MultiMed: Multilingual Medical Speech Recognition via Attention Encoder Decoder

Khai Le-Duc[1,2,3], **Phuc Phan**[4], **Tan-Hanh Pham**[5], **Bach Phan Tat**[6], **Minh-Huong Ngo**[7], **Chris Ngo**[3], **Thanh Nguyen-Tang**[8], **Truong-Son Hy**[9]
[1]University of Toronto, Canada [2]University Health Network, Canada
[3]Knovel Engineering Lab, Singapore [4]FPT University, Vietnam
[5]Florida Institute of Technology, USA [6]KU Leuven, Belgium
[7]VNU University of Engineering and Technology, Vietnam
[8]Johns Hopkins University, USA [9]University of Alabama at Birmingham, USA
duckhai.le@mail.utoronto.ca

https://github.com/leduckhai/MultiMed/tree/master/MultiMed

Abstract

Multilingual automatic speech recognition (ASR) in the medical domain serves as a foundational task for various downstream applications such as speech translation, spoken language understanding, and voice-activated assistants. This technology improves patient care by enabling efficient communication across language barriers, alleviating specialized workforce shortages, and facilitating improved diagnosis and treatment, particularly during pandemics. In this work, we introduce *MultiMed*, the first multilingual medical ASR dataset, along with the first collection of small-to-large end-to-end medical ASR models, spanning five languages: Vietnamese, English, German, French, and Mandarin Chinese. To our best knowledge, *MultiMed* stands as the world's largest medical ASR dataset across all major benchmarks: total duration, number of recording conditions, number of accents, and number of speaking roles. Furthermore, we present the first multilinguality study for medical ASR, which includes reproducible empirical baselines, a monolinguality-multilinguality analysis, Attention Encoder Decoder (AED) vs Hybrid comparative study and a linguistic analysis. We present practical ASR end-to-end training schemes optimized for a fixed number of trainable parameters that are common in industry settings. All code, data, and models are available online.

1 Introduction

Automatic speech recognition (ASR) in the medical domain is a critical foundational task, serving a wide range of downstream tasks and applications, including speech translation (Mutal et al., 2020), electronic health record (Kumah-Crystal et al., 2018), information extraction (Selvaraj and Konam, 2020), speech summarization (Le-Duc et al., 2024a). This technology improves patient care by automating clinical documentation (Hodgson and Coiera, 2016), mitigating shortages of specialized healthcare personnel (Latif et al., 2020), and contributing to more accurate diagnosis and treatment (Luo et al., 2024), particularly under the increased demands observed during pandemic scenarios. Furthermore, the size of the ASR market is projected to reach USD 7.14 billion in 2024, with an anticipated compound annual growth rate (CAGR) of 14.24% from 2024 to 2030, resulting in a market volume of USD 15.87 billion by 2030 (Insights, 2024).

Recent research on ASR in the medical domain has been hindered by the lack of publicly available datasets, mainly due to privacy concerns. Existing datasets (see Table 9), such as the English medical ASR dataset by Fareez et al. (2022), are limited to simulated data on respiratory diseases, restricting research to this category and reducing applicability to diverse accents. The *PriMock57* dataset, containing 57 simulated primary care consultations (9 hours of recordings), also lacks generalizability (Korfiatis et al., 2022). The *AfriSpeech-200* dataset (Olatunji et al., 2023) mixes general and medical-domain speech, while the *myMediCon* dataset (Htun et al., 2024) includes Burmese read speech, both of which lack real-world applicability. The *VietMed* dataset (Le-Duc, 2024) is a real-world dataset focused on the Vietnamese language.

Furthermore, commercial medical ASR APIs, such as Google Cloud Healthcare, IBM Watson, Microsoft Azure Speech Service, Deepgram, and Nuance Dragon Medical One, are not free and do not provide publicly available models for fine-tuning or deployment, nor do they disclose training

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track), pages 1113–1150
July 28-30, 2025 ©2025 Association for Computational Linguistics

details.

This work aims to democratize medical ASR, making it freely accessible to everyone. Our key contributions are as follows.

- We present the *MultiMed* dataset - the first multilingual medical ASR dataset - which includes human-annotated high-quality real-world medical domain speech in 5 languages. To our best knowledge, *MultiMed* is the world's largest medical ASR dataset on all major diversity benchmarks: total duration (150 hours), number of recording conditions (10), number of accents (16) and number of speaking roles (6).
- We release the first publicly available multilingual medical ASR models, spanning small to large end-to-end configurations.
- We present the first multilinguality study for medical ASR, which includes: reproducible empirical baselines, a monolinguality-multilinguality analysis, Attention Encoder Decoder (AED) vs Hybrid study and a linguistic analysis
- We present practical ASR end-to-end training schemes optimized for a fixed number of trainable parameters that are common in industry settings

All code, data, and models are published online.

2 Data

2.1 Data Collection

Speech data with human-annotated transcripts were initially collected from real-world medical conversations published by professional medical channels on YouTube. In contrast to simulated datasets in the literature where doctors and patients play roles, our real-world dataset encompasses natural conversations of 10 distinct recording conditions (Documentary, Interview, Lecture, News, Podcast, Webinar, Speech, Talk, Vlog, Workshop) and 6 speaker roles (Lecturer, Doctor, Host, Patient, Podcaster, Broadcaster). Details of data collection for each language to ensure diversity are described in the Appendix C.1.

Our adherence to the Fair Use Policy and regulations regarding data consent, privacy, and anonymization of speaker identities in medical research is detailed in the Appendix B.

2.2 Data Quality Control

Quality control of the initial human-annotated transcripts from professional YouTube channels was carried out through manual review by our annotators, involving the correction of small inaccuracies or the exclusion of too erroneous transcripts. All transcripts were reviewed by medical experts with a certified linguistic level, which ensured, to the best of our knowledge, the final high-quality transcripts. Details of our annotators are described in the Appendix C.2. Data processing was also performed to further enhance the quality of the transcripts, as described in the Appendix C.3.

2.3 Data Statistics

Table 1 shows the dataset statistics of our *MultiMed* dataset in comparison with all existing publicly available medical ASR datasets, to the best of our knowledge. As shown in the table, our *MultiMed* dataset is the world's largest medical ASR dataset across all major diversity benchmarks: total duration (150 hours of recordings), number of recording conditions (10), number of accents (16) and number of speaking roles (6).

The statistics for the dataset split for each language are also shown in Table 2.

3 Problem Definition

An ASR model transcribes an audio signal into text by mapping an audio signal $x_1^T := x_1, x_2, ..., x_T$ of length T to the most likely word sequence w_1^N of length N. The relation w^* between the acoustic and word sequence is defined as the probability p:

$$w^* = \arg\max_{w_1^N} p(w_1^N | x_1^T) \tag{1}$$

In beam search process, the auxilary quantity Q for each unknown partial string (tree of partial hypotheses) w_1^n is described as:

$$\begin{aligned} Q(n; w_1^n) : &= \prod_{n'=1}^{n} p(w_{n'} | w_0^{n'-1}, x_1^T) \\ &= p(w_n | w_0^{n-1}, x_1^T) \cdot Q(n-1, w_1^{n-1}). \end{aligned} \tag{2}$$

After eliminating the less likely hypotheses in the beam search process, the word sequence probability is determined by the most optimal hypothesis:

$$p(w_1^N | x_1^T) = Q(N; w_1^N). \tag{3}$$

The complete mathematical formulation of AED is shown in Appendix D, while the formulation for the hybrid model is presented in Appendix F.1.

Dataset	Venue	Dur.	Language	Nature	#Rec. Cond.	#Spk	#Acc	#Roles
MultiMed (ours)	-	150h	Multiling.	Real-world	10	198	16	6
VietMed (Le-Duc, 2024)	LREC-COLING	16h	Vietnamese	Real-world	8	61	6	6
PriMock57 (Korfiatis et al., 2022)	ACL	9h	English	Simulated	1	64	4	2
Work by Fareez et al. (2022)	Nature	55h	English	Simulated	1	N/A	1	2
AfriSpeech-200 (Olatunji et al., 2023)	TACL	≈123h	African English	Read speech	1	N/A	N/A	1
myMediCon (Htun et al., 2024)	LREC-COLING	11h	Burmese	Read speech	1	12	5	2

Table 1: Dataset statistics in comparison with all existing works from left to right: Total duration in hours (h), language, nature of speech, number of recording conditions, number of speakers, number of accents, speaking roles. Full details are in Table 9 in the Appendix.

Language	Set	Samples	Total Dur. (h)	Avg. length (s)
Vietnamese	Train	4548	7.81	6.19
	Dev	1137	1.94	6.15
	Test	3437	6.02	6.31
English	Train	27922	83.87	10.81
	Dev	3082	8.96	10.46
	Test	5016	15.91	11.42
French	Train	1725	5.46	11.41
	Dev	52	0.18	12.13
	Test	358	1.15	11.57
Chinese	Train	1346	5.02	13.43
	Dev	97	0.34	12.75
	Test	231	0.85	13.21
German	Train	1551	5.37	12.46
	Dev	310	1.05	12.15
	Test	1242	4.32	12.53

Table 2: Statistics of our data samples: Total duration in hours (h) and average audio length in seconds (s).

4 Experimental Setups

4.1 Model Selection and Training

We opted to evaluate the performance of four pre-trained Whisper models (Radford et al., 2023) with varying sizes: Tiny, Base, Small, and Medium. These models, pre-trained on 680,000h of labeled multilingual data, offered a trade-off between accuracy and computational cost, allowing us to explore the impact of model size on performance. Details of hyperparameter tuning are shown in the Appendix E.1.

To investigate the impact of different fine-tuning strategies, we explored two main fine-tuning approaches for each model size: **Decoder-only fine-tuning** (encoder freezing) and **Fully encoder-decoder fine-tuning**. In the first approach, we focused on fine-tuning only the decoder of the pre-trained Whisper model. The encoder, responsible for aligning audio features, remained frozen during fine-tuning. This strategy aimed to leverage the previously learned representations of the pre-trained encoder for efficient time-frame alignment while adapting learnable parameters in the decoder for vocabulary generation. Otherwise, in the second approach, all parameters in both the encoder and decoder components of the pre-trained Whisper models were learnable. This approach allowed the model to proactively align time-frames in our dataset, potentially leading to better overall performance. The number of parameters for two fine-tuning settings is shown in Table 3.

Model	Fully encoder-decoder ft.	Decoder-only ft.
Tiny	37.76M	29.55M
Base	72.59M	52.00M
Small	241.73M	153.58M
Medium	763.86M	456.64M

Table 3: Statistics of total trainable parameters in the Whisper models for 2 settings: Fully encoder-decoder fine-tuning and decoder-only fine-tuning.

4.2 Evaluation Metrics

To assess the performance of the ASR models, we employed two standard evaluation metrics: Word Error Rate (WER) and Character Error Rate (CER). The description of the two metrics is shown in the Appendix E.2.

5 Experimental Results

5.1 Monolingual Fine-tuning

We fine-tuned various variants of the Whisper model in each language separately (known as monolingual fine-tuning) and analyzed the impact of model size and transfer learning (decoder-only vs. full encoder-decoder fine-tuning) on recognition accuracy, as shown in Table 4 and 5.

A clear correlation was observed between the size and performance of the model. As the model size increased from Tiny to Medium, WER and CER generally decreased across all languages, indicating that larger models better capture complex audio-text representations, improving accuracy.

The best results for most languages were obtained by fine-tuning only the decoder of the Medium model: Vietnamese achieved 20.05% and

Language	Tiny				Base				Small				Medium			
	WER		CER		WER		CER		WER		CER		WER		CER	
	dev	test	dev	test	dev	test	dev	test	dev	test	dev	test	dev	test	dev	test
Vietnamese	34.23	46.98	26.88	33.04	27.16	37.74	21.20	27.34	21.82	28.77	17.97	21.81	20.05	25.43	16.77	19.87
English	29.30	29.73	23.70	19.51	24.26	25.43	18.71	18.23	19.76	20.52	15.36	17.56	19.01	19.41	14.49	15.91
French	54.17	52.89	34.86	34.27	43.91	42.57	27.47	27.88	35.99	33.02	24.52	22.18	34.89	31.05	24.12	21.24
German	29.38	28.22	17.29	20.00	24.27	23.09	14.65	17.16	21.68	19.91	13.58	15.96	18.90	17.92	12.07	14.57
Chinese	91.36	95.97	34.20	43.71	85.66	89.73	27.63	38.02	80.35	88.50	23.95	34.28	79.17	86.52	26.11	35.82

Table 4: Main baselines - WERs and CERs of **decoder-only fine-tuning** (freezing the entire encoder) using different Whisper models on each separate language (**monolingual fine-tuning**)

Language	Tiny				Base				Small				Medium			
	WER		CER		WER		CER		WER		CER		WER		CER	
	dev	test	dev	test	dev	test	dev	test	dev	test	dev	test	dev	test	dev	test
Vietnamese	26.79	43.32	20.18	31.06	23.69	36.48	18.73	26.18	20.61	30.27	16.94	22.55	20.73	29.81	17.25	22.59
English	32.14	29.73	21.50	19.41	27.98	25.09	18.92	16.42	25.88	23.25	17.51	15.21	27.05	25.65	18.12	16.64
French	55.79	55.39	34.31	35.77	45.52	44.15	27.81	28.92	43.18	42.92	30.45	29.04	44.21	41.40	29.57	28.02
German	30.81	31.29	18.72	18.43	27.93	25.25	17.15	15.11	26.16	24.64	15.74	15.46	26.22	24.13	16.02	14.68
Chinese	92.93	98.85	34.00	50.94	86.05	94.58	30.64	42.75	86.44	92.44	27.85	39.71	89.78	94.08	30.19	40.97

Table 5: Main baselines - WERs and CERs of **fully encoder-decoder fine-tuning** using different Whisper models on each separate language (**monolingual fine-tuning**)

25.43% WERs on the dev and test sets, respectively; English reached 19.01% and 19.41% WERs; and French yielded 34.89% and 31.05% WERs.

