UNIVERSITY OF TORONTO

ACL 2025 VIENNA

MultiMed: Multilingual Medical Speech Recognition via Attention Encoder Decoder

Khai Le-Duc, Phuc Phan, Tan-Hanh Pham, Bach Phan Tat, Minh-Huong Ngo, Chris Ngo, Thanh Nguyen-Tang, Truong-Son Hy

duckhai.le@mail.utoronto.ca

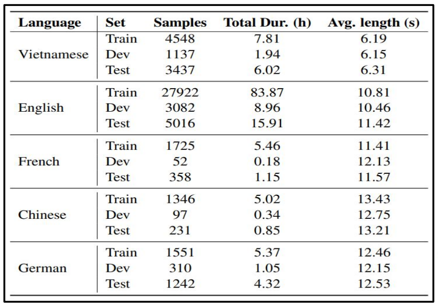
https://github.com/leduckhai/MultiMed/tree/master/MultiMed

Motivation

1. Multilingual medical ASR enables cross-lingual communication for healthcare applications, but remains unexplored.
2. Attention Encoder Decoder (AED) is easier to train and deploy than Hybrid ASR.

Contributions

1. MultiMed - the first multilingual medical ASR dataset, supporting 5 lang.: Vietnamese, English, German, Chinese, and French
2. The first publicly available multilingual medical ASR models, spanning small to large end-to-end configs
3. The first multilingual. study for medical ASR: monoling. - multiling. analysis, AED vs Hybrid and linguistic analysis
4. A practical ASR end-to-end training schemes optimized for a fixed number of trainable params in industry settings



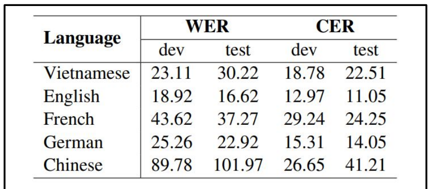
Language	Set	Samples	Total Dur. (h)	Avg. length (s)
Vietnamese	Train	4548	7.81	6.19
	Dev	1137	1.94	6.15
	Test	3437	6.02	6.31
English	Train	27922	83.87	10.81
	Dev	3082	8.96	10.46
	Test	5016	15.91	11.42
French	Train	1725	5.46	11.41
	Dev	52	0.18	12.13
	Test	358	1.15	11.57
Chinese	Train	1346	5.02	13.43
	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

Experimental Setups

- 4 Whisper models: Tiny, Base, Small, and Medium
- Decoder-only fine-tuning (encoder freezing) and Fully encoder-decoder fine-tuning

Key findings

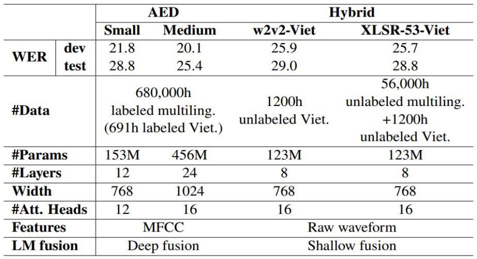
1. Multiling. fine-tuning improves accuracy over monoling., despite potential limitations from dispersed cross-lingual latent speech clusters.



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

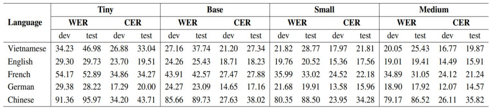
Table 6: Main baselines - WERs and CERs of **fully encoder-decoder fine-tuning** using Small Whisper model on all languages (**multilingual fine-tuning**)

2. Hybrid ASR is more data- and computation-efficient than AED ASR.



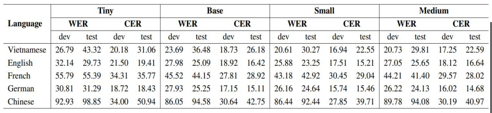
		AED		Hybrid	
		Small	Medium	w2v2-Viet	XLSR-53-Viet
WER	dev	21.8	20.1	25.9	25.7
	test	28.8	25.4	29.0	28.8
#Data		680,000h labeled multiling. (691h labeled Viet.)		1200h unlabeled Viet.	56,000h unlabeled multiling. +1200h unlabeled Viet.
#Params		153M	456M	123M	123M
#Layers		12	24	8	8
Width		768	1024	768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

