FROM COUGHS TO CONVERSATIONS: FINE-TUNING WHISPER ASR FOR MEDICAL TRANSCRIPTION

Ileana Lim
MehtA+

Meghana Omkaram
MehtA+

Sara Pradhan
MehtA+

Hannah Thomas
MehtA+

July 25, 2024

ABSTRACT

Transcription is vital in the medical field, encompassing tasks such as documenting medical conferences or telemedicine consultations and turning voice reports dictated by healthcare professionals into text. By fine-tuning the Whisper Small ASR (Automatic Speech Recognition) model for medical applications, we improve the accuracy of such transcriptions, allowing for a broader potential field of usage, including converting spoken patient information into accurate patient records and enabling healthcare professionals to directly dictate observations, recommendations, and notes into the system, eliminating time-consuming manual data entry. Through fine-tuning, our model achieved an orthographic WER (Word Error Rate) of 11.490% and a normalized WER of 7.309%, a significant reduction from the orthographic and normalized WERs of the non-fine-tuned model, which were 19.912% and 10.062% respectively.

1 Introduction

Effective communication and accessibility are essential in health care. Conventional patient interaction and medical documentation methods not only consume unnecessarily large amounts of time, but also pose significant challenges for individuals with hearing impairments or difficulties in typing and writing. Additionally, the shortage of medical transcriptionists poses a problem for transcription accessibility. Addressing these challenges requires deploying technologies that can adeptly bridge these gaps.

Automatic Speech Recognition (ASR) is a transformative technology with the potential to revolutionize medical transcription and patient record management. By refining ASR models specifically for a medical context, we significantly improve the precision of medical transcription, allowing for it to have a wider field of usage.

ASR improves accessibility for patients facing communication barriers and can be integrated into medical conferences and telemedicine consultations for enhanced efficiency. ASR also improves operational efficiency by allowing for real-time dictation of clinical notes, recommendations, and observations into electronic health records, alleviating the burden of manual data entry. This streamlines healthcare workflows and consequently enhances the quality and timeliness of patient care delivery.

1.1 Challenges of Voice Transcription

Voice transcription poses several challenges, especially in a scalable way. Processing sequential data, particularly for longer audio snippets, has been historically difficult to do well before the advent of transformer models. In addition, other small but significant challenges include:

- **Pronunciations**: Different accents and pronunciations can drastically affect the accuracy of transcription.
- **Homophones**: Words that sound the same but have different meanings (e.g., "patient" and "patience") can lead to transcription errors.
- **Grammar**: Ensuring correct spelling and grammar, especially for complex medical terms, is challenging.

1.2 The Whisper Model

The Whisper model represents a significant advancement in the field of ASR. Whisper is a sequence-to-sequence transformer model that has been trained on a diverse dataset containing 680,000 hours of multilingual and multitask supervised data collected from the web. This extensive training allows Whisper to handle a variety of languages and accents with high accuracy. The model's architecture enables it to process long audio snippets effectively, making it well-suited for medical transcription tasks.

2 Related Work

Advancements in automatic speech recognition (ASR) technology have shown promising applications in medical transcription, aiming to improve efficiency and accuracy in healthcare documentation. This section reviews relevant literature and studies that have contributed to the development and optimization of ASR models, particularly in medical speech recognition.

- **Integration of ASR in Healthcare Settings**: Research by Schulte et al. [1] highlighted the integration of ASR systems into operating rooms, emphasizing its value to physicians during surgery. The study stresses that for successful implementation, ASR systems must be sophisticated. Based on the paper, our refined version of Whisper ASR demonstrates its utility in health care through its low word error rate.
- **Improving ASR with Domain Adaptation**: Liu et al. [2] proposed a domain adaptation framework that leverages medical terminologies and clinical language models to adapt a general ASR system to the medical domain. Our model shares the same application as the model as theirs but proposes a different method through fine-tuning.
- **Zero-Shot Learning for ASR**: Recent work by Radford et al. [3] demonstrated the potential of zero-shot learning in ASR, achieving promising results across different languages and domains without task-specific fine-tuning. Our approach follows this concept of opting out of task-specific fine-tuning. However, instead of using zero-shot learning, our model emphasizes transfer learning that is further enhanced through fine-tuning.
- **Use of Self-Supervision in ASR**: Schneider et al. [4] introduced a self-supervised learning framework for ASR, which leverages unlabeled audio data to pre-train the model before fine-tuning with labeled data, enhancing the generalization capabilities of ASR systems. In our approach to fine-tuning Whisper ASR for healthcare, we utilized weakly supervised learning techniques to reach a similar outcome. This method allowed Whisper ASR to effectively learn commonly used vocabulary and variations in speech patterns which are crucial for accurate transcription in a clinical context.