An exception was Chinese, where fine-tuning the Small model's decoder produced the best results: 23.95% and 34.28% CERs on the dev and test sets. Since Chinese uses characters as fundamental units of meaning, CER is a more accurate measure of recognition than WER (Wang et al., 2016; Gao et al., 2006), unlike alphabetic languages.

5.2 Multilingual Fine-tuning

In addition to fine-tuning each language separately, we also combined all languages for experimentation, known as multilingual fine-tuning, as shown in Table 6. In multilingual fine-tuning, we achieved superior performance in most languages, though there was a slight performance degradation for Chinese, compared to monolingual fine-tuning in Table 5. Both high-resource languages, such as English, and lower-resource languages, including Vietnamese, French, and German, showed improvement under the multilingual fine-tuning regime. This outcome is noteworthy, as previous studies on multilingual fine-tuning observed that shared discrete latent speech representations across languages such as Vietnamese, English, Chinese, French, and German tend to cluster at large distances, and therefore usually affect accuracy in the multilingual setting (Baevski et al., 2020a; Conneau et al., 2021a; Vieting et al., 2023; Tüske et al., 2014; Chuangsuwanich, 2016).

Language	WER		CER	
	dev	test	dev	test
Vietnamese	23.11	30.22	18.78	22.51
English	18.92	16.62	12.97	11.05
French	43.62	37.27	29.24	24.25
German	25.26	22.92	15.31	14.05
Chinese	89.78	101.97	26.65	41.21

Table 6: Main baselines - WERs and CERs of **fully encoder-decoder fine-tuning** using Small Whisper model on all languages (**multilingual fine-tuning**)

5.3 AED vs Hybrid

End-to-end ASR, with the AED approach, and Hybrid ASR models (Hidden Markov Models) are two key paradigms in ASR research. This section compares AED and Hybrid ASR models. For a fair comparison, we use wav2vec 2.0 (Baevski et al., 2020a) as the acoustic model for Hybrid ASR, as it is a Transformer-based encoder, similar to the Transformer-based encoder-decoder of Whisper.

Table 7 presents a comparison between the AED and Hybrid models. The AED models were pre-trained on 680,000 hours of labeled multilingual data, including 691 hours of Vietnamese, while the Hybrid models were pre-trained on unlabeled data. Despite having fewer parameters and less labeled data, Hybrid models achieve comparable WERs on the Vietnamese test set. AED models only outperform Hybrid models significantly when

		AED		Hybrid	
		Small	Medium	w2v2-Viet	XLSR-53-Viet
WER	dev	21.8	20.1	25.9	25.7
	test	28.8	25.4	29.0	28.8
#Data		680,000h labeled multiling. (691h labeled Viet.)		1200h unlabeled Viet.	56,000h unlabeled multiling. +1200h unlabeled Viet.
#Params		153M	456M	123M	123M
#Layers		12	24	8	8
Width		768	1024	768	768
#Att. Heads		12	16	16	16
Features		MFCC		Raw waveform	
LM fusion		Deep fusion		Shallow fusion	

Table 7: Comparison between AED and Hybrid experiments. WERs are reported on our Vietnamese dev and test set. All models were fine-tuned on the same Vietnamese set. Hybrid models employ wav2vec 2.0 as acoustic model (Baevski et al., 2020a). Full details of experiments are shown in Appendix F and the breakdown per speaker is shown in Table 11 in the Appendix.

scaled three times. This finding supports prior research on the data and computational efficiency of Hybrid models in general-domain ASR (Lüscher et al., 2019a; Zeyer et al., 2018b,c, 2019), and is the first confirmation of this trend in the medical domain.

6 Ablation Study: Freezing Schemes

This section presents the results of our ablation study. The Small Whisper models fit within a 24GB GPU without out-of-memory issues. We evaluated the impact of freezing various layers on performance, focusing on test set WERs for most languages and CERs for Chinese. The tested freezing configurations are shown in Table 8.

In both the *0-8 encoder* and *3-11 encoder* settings, model performance on test sets is worse than when the entire encoder is frozen in Table 5. This suggests that, within a fixed budget, freezing the entire encoder, which aligns time frame features with language representations, is crucial to achieve high accuracy and computational efficiency, as seen in the general domain AED ASR (Ueno et al., 2018).

We also explored the effect of freezing Whisper's decoder, focusing on fine-tuning only the last three layers of both the encoder and decoder (*0-8 encoder & 0-8 decoder*). As shown in Table 8, this setup resulted in worse performance compared to fine-tuning only the decoder while freezing layers 0-8 of the encoder (*0-8 encoder*). Performance degradation likely results from a significant reduction in trainable parameters in the decoder, which is responsible for generating subword units. Given the fixed vocabulary, out-of-vocabulary (OOV) words, and context length in Whisper's Byte-Pair Encoding (BPE) tokenizer (Gage, 1994), the decrease in trainable autoregressive parameters likely hinders the decoder's ability to effectively separate subword tokens, leading to reduced decoding accuracy (Ho et al., 2024; Bapna et al., 2020).

We fine-tuned the first three decoder layers and the last three encoder layers (*0-8 encoder & 3-11 decoder*), which generally resulted in higher test set accuracy for most languages compared to *0-8 encoder & 0-8 decoder*. This suggests that freezing a contiguous set of layers is the key to achieving high accuracy with an equivalent number of trainable parameters in the decoder.

Fine-tuning the last three decoder layers (*0-11 encoder & 0-8 decoder*) also outperformed *0-8 encoder & 0-8 decoder* in test accuracy and was competitive with *0-8 encoder & 3-11 decoder*, despite fewer trainable parameters. Likewise, the *3-11 encoder & 3-11 decoder* configuration yielded the worst performance in all languages. These findings support the hypothesis that consistent freezing of contiguous layer groups is critical for high accuracy within a fixed parameter budget.

7 Error Analysis

To our best knowledge, there has been no error analysis based on the linguistic perspective for languages other than English. Therefore, we used the English literature to compare with our findings.

We manually analyzed the errors in 50 randomly collected samples from each language. Generally, the errors observed in medical ASR systems are

Language	0-8 encoder				3-11 encoder				0-8 encoder & 0-8 decoder			
	WER		CER		WER		CER		WER		CER	
	dev	test	dev	test	dev	test	dev	test	dev	test	dev	test
Vietnamese	21.27	29.32	17.60	22.07	21.28	30.74	17.60	22.97	23.44	33.30	19.33	24.78
English	25.68	26.50	14.87	17.84	22.68	25.20	14.73	16.90	16.78	32.11	12.78	22.42
French	39.36	35.50	27.48	23.70	38.71	35.03	26.32	23.59	37.68	35.93	25.69	24.02
German	23.65	21.49	15.04	13.64	22.82	20.94	14.30	13.29	22.64	23.04	14.54	15.14
Chinese	78.97	88.33	23.37	35.72	83.49	89.48	25.20	37.07	80.75	94.91	28.32	38.80
	0-8 encoder & 3-11 decoder				**0-11 encoder & 0-8 decoder**				**3-11 encoder & 3-11 decoder**			
Vietnamese	34.98	32.81	29.34	24.65	24.75	32.11	20.86	25.06	40.87	32.10	36.06	24.30
English	20.61	28.31	15.55	19.56	16.06	31.32	12.68	22.34	21.53	34.81	17.09	22.96
French	35.04	40.70	23.32	32.96	37.97	37.39	27.25	26.60	57.26	40.10	44.82	28.83
German	22.22	21.02	13.83	13.35	22.11	22.26	14.65	14.98	22.86	22.47	15.01	15.23
Chinese	79.76	93.51	23.93	35.34	84.67	87.84	26.24	34.36	132.80	103.04	53.74	41.21

Table 8: Ablation study - WERs and CERs of various freezing schemes using Small Whisper model on each separate language (**monolingual fine-tuning**). Small Whisper model has 12 layers in the encoder and 12 layers in the decoder. For example, *0-8 encoder* means freezing all layers from layer 0 to layer 8 in the encoder, the rest layers are fine-tuned.

diverse and cover a wide range of issues. For all 5 languages, these typically include misrecognition of drug names and dosages, incorrect medical institutions, anatomical discrepancies (e.g., left-right confusion), medical terms' inconsistencies, mismatches in patient age and gender, incorrect identification of physician names, and inaccuracies in dates. These findings are consistent with the study by Hodgson and Coiera (2016) in the English medical ASR dataset. Additionally, the misrecognition is exacerbated by the generation of non-existent terms, which is also known as hallucination in the Large Language Models (LLMs) era, as well as omissions (e.g. deletion errors) and duplications (e.g., insertion errors) within the ASR output (see Figure 12 in the Appendix). These findings are also confirmed by McGurk et al. (2008) in English ASR for radiology reports.

Furthermore, ASR errors typically arise from the proximity of vowels in the phonological space for Vietnamese, English, German, and French, while for Chinese, confusion predominantly stems from minimal pairs with distinct tones and homophones. Detailed error analysis based on the linguistic perspective for each language is in Appendix G.

8 Conclusion

In this work, we present *MultiMed*, a real-world dataset for ASR in the medical domain, accompanied by a collection of small-to-large end-to-end ASR models, covering five languages: Vietnamese, English, German, French, and Mandarin Chinese. To our best knowledge, *MultiMed* stands as the world's largest medical ASR dataset across all major benchmarks.

As the first study of multilingual ASR in the medical domain, our findings demonstrate that **(1):** multilingual fine-tuning produces superior accuracy compared to monolingual fine-tuning, although shared discrete latent speech representations across languages, such as Vietnamese, English, Chinese, French and German, exhibit clustering at large distances, which could potentially reduce accuracy in a multilingual fine-tuning setting. Furthermore, in the AED vs Hybrid study, we showed that **(2):** Hybrid models remain more efficient in terms of data utilization and computational performance compared to AED models. In the layer-wise ablation study of AED models, we found that **(3):** on a fixed budget, freezing the entire encoder is important for achieving both high accuracy and computational efficiency. Additionally, **(4):** maintaining the consistent freezing of a contiguous group of layers is important for achieving high accuracy. Finally, as shown in the linguistic analysis for multilingual medical ASR, we observed that **(5):** medical ASR errors often involve misrecognitions of drug names, dosages, institutions, anatomical details, demographics of patients, physician names, etc., along with hallucinated terms, omissions, and duplications. **(6):** Errors also often arise from the proximity of vowels in the phonological space for Vietnamese, English, German and French, while for Chinese, confusion predominantly stems from minimal pairs with distinct tones and homophones.

9 Limitations

Open research questions: Several research questions about the impact of multilinguality on medical ASR remain unaddressed and fall outside the scope of this study.

- Cross-language transfer learning: How can transfer learning be optimized to leverage data from high-resource languages to improve medical ASR performance in low-resource languages? Can shared acoustic and linguistic representations (e.g., from hospitals' recording conditions and shared medical terms across languages) effectively bridge the gap between typologically different languages?
- Zero-shot and few-shot medical ASR: What are the best methods for enabling general-domain ASR models to understand unseen medical-domain test set (zero-shot learning) or to adapt with minimal medical-domain data (few-shot learning)? How can medical-domain models be trained to generalize effectively across languages without overfitting to dominant languages (e.g., English) in the dataset?
- Code-Switching Challenges: How does each ASR module handle code-switching, where speakers switch between two or more languages within the same sentence, especially for medical terms?
- Bias and Fairness in Multilingual Medical ASR: How can we address biases in multilingual medical ASR models that disproportionately affect minority languages or speakers with diverse accents, especially when patients and doctors are not of major ethnicity? What metrics and evaluation protocols should be established to assess fairness and inclusivity in multilingual medical ASR systems?

Clinical impact: The primary objective of our study is to establish baselines rather than introduce novel techniques to minimize WER in medical ASR systems. Given the critical nature of medical transcription, inaccuracies in ASR output can have serious implications, potentially affecting patient diagnoses and treatment decisions (Adane et al., 2019). Thus, real-world deployment of our systems should be preceded by pilot testing in clinical environments to ensure reliability prior to full-scale implementation.

References

Kasaw Adane, Mucheye Gizachew, and Semalegne Kendie. 2019. The role of medical data in efficient patient care delivery: a review. *Risk management and healthcare policy*, pages 67–73.

Ayo Adedeji, Sarita Joshi, and Brendan Doohan. 2024. The sound of healthcare: Improving medical transcription asr accuracy with large language models. *arXiv preprint arXiv:2402.07658*.

Tejumade Afonja, Tobi Olatunji, Sewade Ogun, Naome A Etori, Abraham Owodunni, and Moshood Yekini. 2024. Performant asr models for medical entities in accented speech. *arXiv preprint arXiv:2406.12387*.

Tasos Anastasakos, John McDonough, Richard Schwartz, and John Makhoul. 1996. A compact model for speaker-adaptive training. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, volume 2, pages 1137–1140. IEEE.

