KinyarMedASR: Enhancing Kinyarwanda Medical Speech Recognition with Transformer-Based Spell Correction

Igiraneza Ishimwe Nicole
Information Technology
Carnegie Mellon University
nii@andrew.cmu.edu

Nishimwe Rebecca
Carnegie Mellon University
Electrical and Computer Engineering
nrebecca@andrew.cmu.edu

Tuyishimire Neza David
Information Technology
Carnegie Mellon University
dtuyishi@andrew.cmu.edu

Uwamahoro Joyeuse
Information Technology
Carnegie Mellon University
juwamaho@andrew.cmu.edu

Abstract

Healthcare workers in Africa face significant documentation challenges, particularly in regions with severe staffing shortages where there are only 1.55 health workers per 1,000 people—far below the WHO-recommended 4.45. Automatic Speech Recognition (ASR) technology offers a promising solution to streamline medical documentation, but its application in African languages remains limited. This paper explores the adaptation of the Whisper-Small-Kinyarwanda model for medical documentation in Kinyarwanda healthcare settings. While the baseline model showed a 31.36% Word Error Rate (WER) and 5.15% Character Error Rate (CER) on a specialized medical dataset, our enhanced model, KinyarMedASR, reduced the WER to 24.34% through the integration of a medical spell correction layer. Despite this improvement, the CER increased to 10.62%, highlighting challenges in character-level precision for medical terminology. These results underscore the potential for domain-specific adaptation to improve ASR systems for African languages, particularly in specialized fields such as healthcare, while also emphasizing the need for further refinements to enhance transcription accuracy. This work highlights critical gaps in current ASR technology for African healthcare settings and provides valuable insights for developing robust, domain-specific speech recognition solutions to support healthcare workers in their daily documentation tasks.

1 Introduction

The medical field is critical to human life, and healthcare professionals often face overwhelming workloads, especially in regions with a shortage of trained staff In Africa, where there are only 1.55 health workers per 1,000 people—far below the WHO-recommended 4.45[1]—this issue is even more pronounced. ASR technology offers a promising opportunity for healthcare professionals to streamline their workflows. For example, doctors and nurses could dictate their notes while examining patients, allowing the system to automatically transcribe spoken words into text. This hands-free approach would save time, reduce the administrative burden of manual documentation, and enable healthcare providers to focus more on patient care. ASR has already demonstrated its ability to increase efficiency and productivity in industries such as call centers and customer service, and its

11-785 Introduction to Deep Learning (Fall 2024) Final Project Report.

potential to transform healthcare could be significant [2]. However, these systems exhibit marked degradation in performance when confronted with African languages and accents[2] in the healthcare settings. This discrepancy stems from the lack of diverse, high-quality health-care datasets for African languages and accents, which leaves existing ASR models inadequately trained for these contexts. Our research seeks to address this gap.

2 Literature Review

Recent advancements in Automatic Speech Recognition (ASR) technology have revolutionized various industries, yet their application in African healthcare settings remains challenging. As healthcare workers in Africa face overwhelming workloads with only 1.55 health workers per 1,000 people far below the WHO-recommended 4.45 there is a pressing need for efficient documentation solutions [3].

A significant breakthrough in multilingual ASR came with OpenAI's Whisper model [4], which demonstrated exceptional performance across multiple languages through its innovative "large-scale weak supervision" training approach. Of particular relevance is the Whisper-Small-Kinyarwanda adaptation, which achieved a promising 24% Word Error Rate (WER) in transcribing Kinyarwanda audio to text. This adaptation, fine-tuned on the Kinyarwanda common-voice dataset, represents a significant step forward in making ASR technology accessible for Kinyarwanda speakers. However, its performance in specialized medical contexts remains unexplored.

Building on this foundation, Chanie et al. [5] developed a comprehensive multilingual ASR system specifically for East African languages, including Kinyarwanda. Their work demonstrated impressive results with a WER of 25.48% for Kinyarwanda through careful data curation and validation. While promising, their research focused on general language usage rather than domain-specific applications like healthcare.