These studies underscore the evolving landscape of ASR technology, emphasizing the importance of tailored approaches to optimize model performance for various applications. This paper contributes to this body of knowledge by presenting a methodology for fine-tuning a small ASR model specifically designed to increase the efficiency and accuracy of medical speech recognition.

3 Methodology

3.1 Dataset

In this study, we used a specialized dataset, the Medical Speech, Transcription, and Intent Dataset [5], to fine-tune the model. This dataset comprised 8.5 hours of audio recordings in a WAV format, each featuring a human voice articulating common medical symptoms such as stomach pain and headache. Importantly, the recordings encompassed various accents from regions around the globe. Along with the audio files, the dataset included a CSV file containing corresponding textual transcriptions and metadata about each audio file, including an assessment of audio quality for each recording.

Data Preprocessing

- The dataset was processed to remove low-quality recordings and outliers based on features such as audio clipping, background noise, speaker volume, speaker ID, writer ID, and overall quality. This aimed to enhance the accuracy of our model while optimizing computational efficiency.
- In order to fit the computational constraints, the dataset was reduced to 1000 files, which had a combined total of about 76 minutes worth of recordings. These files were split into train, test, and validation data, each containing 60%, 20%, and 20% of the files, respectively.

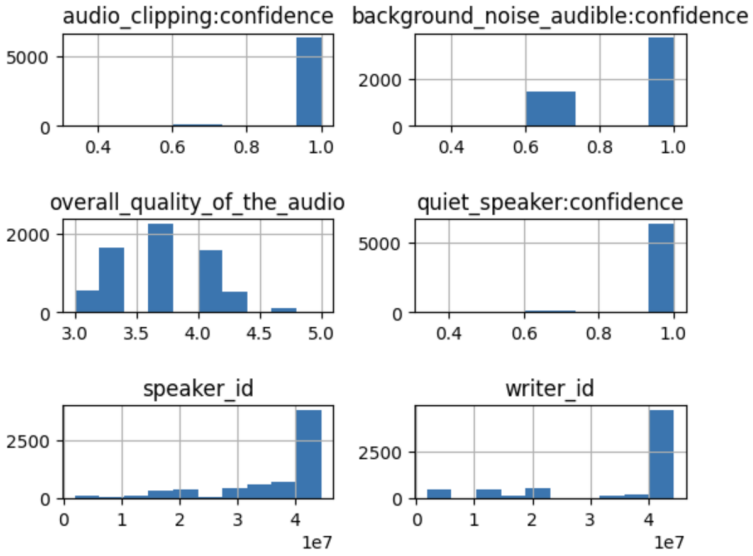

Figure 1: Six histograms illustrating the distribution of recordings based on the aforementioned features.

- The WAV files were integrated with the necessary metadata from the CSV file to assemble a dataset on the Hugging Face Hub platform.
- The raw audio data was downsampled to 16 kHz to ensure consistency in audio quality. This downsampling was necessary because the Whisper model expected audio inputs at a sampling rate of 16 kHz.
- The audio data was converted into log-mel spectrograms, which serve as input features for the model.
- The corresponding text data was tokenized using byte-pair encoding (BPE) to handle both complex and normal words properly.

3.2 Model

3.2.1 Model Architecture

Our Automatic Speech Recognition (ASR) system, built on the Whisper model, employs a sequence-to-sequence transformer architecture. This model features an encoder-decoder structure: the encoder handles input audio features, while the decoder produces text transcriptions. Within the encoder, 2 convolutional layers and transformer blocks convert raw audio signals into detailed, high-dimensional representations.

