Revolutionizing Radiological Analysis: The Future of French Language Automatic Speech Recognition in Healthcare

Mariem Jelassi , Oumayma Jamai , Jacques Demongeot *

Posted Date: 11 March 2024

doi: 10.20944/preprints202403.0619.v1

Keywords: Automatic Speech Recognition (ASR), Medical Transcription, Radiology, Whisper Large-v2 Model, Language-Specific ASR Systems, French Language Processing, AI in Healthcare

Article

Revolutionizing Radiological Analysis: The Future of French Language Automatic Speech Recognition in Healthcare

Mariem Jelassi [1,2], Oumayma Jemai [2,3] and Jacques Demongeot [3,*]

[1] RIADI Laboratory, ENSI, Manouba University, 2010, La Manouba, Tunisia
[2] Health Tech Innovation Systems Inc., ENSI Innovation Hub, 2010, La Manouba, Tunisia
[3] SUP'COM, Carthage University, 2083, Ariana, Tunisia
[4] AGEIS Laboratory, UGA, 38700, La Tronche, France
* Correspondence: jacques.demongeot@univ-grenoble-alpes.fr

Abstract: This study introduces a specialized Automatic Speech Recognition (ASR) system, leveraging the Whisper Large-v2 model, specifically adapted for radiological applications in French language. Our methodology focused on adapting the model to accurately transcribe medical terminology and diverse accents within the French language context, achieving a notable Word Error Rate (WER) of 17.121%. The research involved extensive data collection and preprocessing, utilizing a wide range of French medical audio content. The results demonstrate the system's effectiveness in transcribing complex radiological data, underscoring its potential to enhance medical documentation efficiency in French-speaking clinical settings. The discussion extends to the broader implications of this technology in healthcare, including its potential integration with Electronic Health Records (EHRs) and its utility in medical education. The study also explores future research directions, such as tailoring ASR systems to specific medical specialties and languages. Overall, this research contributes significantly to the field of medical ASR systems, presenting a robust tool for radiological transcription in the French language and paving the way for advanced technology-enhanced healthcare solutions.

Keywords: Automatic Speech Recognition (ASR); medical transcription; radiology; Whisper Large-v2 model; language-specific ASR systems; French language processing; AI in healthcare

I. Introduction

The integration of Artificial Intelligence (AI) in healthcare, particularly through Automatic Speech Recognition (ASR) systems, has been a subject of increasing interest in recent years. These systems, demonstrating significant potential in various medical applications, have revolutionized the way patient-physician interactions are transcribed and clinical documentation is managed [1]. In the field of radiology, where accuracy and efficiency in reporting are paramount, the application of ASR technology can be particularly transformative, offering a new paradigm in the way radiologists work and interact with diagnostic data [2].

Despite the global advancements in ASR technology, its application within the French medical context has been limited. This gap is primarily due to the linguistic and terminological specificity required in medical ASR systems, which are often not met by general-purpose ASR tools [3]. The development of a French-language medical ASR system is thus a technological and linguistic challenge, requiring a deep understanding of medical terminologies and the nuances of spoken French in a clinical setting [4]. The need for extensive and specialized datasets, encompassing a wide range of medical terminologies, accents, and speech patterns specific to the medical profession, poses a significant hurdle [5]. However, this also presents an opportunity to develop tailored solutions that can significantly benefit the medical community, particularly in non-English speaking regions.

 doi:10.20944/preprints202403.0619.v1

Recent studies have shown promising results in cross-lingual applications of ASR, adapting systems to work with low-resource languages [6]. The use of advanced neural network models and language processing techniques has been explored to enhance the accuracy and reliability of medical ASR systems [7]. These advancements are not only technical but also encompass a broader understanding of the medical field, ensuring that the developed systems are finely tuned to the specific needs of healthcare professionals.

The primary objective of this research is to develop a specialized French-language ASR system, tailored for radiological applications. This system aims to facilitate radiologists in efficiently generating medical images reports, thereby enhancing the overall workflow in diagnostic procedures [8]. The novelty of this project lies in its focus on creating a dedicated ASR tool for radiology, addressing the scarcity of French-language audio datasets in the medical domain. By leveraging machine learning techniques, specifically tailored for medical jargon and radiological terms, this tool aims to provide accurate and efficient transcription services [9]. The potential of ASR in medicine is vast, ranging from automated transcription of medical reports to assisting in the drafting process, indexing medical data, and enabling voice-based queries in medical databases [10].

