ENHANCING ASR FOR GERMAN MEDICAL DOMAIN WITHOUT FINE-TUNING

Abdullah Al Foysal and Ronald Böck
Genie Enterprise Deutschland GmbH
{firstname.lastname}@genie-enterprise.com

Abstract: Speech recognition in medical context is important but also challenging. Especially the adaptation of speech models is a concern directly influencing the performance of models and thus, the application of such technology in medical working processes. This issue is related to the availability of speech samples for fine-tuning the systems, which is often problematic to regulatory aspects. Since, however, speech processing provides benefits for medical personnel to optimise working processes, we propose a pipeline, allowing adaption of speech processing as well as automatic output formatting. We decided to establish a post-processing approach, using pre-trained (not necessarily medically updated) speech models, being combined with lexicon- and processing techniques to allow adaptation to medical technical terms. Furthermore, the pipeline comprises handling of spoken formatting commands. The entire system is working (close to) real-time. In the paper, we also demonstrate our approach in a first study.

1 Introduction

Modern Automatic Speech Recognition (ASR) models and systems (cf. e.g. [1, 2, 3, 4]) work well for most general audio or speech. They show extremely good results also in a broad range of languages and topics. However, when it comes to more specific domains, the models tend to struggle and often rather prefer alternative vocabulary, being more present in everyday conversation. In our work, we observed this issue mainly in relation to technical terms; especially, in medical speech.

Clinical communication often uses rare or complex terminology, long compound nouns, and abbreviations. Furthermore, medical personnel use technical terms in German, English, and Latin, or a mix of multiple languages, which can be experienced in almost any doctor visit. Additionally, due to situations, which are rather time critical, a fast-speaking style is often seen in the medical context [5], which is not well covered in typical ASR systems. One reason may be seen in the under-representation of such fast speaking style in the training data of common speech models.

So, the naïve approach is a fine-tuning of the intended speech model. However, we are often be faced with two challenges in this case: 1) the complexity of the model itself, longing for reasonable number of tuning samples, and 2) the regulatory issues; (strict) privacy regulations (e.g. GDPR) limit the access to clinical recordings (cf. e.g. [6]). In both challenges, appropriate annotations are necessary, resulting in high cost of expert medical transcriptions. As a result, there is a need for an adaptation method that improves accuracy without relying on sensitive patient-related audio, expert transcripts, and/or fine-tuning.

In this paper, we present preliminary results, being part of a larger study in the project "KIRAL". One of our project subtasks are investigations of acoustic-based communication during surgeries as well as in the corresponding planning phase. As a first step, we need to analyse diagnostic findings, usually provided in (summarised) text form, but also as dictated report (e.g. computer tomographic (CT) findings). As stated in, for instance [7], the entire process of data analysis, in that case medical texts (medical Natural Language Processing (NLP)), is a challenging issue. Especially the data preparation for model training is not trivial, regarding privacy regulation (cf. Figure 1 in [7]). Considering spoken medical contributions, we are faced with same aspects, being even harder in terms of privacy and anonymisation.

Therefore, we were looking for a solution that allows an easy adaptation, circumventing, on the one hand, the challenges of regulations, and on the other hand, providing handles for medical personnel to update the system without external (expert-based) model adaptation. In this sense, we present our approach in Section 3 and results on a first study in Section 4.2.

2 Related Work

2.1 General ASR

Over time, ASR has improved dramatically, shifting from traditional, statistical techniques to modern neural network-based strategies. Early ASR systems mainly relied on Hidden Markov Models combined with Gaussian

Mixture Models [8]. Although these techniques serve as a foundation for speech recognition research, they struggle with low-quality and complex situations like background noise, large vocabularies, and complex real-world conditions (e.g. [9]).

Deep Neural Networks replace conventional models in early ASR systems to increase accuracy. Afterwards, Recurrent Neural Networks and Long-Short Term Memory Networks are used to better model speech over time [10, 11]. These models work well but process speech sequentially, which makes them slow and hard to scale. Additionally, they often perform poorly in noisy or domain-specific environments such as medical speech, and they require large labeled datasets to train (e.g. [12]). Therefore, real-time usage of these early ASR systems was restricted since they could only work appropriately after the full audio had been captured [13]. These circumstances are nowadays rather solved since more and more large data sets are available.