Brian G Arndt, John W Beasley, Michelle D Watkinson, Jonathan L Temte, Wen-Jan Tuan, Christine A Sinsky, and Valerie J Gilchrist. 2017. Tethered to the ehr: primary care physician workload assessment using ehr event log data and time-motion observations. *The Annals of Family Medicine*, 15(5):419–426.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020a. 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.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020b. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Ankur Bapna, Naveen Arivazhagan, and Orhan Firat. 2020. Controlling computation versus quality for neural sequence models. *arXiv preprint arXiv:2002.07106*.

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

Leo Breiman. 2017. *Classification and regression trees*. Routledge.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

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.

Ekapol Chuangsuwanich. 2016. *Multilingual techniques for low resource automatic speech recognition*. Ph.D. thesis, Massachusetts Institute of Technology.

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

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Patrick Doetsch, Albert Zeyer, Paul Voigtlaender, Ilia Kulikov, Ralf Schlüter, and Hermann Ney. 2017. Returnn: The rwth extensible training framework for universal recurrent neural networks. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5345–5349. IEEE.

Lane F Donnelly, Robert Grzeszczuk, and Carolina V Guimaraes. 2022. Use of natural language processing (nlp) in evaluation of radiology reports: an update on applications and technology advances. In *Seminars in Ultrasound, CT and MRI*, volume 43, pages 176–181. Elsevier.

Mohit Dua, Akanksha, and Shelza Dua. 2023. Noise robust automatic speech recognition: review and analysis. *International Journal of Speech Technology*, 26(2):475–519.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. 2024. The llama 3 herd of models.

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*, pages 512–524, Cham. Springer International Publishing.

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.

G.D. Forney. 1973. The viterbi algorithm. *Proceedings of the IEEE*, 61(3):268–278.

Philip Gage. 1994. A new algorithm for data compression. *The C Users Journal*, 12(2):23–38.

Qin Gao, Xiaojun Lin, and Xihong Wu. 2006. Just-in-time latent semantic adaptation on language model for chinese speech recognition using web data. In

2006 IEEE Spoken Language Technology Workshop, pages 50–53. IEEE.

John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Acoustics, speech, and signal processing, ieee international conference on*, volume 1, pages 517–520. IEEE Computer Society.

Namgyu Ho, Sangmin Bae, Taehyeon Kim, hyunjik.jo, Yireun Kim, Tal Schuster, Adam Fisch, James Thorne, and Se-Young Yun. 2024. Block transformer: Global-to-local language modeling for fast inference. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Tobias Hodgson and Enrico Coiera. 2016. Risks and benefits of speech recognition for clinical documentation: a systematic review. *Journal of the american medical informatics association*, 23(e1):e169–e179.

Laurence Horn and Andrea Hoa Pham. 2004. *Vietnamese tone: A new analysis*. Routledge.

Wei-Chen Hsu, Pei-Xu Lin, Chi-Jou Li, Hao-Yu Tien, Yi-Huang Kang, and Pei-Ju Lee. 2024. An enhanced model for asr in the medical field. In *2024 IEEE International Conference on Information Reuse and Integration for Data Science (IRI)*, pages 43–48. IEEE.

Hay Man Htun, Ye Kyaw Thu, Hutchatai Chanlekha, Kotaro Funakoshi, and Thepchai Supnithi. 2024. mymedicon: End-to-end burmese automatic speech recognition for medical conversations. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12032–12039.

Statista Market Insights. 2024. Artificial intelligence: in-depth market analysis.

Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. pmlr.

Yu Jiang and Christian Poellabauer. 2021. A sequence-to-sequence based error correction model for medical automatic speech recognition. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 3029–3035. IEEE.

Maree Johnson, Samuel Lapkin, Vanessa Long, Paula Sanchez, Hanna Suominen, Jim Basilakis, and Linda Dawson. 2014. A systematic review of speech recognition technology in health care. *BMC medical informatics and decision making*, 14:1–14.

Allard Jongman, Yue Wang, Corinne B Moore, and Joan A Sereno. 2006. *Perception and production of Mandarin Chinese tones*. na.

Snigdhaswin Kar, Prabodh Mishra, Ju Lin, Min-Jae Woo, Nicholas Deas, Caleb Linduff, Sufeng Niu, Yuzhe Yang, Jerome McClendon, D. Hudson Smith, Melissa C. Smith, Ronald W. Gimbel, and Kuang-Ching Wang. 2021. Systematic evaluation and enhancement of speech recognition in operational medical environments. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Alex Papadopoulos Korfiatis, Francesco Moramarco, Radmila Sarac, and Aleksandar Savkov. 2022. Primock57: A dataset of primary care mock consultations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 588–598.

Yaa A Kumah-Crystal, Claude J Pirtle, Harrison M Whyte, Edward S Goode, Shilo H Anders, and Christoph U Lehmann. 2018. Electronic health record interactions through voice: a review. *Applied clinical informatics*, 9(03):541–552.

Siddique Latif, Junaid Qadir, Adnan Qayyum, Muhammad Usama, and Shahzad Younis. 2020. Speech technology for healthcare: Opportunities, challenges, and state of the art. *IEEE Reviews in Biomedical Engineering*, 14:342–356.

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.

Khai Le-Duc. 2024. Vietmed: A dataset and benchmark for automatic speech recognition of vietnamese in the medical domain. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17365–17370.

Khai Le-Duc, Khai-Nguyen Nguyen, Long Vo-Dang, and Truong-Son Hy. 2024a. Real-time speech summarization for medical conversations. *arXiv preprint arXiv:2406.15888*.

Khai Le-Duc, David Thulke, Hung-Phong Tran, Long Vo-Dang, Khai-Nguyen Nguyen, Truong-Son Hy, and Ralf Schlüter. 2024b. Medical spoken named entity recognition.

Xiao Luo, Le Zhou, Kathleen Adelgais, and Zhan Zhang. 2024. Assessing the effectiveness of automatic speech recognition technology in emergency medicine settings: A comparative study of four ai-powered engines. *Research square*, pages rs–3.

Christoph Lüscher, Eugen Beck, Kazuki Irie, Markus Kitza, Wilfried Michel, Albert Zeyer, Ralf Schlüter, and Hermann Ney. 2019a. Rwth asr systems for librispeech: Hybrid vs attention. In *Interspeech*, pages 231–235, Graz, Austria.

Christoph Lüscher, Eugen Beck, Kazuki Irie, Markus Kitza, Wilfried Michel, Albert Zeyer, Ralf Schlüter, and Hermann Ney. 2019b. RWTH ASR Systems for LibriSpeech: Hybrid vs Attention. In *Proc. Interspeech 2019*, pages 231–235.

Christoph Lüscher, Mohammad Zeineldeen, Zijian Yang, Tina Raissi, 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*.

Anirudh Mani, Shruti Palaskar, and Sandeep Konam. 2020a. Towards understanding asr error correction for medical conversations. In *NLPMC*.

Anirudh Mani, Shruti Palaskar, Nimshi Venkat Meripo, Sandeep Konam, and Florian Metze. 2020b. Asr error correction and domain adaptation using machine translation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6344–6348. IEEE.

Simon McGurk, Katrin Brauer, TV Macfarlane, and KA Duncan. 2008. The effect of voice recognition software on comparative error rates in radiology reports. *The British journal of radiology*, 81(970):767–770.

Mehryar Mohri, Fernando Pereira, and Michael Riley. 2002. Weighted finite-state transducers in speech recognition. *Computer Speech & Language*, 16(1):69–88.

Matthew Mulholland, Melissa Lopez, Keelan Evanini, Anastassia Loukina, and Yao Qian. 2016. A comparison of asr and human errors for transcription of non-native spontaneous speech. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5855–5859. IEEE.

Jonathan Mutal, Johanna Gerlach, Pierrette Bouillon, and Hervé Spechbach. 2020. Ellipsis translation for a medical speech to speech translation system. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 281–290.

Revathy Nayar. 2017. Towards designing speech technology based assistive interfaces for children's speech therapy. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pages 609–613.

Hermann Ney, Ute Essen, and Reinhard Kneser. 1994. On structuring probabilistic dependences in stochastic language modelling. *Computer Speech & Language*, 8(1):1–38.

Tobi Olatunji, Tejumade Afonja, Aditya Yadavalli, Chris Chinenye Emezue, Sahib Singh, Bonaventure FP Dossou, Joanne Osuchukwu, Salomey Osei, Atnafu Lambebo Tonja, Naome Etori, et al. 2023. Afrispeech-200: Pan-african accented speech dataset for clinical and general domain asr. *Transactions of the Association for Computational Linguistics*, 11:1669–1685.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth

Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report.

Stefan Ortmanns, Hermann Ney, and Xavier Aubert. 1997. A word graph algorithm for large vocabulary continuous speech recognition. *Computer Speech & Language*, 11(1):43–72.

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. 2019a. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. In *Proc. Interspeech 2019*, pages 2613–2617.

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

Kimberly D Pelland, Rosa R Baier, and Rebekah L Gardner. 2017. 'it is like texting at the dinner table': a qualitative analysis of the impact of electronic health records on patient–physician interaction in hospitals. *BMJ Health & Care Informatics*, 24(2).

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. 2011. The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society.

Daniel Povey, Dimitri Kanevsky, Brian Kingsbury, Bhuvana Ramabhadran, George Saon, and Karthik Visweswariah. 2008. Boosted mmi for model and feature-space discriminative training. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4057–4060. IEEE.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

J Ruby, Susan Diana, Yanmin Yuan, William Harry, J Tisa, J Nedumaan, Yang Yung, J Lepika, Thomas Binford, PS Jagadeesh Kumar, et al. 2020. Automatic speech recognition and machine learning for robotic arm in surgery. *Trends in Technical & Scientific Research*, 4(1):5–9.

David Rybach, Stefan Hahn, Patrick Lehnen, David Nolden, Martin Sundermeyer, Zoltan Tüske, Simon Wiesler, Ralf Schlüter, and Hermann Ney. 2011. Rasr-the rwth aachen university open source speech recognition toolkit. In *Proc. ieee automatic speech recognition and understanding workshop*.

Sakriani Sakti, Keigo Kubo, Sho Matsumiya, Graham Neubig, Tomoki Toda, Satoshi Nakamura, Fumihiro Adachi, and Ryosuke Isotani. 2014. Towards multi-lingual conversations in the medical domain: Development of multilingual medical data and a network-based ASR system. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2639–2643, Reykjavik, Iceland. European Language Resources Association (ELRA).

Askars Salimbajevs and Jurgita Kapočiūtė-Dzikienė. 2022. Automatic speech recognition model adaptation to medical domain using untranscribed audio. In *International Baltic Conference on Digital Business and Intelligent Systems*, pages 65–79. Springer.

Kshitij Saxena, Robert Diamond, Reid F Conant, Terri H Mitchell, Guido Gallopyn, and Kristin E Yakimow. 2018. Provider adoption of speech recognition and its impact on satisfaction, documentation quality, efficiency, and cost in an inpatient ehr. *AMIA Summits on Translational Science Proceedings*, 2018:186.

R. Schluter, I. Bezrukov, H. Wagner, and H. Ney. 2007. Gammatone features and feature combination for large vocabulary speech recognition. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, volume 4, pages IV–649–IV–652.

Antonia Schulte, Rodrigo Suarez-Ibarrola, Daniel Wegen, Philippe-Fabian Pohlmann, Elina Petersen, and Arkadiusz Miernik. 2020. Automatic speech recognition in the operating room–an essential contemporary tool or a redundant gadget? a survey evaluation among physicians in form of a qualitative study. *Annals of Medicine and Surgery*, 59:81–85.

Sai P Selvaraj and Sandeep Konam. 2020. Medication regimen extraction from medical conversations. In *Explainable AI in Healthcare and Medicine: Building a Culture of Transparency and Accountability*, pages 195–209. Springer.

Christine Sinsky, Lacey Colligan, Ling Li, Mirela Prgomet, Sam Reynolds, Lindsey Goeders, Johanna Westbrook, Michael Tutty, and George Blike. 2016. Allocation of physician time in ambulatory practice: a time and motion study in 4 specialties. *Annals of internal medicine*, 165(11):753–760.

Andreas Stolcke and Jasha Droppo. 2017. Comparing human and machine errors in conversational speech transcription. *Interspeech*.

Monica Sunkara, Srikanth Ronanki, Kalpit Dixit, Sravan Bodapati, and Katrin Kirchhoff. 2020. Robust prediction of punctuation and truecasing for medical asr. In *Proceedings of the First Workshop on Natural Language Processing for Medical Conversations*, pages 53–62.

Hanna Suominen, Liyuan Zhou, Leif Hanlen, Gabriela Ferraro, et al. 2015. Benchmarking clinical speech recognition and information extraction: new data, methods, and evaluations. *JMIR medical informatics*, 3(2):e4321.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Zoltán Tüske, Pavel Golik, David Nolden, Ralf Schlüter, and Hermann Ney. 2014. Data augmentation, feature combination, and multilingual neural networks to improve asr and kws performance for low-resource languages. In *Interspeech*, pages 1420–1424.

Sei Ueno, Takafumi Moriya, Masato Mimura, Shinsuke Sakai, Yusuke Shinohara, Yoshikazu Yamaguchi, Yushi Aono, and Tatsuya Kawahara. 2018. Encoder transfer for attention-based acoustic-to-word speech recognition. In *INTERSPEECH*, pages 2424–2428.

Marieke M van Buchem, Hileen Boosman, Martijn P Bauer, Ilse MJ Kant, Simone A Cammel, and Ewout W Steyerberg. 2021. The digital scribe in clinical practice: a scoping review and research agenda. *NPJ digital medicine*, 4(1):57.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

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)*.

