Interspeech 2024
1-5 September 2024, Kos, Greece

Performant ASR Models for Medical Entities in Accented Speech

Tejumade Afonja[†1,6,*], *Tobi Olatunji*[†2,*], *Sewade Ogun*[3,*],
Naome A. Etori[4,*], *Abraham Owodunni*[2,*], *Moshood Yekini*[5,*]

[1]CISPA Helmholtz Center for Information Security [2]Intron Health [3]Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France [4]University of Minnesota - Twin Cities, USA [5]African Masters of Machine Intelligence, AIMS/AMMI
[6]AI Saturdays Lagos [*]Masakhane NLP

tejumade.afonja@cispa.de, tobi@intron.io

Abstract

Recent strides in automatic speech recognition (ASR) have accelerated their application in the medical domain where their performance on accented medical named entities (NE) such as drug names, diagnoses, and lab results, is largely unknown. We rigorously evaluate multiple ASR models on a clinical English dataset of 93 African accents. Our analysis reveals that despite some models achieving low overall word error rates (WER), errors in clinical entities are higher, potentially posing substantial risks to patient safety. To empirically demonstrate this, we extract clinical entities from transcripts, develop a novel algorithm to align ASR predictions with these entities, and compute medical NE Recall, medical WER, and character error rate. Our results show that fine-tuning on accented clinical speech improves medical WER by a wide margin (25-34 % relative), improving their practical applicability in healthcare environments.
Index Terms: speech recognition, medical documentation, medical named-entity recognition, African-accented speech

1. Introduction

In recent years, significant advances have been made in accented speech recognition with state-of-the-art (SOTA) automatic speech recognition (ASR) models proficiently transcribing diverse linguistic interactions [1, 2, 3]. However, the effectiveness of these models in clinical or medical settings,[1] where nuanced communication is paramount, remains a challenge [4]. This becomes particularly evident when clinicians with non-western accents document critical medical information using ASR technology. While these SOTA models achieve low word error rates (WER) on general speech, they commonly struggle with accurately transcribing clinical named entities (NE), e.g., see Table 1. The domain-specific nature of clinical documentation introduces a vulnerability that could have severe consequences for patient well-being – minor inaccuracies in essential elements like drug names, diagnoses, lab results, and lesion measurements (for example, writing *renal* instead of *adrenal*, or *hyper-* instead of *hypo-*) could potentially risk patient safety and expose clinicians to avoidable litigation [5]. To empirically expose this problem, we analyze the performance of several SOTA open-source and commercial ASR models on medical NEs (MNEs). Our investigation reveals that current SOTA general-purpose multilingual ASR models while excelling in cross-domain scenarios, exhibit sub-optimal Recall rates for MNEs in accented speech. This limitation diminishes the practical utility of these models in healthcare settings, underscoring the need for specialized solutions.

†Equal contribution.

[1]We use the terms 'clinical' and 'medical' interchangeably to encompass all aspects related to the practice of medicine and patient care.

Our contributions are as follows:

1. We benchmark 19 open-source and commercial ASR models on African accented clinical speech highlighting the deficiencies of existing architectures in accurately recognizing accented MNEs.
2. We introduce metrics for evaluating medical NER performance in the context of accented speech, including medical named entity Recall, medical WER (M-WER), and medical character error rate (M-CER).
3. We develop a novel fuzzy string matching algorithm to better align ASR-predicted noisy NEs to ground truth NEs for more nuanced analysis.
4. We demonstrate that supervised fine-tuning substantially enhances accented medical NER, making ASR models more applicable and reliable in real-world clinical scenarios.

2. Related Work

Recently, authors in [6] highlighted the challenges faced by popular ASR models in recognizing African named entities like persons, locations, and organizations from accented speech. They improved entity WER through data augmentation techniques. In the medical domain, the authors in [7] relied on large language models for correcting medical ASR transcription errors. The work of [8] also used a sequence-to-sequence model to correct clinical ASR errors. Accurately detecting and classifying medical named entities from text has been explored in [9, 10] where [9] identified five key MNEs, and employed deep learning and multi-task learning approaches to extract crucial information from clinical narratives. Also, the authors in [10] developed an ensemble of deep contextual models trained on clinical corpora from PubMed to enhance clinical NE recognition. The work of [11] separately benchmarked clinical speech recognition and entity extraction. Additionally, a production-scalable BiLSTM-CNN-Char framework with pre-trained embedding was designed by [12], which was shown to achieve better performance in speed and prediction compared to the SOTA models and commercial clinical NE recognition solutions. The authors in [13] proposed a clinical task-specific prompting framework that adopts entity definitions, annotation guidelines and samples, and error analysis-based instructions. However, research benchmarking SOTA ASR models on accented medical NE transcription or recognition is still lacking.