Recent systems use Transformer-based models, improving speech recognition by using attention mechanisms, which help ASR to understand long and complex speech patterns more accurately. Large models such as Whisper [14] show strong performance across multiple languages and tasks, often working well without additional fine-tuning [13], but struggle in domain specific tasks[15].

Self-supervised learning is used by more recent models, such as Wav2Vec 2.0 and HuBERT, to perform well even in noisy environments and with little labeled data [2, 16]. However, they still make errors on rare domain-specific terms, and adapting them to medical speech is difficult due to privacy constraints. In addition, Transformer-based models require substantial computational resources, which limits their suitability for real-time speech recognition.

2.2 Medical Related ASR

The research paper [17] investigates ASR in the medical domain, particularly in doctor–patient conversations. It highlights several challenges, including overlapping speech, interruptions, and spontaneous dialogues. The work shows that conversational clinical speech is substantially harder than dictated speech. Their approach relies on domain-specific training data and does not explore non-training-based adaptation methods.

MultiMed [18] presented a large multilingual medical ASR dataset including German. Their results show that even strong multilingual models perform worse in medical contexts due to terminology and speaking-style differences. While the data set improves coverage, effective adaptation still depends on resource-intensive domain training.

United-MedASR [19] proposed combining synthetic data generation and semantic correction to improve medical ASR based on Whisper [14]. The study shows that vocabulary-aware post-processing improves accuracy but still requires model adaptation using domain-specific data, which can be difficult under privacy constraints.

Large language models were investigated as a post-processing layer to enhance clinical ASR output [20]. This approach improves medical concept accuracy but introduces additional computational complexity and dependency on large models, and it does not focus on language-specific issues such as German compound words.

In [7], German medical NLP is surveyed, highlighting the lack of publicly available clinical datasets and models. For German medical ASR, the lack of audio data is an even bigger problem, since there are (very) few recorded medical speech samples, making it hard to build accurate domain-specific speech recognition systems.

The authors of [21] investigate how to improve German language models for medical NLP by continuing their training on clinical texts and translated English medical data. The results show that training on domain-specific texts helps models work well in medical contexts. However, the computational cost of training language models is often (very) high.

3 Methods

Our method improves German medical speech transcription using a lexicon-based post-processing framework. It works on the output of an existing ASR system. Instead of fine-tuning the ASR model, it corrects recognition errors in clinical terms as well as aligns commands often seen in clinical speech. This includes mistakes with medical terms, spelling variants, punctuation, and spoken formatting commands.

The system is designed for near real-time operation and can be integrated with any modern ASR engine. During live transcription, the ASR output is displayed largely unchanged to ensure low latency and stable streaming. Once the recording ends, a final post-processing step is applied to produce a clean, structured, and medically accurate transcript. This approach allows a low-effort update of medical speech recognition, providing an easy-to-handle adaptation process to a specific sub-domain (e.g., neurology, cardiology, etc.). Furthermore, it also enables non-speech-processing-experts to contribute to adaptations, ideally directly in clinics.

The method consists of four main components, which will be discussed in detail in the following sections:

1. Medical Vocabulary Lexicon.
2. Terminology correction using exact and fuzzy matching.

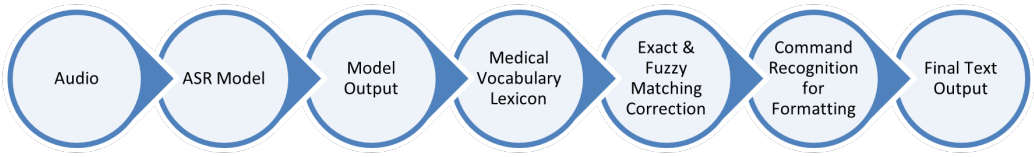

Figure 1 – Workflow of the proposed method, indicating the interplay of the components and modules.