Dong Wang, Zhiyuan Tang, Difei Tang, and Qing Chen. 2016. Oc16-ce80: A chinese-english mixlingual database and a speech recognition baseline. In *2016 Conference of The Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Techniques (O-COCOSDA)*, pages 84–88. IEEE.

Michael J Wargo. 1967. Human operator response speed, frequency, and flexibility: A review and analysis. *Human factors*, 9(3):221–238.

Oliver Wendt, Raymond W Quist, and Lyle L Lloyd. 2011. *Assistive technology: Principles and applications for communication disorders and special education*, volume 4. Brill.

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

Albert Zeyer, Parnia Bahar, Kazuki Irie, Ralf Schlüter, and Hermann Ney. 2019. A comparison of transformer and lstm encoder decoder models for asr. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 8–15. IEEE.

Albert Zeyer, Kazuki Irie, Ralf Schlüter, and Hermann Ney. 2018b. Improved training of end-to-end attention models for speech recognition. In *Interspeech*, Hyderabad, India.

Albert Zeyer, André Merboldt, Ralf Schlüter, and Hermann Ney. 2018c. A comprehensive analysis on attention models. In *Interpretability and Robustness in Audio, Speech, and Language (IRASL) Workshop, Conference on Neural Information Processing Systems (NeurIPS)*, Montreal, Canada.

Jun Zhang, Jingyue Wu, Yiyi Qiu, Aiguo Song, Weifeng Li, Xin Li, and Yecheng Liu. 2023. Intelligent speech technologies for transcription, disease diagnosis, and medical equipment interactive control in smart hospitals: A review. *Computers in Biology and Medicine*, 153:106517.

Contents

1 Introduction **1**

2 Data **2**
- 2.1 Data Collection 2
- 2.2 Data Quality Control 2
- 2.3 Data Statistics 2

3 Problem Definition **2**

4 Experimental Setups **3**
- 4.1 Model Selection and Training 3
- 4.2 Evaluation Metrics 3

5 Experimental Results **3**
- 5.1 Monolingual Fine-tuning 3
- 5.2 Multilingual Fine-tuning 4
- 5.3 AED vs Hybrid 4

6 Ablation Study: Freezing Schemes **5**

7 Error Analysis **5**

8 Conclusion **6**

9 Limitations **7**

A Related Works **17**

B Ethical Statements **19**
- B.1 Fair Use 19
- B.2 Data Consent 20

C Details of Data Creation **23**
- C.1 Details of Data Collection per Language 23
  - C.1.1 English 23
  - C.1.2 French 23
  - C.1.3 Chinese 23
  - C.1.4 German 23
- C.2 Details of Data Quality Control 23
- C.3 Data Processing 23
- C.4 Full Data Statistics 26

D Attention Encoder Decoder (AED) **28**

E Full Experimental Setups **29**
- E.1 Hyperparameter Tuning 29
- E.2 Details of Evaluation Metrics 29

F Details of Hybrid ASR Experiments **29**
- F.1 Hybrid wav2vec 2.0 29
  - F.1.1 Hybrid ASR 29
  - F.1.2 Modified wav2vec 2.0 30

F.2 Experimental Setups 31
F.3 Extra Experimental Results 32

G Full Error Analysis **35**
G.1 English 35
G.2 Vietnamese 35
G.3 Chinese 36
G.4 French 36
G.5 German 37

A Related Works

Among the limited existing studies below, to the best of our knowledge, none of them have made their datasets or pre-trained models publicly available, nor have they been conducted on publicly accessible datasets due to privacy concerns, which poses a significant challenge for the reproducibility and deployment of medical ASR research[1].

Multilingual medical ASR: The research carried out by Lüscher et al. (2023) has focused on the development of Hybrid ASR systems (Lüscher et al., 2019b) in the RETURNN framework (Zeyer et al., 2018a; Doetsch et al., 2017) to transcribing multilingual telephone conversations between patients and physicians, using Gammatone features (Schluter et al., 2007) as input in a supervised only approach. Vieting et al. (2023) examine the efficient utilization of the large multilingual acoustic pre-trained model XLSR-53 (Conneau et al., 2021b) for medical ASR in three languages, focusing on resolving the issue of sampling rate mismatch using wav2vec 2.0 (Baevski et al., 2020b) as encoder and RASR (Rybach et al., 2011) as decoding framework. In (Sakti et al., 2014), a multilingual acoustic model fine-tuned on in-house medical domain data is presented utilizing weighted finite-state transducers (Mohri et al., 2002), speaker adaptive training (Anastasakos et al., 1996), and boosted maximum mutual information (Povey et al., 2008) in conjunction with Kaldi decoding (Povey et al., 2011) of n-gram (Ney et al., 1994) language models for every specific language. However, to the best of our knowledge, in all studies separate monolingual models are typically used for each respective language, rather than utilizing a unified multilingual model capable of transcribing multilingual conversations seamlessly. Therefore, we are the first study to present a unified multilingual model that can dynamically adapt to different languages in medical conversations without the need for separate models.

Acoustic challenges for medical ASR: In this context, several challenges arise, including variability in acoustic and recording conditions, the mismatch in telephony bandwidth, the impact of medical mask usage, and the presence of background noise from various devices and dynamic environmental factors (Lüscher et al., 2023). In addition, a bidirectional input issue is observed, as a single recording channel is shared between the physician and the patient in emergency room and hospital settings. Studies such as (Edwards et al., 2017; Chiu et al., 2018; Kar et al., 2021; Dua et al., 2023) have addressed the challenges related to difficult acoustic conditions by modifying model components like feature extractor, acoustic model, and so on. Furthermore, studies like (Salimbajevs and Kapočiūtė-Dzikienė, 2022) address the robustness in noisy acoustic environments using a large amount of unlabeled medical ASR data. In addition, Luo et al. (2024) uses emergency medical services or prehospital care as a research context to generate data in the domain, as it represents a prototypical example of dynamic and variable medical environments, involving numerous participants, such as healthcare professionals, patients, bystanders, and family members.

Language modeling for medical ASR: The specialized medical terminology in each language presents an additional challenge. A simple method to address these challenges involves correcting ASR errors at the output level (Mani et al., 2020a; Hsu et al., 2024; Mani et al., 2020b) or focusing on medically named entities (Afonja et al., 2024; Le-Duc et al., 2024b; Suominen et al., 2015). Another approach is to train a domain-specific language model to decode the ASR encoder (Jiang and Poellabauer, 2021). Furthermore, LLMs, such as the GPT series (OpenAI et al., 2024; Brown et al., 2020), Gemini (Team et al., 2024, 2023), and Llama (Touvron et al., 2023; Dubey et al., 2024) for example, have potential utility in rectifying medical ASR errors (Adedeji et al., 2024). Another study by Sunkara et al. (2020) involves the joint modeling of punctuation and truecasing in medical ASR transcripts utilizing pre-trained language models, such as BERT (Devlin et al., 2019).

Application of medical ASR: One of the most common use cases is for clinical documentation (Latif et al., 2020). It is a laborious and complex task that could lead to burnout of the clinician (Arndt et al., 2017), inefficiency of the doctor-patient time (Sinsky et al., 2016), and lower patient satisfaction (Pelland et al., 2017). The adoption of electronic health records (EHRs) has been pro-

[1]Medical ASR, also known as Medical-domain ASR, focuses on developing ASR systems specifically tailored for healthcare environments, such as hospitals, clinics, and telemedicine. It aims to transcribe medical dictations, conversations between healthcare providers and patients, or interactions with electronic health records (EHRs). The term medical ASR does not refer to the "ASR of pathological speech", which focuses on developing models capable of recognizing and transcribing speech from individuals with speech impairments or disorders. These impairments can be due to conditions such as dysarthria, aphasia, stuttering, or neurological diseases such as Parkinson

gressively implemented to optimize this process, leveraging medical ASR technology (van Buchem et al., 2021; Zhang et al., 2023; Johnson et al., 2014; Saxena et al., 2018). Secondly, in the context of emergency medical services, medical ASR has been evaluated for its influence on stroke detection, demonstrating potential to enhance response times and diagnostic accuracy (Donnelly et al., 2022). Thirdly, ASR can be employed in surgical environments to enhance communication between the surgeon and both human assistants (e.g., surgical nurses) and digital systems (e.g., robotic arms) (Ruby et al., 2020; Schulte et al., 2020). Fourthly, in pediatric healthcare, medical ASR systems have been investigated for their potential application in remote care management (Nayar, 2017). Fifthly, medical ASR can be employed to support individuals with hearing impairments or disorders related to voice, speech, or language, facilitating more effective communication (Wendt et al., 2011).

B Ethical Statements

Speech data accompanied by high-quality human-annotated transcripts was obtained from YouTube in compliance with the Fair Use Policy and Vietnamese regulations governing data consent, privacy, and medical research, as detailed in this section.

According to Vietnamese law, which is applicable to the location of the hosting of the data and the site of all research activities, international and local researchers are authorized to collect and use the data exclusively for scientific purposes. To further ensure data privacy, segments of the dataset with the potential to reveal speaker identities were anonymized.

B.1 Fair Use

The research adhered rigorously to the principles of Fair Use as defined by the U.S. Copyright Office[2], which are also applicable to content on the YouTube platform. Fair Use is governed by Section 107 of the Copyright Act, which provides a legal framework for evaluating whether a specific use of copyrighted material qualifies under this doctrine. The statute identifies several examples of permissible uses, including criticism, commentary, news reporting, teaching, scholarship, and research, which are particularly relevant in the context of academic and scientific work.

Section 107 outlines a multifactorial approach to determining fair use, which requires an assessment of four key factors. These include:

- (1) **Purpose and character of the use, including whether the use is of a commercial nature or is for nonprofit educational purposes:** This factor evaluates how the copyrighted work is being utilized, particularly whether the use serves a commercial purpose or is directed toward nonprofit educational objectives. Courts tend to favor claims of fair use when the purpose is educational and nonprofit rather than commercial. Furthermore, the concept of "transformative use" plays a significant role in this determination. Transformative uses are characterized by their ability to add new meaning, insight, or purposes to the original work, altering its character in a way that differentiates it from the initial intention. Transformative uses that do not replace or compete with the original purpose of the work are more likely to qualify as fair use.

- (2) **Nature of the copyrighted work:** This factor examines the type of work involved and its relationship to the copyright's goal of fostering creative expression. Works that are highly imaginative or creative, such as novels, films, or songs, receive stronger copyright protection, making their use less likely to be considered fair. In contrast, factual or informational works, such as technical articles or news reports, are less stringently protected, and their use may more readily align with fair-use principles. Additionally, unpublished works are generally given greater protection and their unauthorized use is less likely to meet fair use criteria.

- (3) **Amount and substantiality of the portion used in relation to the copyrighted work as a whole:** This factor assesses both the quantitative and qualitative aspects of the material used in relation to the entire copyrighted work. The use of larger portions of a work typically weighs against fair use, though exceptions exist in cases where the entirety of the work is used for a justified purpose. Conversely, even the use of a small excerpt may be deemed unfair if it constitutes the "heart" or most significant and recognizable aspect of the original work. In this evaluation, the balance between the necessity of the portion used and its impact on the original work is critical.

- (4) **Effect of the use on the potential market for or value of the copyrighted work:** This factor considers the economic impact of the use without a license on both the existing market and the potential future markets for the copyrighted work. Courts analyze whether the unauthorized use undermines the market value or competes with the copyright holder's ability to monetize their work. If the unlicensed use causes substantial harm to the market or diminishes the value of the original work, it is less likely to qualify as fair use.

These factors collectively inform the determination of whether the usage is lawful under the doctrine of Fair Use, providing a nuanced and case-specific analysis.

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

In accordance with applicable legal frameworks, our work is justified under the provisions of the Fair Use doctrine. This assertion is supported by a detailed interpretation of the Fair Use principles [3] by copyrightalliance.org and the ELRC Report on legal issues in web crawling [4] by Pawel Kamocki, which emphasize the transformative nature of our research, its non-commercial scientific purpose and its minimal impact on the market value of the original content. These considerations collectively align with the statutory factors outlined in copyright law, underscoring the legitimacy of our approach. Our detailed interpretation of the Fair Use principles is as follows:

- (1) **Purpose and Character of Use**: The data were collected and utilized strictly for non-commercial and research purposes, aligning with the principles of Fair Use. Rather than directly using the videos obtained from YouTube, we transformed them into audio files at a predefined sampling rate. Long audio files, typically around an hour in duration, were segmented into shorter clips of 10 to 30 seconds. The segments were then randomly shuffled to ensure that they could not be reconstructed to form the original videos. This transformation and randomization process render the dataset distinctly different from the original content, thus qualifying as transformative use. Furthermore, this approach does not substitute for the original purpose or value of YouTube videos.

- (2) **Nature of the Copyrighted Work**: The extracted content primarily consists of factual, non-fictional medical conversations, which further supports its qualification as Fair Use. In addition, YouTube videos are publicly accessible throughout the world, fulfilling the criterion related to the publication status of the copyrighted material.

- (3) **Amount and Substantiality of the Portion Used**: Although there is no quantitative metric to precisely assess the fairness of a specific use, the randomly shuffled 10- to 30-second audio segments do not provide the full context or meaning of the original videos. These short segments are incapable of reproducing or capturing the core or "heart" of the copyrighted works.

- (4) **Effect on the Potential Market**: Our dataset does not serve as a competitor to the original content on YouTube. The 10- to 30-second audio segments do not detract from the YouTube viewership or impact the commercial interests of copyright owners. As a result, our work does not interfere with the potential market value of the original videos or undermine the business of copyright owners.