The implications of ASR technology extend beyond radiology to other medical fields. For instance, in emergency medical services, ASR has been assessed for its impact on stroke detection, showing potential for improving response times and diagnostic accuracy [11]. In long-term care for older adults, ASR models have been used to facilitate interview data transcription, saving time and resources [12]. Even in operation room, ASR techniques can be used to improve the dialogue between the surgeon and human (surgical nurse) or digital (robotic arm) assistants [13,14]. Additionally, in paediatric care, ASR and voice interaction technologies have been explored for remote care management, demonstrating feasibility and effectiveness in tracking symptoms and health events [15].

Recent reviews in the field of healthcare have highlighted significant advancements in Automatic Speech Recognition (ASR) technology and its diverse applications. These advancements underscore the transformative potential of ASR in healthcare, paving the way for more efficient, accurate, and patient-centred medical practices.

A comprehensive review of state-of-the-art approaches in ASR, speech synthesis, and health detection using speech signals has shed light on the current capabilities and future directions of speech technology in healthcare [16]. This review emphasizes the growing importance of ASR in various healthcare settings, from clinical documentation to patient monitoring.

Another study explores the potential of real-time speech-to-text and text-to-speech converters, using Natural Language Grammar (NLG) and Abstract Meaning Representation (AMR) graphs, to enhance healthcare communication and documentation [17]. This technology could revolutionize how medical professionals interact with electronic health records, making the process more intuitive and efficient.

The robustness of ASR systems in noisy environments, a common challenge in medical settings, has also been a focus of recent research [18]. Enhancing the noise robustness of ASR systems is crucial for their effective deployment in diverse healthcare environments, from busy emergency rooms to outpatient clinics.

Furthermore, a systematic literature review on various techniques within the domain of Speech Recognition provides a comprehensive understanding of the advancements and challenges in this field [19]. This review highlights the rapid evolution of ASR technology and its increasing relevance in healthcare.

In addition to these technical advancements, the integration of ASR with patient-reported outcomes and value-based healthcare has been explored [20]. This integration signifies a shift towards more personalized and patient-centred healthcare models, where patient voices and experiences are directly captured and analysed through advanced speech recognition technologies.

These reviews and studies collectively illustrate the significant strides made in ASR technology and its increasing applicability in healthcare. From enhancing clinical workflows to improving patient engagement, ASR technology is set to play a pivotal role in the future of healthcare delivery.

 doi:10.20944/preprints202403.0619.v1

II. Methods

In our study on developing an Automatic Speech Recognition (ASR) system tailored for radiological applications, we meticulously document the methods and processes integral to our research. This section begins with a detailed description of the data preprocessing techniques and datasets foundational to our ASR system. We then describe the model selection criteria, training processes, and the deployment of our speech recognition application. The subsequent sections delve into the tasks and design of our system, followed by an outline of the evaluation metrics that quantify the performance of our system.

A. Data Preprocessing

1. Data Source Selection and Collection Methodology

Our research utilized a diverse array of audio content, with a primary focus on YouTube, which constituted approximately 90% of our data source. This was supplemented by audiobooks and podcasts. The selection strategy was driven by the need to cover a broad spectrum of radiological topics. YouTube, as a rich repository, provided access to a wealth of relevant material including radiology conferences, online courses, and medical descriptions. The integration of audiobooks and podcasts, forming about 10% of our dataset, enriched it with detailed presentations on radiological themes, ensuring a rich variety of accents and tonalities crucial for the development of a robust ASR system.

In our comprehensive approach to data collection, we employed a multi-tiered methodology. This involved systematic categorization based on human body systems, targeted keyword analysis for each organ and imaging type, and the inclusion of diverse pedagogical voices. A critical component of our methodology was the technical extraction of audio from YouTube videos and podcasts, using sophisticated software tools to isolate the audio track from visual elements and to extract high-quality audio. This process created an audio-centric dataset, focusing on the auditory dimensions of medical instruction.

The dataset we compiled boasts over 240 hours of audio content, representing a vast educational repository.

2. Data Collection Methodology

We adopted a systematic approach, exploring various human body systems to ensure comprehensive coverage of radiological topics. Each system and its constituent organs were thoroughly investigated. The metadata for each video, such as imaging type, speaker accent, and duration, were meticulously organized and cataloged in a .csv file, facilitating efficient data management and retrieval. The Table 1 details the distribution of this material among various body systems, highlighting the dataset's depth and scope:

Table 1. Distribution of Audio Content Duration by Body System.