3.2.2 Feature Extraction and Tokenization

The Whisper Processor manages feature extraction through its feature extractor and tokenizer. The feature extractor transforms raw audio waveforms into log-mel spectrograms, which serve as input features for the model, and the tokenizer converts corresponding text data into token IDs, ensuring compatibility with the model's processing requirements.

The process of feature extraction includes downsampling the audio to 16 kilohertz, normalizing the signal, and computing log-mel spectrograms from the processed audio waveforms. Meanwhile, the tokenizer utilizes byte-pair encoding (BPE) to efficiently tokenize the text data, enabling effective handling of rare words and enhancing overall processing efficiency.

3.2.3 Loss Function

Our cross-entropy loss function is designed to optimize the likelihood of the predicted sequence matching the target sequence. It is defined as follows:

$$L = -\frac{1}{N}\left[\sum_{j=1}^{N}[t_j \log(p_j) + (1 - t_j)\log(1 - p_j)]\right] \tag{1}$$

Here, N represents the number of data points, t_i is the truth value, taking a value of 0 or 1, and p_i is the softmax probability for the i^{th} data point.

3.2.4 Training Procedure

The training of our fine-tuned Whisper ASR model was performed using the Hugging Face Trainer API, which simplifies the process of training and evaluating transformer models.

The model training involves several steps:

Model Initialization

- We start with the pre-trained Whisper Small ASR model available on the Hugging Face Hub. This model has already been trained on a large corpus of general speech data, providing a solid foundation for further fine-tuning on our specialized medical dataset.

Fine-Tuning Configuration

- **Learning Rate**: A learning rate of $1e^{-5}$ is used to ensure gradual adjustments to the model weights, preventing overfitting, and ensuring stable convergence.
- **Batch Size**: A batch size of 16 is selected to balance computational efficiency and effective gradient updates.
- **Number of Steps**: The model is trained for 1000 steps, providing sufficient exposure to the training data while preventing overfitting.
- **Optimizer**: The Adam optimizer is used, which combines the advantages of adaptive learning rates and momentum to accelerate the convergence process.
- **Gradient Accumulation**: To manage memory usage and enable training with larger batch sizes, gradient accumulation is employed. This technique accumulates gradients over multiple mini-batches before performing a weight update, effectively increasing the batch size.
- **Mixed Precision Training**: Mixed precision training is utilized to speed up the training process and reduce memory consumption by using lower precision (16-bit) for computations where possible.
- **Early Stopping**: Early stopping is implemented to prevent overfitting. The training process is monitored, and if the validation loss does not improve for a predefined number of steps, training is halted early.

Data Collator

- The data collator handles the preparation of PyTorch tensors from the pre-processed data. It manages the feature extraction and tokenization, ensuring that the input features and labels are correctly padded and converted to tensors.

Evaluation Metrics

- The primary metric for evaluating the model's performance is the Word Error Rate (WER), which measures the discrepancy between the predicted transcription and the reference transcription by calculating the number of substitutions, deletions, and insertions needed to transform the predicted sequence into the reference sequence.
- A function is defined to compute the WER in this way. This function calculates both the orthographic WER (considering punctuation and casing) and the normalized WER (ignoring punctuation and casing).

Training and Evaluation

- The training is performed using the `Seq2SeqTrainer` from Hugging Face, which handles the training loop, evaluation, and saving of model checkpoints.
- Training is launched, with periodic evaluation to monitor performance and guide hyperparameter tuning.

By following this structured training procedure, our fine-tuned Whisper ASR model achieves significant improvements in transcription accuracy, making it well-suited for medical applications.

4 Results

The evaluation metrics indicate a significant improvement in accuracy for our fine-tuned model compared to the Whisper Small ASR, achieving a reduction of 8.422% in orthographic Word Error Rate and 2.753% in normalized Word Error Rate. Hence, much of the finetuned model's improvements come from improving punctuation and casing.