3. Approach

We investigated this problem by evaluating 19 open-source and commercial ASR systems on a dataset of African-accented clinical speech. A schematic is shown in Figure 1. The dataset,

10.21437/Interspeech.2024-2261

Table 1: *Predicted sentences from selected ASR models compared to the reference sentence.*

Model	Sentence
Reference sentence	**lungs clear** but dim scattered **rhonchi** nonproductive **cough**.
Xlsr-53-en	longscler bout deim scattered rong i non-productive hol
Whisper-medium	non-scler, but dim-scattered ronchi, non-productive hub.
GCP [Medical]	lungs , clear . budan scattered **rhonchi** . nonproductive
AWS [Medical] (Primary Care)	last clear but deems scattered **rhonchi** nonproductive.
Whisper-medium-clinical	**lungs clear** but dim scattered **rhonchi** nonproductive **cough**.

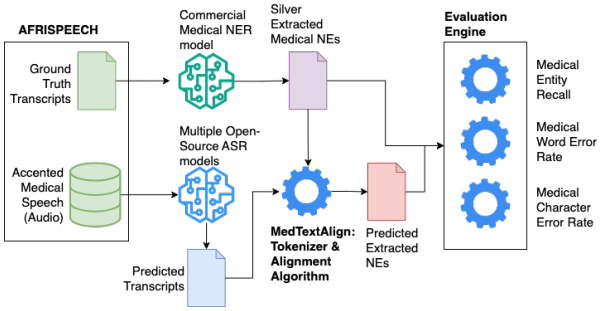

Figure 1: *Methodology: Ground truth transcripts are passed to a commercial medical NER model, and audios are passed through multiple ASR models. Predicted medical entities are extracted using the MedTextAlign algorithm. Metrics are computed over silver NEs*

medical NE extraction approach, ASR models, and evaluation methods are described below.

3.1. Data

For this analysis, we leveraged AfriSpeech-200 [4], a 200-hour Pan-African accented English speech corpus for clinical and general domain ASR with 120 accents, and over 2,300 unique speakers from over 10 African countries. The dataset statistics and NE categories are shown in Table 2. Our evaluation focuses on the clinical domain test subset. After filtering out texts lacking sufficient medical context, we retained a total of 2,844 samples, encompassing 93 different accents.

Table 2: *Dataset splits showing the number of speakers, the number of clips, speech duration, and medical named entity category counts in train/dev/test splits.*

Item	Train	Dev	Test
Number of speakers	1466	247	750
Duration (in hours)	173.4	8.74	18.77
Number of accents	71	45	108
Number of clips/speaker	39.56	13.08	8.46
Number of speakers/accent	20.65	5.49	6.94
# clinical domain clips (61.80 %)	36318	1824	3623
# general domain clips (38.20 %)	21682	1407	2723
Medical named entity category count			
Medication (MED)	4164	132	276
Medical condition (COND)	18804	901	1414
Anatomy (ANA)	13650	645	927
Test treatment procedure (TTP)	10713	428	893
Protected health information (PHI)	3449	105	253

3.2. ASR Models

We evaluated several open-source and commercial (general-purpose and medical) ASR systems covering multiple SOTA ASR architectures shown in Table 4. In addition, we selected two models for fine-tuning based on the model performance reported in [4] and our computational constraints.

3.3. Named Entity Extraction

3.3.1. Extracting Ground Truth Entities

Since the dataset was not annotated with MNEs, we leveraged a commercially available medical NER model, Amazon Comprehend Medical [9],[2] to automatically extract medical NEs from ground truth transcripts. This service has been publicly benchmarked against other NER systems by GigaOm[3] and [12], and has good accuracy in predicting multiple medical NE categories. We call these silver annotations as these are not human annotations.