Body System	Duration (hh:mm:ss)
Nervous System	53:32:04
Musculoskeletal System	61:27:22
Endocrine System	23:42:37
Respiratory System	22:54:55
Cardiovascular System	26:11:11
Digestive System	19:43:15
Reproductive System	17:00:24
Urinary System	6:17:28
Auditory System	5:10:49

 doi:10.20944/preprints202403.0619.v1

Lymphatic and Blood System	4:21:04

Additionally, we calculated the number of hours for each speaker accent, as demonstrated in Table 2, to ensure a diverse representation of accents in our dataset.

Table 2. Distribution of Audio Content Duration by Accent.

Accent	Duration (hours)
African	16h 33m
Algerian	25h 19m
Canadian	2h 38m
Native French	155h 22m
Moroccan	29h 17m
Tunisian	11h 13m

3. Transcription Generation

The transcription generation process formed a significant aspect of our study, requiring an extensive evaluation of various transcription tools and models. This evaluation was meticulously designed to assess not only the performance of these tools, but also their compatibility with the unique requirements of radiological audio content. We embarked on an extensive exploration of transcription tools, including state-of-the-art models such as Whisper, Conformer-CTC large, and wav2vec2-large. These models were specifically chosen for their cutting-edge capabilities in speech recognition and their potential applicability in the medical imaging domain. In addition to these advanced models, we incorporated Python's SpeechRecognition library and specialized online transcription services like oTranscribe, allowing for a comprehensive comparison between traditional transcription methods and contemporary machine learning-based approaches.

To objectively assess the effectiveness of each transcription tool, we employed the Word Error Rate (WER) metric as our primary evaluation criterion. For this purpose, a reference transcription was meticulously crafted for a specific 12-second audio clip. WER, a widely recognized standard in speech recognition research, was calculated using Python's Jiwer library. This approach provided a quantitative basis for comparing the transcription accuracy of each tool, ensuring an unbiased evaluation of their performance. The evaluation process was specifically tailored to address the challenges inherent in transcribing medical imaging content, which frequently includes specialized terminology and a variety of accents. Consequently, the performance of each tool was evaluated not only in terms of overall accuracy, but also for its proficiency in handling the distinct linguistic and acoustic characteristics present in our dataset.

The outcomes of this exhaustive evaluation are encapsulated in Table 3, which presents a detailed comparison of the WER scores for each transcription tool. This table offers valuable insights into the relative strengths and weaknesses of each tool within the specific context of our research. The findings from this analysis played a crucial role in guiding our selection of the most suitable transcription tool for our dataset.

Table 3. Comparative Analysis of Transcription Tool Performance Based on Word Error Rate.

Transcription Tools	**WER**
SpeechRecognition (Python library supporting various engines including Google Web Speech API, Sphinx, etc.)	0.228
oTranscribe (Free online tool designed for audio transcription with features like playback speed control, bookmarking, etc.)	0.628

 doi:10.20944/preprints202403.0619.v1

Conformer-CTC Large (Voice recognition model based on the Conformer-CTC architecture, optimized for automatic speech transcription)	0.2
AssemblyAI (Automatic speech transcription service using machine learning models)	0.100
Whisper Large-v2 (Voice recognition model developed by OpenAI, designed for automatic speech transcription)	0.142
Wav2Vec 2.0 (Model developed by Facebook AI, excels in automatic speech recognition using self-supervised learning approach)	0.257

In the selection of an appropriate transcription tool for our ASR system, we conducted a comparative analysis focusing on both accuracy and practical applicability to our extensive dataset. The Whisper Large-v2 model, developed by OpenAI, emerged as the most suitable choice for our research needs. While the Assembly AI tool exhibited the lowest WER, its usage limitation of three hours per month was not compatible with the scale of our dataset. Conversely, Whisper Large-v2 demonstrated a competitive WER of 0.142 and offered the necessary flexibility for processing extensive audio inputs, accommodating up to two hours per session.