3. On a fixed budget, freezing the entire encoder ensures both high accuracy and computational efficiency.



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

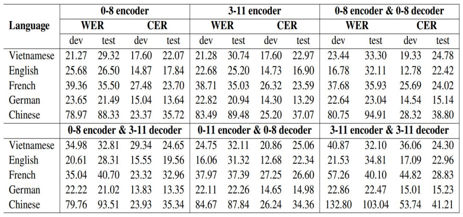
Table 4: Main baselines - WERs and CERs of **decoder-only fine-tuning** (freezing the entire encoder) using different Whisper models on each separate language (**monolingual fine-tuning**)



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**)

4. Maintaining the consistent freezing of a contiguous group of layers ensures high accuracy.



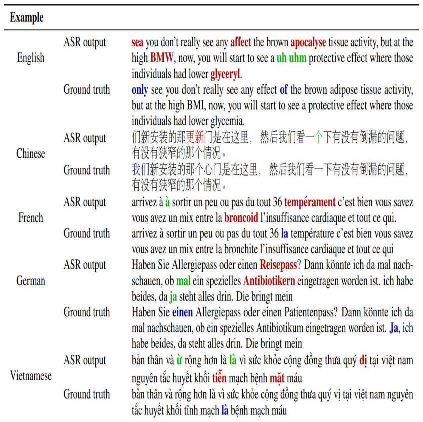
Language	0-8 encoder				3-11 encoder				0-8 encoder & 0-8 decoder			
	WER		CER		WER		CER		WER		CER	
	dev	test	dev	test	dev	test	dev	test	dev	test	dev	test
Vietnamese	21.27	29.32	17.60	22.07	21.28	30.74	17.60	22.97	23.44	33.30	19.33	24.78
English	25.68	26.50	14.87	17.84	22.68	25.20	14.73	16.90	16.78	32.11	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.

Linguistic Analysis

5. Medical ASR errors commonly include misrecognized clinical terms, hallucinations, omissions, and duplications.

6. Errors often arise from vowel proximity in Vietnamese, English, German, and French, and from tonal minimal pairs and homophones in



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

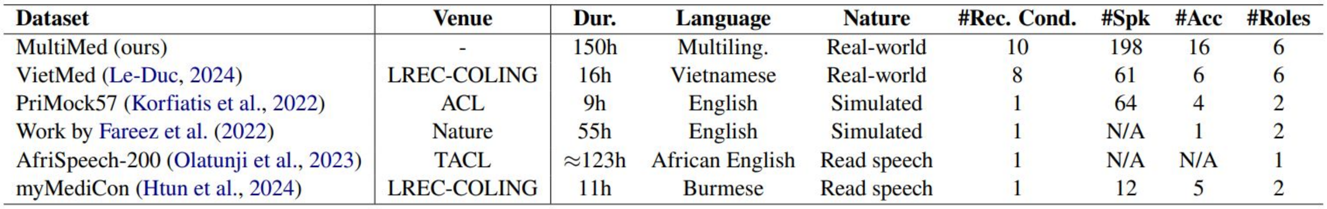
Table 12: An example of ASR errors from ASR output (top) compared to the corresponding ground truth transcript (bottom). Errors are annotated as: substitutions in **red**, deletions in **blue**, and insertions in **green**.

Limitations

Clinical impact: Our study aims to establish baselines for medical ASR, emphasizing the need for clinical pilot testing due to the high stakes of transcription accuracy..

Data

MultiMed is the world's largest medical ASR dataset across all major benchmarks, to the best of our knowledge: total duration, number of recording conditions, number of accents, and number of speaking roles.



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