3.3.2. Selected Named Entities

We focused on five key medical named entity categories: medication (MED), medical condition (COND), anatomy (ANA), test treatment procedure (TTP), and protected health information (PHI). These categories cover a wide range of entities, including medication names, dosages, diagnoses, signs, symptoms, and protected health information such as names, addresses, ID numbers, etc. The distribution of entities across these categories for each dataset split is detailed in Table 2.

3.3.3. MedTextAlign: Extracting Predicted Named Entities

To evaluate predicted MNEs, a naive method is to use an NER model to identify the ASR-predicted MNEs. However, given that the ASR predictions are noisy, often having different lengths and spellings than ground-truth NEs, and single-word to multi-word entity mismatches exist, these issues pose a challenge for most NER models, making them inadequate. For example, "analgesic properties" is misspelled as "anagesic propatis", "digoxin" wrongly transcribed as "dikod sin", and "spironolactone" as "spiro no lactone". An alignment algorithm was thus needed to better match ground-truth MNEs. Therefore, we developed MedTextAlign, a solution that uses a fuzzy string-matching algorithm to better align the predicted to the ground truth (silver) MNEs.

MedTextAlign first tokenizes the predicted transcript, creates a candidate list of unigrams, bigrams, and trigrams from the predicted transcript, then leverages a fuzzy string matching algorithm[4] to compare with each MNE from the ground truth transcript to find the closest match. The fuzzy match, akin to measuring the longest common character subsequence (e.g., in ROUGE-L [14]), produces a score between 0 and 1 for each string pair, with 1 indicating a perfect match, enabling effective

[2]Amazon Comprehend Medical at https://aws.amazon.com/comprehend/medical/

[3]GigaOm Clinical NLP Benchmark at https://gigaom.com/report/healthcare-natural-language-processing/

[4]We used the python SequenceMatcher ratio() method at https://docs.python.org/3/library/difflib.html and set the cut-off threshold to 0.5.

Table 3: *Output from the Wavlm-libri-clean-100h-base model, entities matched exactly are highlighted in bold, while near matches identified through our MedTextAlign strategy are underlined.*

Reference	Prediction
unlike **quinidine**, **disopyramide** does not increase the plasma concentration of **digoxin** in patients	anlike quinidan, disopiramid dos not incruse the plasma concentration of dikod sin in pesion
except for **ketamine**, the following agents have no **analgesic properties** and do not cause **paralysis** or **muscle relaxation**	except for ketami, befullin agents have no anagesic propatis and do not cose **paralysis** o mozul relaxition

Table 4: *Performance evaluation of benchmarked models on AfriSpeech-200 clinical domain test dataset. We report WER comparing transcript and prediction, alongside specific metrics including the medical WER (M-WER), the medical CER (M-CER), and the Recall for the different entities, medication (`MED`), anatomy (`ANA`), medical condition (`COND`), test treatment procedure (`TTP`), and protected health information (`PHI`).*