The robustness of the Whisper model in handling diverse and challenging audio environments was a decisive factor in our selection process. This capability is particularly pertinent for our dataset, which includes YouTube videos and podcasts often embedded with background noise and music. Whisper's proficiency in such scenarios has been substantiated through rigorous evaluations against leading commercial and open-source ASR systems [21]. These evaluations highlighted Whisper's ability to surpass the performance of the best open-source model, NVIDIA STT, across multiple datasets. Furthermore, it competes effectively with commercial ASR systems, showcasing its versatility and reliability in various transcription contexts. This comprehensive performance assessment, as evidenced in Figure 1, affirmed our decision to utilize Whisper Large-v2 for our dataset's transcription needs.

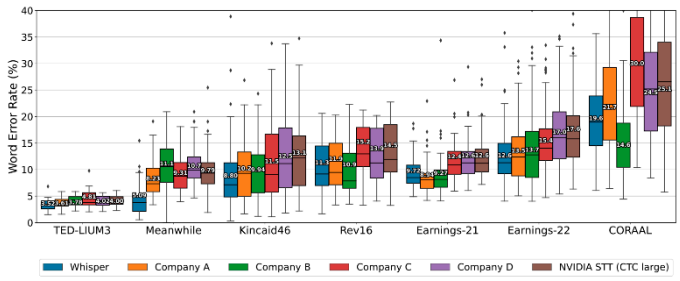

Figure 1. Comparative Performance of Whisper in Long-Form Transcription Against Leading ASR Systems [21].

4. Feature Extraction

The preparation of audio data for our ASR system involved comprehensive preprocessing and feature extraction. Using the librosa library, we implemented steps such as noise reduction, silence trimming, resampling, and compression to ensure the audio quality met in transcription requirements. The process also incorporated the Whisper Feature Extractor class from Hugging Face's transformers module. This stage focused on adjusting parameters like chunk length, feature size, hop length, and sampling rate, aligning them with the specific demands of our research. These modifications were essential for transforming raw audio into a format suitable for machine learning models, laying the groundwork for model training and evaluation in our study.

 doi:10.20944/preprints202403.0619.v1

B. Implementation

1. Fine Tuning and Evaluations

Prior to training the Whisper Large-V2 model, a thorough assessment of available hardware resources was conducted. This assessment focused on the GPU's memory capacity (14.61 GB), computing power, and system RAM. The GPU memory, in particular, was closely monitored to mitigate any potential constraints during the training phase.

2. Loading the Model with 8-Bit Precision

To enhance memory efficiency, the Whisper model was loaded using an 8-bit quantization for weights. This method significantly reduced memory requirements while maintaining a balance between memory conservation and numerical precision.

3. Trainable Parameters and LORA Optimization

The initial model configuration presented approximately 74.5 million trainable parameters. To adapt to our hardware limitations, Low-Rank Adaptation (LORA) techniques were applied, effectively reducing the number of trainable parameters to approximately 15.7 million. This adaptation was crucial for managing the model's complexity and ensuring efficient training.

4. Text Normalization

The BasicTextNormalizer module from the Transformers library was utilized for text normalization. This process involved standardizing the text through lowercasing, punctuation removal, and space normalization. Such normalization was essential for consistent evaluation and comparison of Word Error Rate (WER) across different datasets.

5. Fine-Tuning Process

The Seq2Seq model's training parameters were finely tuned using the Seq2SeqTrainingArguments from the Transformers library. Parameters including batch size, learning rate, and warmup steps were strategically selected to optimize the model's training performance.

6. Deployment of a Speech Recognition Application with Flask and Nginx

The speech recognition application was deployed on an Amazon Web Services (AWS) server. This deployment involved setting up the server environment, installing necessary dependencies, and establishing a secure SSH connection for file transfer. The application utilized Flask for backend management and Nginx for frontend integration.

7. Application Architecture

The application's architecture facilitated audio recording via a web browser, processing on a remote server, and real-time display of transcriptions. Key features included an interactive user interface, seamless integration of Flask and Nginx, and a custom transcription model. The design focused on user-friendliness, allowing for straightforward audio recording and immediate transcription visualization.

III. Results

Our comprehensive fine-tuning process of the Whisper Large-v2 model revealed significant insights into the model's performance under various training configurations. Initially, we meticulously configured the training parameters using the Seq2SeqTrainingArguments object from the Transformers library. Key parameters such as batch size, learning rate, and warmup steps were strategically selected to optimize the model's performance. Notably, we employed a learning rate of 10^{-3}, which emerged as the most effective in enhancing the model's accuracy, as evidenced by a

 doi:10.20944/preprints202403.0619.v1

notable decrease in both Word Error Rate (WER) and Normalized Word Error Rate (Normalized_WER).

