work is less likely to be considered fair.

- (3) **Amount and substantiality of the portion used in relation to the copyrighted work as a whole:** Under this factor, courts look at both the quantity and quality of the copyrighted material that was used. That said, some courts have found use of an entire work to be fair under certain circumstances. And in other contexts, using even a small amount of a copyrighted work was determined not to be fair because the selection was an important part—or the “heart”—of the work.

- (4) **Effect of the use upon the potential market for or value of the copyrighted work:** Here, courts review whether, and to what extent, the unlicensed use harms the existing or future market for the copyright owner’s original work.

According to the law, we assert our defense under the Fair Use doctrine with the help of Fair Use explanation[18] by copyrightalliance.org and ELRC Report on legal issues in web crawling [19] by Pawel Kamocki as follows:

- (1) Obviously we crawled the data and published only for non-commercial and research purposes.

- (1) We did not directly use videos crawled from YouTube. Instead, we transformed them into audio files with a predefined sampling rate. Additionally, we divided lengthy audio files, approximately one hour in duration, into shorter segments lasting between 10 to 30 seconds. These segments were then randomly shuffled, making it impossible for users to piece them together to comprehend the entirety of the originally crawled videos. Therefore, our work is transformative and we do not substitute the original use of the crawled videos.

- (2) Our medical conversations are factual (nonfiction) and hence qualified as fair.

- (2) Videos on YouTube platform are universally accessible around the world, therefore we satisfy the criteria for the copyrighted work’s publication status.

- (3) There is no quantitative test to evaluate whether a given use is fair. The randomly shuffled 10-30 second segments we have created do not provide the complete context and meaning of each video, thus making them incapable of representing the ”heart” of the copyrighted work.

- (4) We don’t utilize our publicly available data to compete with the copyright owners’ business. Furthermore, our 10-30 second segments have no impact on the viewership count on YouTube. As a result, our efforts do not undermine the potential market being pursued by the copyright owners.

Besides our work, several similar works exist that involve the extraction of YouTube videos and their conversion into audio files for research and non-commercial intentions, such as GigaSpeech[20] (China & USA), VoxCeleb[21] (UK), VoxLingua107[22] (UK).

---

[17] https://www.copyright.gov/fair-use/

[18] https://copyrightalliance.org/faqs/what-is-fair-use/

[19] http://www.elra.info/media/filer_public/2021/02/12/elrc-legal-analysis-webcrawling_report-v11.pdf

[20] https://github.com/SpeechColab/GigaSpeech

[21] https://www.robots.ox.ac.uk/ vgg/data/voxceleb/

[22] https://bark.phon.ioc.ee/voxlingua107/

A.2. Data Consent

According to the existing law on the data consent, we are allowed to publish research data. We describe in short as follows:

- First of all, 137/194 countries signed Data Protection and Privacy Legislation Worldwide[23] by the United Nations, including USA, EU, Germany, Vietnam. So Vietnamese law on data protection complies with international law, as Article 6 of the Personal Data Protection Act by the Vietnamese government says: “The protection of personal data is carried out in accordance with international treaties to which the Socialist Republic of Vietnam is a member”.
- Researchers have the right to freely publish sensitive medical data for research without the consent of the data subject (speakers in speech data), as Article 20, Section 4 says: “The party processing personal data is not required to register for processing sensitive personal data in the case of research purposes.”
- Once more, researchers do not need direct or indirect consent from the data subject to publish research papers, as the Article 16 says: “Data deletion will not apply at the request of the data subject in the following cases: Personal data is processed to serve legal requirements, scientific research, and statistics.”
- Again, researchers do not need consent, as Article 9 of the European General Data Protection Regulation (GDPR) permits researchers in Member States to publish personal data for scientific research purposes without consent.
- Researchers are strongly encouraged to publish research on sensitive medical data, according to Law on Medical Examination and Treatment, Constitution of the Socialist Republic of Vietnam, Article 22: “Practitioners (…) are responsible for updating relevant medical knowledge (...) including (...) c) Publish scientific research (...).”
- In case of unexpected issues during publishing research, researchers are “Protected by the law and not responsible when a medical incident still occurs after complying with regulations.”, as stated in Article 42.
- We crawled generated-by-Vietnam data using Vietnamese IP address and a crawler from a Vietnamese company authorized by Vietnamese government, and the right to publish this data for research purposes is protected under Vietnamese Law (shown above), since Google (Youtube) must comply with Vietnamese law on content in Vietnamese cyberspace, as shown in Article 26, Cybersecurity Law, Constitution of the Socialist Republic of Vietnam: “Domestic and foreign enterprises providing services on telecommunications networks, the Internet, and value-added services in cyberspace in Vietnam have activities of collecting, exploiting, analyzing, and processing information data (...) created by service users in Vietnam must store this data in Vietnam (...) as prescribed by the Government.”
- International researchers have the right to publish and process Vietnamese personal data without consent. Also they are both encouraged to publish Vietnamese research data and are protected under Vietnamese law because they must comply with Vietnamese law on generated-by-Vietnam data, according to Article 2 and 10, the Vietnamese Civil Code on Civil Relations with Foreign Elements: “The provisions of Vietnamese civil law apply to civil relations involving foreign elements (...). In case the application or consequences of the application of foreign law are contrary to (...) the Vietnam Civil Code and other basic principles of Vietnamese law , then Vietnamese law applies.”

The YouTube content in our dataset is about medical shows, interviews, lectures, etc., where all participants talked to camera and were aware that the videos are publicly accessible in an attempt to provide medical knowledge to YouTube users. These videos are published by national TV channels, not by some amateur content creators. There are some YouTube videos that speakers are not aware of being recorded, published by amateurs, but we did not include them in our dataset.