				Recall(↑)				
Models	WER	M-WER	M-CER	MED	ANA	COND	TTP	PHI
Pretrained								
Wavlm-libri-clean-100h-base	0.902	0.944	0.504	0.019	0.069	0.063	0.056	0.034
Wavlm-libri-clean-100h-large	0.784	0.852	0.307	0.029	0.177	0.142	0.110	0.043
Hubert-large-ls960-ft	0.712	0.758	0.279	0.070	0.282	0.258	0.168	0.069
Hubert-xlarge-ls960-ft	0.722	0.770	0.275	0.067	0.284	0.262	0.166	0.075
Wav2vec2-large-robust-ft-swbd-300h	0.907	0.919	0.367	0.040	0.129	0.139	0.112	0.051
Wav2vec2-large-960h	0.796	0.846	0.345	0.032	0.189	0.171	0.109	0.049
Wav2vec2-large-960h-lv60-self	0.694	0.753	0.277	0.064	0.309	0.254	0.173	0.087
Wav2vec2-xls-r-1b-english	0.666	0.729	0.266	0.081	0.251	0.249	0.227	0.138
Wav2vec2-large-xlsr-53-english	0.646	0.710	0.272	0.072	0.261	0.256	0.201	0.095
Whisper-small-en	0.486	0.566	0.225	0.215	0.571	0.536	0.475	0.300
Whisper-small	0.451	0.567	0.216	0.235	0.566	0.541	0.486	0.301
Whisper-medium-en	0.415	0.504	0.188	0.330	0.636	0.601	0.532	0.300
Whisper-medium	0.392	0.487	0.174	0.343	0.680	0.627	0.568	0.335
Whisper-large	0.373	0.454	0.154	0.425	0.717	0.667	0.597	0.331
Commercial								
Azure	0.442	0.491	0.216	0.611	0.660	0.623	0.515	0.261
AWS	0.540	0.660	0.249	0.212	0.523	0.485	0.382	0.246
AWS [Medical] (Primary Care)	0.516	0.553	0.218	0.572	0.644	0.567	0.494	0.204
GCP	0.622	0.634	0.391	0.386	0.425	0.380	0.332	0.177
GCP [Medical]	0.527	0.434	0.211	0.568	0.701	0.565	0.513	0.184
Fine-tuned on AfriSpeech-200								
Wav2vec2-large-xlsr-53-english-general	0.473	0.680	0.235	0.144	0.294	0.300	0.297	0.300
Wav2vec2-large-xlsr-53-english-both	0.308	0.467	0.135	0.451	0.658	0.576	0.567	0.356
Wav2vec2-large-xlsr-53-english-clinical	0.307	0.465	0.133	0.496	0.689	0.584	0.588	0.291
Whisper-medium-general	0.532	0.711	0.347	0.114	0.314	0.279	0.282	0.325
Whisper-medium-clinical	0.264	0.388	0.136	0.659	0.806	0.712	0.706	0.405
Whisper-medium-both	**0.241**	**0.365**	**0.118**	**0.731**	**0.822**	**0.725**	**0.726**	**0.490**

matching of nearly correct spellings in the transcript, e.g., matching wrongly spelled "quinidan" or "disopiramid" (see Table 3).

Although not perfect, this strategy proved to be very effective. In Table 3, we underline approximate entity matches and put in bold-face exact matches given by the Wavlm-libri-clean-100h-base model.

3.4. Evaluation Metrics

Given the challenges with ASR alignment, conventional ASR or NER metrics fail to effectively measure the model's ability to transcribe medical entities. Consequently, to comprehensively assess the performance of these ASR models, we opted for a broad range of metrics that cover various evaluation dimensions.

1. Recall: An information retrieval metric that computes the proportion of recovered correct (exact match) entities in the prediction. Precision and F1 score were not computed because they are overly sensitive to ASR noise or errors. Higher Recall is better.
2. Word error rate (WER): a word-level metric that evaluates insertions, deletions, and substitutions in the predicted sequence. Lower is better.
3. Medical WER (M-WER): WER computed between the ground truth MNEs and their aligned MNEs in the prediction alone. This isolates WER on MNEs of interest while ignoring all other words. All ground truth MNEs in each sample are concatenated with intervening spaces. The MNEs recovered by MedTextAlign are also concatenated in the same way. WER is then calculated between resulting sequences. Lower is better.
4. Medical CER (M-CER): Similar to medical WER, but at the character level. M-CER measures the severity of ASR misspellings. Lower is better.

4. Experiments

4.1. Benchmarking

We compared SOTA open-source pre-trained ASR models: Whisper [1], Wav2vec2 [3], XLSR [15], Hubert [2], WavLM [16], alongside commercial clinical and non-clinical ASR systems; Azure [17], AWS [18], and GCP [19]. For all open-source pre-trained ASR models, we refer readers to read their respective papers for details on pretraining corpora, model architecture, and hyperparameters. In addition, we used the Hugging Face transformer library [20] for inference. For each model, we show results on the AfriSpeech clinical domain test set in Table 4.

4.2. Fine-tuning

For the fine-tuning experiments, we fine-tuned the ASR models on three domains: (1) *general* domain (21,682 clips), (2) *clinical* domain (36,318 clips), and (3) *both* domains (58,000 clips). We fine-tuned the models using each domain's training set and tested on the clinical domain test set to investigate the effect of out-of-domain accented data on model performance. Additionally, based on the benchmark results in Table 4 and GPU memory constraints, two top performing open-source model architectures, Whisper-medium [1] and Wav2vec-large-xlsr-53 (XLSR-53) [21], were selected for fine-tuning.