The impact of different learning rates on the model's performance was systematically evaluated (Table 4). Our findings indicated that a learning rate of 10^{-3} led to promising results, particularly with a Normalized_WER of 18.774% and a Normalized Character Error Rate (CER) of 12.103%. Additionally, the application of text normalization techniques significantly reduced the WER, underscoring the importance of this preprocessing step.

Table 4. Learning Rate Performance Table.

Learning Rate	WER (%)	Normalized WER (%)	CER (%)	Normalized CER (%)	Train Loss	Validation Loss
10^{-2}	50.634	45.321	26.198	22.566	5.824	5.533
10^{-3}	25.781	18.774	15.337	12.103	0.378	0.510
10^{-4}	43.634	35.321	26.198	22.566	2.824	2.533

Further experiments were conducted to assess the influence of warmup steps on model performance. The results showed that a warmup step of 1000 yielded the best performance, achieving a Normalized_WER of 18.504% (Table 5). This optimal configuration was subsequently adopted for further experimentation.

Table 5. Warmup Steps Performance Table.

Warmup Steps	Normalized WER (%)	Normalized CER (%)	Train Loss	Eval Loss
250	21.563	18.235	0.499	0.667
500	18.774	15.337	0.288	0.510
750	18.720	15.349	0.270	0.499
1000	18.504	15.320	0.235	0.487
1250	23.757	19.813	0.507	0.689

In addition to these parameters, we delved into the optimization settings, particularly focusing on the Adam epsilon parameter. The optimizer's configuration plays a crucial role in the model's ability to converge to an optimal solution. Our experiments with different Adam epsilon values revealed significant variations in performance, as summarized in Table 6. This table illustrates how subtle changes in optimizer settings can markedly influence the model's effectiveness, guiding us to select the most suitable configuration for our specific needs.

Table 6. Optimizer Parameter Performance.

Adam_epsilon	Normalized WER (%)	Train Loss	Eval Loss
$\mathbf{10^{-7}}$	18.273	0.229	0.477
$\mathbf{10^{-9}}$	17.640	0.219	0.457

The exploration of the LORA model's 'R' parameter further demonstrated the impact of model configuration on performance. The configuration with R=42 improved performance, indicating its effectiveness in enhancing the model's transcription accuracy (Table 7).

 doi:10.20944/preprints202403.0619.v1

Table 7. LORA Configuration Performance.

LORA Configuration (R)	Normalized WER (%)	Train Loss	Eval Loss
R=42	17.660	0.223	0.467
R=52	20.298	0.491	0.649

The fine-tuning of the Whisper Large-v2 model yielded significant improvements in transcription accuracy. The process culminated in a final Word Error Rate (WER) of 17.121%, accompanied by a training loss of 0.210 and a validation loss of 0.448.

These metrics demonstrate the effectiveness of our fine-tuning strategy in enhancing the model's performance for medical radiology term transcription. The progression of the training and validation loss over the fine-tuning period is illustrated in the Loss Curve Figure (Figure 2), providing a visual representation of the model's learning trajectory.

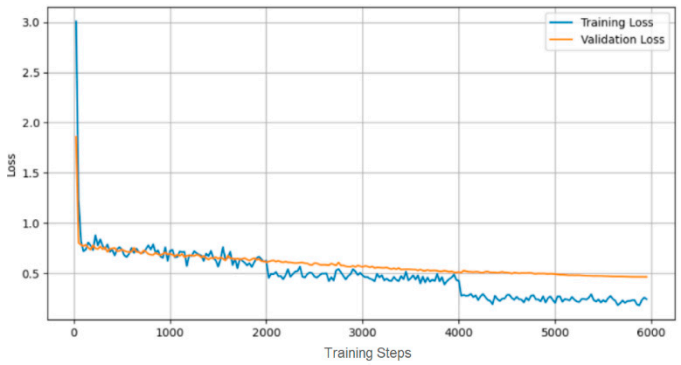

Figure 2. Loss Curve.

Concurrently, we deployed the "WhisperMed Radiology Transcriber," a speech recognition application, on an Amazon Web Services (AWS) server. This application utilizes the fine-tuned Whisper Large-v2 model to provide high-accuracy transcription of medical radiology terms. Key features of the application include real-time transcription capabilities and an intuitive user interface, designed to meet the specific needs of medical professionals.