3. Recognition of spoken formatting commands.
4. Final clean-up and normalization.

Taking these components into account, the overall workflow of our proposed method is as follows, which is also visualised in Figure 1:

1. Audio is transcribed by a streaming ASR system.
2. Raw transcript is displayed during recording.
3. Post-processing is applied immediately after recording.
4. Medical terminology is corrected using the lexicon and matching methods.
5. Spoken punctuation and formatting commands are converted.
6. The final medical transcript is generated.

3.1 Medical Vocabulary Lexicon

The system relies on a medical vocabulary lexicon, which is stored in a dedicated database. It is enriched with German medical terms, including ICD-10 [22] codes. The lexicon stores standard terms, spelling variants, spacing differences, and common ASR errors. It covers disease names, drug names, procedures, clinical abbreviations, and other terminology used in medical reports. In particular, each medical concept has a single canonical form. Alternative forms capture common ASR errors, like wrong capitalization, missing letters, or incorrect word boundaries. However, we mention that the current version of our lexicon does not cover all medical terms, but is rather constructed and will be updated over time in relation to the project's use cases.

The lexicon can be expanded, so new terms can be added without retraining the ASR. This allows adaptation to different specialities, such as cardiology or intensive care, and to institution-specific terms. The design helps to cover rare and compound medical terms that general ASR systems often miss. Furthermore, the concept of lexicon-based post-processing enables non-experts to directly contribute to the adaptation of ASR output. Medical personnel can thus provide new terms but also, and this might be more important in terms of personalization, adapt towards their "common mispronunciation" (e.g. "Bleura" instead of "Pleura"; soft plosive vs hard plosive).

3.2 Terminology Correction Using Exact and Fuzzy Matching

After ASR, the text is processed by a terminology correction module operating in two stages, effectively correcting a wide range of ASR errors:

1. Exact matching: It identifies frequently occurring incorrect forms, including capitalization errors, predictable spacing issues, and known spelling variants. This stage is efficient and well-suited for real-time or near real-time processing. For this, we applied the Flashtext framework [23]. Flashtext is a Python library designed for fast keyword search and replacement in text.
2. Fuzzy matching: Fuzzy matching, based on fuzzy similarity scores implemented in the RapidFuzz library [24], is applied carefully to avoid wrong corrections. It is used only for long words, considering a similarity threshold. Further, we used this only in cases where the replacement is a known medical term. This keeps corrections reliable while preserving non-medical words.

3.3 Spoken Punctuation and Formatting Commands

Medical professionals often use punctuation and structural commands verbally, especially in diagnostic finding dictations, mentioning expressions such as “Komma (English: comma)”, “Punkt (English: full stop)”, “neue Zeile (English: new line)”, or “Absatz (English: new paragraph)”. Standard ASR systems frequently misinterpret these commands, especially at high speaking rates.

The proposed method includes a rule-based module detecting spoken punctuation and formats words and turns them into the correct symbols or layout elements. This includes replacing punctuation words with symbols, converting line and paragraph commands into breaks, and handling compound formatting expressions. This allows hands-free report formatting, which is useful when medical staff dictate while doing other tasks. In our experiments, we figured out that by handling formatting in post-processing, the system works well even if the ASR output has imperfect word boundaries.

3.4 Final Clean-up and Normalization

After terminology correction and command recognition, the system performs a final clean-up step to improve readability and consistency. This includes:

1. Removing duplicated punctuation,
2. Normalizing spacing around symbols,
3. Eliminating known non-informative phrases that occasionally appear in ASR output.

4 Experiments

This section describes the data used, the particular models, as well as the experimental results achieved in this first study.

4.1 Experimental Setup

4.1.1 Data Set

In the study, we consider spoken diagnostic findings and thus, were searching for respective data sets. However, in the medical domain, sharing data is rather complicated (cf. also discussion in Section 1) and therefore, (unfortunately) “clinical corpora are only accessible to the research staff within the lifetime of a project and remain inaccessible forever for the outside world” [6]. This made it difficult to obtain suitable samples for experiments. However, we were able to convince two colleagues to provide recordings for a preliminary study:

- Colleague 1 (native German, no accent, male) provided three acoustic samples of diagnostic findings in German. The main topics are full-body scans in CT. The recording happened with a headset in a calm lab environment.
- Colleague 2 (native English, Indian English accent, male) contributed one sample. The diagnostic report comprises a CT analysis of brain regions. The recordings were conducted with a close-talk microphone in a lab environment, however, comprising also same background noise.