B. Additional Details of *VietMed* Dataset

B.1. Description of ICD-10 Codes

Table 7 shows the detailed description of ICD-10 codes. The audio files in our dataset are classified based on these ICD-10 codes.

B.2. Real Distribution of Accents in Vietnam

Table B.2 shows the real distribution of accents in Vietnam, which our *VietMed* dataset follows.

---

[23] https://unctad.org/page/data-protection-and-privacy-legislation-worldwide

ICD-10 Code	Description of diseases
A00-B99	Certain infectious and parasitic diseases
C00-D49	Neoplasms
D50-D89	Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism
E00-E89	Endocrine, nutritional and metabolic diseases
F01-F99	Mental, Behavioral and Neurodevelopmental disorders
G00-G99	Diseases of the nervous system
H00-H59	Diseases of the eye and adnexa
H60-H95	Diseases of the ear and mastoid process
I00-I99	Diseases of the circulatory system
J00-J99	Diseases of the respiratory system
K00-K95	Diseases of the digestive system
L00-L99	Diseases of the skin and subcutaneous tissue
M00-M99	Diseases of the musculoskeletal system and connective tissue
N00-N99	Diseases of the genitourinary system
O00-O9A	Pregnancy, childbirth and the puerperium
P00-P96	Certain conditions originating in the perinatal period
Q00-Q99	Congenital malformations, deformations and chromosomal abnormalities
R00-R99	Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified
S00-T88	Injury, poisoning and certain other consequences of external causes
U00-U85	Codes for special purposes
V00-Y99	External causes of morbidity
Z00-Z99	Factors influencing health status and contact with health services

Table 7: Description of ICD-10 codes which our dataset follows, according to the 2024 version by World Health Organziation. Each ICD-10 Code, e.g. A00-B99, could be in smaller codes partitioned. However, in our dataset we only used 22 ICD-10 Codes since partitioning into smaller codes makes the annotation too complicated and unnecessary.

Region	Subregion	Typical Provinces	Population
North	Northest	Cao Bằng Hà Giang ...	8M
	Northwest	Điện Biên Hòa Bình ...	4M
	Red River Delta	Hà Nội Hải Phòng ...	20M
Central	North Central Coast	Hà Tĩnh Nghệ An ...	10M
	South Central Coast	Đà Nẵng Bình Thuận ...	9M
	Central Highland	Gia Lai Kon Tum ...	5M
South	Southeast	TP. Hồ Chí Minh Đồng Nai ...	16M
	Southwest	Long An Cần Thơ ...	18M

Table 8: Real distribution of Vietnamese accents. The statistics was retrieved in 2015 from Vietnamese General Statistics Office. In our dataset, we did not split the North accent into subregional accents since it was too difficult for our annotators to correctly recognize subregional accents of the North region.

B.3. Concerns about Noisy Speech in *VietMed*

Real-world speech data should contain real-world acoustic conditions (e.g. background noises, music, etc.). To enhance the quality of a speech dataset, especially for a read speech dataset, people often use a Signal-to-Noise Ratio (SNR) to measure the background noises and discard segments with a high level of SNR. However, using an SNR threshold to obtain only good speech signals, discarding noisy segments, would violate real-world scenarios, making our *VietMed* dataset no longer real world but rather ”simulated”.

Actually, we only removed audio segments that have no speech. We still kept overlapped speech segments, as long as the main speaker's speech is still comprehensible. The quality assurance for real-world ASR datasets should focus on transcription, which we have already addressed in the paper, instead of focusing on the quality of the speech signal.

B.4. Extra Data Statistics for Labeled Medical Data *VietMed-L*

Table 9 shows the statistics of 3 train-dev-test subsets in *VietMed-L*. We split these 3 subsets in a way that made *VietMed-Train* the least generalizability by having the least number of speakers, recording conditions, accents and roles, while prioritizing *VietMed-Dev* and *VietMed-Test* more generalizability. Note that no speaker overlap occured in the 3 subsets.

B.5. Extra Data Statistics for Unlabeled Medical Data *VietMed-U*

Figure 2 shows the distribution of ICD-10 code and Figure 3 shows the distribution of accents in *VietMed-U*. We collected *VietMed-U* in a manner similar to *VietMed-L*, assuring a comparable generalizability as in *VietMed-L*.

C. ASR Error Analysis

C.1. Error Analysis of Pre-trained Model

Table 10 shows the error analysis of our pre-trained model *XLSR-53* on the *VietMed-Test* set.

Table 11 shows the error analysis of our pre-trained model *w2v2-Viet* on the *VietMed-Test* set.

Table 12 shows the error analysis of our best pre-trained model *XLSR-53-Viet* on the *VietMed-Test* set.

C.2. Error Analysis of Confusion Pairs

Table 13 shows the statistics of confusion pairs in *VietMed-Test* using the best pre-trained model *XLSR-53-Viet*. Closely similar words could lead to the decreased accuracy of an ASR system. Therefore, collecting confusion pairs which the ASR system often misrecognized gives researchers an opportunity to analyze common ASR errors and improve the ASR accuracy.

As shown in the table, words that are parts of medical terms and fillers contribute greatly to the decreased accuracy of the ASR system using the pre-trained model *XLSR-53-Viet*. This difficulty was confirmed by our annotators during the dataset annotation, since it was very hard to correctly transcribe medical terms and fillers in real-world medical conversations.

C.3. Error Analysis of OOV

Table 14 shows the list of OOVs loan words found in *VietMed-Train*. In this table, we used the BABEL project's seed lexicon and automatically augmented it with *VietMed-Train*. We used the toolkit Sequitur Grapheme-To-Phoneme[24] (Bisani and Ney, 2008) - the conversion tool on these pronunciation lexica, to extend the seed lexicon, creating the lexicon for training.