IV. Discussion

Our study's integration of the Whisper Large-v2 model into radiological applications marks a significant advancement in medical Automatic Speech Recognition (ASR) systems. Demonstrating high accuracy in transcribing complex medical terminology, the model's effectiveness across diverse audio environments is a testament to its adaptability in various medical settings. This adaptability is crucial, considering the acoustic complexities inherent in different medical fields. The success of AI-driven speech recognition systems in both general healthcare communication and specialized areas like radiation oncology ([22,23]) underscores their potential to revolutionize medical data processing across a spectrum of clinical contexts [24].

In clinical practice, the application of our ASR system holds immense promise. The traditional process of transcribing diagnostic reports is often fraught with human error and inefficiency. By enhancing the accuracy and efficiency of medical documentation, our system stands to significantly improve the quality of patient care, as accurate records are vital for effective treatment planning [25]. Additionally, integrating ASR systems with Electronic Health Records (EHRs) could transform healthcare data management, reducing the administrative load on medical professionals and enabling a greater focus on patient care [26].

However, the implementation of ASR in healthcare is challenging. The system must navigate a vast array of medical terminologies, accents, and speech nuances. Our research represents progress

 doi:10.20944/preprints202403.0619.v1

in this area, but ongoing refinement is essential to meet the stringent accuracy requirements of medical data transcription [27]. Addressing these challenges, particularly in non-English languages, remains a key area for future development. Studies on language-specific medical ASR solutions, such as those in Korean and French, highlight both the challenges and opportunities in creating effective multilingual medical ASR systems [22–24].

Beyond its clinical applications, our ASR system offers significant benefits in medical education. By facilitating the transcription of educational materials, it enhances accessibility and inclusivity, particularly for non-native speakers. This aligns with the digitalization trend in medical education, where technology is increasingly pivotal in enriching learning experiences [27].

Future research avenues are abundant. Tailoring ASR systems to specific medical specialties or languages could greatly expand their utility. Exploring integration with voice-activated medical devices and telemedicine platforms presents opportunities to further leverage ASR technology in healthcare [28].

Our study, despite its successes, encountered limitations due to resource constraints, which restricted our dataset size and prolonged the training period. Future studies should aim to utilize larger datasets and more robust computational resources to improve accuracy and efficiency. Real-world testing in clinical settings is also crucial to assess the system's practicality and identify areas for improvement.

The findings of our research contributes significantly to the medical ASR field, particularly in radiology transcription. The potential impact of our work on clinical practice, healthcare efficiency, and medical education underscores the vital role of technology in advancing healthcare solutions. Addressing the limitations identified, such as dataset diversity and practical application, will be essential in future research to fully realize the potential of ASR systems in healthcare.

V. Conclusions

Our study's development of an Automatic Speech Recognition (ASR) system, specifically designed for radiological applications, represents a significant advancement in the application of technology within the healthcare sector. The successful integration of the Whisper Large-v2 model into our ASR system has led to a notable achievement: a Word Error Rate (WER) of 17.121%. This achievement underscores the system's proficiency in accurately transcribing complex medical terminology and adapting to diverse accents, which are critical in radiological contexts.

The practical implications of our research are particularly significant in clinical settings. By automating the transcription of diagnostic reports, our ASR system addresses a key challenge in radiology – the need for accurate and efficient documentation. This improvement is not just a matter of convenience; it plays a vital role in enhancing patient care by supporting informed decision-making based on precise and reliable medical records.

Moreover, the potential integration of our ASR system with Electronic Health Records (EHRs) could be a game-changer in healthcare administration. Such integration promises to streamline data entry processes, reduce the administrative burden on healthcare professionals, and improve the accuracy of patient records. This aligns with the broader goal of effective healthcare delivery, where accuracy and efficiency are paramount.

While our study has achieved its primary objectives, it also highlights areas for future exploration. The potential of tailoring ASR systems to specific medical specialties or languages, and integrating them with voice-activated medical devices and telemedicine platforms [29], presents exciting avenues for expanding the utility and impact of ASR in healthcare.

Despite its successes, our study faced limitations, primarily due to resource constraints. These limitations necessitated a training dataset of 20,000 examples and extended the training period to 14 days. Future research could benefit from larger datasets and more advanced computational resources to further enhance the accuracy and efficiency of ASR systems. Real-world testing in clinical environments is also crucial to validate the practical applicability of our system and to identify areas for improvement.