By adhering to these principles, we ensure compliance with Fair Use guidelines while maintaining the scientific and ethical integrity of our research. Numerous related works have been conducted that utilize the extraction of video content from YouTube for academic and noncommercial purposes. These studies typically involve systematic retrieval of publicly available videos, followed by their conversion to audio formats to facilitate various lines of research, such as ASR, NLP, and multimedia analysis. Such approaches often aim to leverage the diverse linguistic, cultural, and acoustic features inherent in the vast repository of YouTube content while adhering to ethical guidelines and copyright regulations to ensure the integrity and legality of the research, such as GigaSpeech[5] (China & USA), VoxCeleb[6] (UK), VoxLingua107[7] (UK).

B.2 Data Consent

Our waiver of data consent for the collection of medical ASR datasets is justified based on ethical and regulatory considerations, particularly when the data are deidentified and pose minimal risk to individuals. In compliance with institutional review board (IRB) guidelines and regulatory frameworks, such as the Common Rule, consent can be waived if it is impractical and research has significant potential to advance medical knowledge or improve healthcare outcomes. Anonymization techniques, including speaker de-identification, ensure that patient confidentiality is maintained, mitigating privacy concerns. Additionally, the dataset is used strictly for research purposes, with safeguards in place to prevent misuse or unauthorized access.

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

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

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

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

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

These measures collectively support the ethical and legal justification for waiving individual data consent while upholding privacy protections.

The publication of research data in this study adheres to relevant legal frameworks concerning data consent and privacy protection, both within Vietnam and internationally. A comprehensive explanation is provided below.

- **Global Data Protection and Privacy Compliance**: Of the 194 countries globally, 137 have adopted Data Protection and Privacy Legislation[8], as documented by the United Nations (UN). This includes key signatories such as the USA, EU member states (e.g., Germany), and Vietnam. In alignment with these international frameworks, Vietnam's Personal Data Protection Act stipulates in Article 6 that "The protection of personal data is carried out in accordance with international treaties to which the Socialist Republic of Vietnam is a member." This establishes that Vietnamese data protection laws comply with international standards, ensuring compatibility and lawful handling of personal data for global collaboration in research.

- **Exemption for Sensitive Data Processing for Research**: Article 20, Section 4 of Vietnam's Personal Data Protection Act explicitly states that "The party processing personal data is not required to register for processing sensitive personal data in the case of research purposes." This provision legally allows researchers to process sensitive data, including medical and speech-related datasets, without the explicit consent of individuals, provided the purpose is confined to scientific inquiry.

- **No Consent Requirement for Data Publication in Research**: Under Article 16 of Vietnam's Personal Data Protection Act, the principle of data deletion is waived for cases involving scientific research, statistics, or legal obligations. The law specifies that: "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." Thus, researchers are exempt from obtaining consent from data subjects for the inclusion of their data in publications, reaffirming the permissibility of this study's data handling practices.

- **Encouragement of Research Publication in Vietnam**: The Law on Medical Examination and Treatment, in conjunction with the Constitution of the Socialist Republic of Vietnam, underscores the importance of scientific dissemination. Article 22 mandates that medical practitioners and researchers "are responsible for updating relevant medical knowledge (...) including (...) c) Publish scientific research (...)." This legal encouragement promotes the proactive sharing of findings, particularly when involving sensitive medical data, as part of advancing public health and scientific understanding.

- **Legal Protections for Researchers**: Article 42 of the Law on Medical Examination and Treatment provides explicit protections for researchers. It states that researchers are "protected by the law and not responsible when a medical incident still occurs after complying with regulations." This ensures that any unforeseen outcomes related to the use or publication of research data, provided it aligns with statutory requirements, do not hold researchers liable.

- **Data Collection and Jurisdictional Compliance**: The dataset utilized in this study was collected using Vietnamese IP addresses and a web crawler authorized by a Vietnamese government-recognized company. This method adheres to Vietnam's Cybersecurity Law, as outlined in Article 26 of the Constitution of the Socialist Republic of Vietnam. It mandates that "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." Consequently, YouTube, as a service provider, must comply with Vietnamese regulations concerning data generated within the country's cyberspace.

- **International Researchers and Cross-Border Legal Alignment**: Articles 2 and 10

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

of the Vietnamese Civil Code on Civil Relations with Foreign Elements assert the application of Vietnamese law to international civil relations involving foreign researchers. Specifically, the Code emphasizes that "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." This ensures that international researchers working with Vietnamese data must adhere to Vietnamese laws, while simultaneously receiving legal protections and encouragement under these frameworks.

The dataset utilized in this study comprises YouTube content predominantly centered on medical themes, including televised medical shows, interviews, and educational lectures. In all cases, the participants in the videos spoke directly to the camera, demonstrating awareness that their content was intended for public dissemination. This awareness comes from the context of these videos, which were explicitly produced with the goal of providing accurate and accessible medical knowledge to YouTube audiences. Importantly, these videos were officially published by reputable national television channels, ensuring a professional standard of production and adherence to broadcasting regulations.

In contrast, YouTube videos created by amateur content creators, where the individuals featured may not have been aware of being recorded or of the eventual publication of the footage, were explicitly excluded from our dataset. This exclusion criterion was implemented to maintain ethical standards, particularly regarding informed consent and privacy. By limiting the dataset to professionally produced content with a clear intention of public dissemination, we aimed to ensure that the data collected adhered to legal and ethical guidelines on participant awareness and data use.

C Details of Data Creation

C.1 Details of Data Collection per Language

C.1.1 English

The data was collected from YouTube using the 2024 ICD-10-CM Codes to ensure diversity. We searched for diseases associated with the first 8 codes: A00-B99 (Infectious diseases), C00-D49 (Neoplasms), D50-D89 (Blood Diseases), etc. Due to time constraints, we searched only for these codes, applying filters for videos longer than 20 minutes and with subtitles to ensure accuracy. The videos were manually selected, prioritizing diverse speakers, accents, and contexts.

For the first 3 codes, we obtained 20 hours of video and subtitles and 10 hours for the rest. Videos and metadata, including recording conditions, speaker roles, genders, and accents, were saved.

C.1.2 French

We collected French medical videos from YouTube using terms such as “urgence”, “consultation médicale”, and “cancer”. The videos required Closed Captions (CC) with timing, either manually annotated or auto-generated by YouTube. We focused on videos over 20 minutes, covering topics such as oncology, cardiology, and pediatrics, from diverse contexts (lectures, interviews, consultations) and recording conditions (clean audio to noisy emergency rooms with multiple speakers) and prosodies (calm narration to distressed cries).

C.1.3 Chinese

We tried collecting Chinese videos using the same method as for French, but found very few Mandarin videos with CC from mainland China. Most Chinese subtitles were hardcoded, and available CC were in English (from Singapore) or Traditional Chinese (from Taiwan or Hong Kong). After attempting to upload Chinese videos to our channel for automatic CC generation, we found YouTube could not generate subtitles due to language complexity. As a result, we mainly used videos from Singapore and Taiwan, with fewer from mainland China.

C.1.4 German

The data was collected from YouTube using 2024 ICD-10 codes for diversity. We searched for videos related to diseases linked to these codes, but found limited human-labeled subtitles. To address this, we included German medical terms such as “Krankenhaus” and “Krankheit” in our searches. All selected videos had manually annotated captions with accurate timestamps. We prioritized videos longer than 20 minutes, then shorter ones, ensuring diversity in speakers by gender, accent, and context (lectures, discussions, interviews).

C.2 Details of Data Quality Control

The French transcript was triple-validated by a native French Literature lecturer and a C1-level linguist - a professional medical expert, ensuring transcription accuracy and alignment with the CC timing. The Chinese transcript was similarly validated by a native speaker and an HSK-5 level linguist professional medical expert. The English transcript was initial reviewed by a TESOL-certified linguist, followed by cross-checking by three C1-level speakers, one of whom is a professional biomedical expert. Due to labor constraints, the German transcript was double-validated by a single C1-level professional biomedical expert.

All of our annotators were instructed to adhere to the following quality control procedures:

1. Listen carefully to the audio recordings
2. Validate the human-annotated transcripts provided by professional YouTube channels by correcting minor inaccuracies or excluding transcripts deemed too erroneous
3. Identify the start and end points of individual utterances
4. Identify the speaker, recording conditions, accents, speaking roles (when applicable)

C.3 Data Processing

Transcription errors often arise from time-stamp mismatches when segmenting long-form audio into shorter segments. Annotators use long-form audio to improve efficiency and capture extended contexts, such as discussions or lectures. Due to GPU memory limitations, training is restricted to short-form audio to prevent out-of-memory (OOM) issues. As a result, annotators split long transcripts, causing time-stamp mismatches, typically within one second. This can lead to missing words at the start or end of recordings, highlighting the limitations of human-labeled datasets, where annotators struggle to capture words occurring faster than one second (Wargo, 1967). For a standard conversational spontaneous ASR English dataset such as

Switchboard (Godfrey et al., 1992), the Word Error Rate (WER) for human annotators ranges from 5% to 15% (Stolcke and Droppo, 2017). In contrast, for a more challenging real-world ASR dataset, the WER for human annotators without ASR support ranges from 17% to 31% (Mulholland et al., 2016).

In contrast, forced alignment can address this issue as machines can "listen" to words in 10ms-20ms intervals. However, forced alignment is limited by the quality of human-provided training data, making no transcript entirely accurate. To achieve "more perfect" transcripts, we employ a human-machine collaboration approach.

To maximize the data quality for the training model. We implemented a tailored data quality control pipeline designed to address specific challenges inherent to multi-audio sources. The transcription process is often manual and can be inaccurate. Dividing audio into very short segments (i.e., less than 5 seconds) frequently results in serious misalignment with the transcriptions, which harms the training process. By concatenating these short segments, we created longer and more coherent training samples. This mitigates the misalignment problem and provides the model with a richer understanding of the patterns and intonation of spoken language. The results of the analysis before and after concatenation are shown in Figure 1.

Additionally, extraneous noise text elements such as silence markers, filled pauses, and HTML tags, while present in raw transcripts, do not contribute to meaningful model learning. We removed these elements to focus the model's attention on relevant speech content. In particular, we chose to retain punctuation marks during the cleaning process. Punctuation plays a crucial role in conveying the nuances of spoken language, and its presence in training data encourages the model to generate transcripts that are not only accurate but also expressive and natural-sounding.

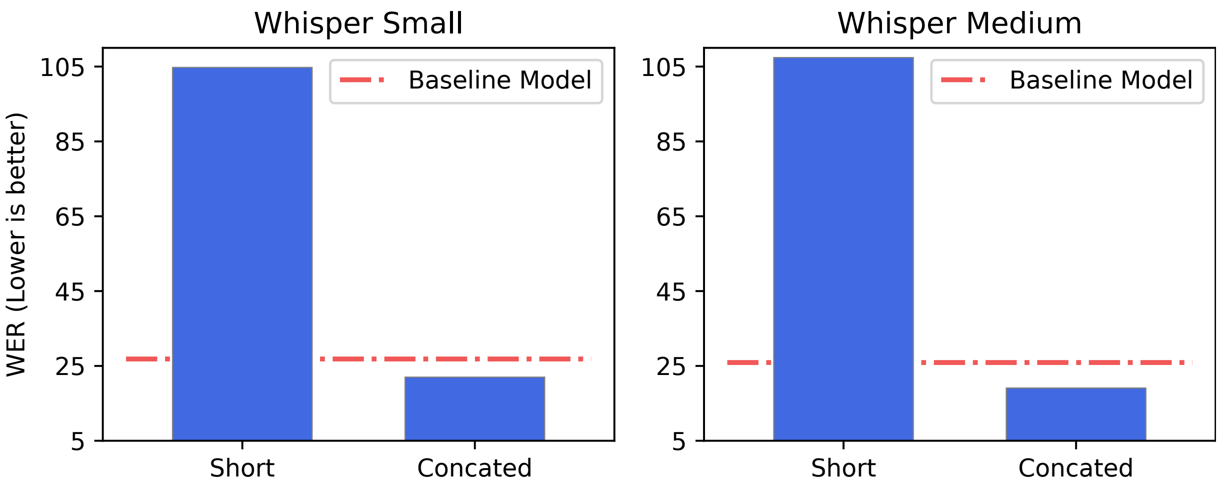

Figure 1: Illustrating the performance comparison of Whisper models trained on two distinct audio segmentation approaches for German language data: human-segmented short audio clips and concatenated continual audio segments of approximately maximum 15 seconds in length. We evaluate performance using both Whisper Small and Whisper Medium model sizes. The results demonstrate a notable improvement in model performance when trained on concatenated audio, highlighting the efficacy of this data preparation technique in enhancing transcription accuracy in the German language context

C.4 Full Data Statistics

MultiMed dataset consists of multilingual audio recordings in five languages: Vietnamese, English, French, German, and Mandarin Chinese. Each audio clip was segmented into short snippets with an average length of approximately 6 seconds for Vietnamese and around 12 seconds for the other languages. This segmentation facilitated efficient training and improved the model's responsiveness to shorter speech segments. The dataset was subsequently uploaded to the Hugging Face platform for further training and analysis. The statistics of our data samples are described in Table 2.

Table 9 shows the statistics of the data set compared to all existing publicly available medical ASR datasets, based on our best knowledge of the current literature. As shown in the table, our *MultiMed* dataset stands as the world's largest dataset in terms of total duration (150 hours of recordings), number of recording conditions (10), number of accents (16) and number of speaking roles (6).