These samples were used for testing our approach and gaining the achievements presented in Section 4.2. Due to regulatory aspects, we are not allowed to share the current data. Indeed, in the “KIRAL” project, we are planning additional data collections, which are intended to be conducted under consents, allowing also (at least partial) dissemination of recordings.

4.1.2 Hardware and Software Setup

All experiments were conducted on a local workstation running Windows 11 as the host operating system. The system was equipped with 16 GB RAM and an NVIDIA GeForce RTX 3060 GPU, which was used to accelerate ASR inference. GPU support was enabled throughout the experiments to ensure near real-time transcription performance.

To ensure reproducibility and portability, the system was deployed using Docker containers. The runtime environment was based on Ubuntu 22.04 with NVIDIA CUDA 12.8.1 and cuDNN runtime support.

4.1.3 Applied Speech Models

During the experiments (for both German and English medical audio), we mainly used two Whisper models [14], namely Whisper Medium and Whisper Turbo. Whisper Medium is a mid-sized speech recognition model with about 769 million parameters. It offers good accuracy with reasonable computational cost. In contrast, Whisper Turbo (around 809 million parameters) is a faster version designed for high-speed transcription. Even though Whisper Turbo has more parameters, it is optimized in the way the parameters are handled internally more efficiently, leading to faster transcripts [25].

As German medical ASR baseline, we considered a fine-tuned Whisper model (Whisper-Small-German) from the MultiMed [18] research work. This model is fine-tuned on medical German data and available in the Hugging Face platform [26].

We applied the MedASR model from Google [27] as English medical ASR baseline. This model is pre-trained on medical domain data, but supports only the English language. It is also available for download in the Hugging Face platform [28].

Finally, we emphasize that we considered two settings: 1) this is the raw or direct output of the particular models, and 2) the post-processing approach as suggested in Section 3.

4.1.4 Evaluation Metric

For evaluation, we relied on the samples provided by the colleagues, using them purely for testing purposes. As ground truth, we generated a manual transcription of the recordings.

To measure the performance of the approaches, namely the direct transcript of the ASR systems as well as the output of our system, we decided for the common Word Error Rate (WER) measure. This indicates the ratio of incorrectly recognised words compared to a human-generated ground truth.

Additionally, we mention that in the proposed approach, a handling of punctuation/formatting indicators is already realised. Therefore, we decided on the following *modification* of WER: if in the ground truth a punctuation/formatting indicator is mentioned and the pipeline reacts appropriately, we do not consider this as a word error since the intended reaction was created by the system.
In particular, the raw output of ASR systems keeps spoken punctuation words as normal text. To create an appropriate text output, our approach immediately change this indicators into the respective format. However, to avoid penalties by WER, we do not consider the following substitutions not as errors:

- “Punkt” or “Full Stop” $\rightarrow$ “.”
- “Komma” or “Comma” $\rightarrow$ “,”
- “Absatz" or "New Paragraph” as well as “neue Zeile" or "Next Line” $\rightarrow$ line breaks
- “Klammer auf/zu” or "open/close Brackets" $\rightarrow$ “()”.

4.2 Experimental Results

Table 1 and Table 2 present the achievements of the current study, using WER as metric (be aware of the hints in Section 4.1.4), comparing the systems raw outputs and the proposed approach against human-generated ground truth.

Although, the direct output of the ASR systems already shows a good performance, especially for Whisper Turbo, they still have trouble with various medical terms. Considering the higher speaking rate of medical personnel, in particular in diagnostic finding use cases, this comprises additional challenges to speech systems. In our study, we saw that, in particular, word boundaries are critical, often resulting in combining or tightening two words. In most cases, this happened with a medical term and the succeeding word.