The challenge of limited data resources has led researchers to explore innovative approaches. Mohamud et al. [6] investigated rapid development of ASR systems for African languages using self-supervised learning. Their research demonstrated that pre-training models on large amounts of raw speech data is crucial for developing efficient ASR systems in low-resource conditions. This finding is particularly relevant for medical ASR in African languages, where labeled medical data is extremely scarce.

In healthcare applications specifically, Johnson et al. [7] provided valuable insights through their systematic review of speech recognition technology in healthcare settings. Their work revealed that while ASR can significantly reduce documentation time and costs, several factors need careful consideration for successful implementation, particularly in specialized medical domains. These include:

- Accurate recognition of medical terminology
- Handling of domain-specific accents and pronunciations
- Integration with medical workflows
- Adaptation to various medical specialties

Recent work by Ritchie et al. [8] further advances the field by combining multilingual modeling with self-supervised learning approaches. Their experiments with 15 African languages demonstrated that pooling available data and using pre-training techniques can significantly improve recognition quality, though they note that high-quality data availability remains a limiting factor.

Gap Analysis: The current literature reveals several critical gaps in medical ASR for African languages:

1. Domain Adaptation: While general-purpose ASR systems like Whisper show promising results for Kinyarwanda, their performance in medical contexts remains untested. Medical terminology, pronunciation patterns, and documentation structures differ significantly from general language use.
2. Medical Data Scarcity: There is a severe lack of medical speech datasets in African languages, particularly Kinyarwanda. This shortage hampers the development of domain-specific ASR systems for healthcare applications.

3. Medical Accuracy Requirements: While current WER rates might be acceptable for general use, medical documentation requires significantly higher accuracy to ensure patient safety and proper care delivery.

3 Model Description

Our KinyarMedASR system builds upon the Whisper-Small-Kinyarwanda model, adding and extra transformer architecture specifically designed for medical domain adaptation. The system consists of two main components: the base ASR and an extra transformer specialized medical spell correction layer.

3.1 Base Architecture

The foundation of our model is the Whisper-Small-Kinyarwanda architecture, which includes:

- An encoder-decoder transformer architecture
- Multi-head attention mechanisms
- Input processing for audio features
- Token generation capabilities for Kinyarwanda language

This baseline model is built upon Open AI Whisper Model [4]. This model as exhibit state of the art capability and tho not performing as the foundational Whisper model in other popular languages such as English, its performance are promising in hard accented language such as Kinyarwanda.

3.2 Model Extension

We borrow ideas from prior work such as [10], where Dinh-Truong et al. (2021)[10] proposed a transformer-based architecture for spell correction in the Vietnamese language. These studies demonstrated the efficacy of transformers in handling sequence-to-sequence tasks, especially in low-resource settings or with domain-specific requirements,highlighting its ability to adapt to various linguistic patterns and error types.

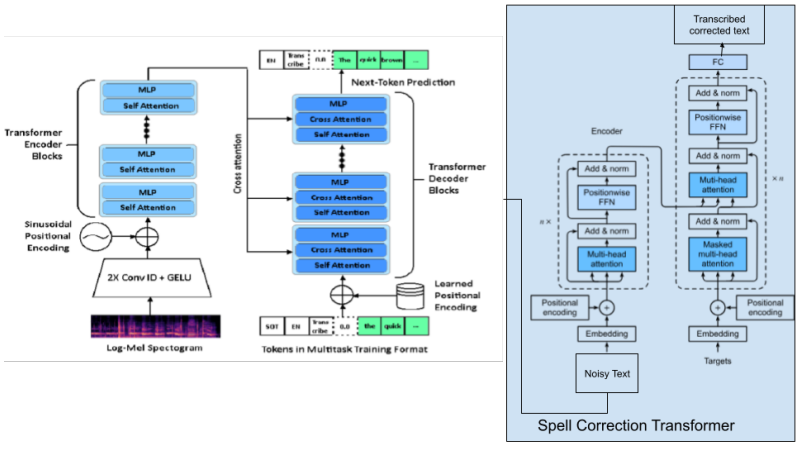


Figure 1: Model Architecture (Extended Whisper)

It is important to note that during training, the whisper model is not trained and the spell correction is trained just on the noisy data explained below.