Dataset	Venue	Dur.	Language	Nature	#Rec. Cond.	#Spk	#Acc	#Roles
MultiMed[1] (ours)	-	150h	Multiling.	Real-world	10	198	16	6
VietMed (Le-Duc, 2024)	LREC-COLING	16h	Vietnamese	Real-world	8	61	6	6
PriMock57[2] (Korfiatis et al., 2022)	ACL	9h	English	Simulated	1	64	4	2
Work by Fareez et al. (2022)[3]	Nature	55h	English	Simulated	1	N/A	1	2
AfriSpeech-200[4] (Olatunji et al., 2023)	TACL	≈123h	African English	Read speech	1	N/A	N/A	1
myMediCon[5] (Htun et al., 2024)	LREC-COLING	11h	Burmese	Read speech	1	12	5	2

Table 9: Dataset statistics in comparison with all existing works from left to right: Total duration in hours (h), language, nature of speech, number of recording conditions, number of speakers, number of accents, speaking roles.
[1]In our dataset, only the number of recording conditions, speakers, accents and speaking roles for Vietnamese and English are identified because of technical and privacy issues. Therefore, the exact number of speakers and accents must be much larger than the currently reported number. 10 recording conditions include: Documentary, Interview, Lecture, News, Podcast, Webinar, Speech, Talk, Vlog, Workshop. 10 English accents include: Main US, Southern US, UK, Australian, Indian, Mexican, European, Japanese, Uzbekistan, Russian. 6 Vietnamese accents include: North, South Central Coast, South East, South West, Central Highland, North Central Coast.
[2]Speech collected by simulated medical conversations between 2 speaking roles - clinicians and actors/actresses. 4 English accents include: British English, European, other English, and other non-English.
[3]Speech was recorded as patient-physician interviews (counted as 1 recording condition and 2 speaking roles) by West England speakers (counted as 1 accent)
[4]AfriSpeech-200 dataset is a mix of general-domain and medical-domain speech. To our best understanding of the paper, we estimate the total duration of medical-domain speech to be around 123 hours. Recordings were collected by crowd-sourced workers to read aloud the medical transcripts (also known as read speech), thus both the number of recording conditions and speaking roles are counted as 1.
[5]myMediCon dataset hired speakers to read aloud the translated medical transcripts from English corpus (thus known as read speech). 5 speakers' accents include: Native Burmese, Pa'O, Kachin, Dawei, and Mon. 2 speaking roles are patients and doctors.

D Attention Encoder Decoder (AED)

An ASR model aims to convert an audio signal into text by mapping an audio signal $x_1^T := x_1, x_2, ..., x_T$ of length T to the most likely word sequence w_1^N of length N. The word sequence probability is defined as:

$$p(w_1^N|x_1^T) = \prod_{n=1}^{N} p(w_n|w_1^{n-1}, x_1^T). \quad (4)$$

In the encoder-decoder architecture, given D as the dimension size of the feature, the input audio signal matrix could be described as $x_1^T \in \mathbb{R}^{T\times D_{input}}$. For the sake of simplicity, downsampling prior to or within the encoder, achieved by a fixed factor, such as striding in a Convolutional Neural Network (CNN) is omitted. Consequently, the encoder output sequence is as follows:

$$h_1^T = Encoder(x_1^T) \in \mathbb{R}^{T\times D_{encoder}}. \quad (5)$$

Using a stack of Transformer ($\mathcal{T}$) blocks (Vaswani et al., 2017), the encoder output sequence is described as function composition:

$$h_1^T = \mathcal{T}_0 \circ ... \circ \mathcal{T}_{N_{EncLayers}}(x_1^T). \quad (6)$$

In the decoder, the probability for every single word is described as:

$$\begin{aligned} p(w_n|w_1^{n-1}, x_1^T) &= p(w_n|w_1^{n-1}, h_1^T(x_1^T)) \\ &= p(w_n|w_1^{n-1}, h_1^T). \end{aligned} \quad (7)$$

Based on Eq. 4, the word sequence probability given the output of encoder is formulated as:

$$p(w_1^N|x_1^T) = \prod_{n=1}^{N} p(w_n|w_1^{n-1}, h_1^T). \quad (8)$$

Decoder hidden state is formulated as:

$$g_n = f(g_{n-1}, w_{n-1}, c_n) \in \mathbb{R}^{D_g}, \quad (9)$$

where f is neural network; D_g is hidden state dimension; and c_n is context vector, e.g. weighted sum of encoder outputs via attention mechanism.

The attention mechanism in the decoder is described by 3 components: context vector c_n, attention weights $\alpha_{n,t}$, and attention energy $e_{n,t}$:

$$\begin{aligned} c_n &= \sum_{t=1}^{T} \alpha_{n,t} h_t \in \mathbb{R}^{D_{encoder}}, \\ \alpha_{n,t} &= \frac{\exp(e_{n,t})}{\sum_{t'=1}^{T} \exp(e_{n,t'})} \\ &= Softmax_T(\exp(e_{n,t})) \in \mathbb{R}, \\ e_{n,t} &= Align(g_{n-1}, h_t) \in \mathbb{R} \\ &= W_2 \cdot \tanh(W_1 \cdot [g_{n-1}, h_t]), \end{aligned} \quad (10)$$

where n is decoder step; t is encoder frame; $\alpha \in \mathbb{R}^{T\times N}$ is attention weight matrix; $\alpha_n \in \mathbb{R}^T$ is normalized probability distribution over t; $Softmax_T$ is Softmax function over spatial dimension T, not feature dimension; $W_1 \in \mathbb{R}^{(D_g+D_{encoder})\times D_{key}}$; $W_2 \in \mathbb{R}^{D_{key}}$.

During decoding, the output probability distribution over vocabulary is described as:

$$\begin{aligned} &p(w_n = *|w_1^{n-1}, h_1^T) \\ &= Softmax(MLP(w_{n-1}, g_n, c_n)) \in \mathbb{R}^N, \end{aligned} \quad (11)$$

where MLP is Multi-layer Perceptron.

For training AED model, sequence-level cross-entropy loss is employed:

$$\begin{aligned} \mathscr{L}_{AED} &= -\sum_{(x_1^T, w_1^N)} \log p(w_1^N|x_1^T) \\ &= -\sum_{(x_1^T, w_1^N)} \sum_{n=1}^{N} \log p(w_n|w_1^{n-1}, x_1^T). \end{aligned} \quad (12)$$

In beam search process, the auxilary quantity for each unknown partial string (tree of partial hypotheses) w_1^n is described as:

$$\begin{aligned} Q(n; w_1^n) :&= \prod_{n'=1}^{n} p(w_{n'}|w_0^{n'-1}, x_1^T) \\ &= p(w_n|w_0^{n-1}, x_1^T) \cdot Q(n-1, w_1^{n-1}). \end{aligned} \quad (13)$$

After eliminating the less likely hypotheses in the beam search process, the word sequence probability is determined by the most optimal hypothesis:

$$p(w_1^N|x_1^T) = Q(N; w_1^N). \quad (14)$$

E Full Experimental Setups

E.1 Hyperparameter Tuning

The training process leveraged powerful A100 SXM4 GPUs. To ensure consistent results, we fixed the random seed at 42 throughout the training runs. For all models, we adopted a common training configuration with a batch size of 8, a learning rate of 0.0001, and 20 training epochs. We applied several data pre-processing techniques during training, including lowercasing text, removing punctuation, and normalizing the audio input. In addition, we trained each model on a language-specific subset of the dataset to optimize its performance for the targeted language.

The optimizer chosen for training was Adam (Kingma and Ba, 2014) with the standard betas configuration (0.9, 0.999) and an epsilon value of 1e-8. We employed a linear learning rate scheduler with a warmup period of 100 steps to gradually increase the learning rate during the initial training phase. No data augmentation such as SpecAugment (Park et al., 2019a) was applied.

E.2 Details of Evaluation Metrics

To assess the performance of the ASR models, we used two standard evaluation metrics: WER and CER. Lower WER and CER scores indicate better model performance in terms of accurately transcribing spoken audio.

WER focuses on the accuracy of recognized words. It calculates the percentage of errors made at the word level, including insertions, deletions, and substitutions compared to the ground truth reference transcript, as described in Equation 15.

$$WER = \frac{S+D+I}{N} = \frac{S+D+I}{S+D+C} \tag{15}$$

where S is the number of word substitutions, D is the number of word deletions, I is the number of word insertions, C is the number of correct words, and N is the number of words in the reference data $(N = S + D + C)$.

Generally speaking, S represents the count of replaced words, D denotes the count of omitted words present in the reference data but absent in the ASR hypothesis, and I indicates the count of inserted words present in the ASR hypothesis but absent in the reference data. The alignment process between the ASR hypothesis and the reference data proceeds sequentially from left to right.

WER measures the number of insertions, deletions, and substitutions made at the word level, while the CER focuses on errors at the character level, as illustrated in Equation 16.

$$CER = \frac{S_c+D_c+I_c}{N_c} = \frac{S_c+D_c+I_c}{S_c+D_c+C_c} \tag{16}$$

where S_c is the number of character substitutions, D_c is the number of character deletions, I_c is the number of character insertions, C_c is the number of correct characters, and N_c is the number of characters in the reference data $(N_c = S_c + D_c + C_c)$.

F Details of Hybrid ASR Experiments

F.1 Hybrid wav2vec 2.0

F.1.1 Hybrid ASR

An ASR model aims to convert an audio signal into text by mapping an audio signal x_1^T of length T to the most likely word sequence w_1^N of length N. The relation w^* between the acoustic and word sequence is:

$$w^* = \arg\max_{w_1^N} p(w_1^N|x_1^T) \tag{17}$$

Bayes theorem: By applying Bayes' Theorem, the probability $p(x)$ can be ignored during the maximization process, as it functions only as a normalization constant and does not influence the final result.

$$\begin{aligned} p(w_1^N|x_1^T) &= \frac{p(x_1^T|w_1^N)p(w_1^N)}{p(x_1^T)} \\ &\propto p(x_1^T|w_1^N)p(w_1^N) \end{aligned} \tag{18}$$

Therefore:

$$w^* = \arg\max_{w_1^N} \underbrace{p(x_1^T|w_1^N)}_{\text{acoustic model}} \cdot \underbrace{p(w_1^N)}_{\text{language model}} \tag{19}$$

Acoustic modeling: First, alignments between the acoustic observations x_1^T and labels w_1^N are obtained by using Gaussian-Mixture-Model/Hidden-Markov-Model (GMM/HMM) as labels for Deep-Neural-Network/Hidden-Markov-Model (DNN/HMM) training (DNN is wav2vec 2.0 encoder (Baevski et al., 2020b) in this case).

$$p(x_1^T|w_1^N) = \sum_{[s_1^T]} \prod_{t=1}^{T} p(x_t, s_t|s_{t-1}, w_1^N)$$
$$= \sum_{[s_1^T]} \prod_{t=1}^{T} \underbrace{p(s_t|s_{t-1}, w_1^N)}_{\text{transition prob.}} \cdot \underbrace{p(x_t|s_t, s_{t-1}, w_1^N)}_{\text{emission prob.}} \tag{20}$$

GMM/HMM: The labels used in the acoustic modeling are context-dependent phonemes (triphones), instead of BPE subword units like in AED. During the GMM/HMM process, a CART (Classification and Regression Tree) (Breiman, 2017) is used to link the states s. The GMM is a weighted sum over K normal distributions and is calculated as:

$$p(x_t|s_t, s_{t-1}, w_1^N) = \sum_{i=1}^{K} c_i \cdot \mathcal{N}(x_t|\mu_i, \sigma_i^2), \tag{21}$$

resulting in a multimodal emission probability with parameters μ_i, σ_i and mixture weights c_i for $i \in [\![1, K]\!]$. The mixture weights are non-negative and sum up to unity.

DNN/HMM: The posterior probability $p(a_{s_t}|x_1^T)$ could be discriminatively modeled using DNN (wav2vec 2.0 encoder), resulting in the DNN/HMM approach. The emission probability in the HMM could be calculated using the Bayes rule:

$$p(x_1^T|a_{s_t}) = \frac{p(a_{s_t}|x_1^T)p(x_1^T)}{p(a_{s_t})}. \tag{22}$$

The probability $p(a_{s_t})$ could be estimated as the relative frequency of a_{s_t}. For a simplified Bayes decision rule, the probability $p(x_1^T)$ is removed.

Decoding: During the ASR decoding process, the acoustic model and n-gram language model (Ney et al., 1994) should be combined based on the Bayes decision rule using Viterbi decoding algorithm (Forney, 1973) which recursively calculates the maximum path to a find best-path in the alignment graph of all possible predicted words to the acoustic observations:

$$w_1^N = \arg\max_{N, w_1^N} p\Big(\prod_{n=1}^{N} p(w_n|w_{n-m}^{n-1}) \cdot \max_{[s_1^T]} \prod_{t=1}^{T} p(x_t, s_t|s_{t-1}, w_1^N)\Big) \tag{23}$$

Afterwards, beam search (acoustic model and n-gram language model pruning) is employed to solely focus on the most promising predicted words at each time step t (Ortmanns et al., 1997).

F.1.2 Modified wav2vec 2.0

The model consists of a multi-layer convolutional neural network feature extractor CNN that receives T time-step raw audio waveform $x_1^T := x_1, x_2, ..., x_T$ (or x for simplification, $x \in R^{T\times1}$) as input and produces latent speech representations $x^{FE} \in R^{T\times1}$. These representations are then pushed into a stack of Transformer ($\mathcal{T}$) layers (Vaswani et al., 2017) which generates contextualized representations for Softmax SM classification.