XLSR-53 (378.9 M parameters) is an encoder-decoder architecture with a convolution-based feature extractor pre-trained using a self-supervised objective. Whisper-medium (789.9 M parameters) is a decoder-only multi-task architecture trained on over 680,000 hours of multilingual and multitask data using a weak supervision objective.

Each model was fine-tuned using mixed-precision training, with the AdamW optimizer [22], a batch size of 16 for 10 epochs, using a linear learning rate decay after a warmup over the first 10 % of iterations. Learning rates of $1e{-}4$ and $2.5e{-}4$ were used for the XLSR-53 and Whisper models respectively. The XLSR-53 models were trained on a single Tesla T4 GPU with 16GB GPU memory while Whisper was trained on a RTX8000 GPU with 48GB GPU memory. In general, fine-tuning took between 24-48 hours for each model.

5. Results and Discussion

Benchmarking results on 19 open-source and commercial ASR systems, as well as our fine-tuning experiments, are presented in Table 4.

5.1. Large multilingual models with web-scale training data generalize better

The overarching trend favors ASR models like Whisper [1] that were trained on vast amounts of multilingual web-scale speech data. Their data diversity and pretraining objective confer better generalization capabilities to accented speech and the clinical domain, as evidenced by their lower WER and higher Recall on MNEs, outperforming ASR models trained on monolingual data by a wide margin.

5.2. WER vs medical WER

As consistently observed across all pre-trained model families, M-WER was relatively worse overall by 4-51 % than WER, empirically validating the performance gap on medical NEs. The only exception was GCP [Medical] where its M-WER was better, demonstrating a trade-off in domain-specific fine-tuning.

5.3. Relative Performance across Entity Categories

Although Whisper-large outperformed other open- and closed-source models on WER, its MNE Recall was still poor overall with 42 % for medications (`MED`), 33 % for protected health information (`PHI`), 59 % for test treatment procedure (`TTP`), and 67 % for medical conditions (`COND`), falling far below its practical applicability in real-world clinical scenarios [23] due to the extent of required editing. Also, its 71 % Recall for anatomy (`ANA`) may have resulted from the relative abundance of body parts like leg, brain, heart, liver, etc., in web-scale text.

5.4. Medical CER: Exact vs Approximate Match

As seen in Table 3, medical WER sometimes unfairly penalized even the most minuscule ASR errors, e.g., quinidan vs quinidine, especially with multi-word entities, e.g., "muscle relaxation" vs "mozul relaxition", treating minor and severe ASR errors alike, a phenomenon that was not investigated in most prior works [7]. M-CER is complimentary in this regard, allowing us to better evaluate the severity of ASR misspellings. Also, lower M-CER helps to select the better of two ASR models with comparable WERs, like GCP [Medical] and AWS [Medical].

5.5. Fine-tuning Results on General vs Clinical Domain

The fine-tuned models significantly improved on MNE Recall, WER, medical WER, and medical CER, as the models were better adapted to accented speech in the clinical domain. However, this is not a silver bullet. Our results show that fine-tuning the ASR models on the general domain accented speech alone, in fact, worsens the WER, M-WER, M-CER, and Recall on clinical speech. XLSR-53 fine-tuned on the clinical subset reduced WER by 35 %, M-WER by 31 %, and M-CER by 43 % relative to the general domain. Fine-tuning Whisper-medium on both domains yielded the best results overall, improving WER by 54 %, M-WER by 48 %, and M-CER by 65 % relative to finetuning on the general domain only, suggesting that the ASR models still benefit from exposure to general-domain accented speech.

6. Limitations

ASR models, while beneficial, can risk patient safety and expose clinicians to liability through minor errors like mistranscribed drug names, doses, or diagnoses. Verification steps and spell-checkers can be integrated into the workflow to mitigate potential errors. Using automatically generated named entities instead of human annotation also introduces errors in entity identification. Automated systems can serve as initial annotation agents, with their outputs refined by domain experts. Lastly, the AfriSpeech-200 dataset often includes medical abbreviations (e.g., "Pt" for "Patient"), therefore, transcripts should be normalized for more accurate benchmarking.