Block	Input	Output
Whisper Encoder	Raw audio signal	Noisy transcription $Y = \{y_1, y_2, \ldots, y_T\}$
Spell Correction Encoder	Noisy transcription $Y = \{y_1, y_2, \ldots, y_T\}$	Contextual embeddings
Spell Correction Decoder	Contextual embeddings	Predicted corrected transcription $Y_{\text{pred}} = \{y_{\text{pred},1}, y_{\text{pred},2}, \ldots, y_{\text{pred},T}\}$
Loss Calculation	Predicted corrected transcription Y_{pred}	Loss based on true ground truth $Y^* = \{y_1^*, y_2^*, \ldots, y_T^*\}$

Table 1: Model Architecture: Input and Output of Each Block

4 Dataset

Since KinyarWhisper Small was fine-tuned on Mozilla Common Voice and does not perform well in the medical context, we opted to generate our own text data. We use the output of Whisper to prompt GPT-4 to generate noisy text data with spelling errors similar to those made by Whisper, but with the added benefit of providing medical context to help guide the generation. As a result, GPT-4 generates both noisy and corrected text data. This data is then used to train the spell correction layer, which will be integrated into the KinyarWhisper Small model to create a complete ASR system with spell correction. For evaluation, we record 60 audio samples in the medical domain and manually transcribe them to assess both the base model and the enhanced model's performance.

5 Evaluation Metric

The performance of automatic speech recognition (ASR) systems is commonly evaluated using two key metrics: Word Error Rate (WER) and Character Error Rate (CER). WER quantifies the proportion of incorrectly transcribed words by calculating the minimum edit distance between the predicted and reference transcripts, normalized by the total number of words in the reference. Similarly, CER measures errors at the character level, providing finer-grained insights, particularly for languages with complex orthographies or where spelling accuracy is critical. Lower values for both metrics indicate better transcription accuracy. These metrics are widely used in ASR research and development, as they effectively capture the system's ability to handle various transcription challenges [4][8]. In this study, we utilize both WER and CER to evaluate the baseline Whisper-Small-Kinyarwanda model and the enhanced KinyarMedASR system, highlighting the impact of domain-specific adaptations on transcription quality.

WER is defined as:

$$\text{WER} = \frac{S + D + I}{N} \tag{1}$$

where:

- S is the number of substitutions,
- D is the number of deletions,
- I is the number of insertions, and
- N is the total number of words in the reference transcript.

WER represents the minimum number of word-level edits required to transform the predicted transcription into the reference transcription, normalized by the length of the reference.

CER is defined as:

$$\text{CER} = \frac{S + D + I}{C} \tag{2}$$

where:

- S is the number of character substitutions,
- D is the number of character deletions,
- I is the number of character insertions, and

- C is the total number of characters in the reference transcript.

CER is similar to WER but operates at the character level, providing finer granularity in error measurement.

6 Training and Optimization

To enhance the performance of the Whisper model for medical transcription, we incorporate a secondary transformer network that performs spell correction. The encoder of this transformer takes the noisy transcription output $Y = \{y_1, y_2, \ldots, y_T\}$ from the Whisper model, while the decoder generates the corrected transcription $Y_{\text{pred}} = \{y_{\text{pred},1}, y_{\text{pred},2}, \ldots, y_{\text{pred},T}\}$, which is the predicted ground truth.

6.1 Objective Function

The training objective is to minimize the negative log-likelihood of the correct token y_t^* from the true ground truth sequence $Y^* = \{y_1^*, y_2^*, \ldots, y_T^*\}$, at each time step, conditioned on the noisy sequence Y and the preceding predicted tokens $y_{\text{pred},<t}$:

$$L = -\sum_{t=1}^{T} \log P(y_t^* \mid y_{\text{pred},<t}, Y; \theta)$$

where θ represents the parameters of the transformer model.