First, we found that the seed lexicon by BABEL was overwhelmed by North and North Central Coast accents, leaving almost no other accents like South Central Coast, Central Highland, Southwest and Southeast. Therefore, this lexicon hurts the accuracy of ASR systems on a generalized dataset like *VietMed*. Second, *VietMed* has a very large number of medical terms, which often come from English loan words. So automatic extension of the seed lexicon without human correction led to wrong phoneme mapping of medical terms, which also hurts the accuracy of ASR systems.

---

[24] https://github.com/sequitur-g2p/sequitur-g2p

	VietMed-Train	*VietMed-Dev*	*VietMed-Test*
Dur. [hours]	5	5	6
#Speakers	13	21	27
#Words	70k	69k	76k
#Rec. cond.	2	4	6
#Accents	3	4	5
#Roles	3	4	6

Table 9: Data statistics of *VietMed-L*, retrieved from file ”Metadata” in the dataset.

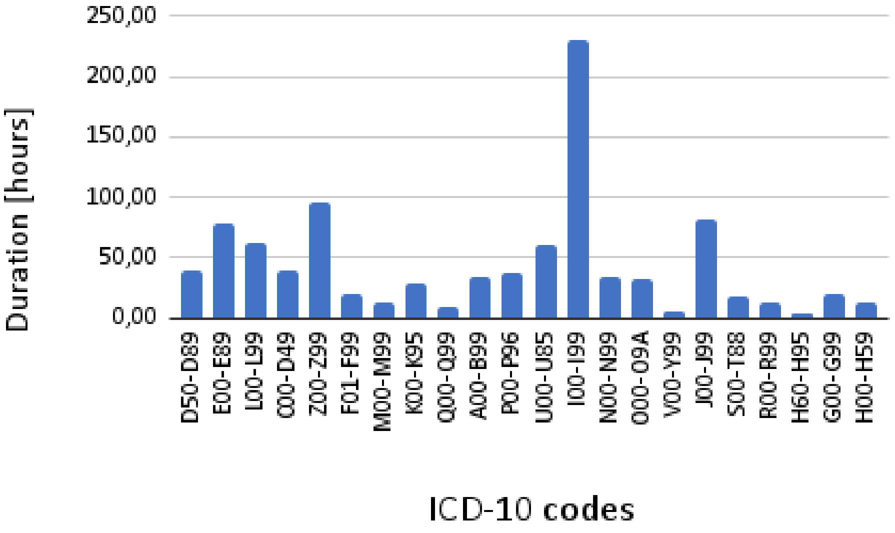

Figure 2: Distribution of ICD-10 code in *VietMed-U*.

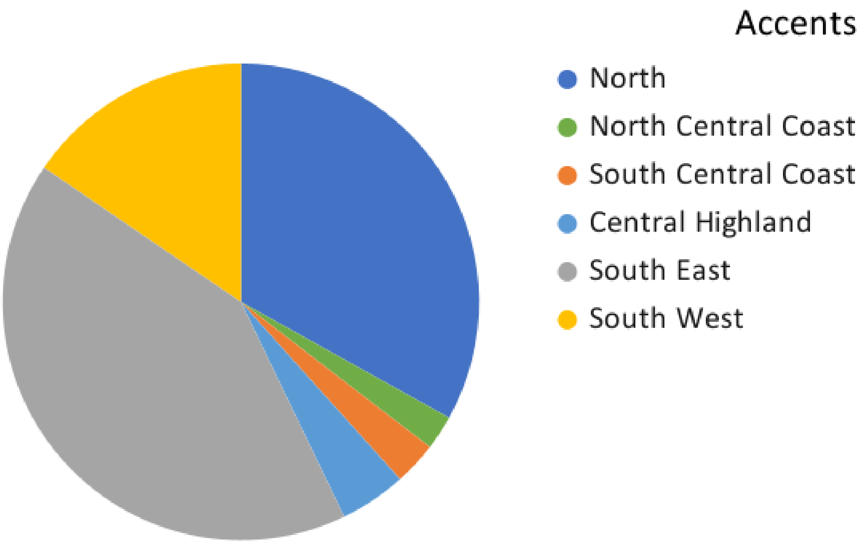

Figure 3: Distribution of accents in *VietMed-U*.