In the scope of this work, we mathematically formulate our modified architecture for Hybrid wav2vec 2.0 as follows.

Wave normalization[9]: The raw audio waveform x is first normalized to the range between 0 and 1 by the wave normalization layer $WaveNorm$ before being pushed into the feature extractor, as shown in Equation 24.

$$x^{WaveNorm} = WaveNorm(x) = LayerNorm(x) \in R^{T\times1} \tag{24}$$

$WaveNorm$ could be either layer normalization $LayerNorm$ (Ba et al., 2016) or batch normalization $BatchNorm$ (Ioffe and Szegedy, 2015).

Feature extractor: The normalized raw audio waveform is pushed into a stack of CNN layers and a feed-forward (FFW) layer.

$$x^{FE} = FeatureExtractor(x) = FFN \circ CNNs \circ WaveNorm(x) \tag{25}$$

Time-downsampling in feature extractor[10]: When there is a sampling rate mismatch, the feature extractor of 16 kHz pre-trained models can

[9] Our modification of wav2vec 2.0 architecture for Hybrid ASR

[10] Our modification of wav2vec 2.0 architecture for sampling rate mismatch between pre-trained models and fine-tuned dataset

be modified to handle 8 kHz sampled data while still producing representations with the same 20 ms frame shift. By halving the stride of a convolutional layer in a stack of CNN layers in the feature extractor, we will receive features at the desired frame rate while reducing the downsampling factor from the waveform to the feature frames by a factor of 2.

$$\begin{aligned} x^{FE} &:= TimeDownsampling(x^{FE})| \\ x^{FE} &\in R^{\frac{1}{2}T_{FE} \times F_{FE}} \end{aligned} \tag{26}$$

In a generalized formulation shown in Equation 27, the time-downsampling could be done given a general time-downsampling factor TDF

$$\begin{aligned} x^{FE} &:= TimeDownsampling(x^{FE})| \\ x^{FE} &\in R^{\frac{1}{TDF}T_{FE} \times F_{FE}} \end{aligned} \tag{27}$$

Transformer as contextualized encoder: In an arbitrary l-th transformer layer, the output $x_l^{\mathcal{T}}$ is briefly defined as:

$$\begin{aligned} x_l^{\mathcal{T}} &= \mathcal{T}(x_{l-1}^{\mathcal{T}}) \\ &= FFW \circ MHA(x_{l-1}^{\mathcal{T}}) \end{aligned} \tag{28}$$

where MHA is multi-head attention which is a function defined by self-attention functions SA:

$$MHA(x_{l-1}^{\mathcal{T}}) = SA(x_{l-1}^{\mathcal{T}}) + x_{l-1}^{\mathcal{T}} \tag{29}$$

Then, we have a full equation for an arbitary l-th Transformer layer:

$$\begin{aligned} x_l^{\mathcal{T}} &= FFW(MHA(x_{l-1}^{\mathcal{T}})) + MHA(x_{l-1}^{\mathcal{T}}) \\ &= FFW(SA(x_{l-1}^{\mathcal{T}}) + x_{l-1}^{\mathcal{T}}) \\ &+ \left[SA(x_{l-1}^{\mathcal{T}}) + x_{l-1}^{\mathcal{T}}\right] \end{aligned} \tag{30}$$

For layer-wise formulation, the 0-th Transformer layer (the first layer) is connected to the feature extractor, which is defined as:

$$x_0^{\mathcal{T}} = \mathcal{T}(x^{FE}) \tag{31}$$

Given an L-Transformer-layer wav2vec 2.0 architecture, the $L-1$-th Transformer layer (the final layer) is defined as a chain function as:

$$\begin{aligned} x_{L-1}^{\mathcal{T}} &= \mathcal{T}(x_{L-2}^{\mathcal{T}}) \\ &= \mathcal{T} \circ \mathcal{T} \circ ... \circ \mathcal{T}(x_0^{\mathcal{T}}) \\ &= \mathcal{T} \circ \mathcal{T} \circ ... \circ \mathcal{T} \circ \mathcal{T}(x^{FE}) \end{aligned} \tag{32}$$

where L is the total number of Transformer layers in the encoder, layer indices start from 0 to $L-1$.

Time-reupsampling: For wav2vec 2.0 architecture, regardless of whether a sampling rate mismatch exists or not, it is necessary to re-upsample the final Transformer layer prior to its input into a Softmax layer for frame-wise classification. Failure to do so would lead to a discrepancy in the number of time frames during the calculation of the frame-wise loss objective function. Consequently, a FFW necessitates upsampling to ensure alignment with the rest of the architecture.

$$\begin{aligned} x_{Reup} &:= TimeReupsampling(x_{L-1}^{\mathcal{T}}) \\ &:= FFW(x_{L-1}^{\mathcal{T}})| \\ x_{Reup} &\in R^{T \times d} \end{aligned} \tag{33}$$

where d is the size of context-dependent states (CDS), or size of CART labels.

Hypothesis (output): Finally, x_{Reup} goes to a Softmax layer SM to produce a matrix of hypotheses $z \in R^{T \times d}$.

$$z := SM(x_{Reup})| z \in R^{T \times d} \tag{34}$$

Loss function: The hypothesis matrix z is compared with the ground truth y to calculate the frame-wise cross-entropy (CE) loss matrix $\mathscr{L}(z, y) \in R^{T \times d}$. The total loss value is the sum of all the elements in the loss matrix $\mathscr{L}(z, y)$.

$$\begin{aligned} \mathscr{L}(z, y) &:= \mathscr{L}_f(z, y) = \|\mathscr{L}(z, y)\|^1 \\ &:= -[y \cdot \log(z)], \quad f = CE \\ &> 0 \quad \forall \log \in \{\log_2, \log_n, \log_{10}\} \end{aligned} \tag{35}$$

F.2 Experimental Setups

For n-gram language modelling and the initialization of GMM-HMM, we used the same configurations and hyperparameters as in (Lüscher et al., 2023). We employed the BABEL project's seed lexicon and augmented it with additional Vietnamese text data. Using the toolkit Sequitur Grapheme-To-Phoneme[11] (Bisani and Ney, 2008) - the conversion tool on pronunciation lexicon, the seed lexicon from BABEL was extended, creating the augmented lexicon for training. The statistics for the n-gram language model and the augmented lexicon are shown in Table 10.

The labels for the acoustic model were generalized triphone states obtained by CART with 4501 labels. During GMM-HMM process, we

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

Trained lexicon		Language model		*dev*		*test*	
#words	#vocab	#words	Size (in MBs)	OOV	PPL	OOV	PPL
17,000	5295	8.5M	98	0.76%	66	0.66%	84

Table 10: Statistics of 4-gram language model and augmented lexicon for hybrid ASR training, including for both GMM-HMM and wav2vec 2.0 training. OOVs and Perplexities (PPLs) are reported on our Vietnamese dev and test set.

found that WERs on the Vietnamese test sets of Speaker Adaptive Training (SAT) was quite comparable to Speaker Adaptive Training + Vocal Tract Length Normalization (SAT+VTLN). So, we fed SAT alignments into wav2vec 2.0 as input for the Hybrid ASR training.

For self-supervised wav2vec 2.0 training (Baevski et al., 2020a) and fine-tuning, we used the same vanilla setups and hyperparameters in (Le-Duc, 2023). All models had 123M parameters including 7 CNN layers and 8 Transformer layers, as shown in Table 7 in the main paper. The last CNN layer had a stride halved for adaptation to the 8kHz data. The pre-training epoch that led to the best WERs on dev was used to fine-tune with framewise CE loss. The SpecAugment (Park et al., 2019b) was employed during 33 fine-tuning epochs.

We employed RETURNN framework (Zeyer et al., 2018a) for supervised training (fine-tuning the wav2vec 2.0 models) and Fairseq (Ott et al., 2019) for self-supervised wav2vec 2.0 training on the unlabeled data. ASR decoding was performed using the RASR toolkit (Rybach et al., 2011). The pre-trained wav2vec 2.0 models from Fairseq (in Pytorch) were converted to RETURNN models (in Tensorflow) with our PyTorch-to-RETURNN toolkit[12].

F.3 Extra Experimental Results

Table 11 shows the breakdown per speaker in the Vietnamese test set of the Hybrid ASR results in Table 7. Two pre-trained wav2vec 2.0 models were used for fine-tuning on the Vietnamese set: XLSR-53-Viet and w2v2-Viet, leading to WERs on test set 28.8%, 29.0% respectively.

[12]https://github.com/rwth-i6/pytorch-to-returnn

Speaker ID	# Snt	# Wrd	Corr	Sub	Del	Ins	Err	S.Err
XLSR-53-Viet								
vietmed_002	363	7631	58.5	30.9	10.6	6.6	48.1	100.0
vietmed_004	446	10575	68.3	18.5	13.2	4.9	36.6	100.0
vietmed_014_a	18	491	88.6	3.1	8.4	5.9	17.3	100.0
vietmed_014_b	164	4034	77.2	11.8	11.1	3.7	26.5	100.0
vietmed_015_a	73	1779	86.1	5.5	8.4	3.9	17.8	98.6
vietmed_015_b	297	5669	83.3	6.9	9.8	4.2	20.9	96.6
vietmed_015_c	55	1010	69.4	14.4	16.2	5.5	36.1	100.0
vietmed_017_a	47	1104	78.3	12.0	9.7	4.6	26.4	100.0
vietmed_017_b	86	2061	81.5	9.8	8.6	5.0	23.5	100.0
vietmed_018_a	63	1527	76.0	11.9	12.2	19.4	43.5	100.0
vietmed_018_b	192	5293	76.7	10.8	12.5	6.9	30.2	100.0
vietmed_018_c	118	2761	76.5	10.9	12.5	8.2	31.7	100.0
vietmed_018_d	20	412	55.1	19.7	25.2	5.6	50.5	100.0
vietmed_018_e	5	76	56.6	19.7	23.7	7.9	51.3	100.0
vietmed_018_f	25	639	64.8	20.7	14.6	6.9	42.1	100.0
vietmed_019_a	58	1490	77.7	10.3	12.0	6.8	29.1	100.0
vietmed_019_b	116	2776	77.5	11.1	11.4	6.6	29.1	100.0
vietmed_023	390	7414	85.5	9.1	5.3	4.6	19.1	97.7
vietmed_024	376	7425	86.6	7.0	6.4	5.4	18.9	98.7
vietmed_025_a	101	2280	80.8	10.1	9.1	5.0	24.2	100.0
vietmed_025_b	91	1838	82.5	9.2	8.3	5.3	22.8	98.9
vietmed_026	21	355	55.8	29.9	14.4	7.3	51.5	100.0
vietmed_027_a	29	710	85.5	6.5	8.0	5.2	19.7	100.0
vietmed_027_b	64	1454	76.3	14.6	9.1	6.2	29.8	98.4
vietmed_028_a	106	2617	83.7	8.7	7.6	4.6	20.9	99.1
vietmed_028_b	21	475	77.7	11.8	10.5	5.9	28.2	95.2
vietmed_029	92	2240	83.8	7.9	8.3	5.3	21.6	100.0
Sum/Avg	3437	76136	76.9	13.0	10.1	5.7	**28.8**	99.2
Mean	127.3	2819.9	75.9	12.7	11.4	6.2	30.3	99.4
S.D.	129.6	2743.3	10.0	6.7	4.6	2.9	11.0	1.2
Median	86.0	1838.0	77.7	10.9	10.5	5.5	28.2	100.0
w2v2-Viet								
vietmed_002	363	7631	56.6	31.0	12.4	6.1	49.5	100.0
vietmed_004	446	10575	65.5	20.6	13.9	4.5	39.0	99.6
vietmed_014_a	18	491	89.0	2.9	8.1	6.1	17.1	100.0
vietmed_014_b	164	4034	77.6	12.7	9.7	4.9	27.3	100.0
vietmed_015_a	73	1779	87.5	5.0	7.5	3.7	16.1	98.6
vietmed_015_b	297	5669	83.3	6.3	10.4	3.7	20.3	96.6
vietmed_015_c	55	1010	68.6	13.8	17.6	4.6	35.9	100.0
vietmed_017_a	47	1104	78.4	12.0	9.5	4.7	26.3	100.0
vietmed_017_b	86	2061	80.4	10.8	8.8	4.8	24.4	100.0
vietmed_018_a	63	1527	75.6	12.7	11.7	19.6	44.0	100.0
vietmed_018_b	192	5293	77.3	10.0	12.7	6.7	29.3	100.0
vietmed_018_c	118	2761	75.4	12.4	12.2	7.4	32.0	100.0
vietmed_018_d	20	412	51.7	20.1	28.2	5.1	53.4	100.0
vietmed_018_e	5	76	48.7	27.6	23.7	5.3	56.6	100.0
vietmed_018_f	25	639	64.6	20.5	14.9	6.9	42.3	100.0
vietmed_019_a	58	1490	77.4	11.2	11.3	7.0	29.6	100.0

vietmed_019_b	116	2776	78.2	10.5	11.3	6.6	28.4	100.0
vietmed_023	390	7414	86.8	7.7	5.5	4.4	17.6	96.7
vietmed_024	376	7425	86.9	6.3	6.7	4.9	18.0	97.6
vietmed_025_a	101	2280	82.3	9.3	8.4	5.1	22.9	98.0
vietmed_025_b	91	1838	83.2	9.0	7.7	6.4	23.1	98.9
vietmed_026	21	355	56.3	27.3	16.3	7.6	51.3	100.0
vietmed_027_a	29	710	86.1	6.6	7.3	5.8	19.7	100.0
vietmed_027_b	64	1454	75.8	14.9	9.4	6.4	30.6	100.0
vietmed_028_a	106	2617	83.5	8.7	7.9	4.6	21.1	100.0
vietmed_028_b	21	475	76.4	14.5	9.1	6.5	30.1	95.2
vietmed_029	92	2240	84.6	7.7	7.7	5.7	21.1	100.0
Sum/Avg	3437	76136	76.5	13.1	10.3	5.5	**29.0**	98.9
Mean	127.3	2819.9	75.5	13.0	11.5	6.1	30.6	99.3
S.D.	129.6	2743.3	11.3	7.2	5.1	2.9	11.9	1.3
Median	86.0	1838.0	77.6	11.2	9.7	5.7	28.4	100.0

Table 11: Breakdown per speaker on the Vietnamese test set of the Hybrid ASR results in Table 7. Two pre-trained wav2vec 2.0 models were used for fine-tuning on the Vietnamese set: XLSR-53-Viet and w2v2-Viet, leading to WERs on test set 28.8%, 29.0% respectively.
Column from left to right is: Speaker ID, Number of sentences, Number of words, Corrections, Substitution Errors, Deletion Errors, Insertion Errors, Word-Error-Rate, Sentence-Error-Rate.