In summary, our research makes a significant contribution to the field of medical ASR systems, particularly in radiology. It offers a robust and efficient tool for medical transcription, with the potential to significantly impact clinical practice and healthcare efficiency. Our findings pave the way for future innovations in technology-enhanced healthcare solutions.

References

1. J. Zapata and A. S. Kirkedal, Assessing the performance of automatic speech recognition systems when used by native and non-native speakers of three major languages in dictation workflows, in *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)*, 2015, pp. 201–210. Accessed: Dec. 18, 2023. [Online]. Available: https://aclanthology.org/W15-1825.pdf
2. Y. Jiang and C. Poellabauer, A Sequence-to-sequence Based Error Correction Model for Medical Automatic Speech Recognition, in *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, IEEE, 2021, pp. 3029–3035. Accessed: Dec. 18, 2023. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9669554/
3. A. Salimbajevs and J. Kapočiūtė-Dzikienė, Automatic Speech Recognition Model Adaptation to Medical Domain Using Untranscribed Audio, in *Digital Business and Intelligent Systems*, vol. 1598, M. Ivanovic, M. Kirikova, and L. Niedrite, Eds., in Communications in Computer and Information Science, vol. 1598. , Cham: Springer International Publishing, 2022, pp. 65–79. doi: 10.1007/978-3-031-09850-5_5.
4. M. Zielonka *et al.*, A survey of automatic speech recognition deep models performance for Polish medical terms, in *2023 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA)*, IEEE, 2023, pp. 19–24. Accessed: Dec. 18, 2023. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10274442/
5. A. Mroz, Seeing how people hear you: French learners experiencing intelligibility through automatic speech recognition, *Foreign Language Annals*, vol. 51, no. 3, pp. 617–637, Sep. 2018, doi: 10.1111/flan.12348.
6. G. Chatzoudis, M. Plitsis, S. Stamouli, A.-L. Dimou, A. Katsamanis, and V. Katsouros, Zero-Shot Cross-lingual Aphasia Detection using Automatic Speech Recognition. arXiv, Apr. 01, 2022. Accessed: Dec. 18, 2023. [Online]. Available: http://arxiv.org/abs/2204.00448
7. M. Sunkara, S. Ronanki, K. Dixit, S. Bodapati, and K. Kirchhoff, Robust Prediction of Punctuation and Truecasing for Medical ASR. arXiv, Jul. 11, 2020. Accessed: Dec. 18, 2023. [Online]. Available: http://arxiv.org/abs/2007.02025
8. M. L. Scholz, H. Collatz-Christensen, S. N. F. Blomberg, S. Boebel, J. Verhoeven, and T. Krafft, Artificial intelligence in Emergency Medical Services dispatching: assessing the potential impact of an automatic speech recognition software on stroke detection taking the Capital Region of Denmark as case in point, *Scand J Trauma Resusc Emerg Med*, vol. 30, no. 1, p. 36, Dec. 2022, doi: 10.1186/s13049-022-01020-6.
9. C. Hacking, H. Verbeek, J. P. Hamers, and S. Aarts, The development of an automatic speech recognition model using interview data from long-term care for older adults, *Journal of the American Medical Informatics Association*, vol. 30, no. 3, pp. 411–417, 2023.
10. E. Sezgin, B. Oiler, B. Abbott, G. Noritz, and Y. Huang, 'Hey Siri, Help Me Take Care of My Child': A Feasibility Study With Caregivers of Children With Special Healthcare Needs Using Voice Interaction and Automatic Speech Recognition in Remote Care Management, *Frontiers in Public Health*, vol. 10, pp. 366, 2022.
11. L. F. Donnelly, R. Grzeszczuk, and C. V. Guimaraes, 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*, Elsevier, 2022, pp. 176–181. Accessed: Dec. 18, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0887217122000191
12. M. Vatandoost and S. Litkouhi, The future of healthcare facilities: how technology and medical advances may shape hospitals of the future, *Hospital Practices and Research*, vol. 4, no. 1, pp. 1–11, 2019.
13. [13] J. Ruby *et al.*, Automatic Speech Recognition and Machine Learning for Robotic Arm in Surgery, *Am. J. Clinical Surgery*, vol. 2, pp. 10-18, 2020.
14. [14] J. Schulte *et al.*, Automatic speech recognition in the operating room, *Ann. Med. Surg. (London)*, vol. 59, pp. 81-85, 2020.
15. J. A. Brink, R. L. Arenson, T. M. Grist, J. S. Lewin, and D. Enzmann, Bits and bytes: the future of radiology lies in informatics and information technology, *Eur Radiol*, vol. 27, no. 9, pp. 3647–3651, Sep. 2017, doi: 10.1007/s00330-016-4688-5.
16. S. Latif, J. Qadir, A. Qayyum, M. Usama, and S. Younis, Speech technology for healthcare: Opportunities, challenges, and state of the art, *IEEE Reviews in Biomedical Engineering*, vol. 14, pp. 342–356, 2020.
17. V. Kumar, H. Singh, and A. Mohanty, Real-Time Speech-To-Text/Text-To-Speech Converter with Automatic Text Summarizer Using Natural Language Generation and Abstract Meaning Representation, *International Journal of Engineering and Advanced Technology (IJEAT)*, vol. 9, pp. 2361–2365, 2020.