Speaker ID	Rec.	ICD-10	Role	Gend	Acc.	# Snt	# Wrd	Corr	Sub	Del	Ins	Err	S.Err
vietmed_002	Tel.	N00-N99	Lec.	F	SCC	363	7631	30.7	54.4	14.9	5.6	74.9	100.0
vietmed_004		M00-M99	Doc.	M	SCC	446	10575	51.7	34.8	13.5	6.8	55.0	100.0
vietmed_014_a		K00-K95	Host	F	N	18	491	63.7	23.4	12.8	3.7	39.9	100.0
vietmed_014_b			Doc.	M	N	164	4034	59.6	28.5	11.9	5.2	45.6	100.0
vietmed_015_a		O00-O9A	Host	F	N	73	1779	68.8	20.3	10.9	4.2	35.4	100.0
vietmed_015_b			Doc.	F	N	297	5669	58.8	28.3	12.9	4.4	45.6	100.0
vietmed_015_c			Pat.	F	N	55	1010	43.0	37.3	19.7	3.7	60.7	100.0
vietmed_017_a	Talk.	U00-U85	Doc.	F	SW	47	1104	50.0	37.2	12.8	5.4	55.4	100.0
vietmed_017_b			Doc.	M	N	86	2061	62.8	26.9	10.2	5.0	42.2	100.0
vietmed_018_a		K00-K95	Host	F	SW	63	1527	54.3	32.5	13.2	19.6	65.3	100.0
vietmed_018_b			Doc.	M	SW	192	5293	59.8	26.2	14.0	7.2	47.4	100.0
vietmed_018_c			Doc.	F	SW	118	2761	55.3	31.4	13.2	8.7	53.3	100.0
vietmed_018_d			Pat.	F	SW	20	412	33.3	36.9	29.9	6.1	72.8	100.0
vietmed_018_e			Pat.	M	SW	5	76	31.6	40.8	27.6	10.5	78.9	100.0
vietmed_018_f			Doc.	M	SW	25	639	41.2	42.9	16.0	5.0	63.8	100.0
vietmed_019_a		L00-L99	Host	F	SW	58	1490	55.1	31.9	13.0	6.9	51.8	100.0
vietmed_019_b			Doc.	F	SW	116	2776	56.5	30.5	13.0	7.7	51.3	100.0
vietmed_023	Pod.	P00-P96	Pod.	F	SW	390	7414	55.4	35.8	8.8	4.9	49.6	99.7
vietmed_024		O00_O99	Pod.	F	SE	376	7425	61.2	28.8	10.0	4.7	43.5	99.7
vietmed_025_a	Diag.	H60-H95	Host	F	SW	101	2280	60.3	29.1	10.7	5.0	44.7	100.0
vietmed_025_b			Doc.	M	SE	91	1838	65.7	24.8	9.5	6.6	40.9	100.0
vietmed_026	Lec.	A00-B99	Lec.	M	NCC	21	355	31.8	47.6	20.6	6.5	74.6	100.0
vietmed_027_a	News	S00-T88	Host	F	SW	29	710	70.8	20.8	8.3	6.2	35.4	100.0
vietmed_027_b			Brc.	M	SE	64	1454	49.5	39.1	11.3	5.6	56.1	100.0
vietmed_028_a		V00-Y99	Host	F	SE	106	2617	52.7	34.7	12.6	4.1	51.5	100.0
vietmed_028_b			Brc.	M	SE	21	475	47.6	41.5	10.9	6.7	59.2	100.0
vietmed_029			Brc.	F	SE	92	2240	60.4	30.0	9.6	5.4	45.1	100.0
Sum/Avg						3437	76136	54.2	33.5	12.3	6.0	**51.8**	99.9
Mean						127.3	2819.9	53.0	33.2	13.8	6.4	53.3	100.0
Standard Deviation						129.6	2743.3	11.4	8.0	5.2	3.1	12.1	0.1
Median						86.0	1838.0	55.3	31.9	12.8	5.6	51.5	100.0

Table 10: Analysis of ASR errors on *VietMed-Test* set using the baseline model *XLSR-53* (WER = 51.8). Column from left to right is: Speaker ID, Recording Condition, ICD-10 Code, Speaker Role, Gender, Accent, Number of sentences, Number of words, Corrections, Substitution Errors, Deletion Errors, Insertion Errors, Word-Error-Rate, Sentence-Error-Rate.
For Recording Condition, there are: Telephone (Tel.), Talkshow (Talk.), Podcast (Pod.), Diagnosis (Diag.), Lectures (Lec.), News.
For Speaker Role, there are: Lecturer (Lec.), Doctor (Doc.), Talkshow Host (Host), Patient (Pat.), Podcaster (Pod.), Broadcaster (Brc.).
For Gender, there are: Male (M) and Female (F).
For Accent, there are: South Central Coast (SCC), North (N), Southwest (SW), Southeast (SE), North Central Coast (NCC).

Speaker ID	Rec.	ICD-10	Role	Gend	Acc.	# Snt	# Wrd	Corr	Sub	Del	Ins	Err	S.Err
vietmed_002	Tel.	N00-N99	Lec.	F	SCC	363	7631	33.8	50.7	15.5	5.6	71.8	100.0
vietmed_004		M00-M99	Doc.	M	SCC	446	10575	52.1	34.1	13.8	6.5	54.3	100.0
vietmed_014_a		K00-K95	Host	F	N	18	491	72.3	15.9	11.8	5.1	32.8	100.0
vietmed_014_b			Doc.	M	N	164	4034	57.8	28.6	13.6	4.6	46.8	100.0
vietmed_015_a		O00-O9A	Host	F	N	73	1779	70.8	18.1	11.1	4.5	33.7	100.0
vietmed_015_b			Doc.	F	N	297	5669	60.1	26.7	13.2	4.7	44.6	99.7
vietmed_015_c			Pat.	F	N	55	1010	44.4	37.5	18.1	5.4	61.1	100.0
vietmed_017_a	Talk.	U00-U85	Doc.	F	SW	47	1104	51.6	36.2	12.1	6.3	54.6	100.0
vietmed_017_b			Doc.	M	N	86	2061	62.4	26.7	10.9	4.9	42.4	100.0
vietmed_018_a		K00-K95	Host	F	SW	63	1527	59.2	27.5	13.3	19.6	60.4	100.0
vietmed_018_b			Doc.	M	SW	192	5293	59.5	26.3	14.3	6.7	47.2	100.0
vietmed_018_c			Doc.	F	SW	118	2761	57.7	29.6	12.7	9.0	51.4	100.0
vietmed_018_d			Pat.	F	SW	20	412	34.7	34.5	30.8	4.9	70.1	100.0
vietmed_018_e			Pat.	M	SW	5	76	42.1	34.2	23.7	7.9	65.8	100.0
vietmed_018_f			Doc.	M	SW	25	639	44.0	38.2	17.8	7.0	63.1	100.0
vietmed_019_a		L00-L99	Host	F	SW	58	1490	58.6	28.7	12.7	6.8	48.2	100.0
vietmed_019_b			Doc.	F	SW	116	2776	58.9	28.4	12.7	7.4	48.5	100.0
vietmed_023	Pod.	P00-P96	Pod.	F	SW	390	7414	63.0	29.6	7.4	4.8	41.8	99.7
vietmed_024		O00_O99	Pod.	F	SE	376	7425	65.4	25.9	8.6	5.8	40.3	99.5
vietmed_025_a	Diag.	H60-H95	Host	F	SW	101	2280	65.3	24.5	10.2	4.6	39.3	100.0
vietmed_025_b			Doc.	M	SE	91	1838	67.2	23.2	9.5	7.1	39.8	100.0
vietmed_026	Lec.	A00-B99	Lec.	M	NCC	21	355	26.5	47.3	26.2	4.8	78.3	100.0
vietmed_027_a	News	S00-T88	Host	F	SW	29	710	68.7	22.5	8.7	5.5	36.8	100.0
vietmed_027_b			Brc.	M	SE	64	1454	41.5	44.6	13.9	5.2	63.7	100.0
vietmed_028_a		V00-Y99	Host	F	SE	106	2617	59.7	28.8	11.5	4.4	44.7	99.1
vietmed_028_b			Brc.	M	SE	21	475	48.8	39.2	12.0	5.1	56.2	100.0
vietmed_029			Brc.	F	SE	92	2240	64.4	26.1	9.6	5.9	41.6	100.0
Sum/Avg						3437	76136	56.5	31.2	12.3	6.0	**49.5**	99.9
Mean						127.3	2819.9	55.2	30.9	13.9	6.3	51.1	99.9
Standard Deviation						129.6	2743.3	12.0	8.3	5.4	2.9	12.2	0.2
Median						86.0	1838.0	58.9	28.7	12.7	5.5	48.2	100.0