In the German language (cf. Table 1), we see slight improvements, which are, of course, being based on the underlying ASR model. Whisper Turbo is by design larger and being trained in a more generalised way. This results in a lower WER itself. However, by additional corrections we still decreased the WER. For Whisper Medium only in the complex sample 3 an improvement was gained.

From the perspective of the MultiMed-German model, we observe (cf. Table 1) that adding the post-processing component from our proposed approach to the model’s raw output leads to a slight improvement in WER performance, particularly for German Audio 1 and German Audio 2.

Regarding the results on the English sample in Table 2, we see mainly a gain in performance (i.e. decrease of WER) in the MedASR system. We analysed the specific outputs in detail: We observed that certain medical terms, such as ’haematoma’ and ’lumbar spine’, were incorrectly transcribed in the MedASR raw output as ’haematuma’ and ’lumber spine’, respectively. These errors likely occurred due to background noise, unclear speech, or fast initial speech from the speaker. Some special symbols also appear in the raw output from MedASR, likely due

Table 1 – Word Error Rate (in percent) of the ASR systems considering the *German* speech samples. Two settings were compared: raw output of the ASR and output of the proposed pipeline (cf. Section 3).

Data	**Whisper Medium**		**Whisper Turbo**		**MultiMed-German**	
	Raw Output	Proposed Pipeline	Raw Output	Proposed Pipeline	Raw Output	Proposed Pipeline
German medical audio 1	0.15	0.15	0.041	0.031	0.28	0.27
German medical audio 2	0.15	0.15	0.061	0.061	0.43	0.40
German medical audio 3	0.25	0.17	0.11	0.10	0.40	0.40
Average WER	0.183	0.157	0.071	0.064	0.37	0.35

to model hallucination. Due to these reasons, MedASR shows a higher WER compared to the other two models. However, such misspellings were successfully corrected during the proposed post-processing stage. That is why we are getting slightly better result in MedASR with our proposed pipeline.

Although we are not able to generalise from one sample in English, we interestingly be faced with higher WER. This is unexpected since from our knowledge, there is more medical speech data in English available than in German. So, we expected a better performance from the English models. Our assumption is that the current WER values might be caused by the speakers accent. However, even with no further adaptation of the lexicon toward accent-based pronunciation, our pipeline helped to slightly reduce the WER.

Finally, we also analysed the formatted output of the pipeline. In the current set of samples, the formatting commands were recognised and considered appropriately, resulting a well structured outputs that can be used in medical patient reports.

Table 2 – Word Error Rate (in percent) of the ASR systems considering the *English* speech sample. Two settings were compared: raw output of the ASR and output of the proposed pipeline (cf. Section 3).

Data	**Whisper Medium**		**Whisper Turbo**		**MedASR**	
	Raw Output	Proposed Pipeline	Raw Output	Proposed Pipeline	Raw Output	Proposed Pipeline
English medical audio	0.18	0.18	0.15	0.14	0.20	0.18

As a final remark, we are aware of the current limited number of speech samples and thus, the limited explanatory power of the results. Therefore, we plan further recordings and another study to evaluate the proposed pipeline. Our focus will be, of course, on the handling of German diagnostic findings.

5 Conclusion and Outlook

This paper presents an approach which 1) allows an adaptation of medical ASR and 2) an automatic formatting based on stated commands.
Regarding the first aspect, we used a post-processing paradigm to establish ASR corrections, in contrast to typical fine-tuning methods, relying on existing speech samples. From our perspective, the suggested pipeline better fits the current situation of available speech samples. Furthermore, it explicitly enables medical personnel to contribute as well as to handle the adaptation process by themselves (cf. particularly Section 3.1).
The second aspect immediately allows a formatting of the ASR transcribed outputs in a hands-free way. This assists also the medical personnel to focus on their main tasks, not being distracted by paper-work-like issues, however, deliver appropriate reports for well-organised patient treatments.

In our preliminary study, we got aware of some aspects to improve our approach. One aspect is the handling of ambiguity of numbers in medical indicators. For example, one colleague stated “degeneration of the l34”, saying l-3-4 as single digits. If not treated in an appropriate way, this might be misinterpreted as l-34 (say: thirty-four).
Furthermore, we are planning to integrate more speech systems into our pipeline as well as running more studies to prove the benefits of our approach.