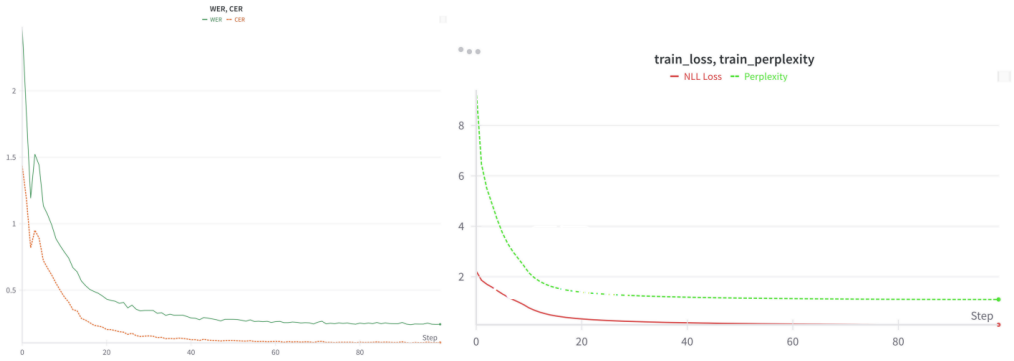
The encoder processes the noisy transcription Y to generate contextual embeddings, while the decoder uses these embeddings to generate the corrected output. During training, the loss is computed with respect to the true ground truth Y^*, and the model learns to map noisy inputs to clean outputs, effectively correcting spelling errors and improving transcription accuracy.

6.2 Role of Transformers in Spell Correction

Transformers excel in sequence-to-sequence tasks such as spell correction due to their attention mechanism [10]. The multi-head self-attention mechanism in the encoder captures long-range dependencies in the noisy transcription, enabling the model to understand context and error patterns. The cross-attention mechanism in the decoder leverages these contextual embeddings to produce accurate corrections. Additionally, the parallelizable architecture of transformers allows for efficient training on large datasets.

7 Results

The results of this study demonstrate the effectiveness of domain-specific adaptation in enhancing automatic speech recognition for medical contexts. The KinyarMedASR model achieved a marked improvement in WER, reducing it from 31.36% in the baseline model to 24.34%. This reduction underscores the model's enhanced capability to accurately transcribe medical speech in Kinyarwanda. Despite this success, the CER increased to 10.62 which may be attributed to the complexities of medical spelling patterns and the limited dataset size. Notably, the integration of a spell correction layer contributed significantly to the improved WER by addressing spelling inconsistencies in the medical domain. These findings indicate the potential of tailored ASR systems to improve transcription quality in low-resource languages, particularly in specialized domains such as healthcare.



Model	Word Error Rate (WER)	Character Error Rate (CER)
KinyarWhisper Small	31.36%	5.15%
KinyarMedASR	24.34%	10.62%

8 Discussion

The improvements in WER achieved by the KinyarMedASR model demonstrate the critical role of domain-specific adaptation in automatic speech recognition for Kinyarwanda medical contexts. The inclusion of a spell correction layer proved instrumental in addressing context-specific spelling errors, thereby enhancing overall transcription accuracy. This innovation, coupled with the use of a custom dataset, resulted in substantial reductions in WER. However, the higher CER observed suggests a trade-off between word-level accuracy and character-level precision, likely influenced by the model's handling of specialized medical terms and pronunciation variations. Additionally, the limited size of the custom medical dataset may have constrained the model's ability to generalize across diverse medical scenarios. These findings highlight the importance of expanding and diversifying datasets to overcome data scarcity challenges and further enhance the model's performance. Overall, the improved transcription capabilities of the KinyarMedASR model have important implications for streamlining medical documentation processes, thereby enabling healthcare workers to focus more on patient care.

9 Future Works

Future efforts will focus on addressing current limitations to further improve the performance and applicability of the KinyarMedASR model. Expanding the dataset with more diverse and extensive medical audio samples from various speakers, accents, and clinical environments is a priority. Another avenue for future work involves integrating the ASR system into healthcare workflows, creating a user-friendly interface to facilitate adoption by medical professionals. Extending the model's capabilities to support additional African languages commonly used in healthcare settings is also a key goal, enhancing accessibility and usability across the continent. Furthermore, incorporating advanced error-correction mechanisms could address current challenges in character-level accuracy, improving the overall transcription quality. These efforts aim to develop a more robust, domain-specific ASR system that can significantly impact medical documentation efficiency in low-resource settings.