Table 11: Analysis of ASR errors on *VietMed-Test* set using the baseline model *w2v2-Viet* (WER = 49.5). Column from left to right is: Speaker ID, Recording Condition, ICD-10 Code, Speaker Role, Gender, Accent, Number of sentences, Number of words, Corrections, Substitution Errors, Deletion Errors, Insertion Errors, Word-Error-Rate, Sentence-Error-Rate.
For Recording Condition, there are: Telephone (Tel.), Talkshow (Talk.), Podcast (Pod.), Diagnosis (Diag.), Lectures (Lec.), News.
For Speaker Role, there are: Lecturer (Lec.), Doctor (Doc.), Talkshow Host (Host), Patient (Pat.), Podcaster (Pod.), Broadcaster (Brc.).
For Gender, there are: Male (M) and Female (F).
For Accent, there are: South Central Coast (SCC), North (N), Southwest (SW), Southeast (SE), North Central Coast (NCC).

Speaker ID	Rec.	ICD-10	Role	Gend	Acc.	# Snt	# Wrd	Corr	Sub	Del	Ins	Err	S.Err
vietmed_002	Tel.	N00-N99	Lec.	F	SCC	363	7631	57.8	31.2	11.0	6.3	48.5	100.0
vietmed_004		M00-M99	Doc.	M	SCC	446	10575	68.8	18.7	12.5	5.4	36.6	100.0
vietmed_014_a		K00-K95	Host	F	N	18	491	87.8	3.5	8.8	4.7	16.9	100.0
vietmed_014_b			Doc.	M	N	164	4034	77.2	12.2	10.5	4.6	27.4	100.0
vietmed_015_a		O00-O9A	Host	F	N	73	1779	85.2	5.8	9.0	3.6	18.4	97.3
vietmed_015_b			Doc.	F	N	297	5669	82.4	7.7	9.8	4.2	21.8	97.3
vietmed_015_c			Pat.	F	N	55	1010	70.1	14.9	15.0	5.8	35.7	100.0
vietmed_017_a	Talk.	U00-U85	Doc.	F	SW	47	1104	76.6	13.1	10.2	4.2	27.5	100.0
vietmed_017_b			Doc.	M	N	86	2061	80.1	10.4	9.6	4.8	24.7	100.0
vietmed_018_a		K00-K95	Host	F	SW	63	1527	73.7	13.2	13.2	18.7	45.1	100.0
vietmed_018_b			Doc.	M	SW	192	5293	75.3	12.1	12.6	6.5	31.2	100.0
vietmed_018_c			Doc.	F	SW	118	2761	74.3	12.4	13.3	7.3	33.0	100.0
vietmed_018_d			Pat.	F	SW	20	412	55.1	20.6	24.3	5.6	50.5	100.0
vietmed_018_e			Pat.	M	SW	5	76	57.9	19.7	22.4	7.9	50.0	100.0
vietmed_018_f			Doc.	M	SW	25	639	64.9	20.3	14.7	6.1	41.2	100.0
vietmed_019_a		L00-L99	Host	F	SW	58	1490	75.2	12.6	12.2	6.7	31.5	100.0
vietmed_019_b			Doc.	F	SW	116	2776	75.7	11.9	12.5	6.2	30.5	100.0
vietmed_023	Pod.	P00-P96	Pod.	F	SW	390	7414	83.3	10.6	6.0	4.1	20.8	97.4
vietmed_024		O00_O99	Pod.	F	SE	376	7425	85.0	8.0	7.1	5.0	20.1	98.4
vietmed_025_a	Diag.	H60-H95	Host	F	SW	101	2280	80.4	10.6	9.0	4.8	24.4	100.0
vietmed_025_b			Doc.	M	SE	91	1838	81.8	10.0	8.3	5.1	23.3	98.9
vietmed_026	Lec.	A00-B99	Lec.	M	NCC	21	355	57.7	27.9	14.4	7.3	49.6	100.0
vietmed_027_a	News	S00-T88	Host	F	SW	29	710	83.5	8.0	8.5	4.6	21.1	100.0
vietmed_027_b			Brc.	M	SE	64	1454	74.8	15.8	9.4	5.2	30.4	100.0
vietmed_028_a		V00-Y99	Host	F	SE	106	2617	82.7	8.8	8.6	4.2	21.6	99.1
vietmed_028_b			Brc.	M	SE	21	475	74.3	14.9	10.7	5.3	30.9	100.0
vietmed_029			Brc.	F	SE	92	2240	83.9	7.5	8.5	5.6	21.7	97.8
Sum/Avg						3437	76136	75.9	13.8	10.3	5.6	**29.6**	99.1
Mean						127.3	2819.9	75.0	13.4	11.6	5.9	30.9	99.5
Standard Deviation						129.6	2743.3	9.3	6.4	4.1	2.8	10.5	1.0
Median						86.0	1838.0	75.7	12.2	10.5	5.3	30.4	100.0