7. Conclusion

This work highlights a noticeable disparity between general and medical WER for many SOTA ASR models, pointing to challenges in accurately recognizing accented medical named entities. Fine-tuning these models with domain-specific data was beneficial in addressing some of these issues, indicating that tailored fine-tuning can enhance ASR performance in healthcare.

8. Acknowledgements

We appreciate the invaluable support from Intron Health for contributing the dataset and compute for experiments. Tejumade Afonja is partially supported by ELSA – European Lighthouse on Secure and Safe AI funded by the European Union under grant agreement No. 101070617. We appreciate the support provided by the BioRAMP researchers, whose collaboration and insights have been fundamental to our research.

9. References

[1] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning*. PMLR, 2023, pp. 28 492–28 518.

[2] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[3] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[4] T. Olatunji, T. Afonja, A. Yadavalli, C. C. Emezue, S. Singh, B. F. Dossou, J. Osuchukwu, S. Osei, A. L. Tonja, N. Etori *et al.*, "AfriSpeech-200: Pan-African accented speech dataset for clinical and general domain ASR," *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 1669–1685, 2023.

[5] S. Ajami, "Use of speech-to-text technology for documentation by healthcare providers," *The National medical journal of India*, vol. 29, no. 3, p. 148, 2016.

[6] T. Olatunji, T. Afonja, B. F. P. Dossou, A. L. Tonja, C. C. Emezue, A. M. Rufai, and S. Singh, "AfriNames: Most ASR Models "Butcher" African Names," in *Proc. Interspeech*, 2023, pp. 5077–5081.

[7] A. Adedeji, S. Joshi, and B. Doohan, "The sound of healthcare: Improving medical transcription ASR accuracy with large language models," 2024.

[8] Y. Jiang and C. Poellabauer, "A sequence-to-sequence based error correction model for medical automatic speech recognition," in *International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2021, pp. 3029–3035.

[9] P. Bhatia, B. Celikkaya, M. Khalilia, and S. Senthivel, "Comprehend Medical: A named entity recognition and relationship extraction web service," in *International Conference On Machine Learning And Applications (ICMLA)*. IEEE, 2019, pp. 1844–1851.

[10] Y. Zhou, C. Ju, J. H. Caufield, K. Shih, C. Chen, Y. Sun, K.-W. Chang, P. Ping, and W. Wang, "Clinical named entity recognition using contextualized token representations," *arXiv preprint arXiv:2106.12608*, 2021.

[11] H. Suominen, L. Zhou, L. Hanlen, G. Ferraro *et al.*, "Benchmarking clinical speech recognition and information extraction: new data, methods, and evaluations," *JMIR medical informatics*, vol. 3, no. 2, p. e4321, 2015.

[12] V. Kocaman and D. Talby, "Accurate clinical and biomedical named entity recognition at scale," *Software Impacts*, vol. 13, p. 100373, 2022.

[13] Y. Hu, Q. Chen, J. Du, X. Peng, V. K. Keloth, X. Zuo, Y. Zhou, Z. Li, X. Jiang, Z. Lu *et al.*, "Improving large language models for clinical named entity recognition via prompt engineering," *Journal of the American Medical Informatics Association*, p. ocad259, 2024.

[14] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 74–81. [Online]. Available: https://aclanthology.org/W04-1013

[15] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. M. Pino, A. Baevski, A. Conneau, and M. Auli, "XLS-R: Self-supervised cross-lingual speech representation learning at scale," in *Proc. Interspeech*, 2022.

[16] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[17] "Cloud computing services: Microsoft Azure. Cloud Computing Services | Microsoft Azure. (n.d.)." https://azure.microsoft.com/, accessed: 2024-03-01.

[18] "Cloud Computing Services - Amazon Web Services (AWS)," https://aws.amazon.com, accessed: 2024-03-01.

[19] "Cloud Computing Services | Google Cloud," https://cloud.google.com/, accessed: 2024-03-01.

[20] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45. [Online]. Available: https://www.aclweb.org/anthology/2020.emnlp-demos.6

[21] J. Grosman, "Fine-tuned XLSR-53 large model for speech recognition in English," https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-english, 2021.

[22] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2018.

[23] E. Luchies, M. Spruit, and M. Askari, "Speech technology in Dutch health care: A qualitative study." in *HEALTHINF*, 2018, pp. 339–348.