10 Conclusion

Our KinyarMedASR model shows promising results for medical speech recognition in Kinyarwanda. By adding a spell correction system to the existing model, we reduced word errors from 31.36% to 24.34%. This improvement is important because it helps doctors and nurses spend less time on paperwork and more time with patients. We created a special dataset of medical terms and conversations in Kinyarwanda, which helped our model better understand medical language. We also made sure to protect patient privacy in all our work by not including any user information. We see the system to be used in strictly reporting of medical observation and not a comprehensive reporting tool which will involve patient information.

We hope to extend the test dataset, as well as test our approach in other contexts. If successful then this approach could be adopted in developing ASR for specific domains without worrying so much about audio data collection which is a tedious task

11 Division of Work

Rebecca, Joyeuse, Nicole, and David:

Shared Responsibilities:

Test Dataset Creation: Each team member recorded and transcribed 15 audio files into Kinyarwanda, ensuring the dataset is rich and varied. Conducted quality checks on each other's transcriptions to maintain accuracy and consistency across the dataset. Each team member did a literature review to understand available models and gaps the team can fill in present ASR.

Specific Contributions:

1. Rebecca:

Special Focus: Literature Review Coordination and Technical Documentation

Responsibilities: Lead the organization and synthesis of the literature review findings. Ensure that all sources are properly cited and relevant insights are integrated into the project documentation. Assist in preparing the final report including the methodological approach, dataset description, model performance, and discussion.

2. Joyeuse and David:

Special Focus: Baseline Model Setup

Responsibilities: Collaboratively set up the baseline model using the existing model framework. Run tests to ensure the model's functionality with the newly created dataset. Iterate on model configurations to optimize performance. Calculated Word Error Rate(WER) and Character Error Rate(CER) to validate the models performance.

3. Nicole:

Special Focus: Data Management and Technical Documentation

Responsibilities: Manage the dataset including storage, versioning, and access. Ensure that data is properly formatted and ready for use in machine learning tasks. Assist in documenting the process and results of the model tests.

Collaboration and Communication:

Team Meetings: Regular team meetings to share updates, discuss findings, and troubleshoot any issues with the dataset or model.

Peer Review: Engage in peer review sessions where each member presents their part of the work for feedback and suggestions from the rest of the team.

References

[1] A. Ahmat et al., 'The health workforce status in the WHO African Region: findings of a cross-sectional study', BMJ Glob. Health, vol. 7, no. Suppl 1, p. e008317, May 2022, doi: 10.1136/bmjgh-2021-008317.

[2] T. Olatunji et al., 'AfriSpeech-200: Pan-African Accented Speech Dataset for Clinical and General Domain ASR', Sep. 30, 2023, arXiv: arXiv:2310.00274. Accessed: Oct. 05, 2024. [Online]. Available: http://arxiv.org/abs/2310.00274.

[3] A. Ahmat et al., "The health workforce status in the WHO African Region: findings of a cross-sectional study," BMJ Glob. Health, vol. 7, no. Suppl 1, p. e008317, 2022.

[4] A. Radford et al., "Robust Speech Recognition via Large-Scale Weak Supervision," arXiv preprint arXiv:2212.04356, 2022.

[5] Y. Chanie et al., "Multilingual Automatic Speech Recognition for Kinyarwanda, Swahili and Luganda: Advancing ASR in Select East African Languages," in AfricaNLP workshop at ICLR2023, 2023.

[6] J. H. Mohamud et al., "Fast Development of ASR in African Languages using Self Supervised Speech Representation Learning," arXiv preprint arXiv:2103.08993, 2021.

[7] M. Johnson et al., "A systematic review of speech recognition technology in health care," BMC Medical Informatics and Decision Making, vol. 14, no. 94, pp. 1-14, 2014.