Acknowledgment

This research was partially funded by the German Federal Ministry of Research, Technology and Space (BMFTR) in the project “KIRAL” (grant number: 16SV9543).

References

[1] KUCHAIEV, O., J. LI, H. NGUYEN, O. HRINCHUK, R. LEARY, B. GINSBURG, S. KRIMAN, S. BELIAEV, V. LAVRUKHIN, J. COOK, P. CASTONGUAY, M. POPOVA, J. HUANG, and J. M. COHEN: *Nemo: a toolkit for building ai applications using neural modules*. 2019. URL `https://arxiv.org/abs/1909.09577`. `1909.09577`.

[2] BAEVSKI, A., H. ZHOU, A. MOHAMED, and M. AULI: *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. Curran Associates Inc., Red Hook, NY, USA, 2020.

[3] RAVANELLI, M., T. PARCOLLET, P. PLANTINGA, A. ROUHE, S. CORNELL, L. LUGOSCH, C. SUBAKAN, N. DAWALATABAD, A. HEBA, J. ZHONG, J.-C. CHOU, S.-L. YEH, S.-W. FU, C.-F. LIAO, E. RASTORGUEVA, F. GRONDIN, W. ARIS, H. NA, Y. GAO, R. D. MORI, and Y. BENGIO: *Speechbrain: A general-purpose speech toolkit*. 2021. URL `https://arxiv.org/abs/2106.04624`. 2106.04624.

[4] RADFORD, A., J. W. KIM, T. XU, G. BROCKMAN, C. MCLEAVEY, and I. SUTSKEVER: *Robust speech recognition via large-scale weak supervision.* In A. KRAUSE, E. BRUNSKILL, K. CHO, B. ENGELHARDT, S. SABATO, and J. SCARLETT (eds.), *Proceedings of the 40th International Conference on Machine Learning*, vol. 202 of *Proceedings of Machine Learning Research*, pp. 28492–28518. PMLR, 2023.

[5] EICHLER WAAGE, A. K. and J. IWARSSON: *The effect of speaking rate on voice and breathing behavior*. *Journal of Voice*, 2024. doi:10.1016/j.jvoice.2024.07.004. In press.

[6] LOHR, C., S. BUECHEL, and U. HAHN: *Sharing copies of synthetic clinical corpora without physical distribution — a case study to get around IPRs and privacy constraints featuring the German JSYNCC corpus*. In N. CALZOLARI, K. CHOUKRI, C. CIERI, T. DECLERCK, S. GOGGI, K. HASIDA, H. ISAHARA, B. MAEGAARD, J. MARIANI, H. MAZO, A. MORENO, J. ODIJK, S. PIPERIDIS, and T. TOKUNAGA (eds.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA), Miyazaki, Japan, 2018.

[7] ZESCH, . B. J., TORSTEN: *German medical natural language processing–a data-centric survey*. In *The Upper-Rhine Artificial Intelligence Symposium UR-AI 2022 : AI Applications in Medicine and Manufacturing*, pp. 137—-145. Furtwangen University, 2022.

[8] PRABHAVALKAR, R., T. HORI, T. N. SAINATH, R. SCHLÜTER, and S. WATANABE: *End-to-end speech recognition: A survey*. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32, pp. 325–351, 2023.

[9] FRIKHA, M., A. B. HAMIDA, and M. LAHIANI: *Hidden markov models (hmms) isolated word recognizer with the optimization of acoustical analysis and modeling techniques*. *International Journal of the Physical Sciences*, 6(22), pp. 5064–5074, 2011.

[10] GRAVES, A., A.-R. MOHAMED, and G. HINTON: *Speech recognition with deep recurrent neural networks*. In *2013 IEEE international conference on acoustics, speech and signal processing*, pp. 6645–6649. Ieee, 2013.

[11] ORUH, J., S. VIRIRI, and A. ADEGUN: *Long short-term memory recurrent neural network for automatic speech recognition*. *IEEE Access*, 10, pp. 30069–30079, 2022.