18. M. Dua, Akanksha, and S. Dua, Noise robust automatic speech recognition: review and analysis, *Int J Speech Technol*, vol. 26, no. 2, pp. 475–519, Jul. 2023, doi: 10.1007/s10772-023-10033-0.
19. A. Rista and A. Kadriu, Automatic Speech Recognition: A Comprehensive Survey, *SEEU Review*, vol. 15, no. 2, pp. 86–112, Dec. 2020, doi: 10.2478/seeur-2020-0019.
20. T. Raclin *et al.*, Combining Machine Learning, Patient-Reported Outcomes, and Value-Based Health Care: Protocol for Scoping Reviews, *JMIR Research Protocols*, vol. 11, no. 7, p. e36395, 2022.
21. 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. 28492–28518. Accessed: Dec. 20, 2023. [Online]. Available: https://proceedings.mlr.press/v202/radford23a.html
22. S.-J. Chun, J. B. Park, H. Ryu, and B.-S. Jang, Development and benchmarking of a Korean audio speech recognition model for Clinician-Patient conversations in radiation oncology clinics, *International Journal of Medical Informatics*, vol. 176, p. 105112, Aug. 2023, doi: 10.1016/j.ijmedinf.2023.105112.
23. T. Zhang and T. Feng, Application and technology of an open source AI large language model in the medical field, *Radiology Science*, vol. 2, pp. 96–104, Dec. 2023, doi: 10.15212/RADSCI-2023-0007.
24. X. Jia, J. A. M. Cunha, and Y. Rong, Artificial intelligence can overcome challenges in brachytherapy treatment planning, *Journal of applied clinical medical physics*, vol. 23, no. 1, 2022, Accessed: Dec. 24, 2023. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8803284/
25. S. Cruz Rivera *et al.*, Patient-reported outcomes in the regulatory approval of medical devices, *Nature medicine*, vol. 27, no. 12, pp. 2067–2068, 2021.
26. M. M. Elsokah and A. R. Zerek, Design and development Intelligent Medical Care Bed Using Voice Recognition, in *2022 IEEE 2nd International Maghreb Meeting of the Conference on Sciences and Techniques of Automatic Control and Computer Engineering (MI-STA)*, May 2022, pp. 299–304. doi: 10.1109/MI-STA54861.2022.9837521.
27. A. Czyzewski, Optimizing medical personnel speech recognition models using speech synthesis and reinforcement learning, *The Journal of the Acoustical Society of America*, vol. 154, no. 4_supplement, pp. A202–A203, 2023.
28. M. H. Davari, T. Kazemi, and M. Saberhosseini, The status of Clinical education in ophthalmology surgery ward of Vali-e-Asr Hospital affiliated with Birjand University of Medical Science before and after intervention, *Journal of Surgery and Trauma*, vol. 6, no. 1, 2018, Accessed: Dec. 24, 2023. [Online]. Available: https://jsurgery.bums.ac.ir/article-1-127-.pdf
29. A. E. Chung, A. C. Griffin, D. Selezneva, and D. Gotz, Health and Fitness Apps for Hands-Free Voice-Activated Assistants: Content Analysis, *JMIR mHealth and uHealth*, vol. 6, no. 9, p. e9705, Sep. 2018, doi: 10.2196/mhealth.9705.