[8] S. Ritchie et al., "Large vocabulary speech recognition for languages of Africa: multilingual modeling and self-supervised learning," arXiv preprint arXiv:2208.03067, 2022. [9] A. Vaswani et al., "Attention is all you need," arXiv preprint arXiv:1706.03762, 2017. [Online]. Available: https://arxiv.org/abs/1706.03762. [10] D.-T. Do, H. T. Nguyen, T. N. Bui, and D. H. Vo, "VSEC: Transformer-based model for Vietnamese spelling correction," arXiv preprint arXiv:2111.00640, 2021. [Online]. Available: https://arxiv.org/abs/2111.00640.

12 Bonus Visualization

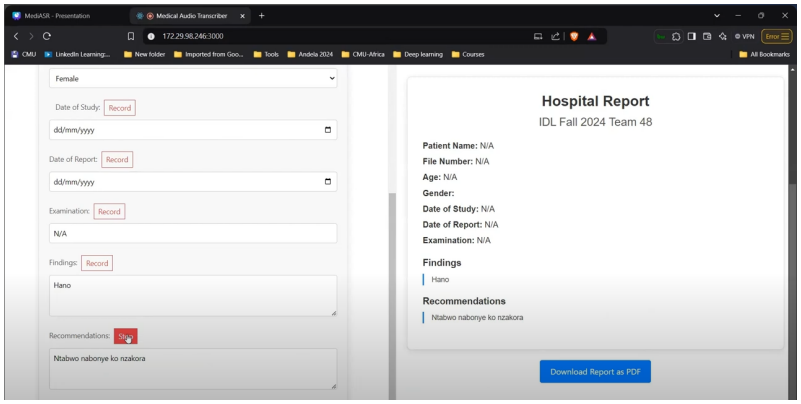


Figure 2: Visualization of Model in Action

13 Bonus Dataset

Index	Correct Text	Noisy Text
1	"Gukora imyitozo ngororamubiri no kwita ku mirire bitera imbaraga ku mubiri, bityo bigatuma umubiri ugira ubuzima bwiza kandi ukomeye. Kwita ku mirire, gukurikiza gahunda y'imyitozo ngororamubiri, no kuruhuka neza bigira ingaruka nziza ku mubiri no ku buzima bw'umutima."	"Gukora imyitozo ngororamubiri kwita no ku mirire bitera imbaraga ku muairi, bityo bigatuma umubiri ugira ubuzima bwiza kandi ukomeye. Kwita ku mirire, gukurikiza gahunda y'imyitozo ngororamubiri, no kuruhuza neza bigira ingaruka nziza ku mubiri no ku buzima bw'umutima."
2	"Kugira umubiri w'umurimo ugira ubuzima bwiza, ni ukwiga uko wita ku buzima bwawe no gukurikiza gahunda y'imyitozo ituma umubiri wawe urushaho gukomera. Imyitozo ngororamubiri igira uruhare runini mu kongera imbaraga mu mubiri no kugabanya ibyago by'indwara."	"Kugira umubiri w'umurimo ugira ubuzima bwiza, ni ukwiga uko wita ku buzima bwawe no gukurikiza y'imyitozo ituma gahunda umubiri wawe urushaho gukomera. Imyitozo ngororamubiri igira uruhare runini mu kongera imbaraga mu mubiri no kugabanya by'indwara. ibyago"
3	"Kwiyitaho no kwita ku buzima bwawe ni urugendo rurerure rwubaka imbaraga z'umubiri n'umutima. Kwitonda no kugenzura imibereho yacu buri munsi bigira uruhare mu guhangana n'indwara n'imihangayiko. Ni ngombwa ko buri wese abona igihe cyo kwiyitaho no kwita ku buzima bw'umubiri n'umutima."	"Kwiyitaho no kwita ku buzima bwawe mi urugendo rurerure rwubaka imbaraga n'umutima. z'umubiri Kwitonda no kugenzura imibereho yacu buri munsi bigira uruhare mu guhangana n'indwara n'imihangayiko. Ni ngzmbwa ko buri wese abona igihe cyo kwiyitaho no kwita buzima ku bw'umubiri n'umutima."

Table 2: Correct vs Noisy Text