[12] TAYE, M. M.: *Understanding of machine learning with deep learning: Architectures, workflow, applications and future directions*. *Computers*, 12(5), 2023. doi:10.3390/computers12050091.

[13] ARRIAGA, C., A. POZO, J. CONDE, and A. ALONSO: *Evaluation of real-time transcriptions using end-to-end asr models*. *arXiv preprint arXiv:2409.05674*, 2024.

[14] RADFORD, A., 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*, pp. 28492–28518. PMLR, 2023.

[15] HWANG, H., E. JORDAN, D.-H. KIM-DUFOR, C. LEMEY, and M. ALRAHABI: *Evaluating ASR in a clinical context : What whisper misses*. In M. ABBAS, T. YOUSEF, and L. GALKE (eds.), *Proceedings of the 8th International Conference on Natural Language and Speech Processing (ICNLSP-2025)*, pp. 374–378. Association for Computational Linguistics, Southern Denmark University, Odense, Denmark, 2025.

[16] HSU, W.-N., 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*, 29, pp. 3451–3460, 2021.

[17] JAUNZEIKARE, D., A. KANNAN, P. NGUYEN, H. SAK, A. SANKAR, J. TANSUWAN, N. WAN, Y. WU, and X. ZHANG: *Speech recognition for medical conversations*. *arXiv preprint arXiv:1711.07274*, 2017.

[18] LE-DUC, K., P. PHAN, T.-H. PHAM, B. P. TAT, M.-H. NGO, T. NGUYEN-TANG, and T.-S. HY: *MultiMed: Multilingual medical speech recognition via attention encoder decoder*. In G. REHM and Y. LI (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pp. 1113–1150. Association for Computational Linguistics, Vienna, Austria, 2025. doi:10.18653/v1/2025.acl-industry.79.

[19] BANERJEE, S., A. AGARWAL, and P. GHOSH: *High-precision medical speech recognition through synthetic data and semantic correction: United-medasr. arXiv preprint arXiv:2412.00055*, 2024.

[20] ADEDEJI, A., S. JOSHI, and B. DOOHAN: *The sound of healthcare: Improving medical transcription asr accuracy with large language models. arxiv 2024. arXiv preprint arXiv:2402.07658*, ????

[21] IDRISSI-YAGHIR, A., A. DADA, H. SCHÄFER, K. ARZIDEH, G. BALDINI, J. TRIENES, M. HASIN, J. BEWERSDORFF, C. S. SCHMIDT, M. BAUER, K. E. SMITH, J. BIAN, Y. WU, J. SCHLÖTTERER, T. ZESCH, P. A. HORN, C. SEIFERT, F. NENSA, J. KLEESIEK, and C. M. FRIEDRICH: *Comprehensive study on German language models for clinical and biomedical text understanding*. In N. CALZOLARI, M.-Y. KAN, V. HOSTE, A. LENCI, S. SAKTI, and N. XUE (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 3654–3665. ELRA and ICCL, Torino, Italia, 2024.

[22] FEDERAL INSTITUTE FOR DRUGS AND MEDICAL DEVICES (BFARM): *Icd-10 codes*. `https://www.bfarm.de/EN/Code-systems/Services/Code-Search-Online/_node.html`, 2026. Last accessed: January 20, 2026.

[23] VIKASH SINGH: *Flashtext*. `https://pypi.org/project/flashtext/1.4/`, 2026. Last accessed: January 20, 2026.

[24] MAX BACHMANN: *Rapidfuzz*. `https://pypi.org/project/RapidFuzz/`, 2026. Last accessed: January 20, 2026.

[25] OPENAI: *Whisper*. `https://github.com/openai/whisper`, 2022. Last accessed: January 20, 2026.

[26] LEDUCKHAI: *Multimed*. `https://huggingface.co/leduckhai/MultiMed`, 2026. Last accessed: January 20, 2026.

[27] WU, K., E. VARIANI, T. BAGBY, and M. RILEY: *Last: Scalable lattice-based speech modelling in jax*. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

[28] GOOGLE HEALTH AI: *Medasr: Medical automated speech recognition*. `https://huggingface.co/google/medasr`, 2025. Last accessed: January 20, 2026.