Table 12: Analysis of ASR errors on *VietMed-Test* set using the best baseline model *XLSR-53-Viet* (WER = 29.6).
Column from left to right is: Speaker ID, Recording Condition, ICD-10 Code, Speaker Role, Gender, Accent, Number of sentences, Number of words, Corrections, Substitution Errors, Deletion Errors, Insertion Errors, Word-Error-Rate, Sentence-Error-Rate.
For Recording Condition, there are: Telephone (Tel.), Talkshow (Talk.), Podcast (Pod.), Diagnosis (Diag.), Lectures (Lec.), News.
For Speaker Role, there are: Lecturer (Lec.), Doctor (Doc.), Talkshow Host (Host), Patient (Pat.), Podcaster (Pod.), Broadcaster (Brc.).
For Gender, there are: Male (M) and Female (F).
For Accent, there are: South Central Coast (SCC), North (N), Southwest (SW), Southeast (SE), North Central Coast (NCC).

Index	**Occurrences**	**Confusion pair**	**Type**
1	75	bé $\Longrightarrow$ béo	Med

Table 13 continued from previous page

Index	Occurrences	Confusion pair	Type
2	75	cung ⟹ công	-
3	49	các ⟹ cái	-
4	34	trẻ ⟹ sẽ	Med
5	33	bú ⟹ bốn	Med
6	31	implant ⟹ lên	Med
7	30	thai ⟹ hai	Med
8	28	cái ⟹ các	Fill
9	26	là ⟹ mà	Fill
10	25	tử ⟹ bệnh	Med
11	25	vì ⟹ thì	Fill
12	24	răng ⟹ đang	Med
13	23	cấy ⟹ cái	Med
14	23	làm ⟹ là	-
15	21	là ⟹ và	Fill
16	20	đó ⟹ nó	Fill
17	19	và ⟹ là	Fill
18	19	và ⟹ mà	Fill
19	19	âm ⟹ ăn	Med
20	18	là ⟹ làm	Fill
21	18	mình ⟹ mà	Fill
22	18	trồng ⟹ trong	Med
23	17	bú ⟹ bố	Med
24	17	chị ⟹ chỉ	-
25	17	có ⟹ cái	-
26	17	là ⟹ lại	Fill
27	17	mà ⟹ và	Fill
28	17	sẽ ⟹ phải	Fill
29	17	đi ⟹ đây	Fill
30	16	nó ⟹ đó	Fill
31	16	tử ⟹ về	Med
32	15	con ⟹ còn	Med
33	15	progesterone ⟹ cholesterol	Med
34	15	rong ⟹ năm	Med
35	15	thủ ⟹ phẫu	Med
36	14	implant ⟹ selen	Med
37	14	que ⟹ quen	Med
38	13	còn ⟹ có	Fill
39	13	có ⟹ các	Fill
40	13	có ⟹ đó	Fill
41	13	lại ⟹ là	Fill
42	12	như ⟹ nhưng	Fill
43	11	bà ⟹ mà	-
44	11	bình ⟹ bệnh	Med
45	11	cung ⟹ trong	Med
46	11	là ⟹ nó	Fill
47	11	mình ⟹ bệnh	-
48	11	răng ⟹ gan	Med
49	11	răng ⟹ ăn	Med
50	11	vào ⟹ và	-
51	10	anh ⟹ ăn	-
52	10	bà ⟹ ba	-
53	10	chú ⟹ chúng	-
54	10	cách ⟹ các	-
55	10	cô ⟹ của	-

Table 13 continued from previous page

Index	Occurrences	Confusion pair	Type
56	10	da ⟹ ra	Med
57	10	khi ⟹ thì	-
58	10	lạ ⟹ là	-
59	10	tóc ⟹ tác	Med
60	10	vòng ⟹ phòng	-
61	10	đo ⟹ đó	Med
62	10	đại ⟹ tại	-
63	9	cổ ⟹ của	Med
64	9	dặm ⟹ giảm	Med
65	9	hay ⟹ hai	-
66	9	ngừa ⟹ là	Med
67	9	nói ⟹ nó	-
68	9	răng ⟹ rằng	Med
69	9	sau ⟹ sao	-
70	9	tai ⟹ tay	Med
71	9	thì ⟹ cái	Fill
72	9	tràng ⟹ trạm	Med
73	9	tóc ⟹ tắt	Med
74	9	ốc ⟹ cái	Med
75	8	chị ⟹ thì	-
76	8	cong ⟹ công	Med
77	8	em ⟹ xem	-
78	8	estrogen ⟹ selen	Med
79	8	kinh ⟹ cân	Med
80	8	nhi ⟹ như	Med
81	8	nè ⟹ này	Fill
82	8	quy ⟹ quá	Med
83	8	ruột ⟹ rồi	Med
84	8	răng ⟹ năng	Med
85	8	tai ⟹ ta	Med
86	8	thật ⟹ thực	-
87	8	thể ⟹ thế	Med
88	8	trồng ⟹ chọn	Med
89	8	tóc ⟹ tốt	Med
90	8	tự ⟹ từ	Med
91	8	và ⟹ vào	Fill
92	8	để ⟹ đến	Fill
93	7	an ⟹ ăn	-
94	7	bạn ⟹ bệnh	-
95	7	canxi ⟹ xây	Med
96	7	cho ⟹ cái	-
97	7	cái ⟹ có	Fill
98	7	có ⟹ tốt	Fill
99	7	cơn ⟹ cân	Med
100	7	dày ⟹ dài	Med
101	7	ghép ⟹ kết	Med
102	7	già ⟹ ra	Med
103	7	kinh ⟹ đến	Med
104	7	kỹ ⟹ cái	-
105	7	là ⟹ ta	Fill
106	7	nữ ⟹ nữa	-
107	7	qua ⟹ quá	-
108	7	siêu ⟹ thức	Med
109	7	thì ⟹ vì	Fill