G Full Error Analysis

Figure 12 shows an example of common ASR errors from the ASR output compared to the corresponding ground truth transcript. Three ASR errors considered are substitutions, deletions, and insertions.

Below is the full error analysis based on the linguistic perspective for all 5 languages: English, Vietnamese, Chinese, French, and German.

G.1 English

Our error analysis of the ASR system revealed several phonological issues that affected the performance of the model. One significant issue involves the minimal phonological distance between certain vowel sounds, particularly in minimal pairs such as "long" vs. "lung" and "pen" vs. "pan". Due to the close proximity of these sounds in the phonetic space, the model often confuses them, leading to clinically significant errors, such as transcribing "lung cancer" as "long cancer".

Another source of error is related to the use of weak forms in speech, where certain words are pronounced in a reduced or less distinct manner. This results in frequent misrecognitions, such as interpreting "our" as "are", "and" as "in", "for" as "very", and even more complex substitutions like "earlierologist" for the phrase "earlier I was just". Additionally, numerical errors are common; for instance, the model may interpret "4 to 5" as "45", which could lead to critical inaccuracies in medical records. This type of substitution also extends to domain-specific terminology, such as transcribing "system that" as "systemic".

In addition, discrepancies were identified at the beginning and end of the transcriptions. This issue is largely attributed to inconsistencies between the training and testing conditions: the dataset was annotated using long-form audio segments, yet the model was trained and evaluated with short-form audio inputs. This mismatch creates boundary errors and negatively affects the model's ability to capture context, leading to truncation or overlap in predictions. In particular, this problem is not limited to English but has also been observed in other languages, indicating a systematic problem in handling different input formats during ASR processing.

G.2 Vietnamese

In the Vietnamese test set, a detailed analysis of ASR errors reveals that several phonological characteristics of the Vietnamese language pose significant challenges for model performance. Vietnamese is a tonal language with a complex phonetic system that includes a variety of vowels, consonants, and six distinct tones, all of which carry meaning and are integral to word differentiation (Horn and Pham, 2004). As a result, the ASR system often struggles with minimal phonetic contrasts, particularly when dealing with similar-sounding phonemes and tones.

Vowel confusion: Vietnamese vowels exhibit subtle distinctions, especially in terms of vowel height and backness. Pairs such as "cái" vs. "cứ" demonstrate this challenge. The model frequently confuses these due to their similar articulatory features and acoustic proximity. For instance, "cái" (meaning "thing" or "classifier for objects") and "cứ" (meaning "to keep doing something") differ primarily in vowel quality, but the ASR system often fails to capture this distinction, leading to misrecognition.

Consonant ambiguity: Consonant sounds in Vietnamese can also present difficulties, particularly when the phonemes are produced with similar places of articulation. An example is "nó" (he/she/it) vs. "nói" (to speak), where the confusion arises due to the similarity in nasal sounds and the rapid articulation of connected speech. Similarly, the pair "bác" (uncle/aunt) and "mắc" (to catch/to be caught) are often misrecognized due to the shared stop consonant sounds, complicated further by the presence of nasal or plosive release.

Tonal ambiguity: Vietnamese tones are particularly problematic for ASR systems, as they are both lexically and syntactically significant. The six tones in Vietnamese include level, rising, falling, broken, creaky, and low tones, which can completely change the meaning of a word. For instance, the pair "năng ngọng" (meaning "slurred speech") and "nặng nhọc" (meaning "laborious") illustrates how the model struggles to distinguish between tones, leading to semantically incorrect transcriptions. The difference between these phrases lies in tone distinctions, which are subtle and can be easily confounded by background noise or speaker variability.

Gender and regional variations: Furthermore, phonological variability due to gender differences (for example, male vs. female voice pitch) and regional dialects (Northern, Central, and Southern accents) further complicates the ability of the ASR system to correctly distinguish words of similar sound. For example, "nử" (variant pronunciation for some speakers, typically Northern) and "nữ" (female) differ mainly in tone and vowel length, which may be pronounced differently across dialects, increasing the error rate.

These types of phonological errors highlight the need for enhanced acoustic modeling that can account for the intricate vowel and consonant distinctions and the tonal nature of Vietnamese, especially in the medical domain. It also underscores the importance of incorporating a diverse set of training data that reflect different regional accents and speech patterns for patients and doctors to improve the robustness of the medical ASR system in Vietnamese language contexts.

G.3 Chinese

A primary source of errors arises from minimal pairs that differ solely in tonal pronunciation or involve homophones, both of which are highly prevalent in Mandarin Chinese. Given that Mandarin is a tonal language with four distinct tones (plus a neutral tone), words that share similar phonetic sounds but differ in tone can easily be confused by ASR systems (Jongman et al., 2006). This tonal ambiguity leads to significant transcription errors, especially in medical contexts where precision is crucial.

For instance in our test set, words like 麻闭 (mábì) and 麻痹 (má bì), or 跟本 (gēnběn) and 根本 (gēn běn), demonstrate how tonal distinctions are critical for differentiating between distinct meanings. Similarly, homophones such as 以 (yǐ) and 已 (yǐ), or 是 (shì) and 适 (shì), present further challenges, as the ASR system struggles to disambiguate words with identical phonetic pronunciation but different meanings. The error is compounded by the context-dependent nature of these terms, which requires a sophisticated understanding of the surrounding text to accurately differentiate them.

Additionally, errors are frequently caused by words that sound alike but differ in their meaning, as seen in examples like 代 (dài) vs 待 (dài) or 其二 (qí èr) vs 妻儿 (qī ér). In the medical domain, such mistakes can lead to severe clinical misinterpretations, affecting patient safety. For example, confusion between 没 (méi) (not) and 霉 (méi) (mold) could result in significant differences in the interpretation of a patient's condition or diagnosis.

Another frequent source of error is phonetic approximation in the sound space, where slight variations in pronunciation result in incorrect word predictions. Examples include 到路 (dào lù) vs 倒漏 (dào lòu) and 一确的 (yī què de) vs 一切都 (yīqiè dōu). These phonetic approximations arise due to the ASR system's inability to distinguish subtle differences, particularly in connected speech where articulation may be less clear. Such approximations can be particularly problematic in medical transcription, where terms like 答案 (dá'àn) (answer) being mistaken for 大碍 (dà ài) (serious problem) could alter the intended meaning of a clinical statement.

G.4 French

The errors encountered in the medical domain's ASR systems can be attributed to various phonological challenges, especially in datasets with languages like French, where the close proximity of certain phonemes in the acoustic space leads to frequent misinterpretations. These challenges typically arise from the inherent acoustic similarity between phonemes or word pairs that sound alike but have different meanings or spelling, often referred to as homophones or near-homophones For instance, in the French language, there are numerous vowel and consonant pairs that share similar acoustic characteristics but differ in meaning, making them susceptible to confusion. Some notable examples include:

- "attention" vs "ah tiens": Both phrases have similar phonetic structures, but the former is a common French word meaning "careful" or "attention", while the latter is a colloquial expression that might refer to a surprise or exclamation. A misinterpretation of these terms could lead to clinical miscommunication in situations requiring urgency or specific instructions.

- "engardré" vs "encadré": These words differ by a single vowel sound, but the first ("engardré") is a non-standard form or a potential misheard word, while the latter ("encadré") means "framed" in French. Such phonetic ambiguity can easily result in incorrect transcription, especially when the ASR model is unable to distinguish between similar-sounding terms within the context of a medical discussion.

- "à mettre" vs "est maître": The phrase "à mettre" (meaning "to put") is often misheard as "est maître" (meaning "is the master"), as both phrases have a similar rhythm and vowel-consonant structure. In medical settings, such confusion could mislead the interpretation of a patient's condition or instructions for care.

- "bonchique" vs "bronchite": A typical error arises when the ASR system confuses "bronchite" (bronchitis) with a distorted form like "bonchique". This could be catastrophic in medical contexts, as bronchitis refers to a serious respiratory condition, and an error here could delay proper diagnosis or treatment.

- "choléraux" vs "cholestérol": The acoustic similarity between "choléraux" (a non-standard or incorrect form) and "cholestérol" (cholesterol) presents another challenge. Cholesterol is a critical term in medical diagnostics, and errors in its transcription could result in the omission of vital health information, leading to inaccurate clinical assessments or interventions.

- "mé" vs "mais": The confusion between "mé" (which can be a shorthand or mispronunciation of "mais" meaning "but") is another example. Such errors are especially significant in medical contexts where subtle linguistic distinctions, even in less formal speech, can alter the meaning of a diagnosis or treatment plan.

G.5 German

In the context of ASR error analysis within the medical domain, our German test set presents distinct challenges that stem from both phonological and orthographic factors, which significantly affect the model's accuracy and performance.

Firstly, the issue of phonological proximity is particularly noticeable in minimal pairs—pairs of words that differ only in one sound. In the German language, small phonological differences between vowels in minimal pairs can cause considerable confusion for ASR models, as these systems often struggle to accurately distinguish between such similar-sounding words. For instance, the words "verschmerzen" (to suffer pain) and "vor Schmerzen" (before pain) have a very slight phonetic distinction, yet they represent entirely different meanings, potentially leading to misinterpretation by the ASR system. Similarly, words like "anestätiger" (anesthetist) and "Lokalanästhetikasalbe" (local anesthetic cream) contain subtle phonetic differences that can cause errors in transcription, especially when such words are transcribed without appropriate context or clarity.

Secondly, the orthographic characteristics of the German language further complicate ASR performance. German has a system of capitalization where nouns and imperative verbs are capitalized, while adjectives, adverbs, verbs, and other parts of speech are written in lowercase. This capitalization rule is not just a grammatical convention, but a semantic one, as it helps distinguish between different parts of speech and the meaning of the sentence. ASR models that fail to accurately capture these distinctions often produce errors that are both semantically and syntactically problematic. For example, "venenzugang" (venous access) vs "Venenzugang" (with proper capitalization) may lead to a loss of meaning or context in the transcribed text. Similarly, confusion between "komme" (come) and "Komme" (I come, in the imperative) can alter the intended message, especially in medical contexts where the clarity of instructions is critical.

Example		
English	ASR output	**sea** you don't really see any **affect** the brown **apocalyse** tissue activity, but at the high **BMW**, now, you will start to see a **uh uhm** protective effect where those individuals had lower **glyceryl**.
	Ground truth	**only** see you don't really see any effect **of** the brown adipose tissue activity, but at the high BMI, now, you will start to see a protective effect where those individuals had lower glycemia.
Chinese	ASR output	们新安装的那更新门是在这里， 然后我们看一个下有没有倒漏的问题，有没有狭窄的那个情况。
	Ground truth	我们新安装的那个心门是在这里， 然后我们看一下有没有倒漏的问题，有没有狭窄的那个情况。
French	ASR output	arrivez à **à** sortir un peu ou pas du tout 36 **tempérament** c'est bien vous savez vous avez un mix entre la **broncoid** l'insuffisance cardiaque et tout ce qui.
	Ground truth	arrivez à sortir un peu ou pas du tout 36 **la** température c'est bien vous savez vous avez un mix entre la bronchite l'insuffisance cardiaque et tout ce qui
German	ASR output	Haben Sie Allergiepass oder einen **Reisepass**? Dann könnte ich da mal nachschauen, ob **mal** ein spezielles **Antibiotikern** eingetragen worden ist. ich habe beides, da **ja** steht alles drin. Die bringt mein
	Ground truth	Haben Sie **einen** Allergiepass oder einen Patientenpass? Dann könnte ich da mal nachschauen, ob ein spezielles Antibiotikum eingetragen worden ist. **Ja**, ich habe beides, da steht alles drin. Die bringt mein
Vietnamese	ASR output	bản thân và **ừ** rộng hơn là **là** vì sức khỏe cộng đồng thưa quý **dị** tại việt nam nguyên tắc huyết khối **tiễn** mạch bệnh **mặt** máu
	Ground truth	bản thân và rộng hơn là vì sức khỏe cộng đồng thưa quý vị tại việt nam nguyên tắc huyết khối tĩnh mạch **là** bệnh mạch máu

Table 12: An example of ASR errors from ASR output (top) compared to the corresponding ground truth transcript (bottom). Errors are annotated as: substitutions in **red**, deletions in **blue**, and insertions in **green**.