Table 13 continued from previous page

Index	Occurrences	Confusion pair	Type
110	7	thì ⟹ để	Fill
111	7	tử ⟹ thành	Med
112	7	vậy ⟹ mà	Fill
113	7	vắcxin ⟹ sĩ	Med
114	7	âm ⟹ tâm	Med
115	7	đó ⟹ nữa	Fill
116	7	để ⟹ cái	Fill
117	6	buồng ⟹ buồn	Med
118	6	bà ⟹ và	-
119	6	cho ⟹ chất	-
120	6	cho ⟹ ra	-
121	6	con ⟹ có	Med
122	6	cung ⟹ không	Med
123	6	cách ⟹ cái	-
124	6	cái ⟹ với	Fill
125	6	có ⟹ của	Fill
126	6	có ⟹ nó	-
127	6	cấy ⟹ thấy	Med
128	6	của ⟹ có	-
129	6	d ⟹ b	-
130	6	dịch ⟹ việc	Med
131	6	f0 ⟹ không	Med
132	6	ghép ⟹ biết	Med
133	6	hợp ⟹ hai	-
134	6	khiếm ⟹ khiến	-
135	6	khá ⟹ khác	-
136	6	lý ⟹ lấy	-
137	6	lạ ⟹ lại	-
138	6	mãn ⟹ mạn	Med
139	6	ngày ⟹ này	-
140	6	nhổ ⟹ nhỏ	Med
141	6	nín ⟹ đến	Med
142	6	nó ⟹ là	Fill
143	6	phải ⟹ cái	-
144	6	ra ⟹ da	-
145	6	rong ⟹ tâm	Med
146	6	sợ ⟹ sở	-
147	6	sữa ⟹ sự	Med
148	6	thì ⟹ bị	Fill
149	6	thì ⟹ chúng	Fill
150	6	thì ⟹ thể	Fill
151	6	thú ⟹ thuốc	Med
152	6	thấy ⟹ cái	-
153	6	thể ⟹ sẽ	Med
154	6	trẻ ⟹ kể	Med
155	6	trẻ ⟹ để	Med
156	6	trồng ⟹ viêm	Med
157	6	u ⟹ ung	Med
158	6	viện ⟹ vị	Med
159	6	với ⟹ cái	Fill
160	6	xơ ⟹ thư	Med
161	6	âm ⟹ vitamin	Med
162	6	đo ⟹ đau	Med

Table 13 continued from previous page

Index	Occurrences	Confusion pair	Type
163	6	đây ⟹ này	Fill
164	6	đấy ⟹ đây	Fill
165	6	đầu ⟹ đau	Med
166	6	đầy ⟹ đây	-
167	6	đủ ⟹ đúng	-
168	5	cho ⟹ cao	-
169	5	cho ⟹ trong	-
170	5	chân ⟹ nhân	Med
171	5	chín ⟹ chính	Med
172	5	chỉ ⟹ cái	-
173	5	covid19 ⟹ chính	Med
174	5	còn ⟹ và	-
175	5	có ⟹ bác	Fill
176	5	có ⟹ là	Fill
177	5	do ⟹ ra	-
178	5	dạng ⟹ giảm	-
179	5	dự ⟹ nhiều	-
180	5	gây ⟹ cái	-
181	5	hoặc ⟹ họ	-
182	5	hư ⟹ hơn	Med
183	5	không ⟹ trong	-
184	5	khỏe ⟹ khoẻ	Med
185	5	kinh ⟹ cái	Med
186	5	kết ⟹ cái	Med
187	5	là ⟹ người	Fill
188	5	là ⟹ này	Fill
189	5	là ⟹ đã	Fill
190	5	mà ⟹ là	Fill
191	5	mái ⟹ máy	Med
192	5	mất ⟹ mức	-
193	5	mặt ⟹ mạch	Med
194	5	nang ⟹ năng	-
195	5	nhân ⟹ nhắn	Med
196	5	nhũ ⟹ nhiều	Med
197	5	này ⟹ ngày	Fill
198	5	nó ⟹ cái	Fill
199	5	nó ⟹ có	Fill
200	5	nền ⟹ nên	Med
201	5	phụ ⟹ phẫu	Med
202	5	que ⟹ quá	Med
203	5	quên ⟹ khuyên	-
204	5	răng ⟹ căn	Med
205	5	sao ⟹ ra	-
206	5	sâu ⟹ sau	Med
207	5	sẽ ⟹ sĩ	-
208	5	sức ⟹ rất	Med
209	5	thanh ⟹ thành	Med
210	5	thuyên ⟹ nguyên	Med
211	5	thì ⟹ người	Fill
212	5	thì ⟹ này	Fill
213	5	thính ⟹ tính	Med
214	5	thể ⟹ để	Med
215	5	tiêm ⟹ tim	Med
216	5	truyền ⟹ trì	Med

Table 13 continued from previous page

Index	Occurrences	Confusion pair	Type
217	5	tránh ⟹ trình	Med
218	5	trên ⟹ chân	-
219	5	trắng ⟹ tháng	Med
220	5	tức ⟹ rất	-
221	5	tử ⟹ công	Med
222	5	và ⟹ giảm	Fill
223	5	vâng ⟹ vân	-
224	5	xơ ⟹ oxy	Med
225	5	áp ⟹ tác	Med
226	5	âm ⟹ năm	Med
227	5	ăn ⟹ anh	Med
228	5	đeo ⟹ đều	-
229	5	đâu ⟹ đau	-
230	5	đó ⟹ đã	-
231	5	đầu ⟹ nào	Med
232	5	để ⟹ thì	-
233	5	để ⟹ đấy	-
234	5	đợt ⟹ được	-
235	5	ở ⟹ của	-

Table 13: Statistics of confusion pairs in *VietMed-Test* using the best pre-trained model *XLSR-53-Viet* (WER = 29.6).
In this table, we divide into 2 types of confusion pairs: Medical (a word that is a part of a medical term) and Filler (a word that is a part of a filler in real-world conversations). Only confusion pairs that have at least 5 occurrences in the recognition of the *VietMed-Test* are included in this table.

OOV	**Phonemes**	**Correct**
acenocoumarol	a:_2 k E_1 n o_1 k a_1 u_1 m a:_1 z O_1 n	N
alo	a:_1 l O_1	Y
amin	a:_1 m i_1 n	Y
amylase	a:_1 m i_1 l a:_1	N
apomorphine	a:_2 p o_1 m o_1 f i_1 n	Y
ascorbic	a:_1 s k O_1 b_&amp;lt; i_2 k	Y
aspirin	a:_1 s p i_1 z i_1 n	N
betacarotene	b_&amp;lt; E_1 t a:_2 k a:_1 z O_1 t E_1 n	N
betaglucan	b_&amp;lt; E_1 t a:_1 l u_1 k a:_1 n	Y
canxi	k a:_1 n s i_1	Y
catecholamine	k a:_2 t E_1 ts\O_1 l a:_1 m i_1 n	N
cbt	k b_&amp;lt; t	N
cholesterol	ts\O_1 l E_1 s t @:_1 O_1 n	N
clohidric	k @:_3 l o_1 a_1 z i_2 k	N
collagen	k o_1 l l a:_1 z E_1 n	Y
cologen	k o_1 l o_1 G E_1 n	Y
corticoid	k O_1 t i_1 k O_1 i_1	Y
cortisol	k O_1 t i_1 s O_1 n	Y
covid	k o_1 v i_1	N
ct	k t	N
dbs	z b_&amp;lt;	N
gen	G E_1 n	Y
google	G O_1 o_1 G o_1	N
gút	G u_2 t	Y
hdl	h d_&amp;lt; n	N
hemoglobin	h E_1 m o_1 G @:_3 l O_1 b_&amp;lt; i_1 n	Y

Table 14 continued from previous page

OOV	Phonemes	Correct
hormone	h O_1 m O_1 n	Y
inr	i_1 n	N
insulin	i_1 n s u_1 l i_1 n	Y
internet	i_1 n t @:_1 n E_2 t	Y
iod	i_1 o_2 t	Y
kcal	k k a:_1 n	N
kilogam	k i_1 l o_1 G a:_1 m	Y
laser	l a:_1 @:_1	N
ldl	l d_&amp;lt; n	N
levodopa	l E_1 v o_1 d_&amp;lt; O_2 p a:_1	Y
liraglutide	l i_1 z a:_1 l u_1 t i_1 d_&amp;lt; E_1	N
livestream	l a:_1 i_1 s ts\i_1 m	Y
mc	m k	N
mililit	m i_1 l i_1 l i_2 t	Y
milimet	m i_1 l i_1 m E_2 t	Y
monitor	m O_1 n i_1 t O_1	Y
mri	m z i_1	N
multivitamin	m u_1 n t i_1 v i_1 t a:_1 m i_1 n	Y
natri	n a:_1 t z i_1	N
niu	n i_1 u_1	Y
noark	n O_1 a:_1 k	Y
orlistat	O_1 l i_2 t a:_2 t	N
pacemaker	p a:_2 k E_1 m a:_1 k @:_1	N
parkinson	p a:_2 k i_1 n s O_1 n	N
pepsin	p E_2 p s i_1 n	Y
phytoncide	f i_1 t O_1 n s i_1 d_&amp;lt; E_1	N
pp	p p	N
protein	p @:_3 z o_1 t i_1 n	N
qr	k	N
radiography	z a:_1 d_&amp;lt; i_1 o_1 G @:_3 z a:_1 f i_1	N
run	z u_1 n	N
selen	s E_1 l E_1 n	Y
show	s @_1 u_1	N
sulfonylurea	s u_1 l f O_1 n i_1 l u_1 i_2	N
sunfuric	s u_1 n f u_1 i_2 k	N
test	t E_2 t	N
umami	u_1 m a:_1 m i_1	Y
vitamin	v i_1 t a:_1 m i_1 n	Y
vitamina	v i_1 t a:_1 m i_1 n a:_1	Y
vắcxin	v a_2 k s i_1 n	Y
ôliu	o_1 l i_1 u_1	Y

Table 14: List of OOVs found in *VietMed-Train*. In this table, only loan words are included together with their corresponding phonemes (in BABEL IARPA format). Since the use of the automatic toolkit Sequitur Grapheme-To-Phoneme (Bisani and Ney, 2008), some OOVs are correctly or incorrectly mapped, which we denote as Yes (Y) or No (N).