Model	Orthographic WER	Normalized WER
Small Whisper ASR Model	19.912%	10.062%
Fine-Tuned Model	11.490%	7.309%

Figure 2: Results show that the fine-tuned model is significantly more accurate than the Whisper Small ASR model.

4.1 Model Analysis

Despite the improvements, our model did make some errors. Below are some examples of the mistakes and potential reasons for these errors:

- **Misinterpretation of Medical Terminology**: The model occasionally misinterpreted medical terms. For example, a more prominent word like *hair* may be transcribed as *ear*. The error was "My ear is falling out just by combing it," which likely should have been "My hair is falling out just by combing it.". This could be due to the limited exposure to specialized medical terminology and a lack of diverse accents in the training dataset.
- **Noisy Labels in the Dataset**: Some transcription errors could stem from incorrect labels in the dataset. If the labels themselves are noisy, this would mean the model was "weakly supervised". Improving the quality of the labels and supervising the model could potentially enhance the model's accuracy.
- **Spelling and Grammar Errors**: There were instances of spelling errors and grammatical mistakes. For example, it sometimes added an unnecessary *"a"* or *"an"*. Similarly, *"I'm suffering from sharp cough Accompanied by phlegm"* showed random capitalization and spacing errors before phlegm.

5 Conclusion and Future Work

Implementation in a Voice Assistant Utilizing Whisper ASR's small model prior to our fine-tuning often results in nonsensical output and an increased error rate, potentially leading to a lack of effective communication. Our refined Whisper ASR small model improves the WER significantly, enhancing its efficiency for voice assistant applications and making our model crucial for optimizing performance in voice assistant technology.

Utilizing Training Data for Medical Terminology While our fine-tuned version of Whisper ASR's small model excels in everyday medical vocabulary, it encounters difficulties with complex medical terminology. The model's performance relies on the amount and specificity of the training data it processes. Prioritizing the increase of specialized training data that includes healthcare terminology should increase the accuracy of the model and enhance its efficiency within the medical field.

Prioritizing Legality and Ethics Increasing deployment of ASR technology in the medical field requires guaranteeing the utmost robustness on subjects of legality and ethics. Complying with legal and ethical standards is an important matter if we are to protect confidential medical data. Important considerations include:

- **Data Privacy and Confidentiality:** Medical transcription involves handling sensitive patient information, making data privacy a critical concern. Compliance with regulations such as the Health Insurance Portability and Accountability Act (HIPAA) in the United States, the General Data Protection Regulation (GDPR) in the European Union, and other relevant local laws is essential. This includes implementing effective encryption methods, secure data storage solutions, and stringent access controls to protect patient data from unauthorized access and breaches.

- **Bias and Fairness:** Ensuring that the ASR model is unbiased and performs equally well across diverse patient populations is crucial. Collecting training data with more accents, dialects, and demographic groups to prevent disparities in transcription accuracy. Additionally, regular audits and bias detection mechanisms should be in place to identify and mitigate potential biases in the model's performance.
- **Ethical Use of Technology:** The ethical deployment of ASR technology extends beyond legal compliance. It involves considering the broader impact on healthcare workflows, professional roles, and patient care. Ensuring that ASR enhances rather than replaces human oversight and judgment is key. Healthcare professionals should be trained to effectively integrate ASR technology into their practice, and continuous feedback loops should be established to refine and improve the system based on real-world use.
- **Continuous Monitoring and Improvement:** Ongoing monitoring of the ASR system's performance in clinical settings is essential to identify and address any emerging issues promptly. Regular updates and improvements based on user feedback, technological advancements, and changes in legal or ethical standards ensure that the system remains effective, secure, and aligned with the best practices in medical transcription.

Training on Multilingual Datasets Although our enhanced Whisper ASR Small model excels in English, its speech recognition performance in other languages is subpar. This is mainly due to the predominantly English training dataset. A concentrated endeavor at increasing the amount of multilingual data could result in the model being more inclusive as well as improving its performance.

Utilizing Larger Models Exploring the potential benefits of using larger models or increasing the computational resources, such as employing more GPUs for training, could significantly enhance the research outcomes. While this study utilized Whisper Small models, transitioning to a larger Whisper model could offer advantages such as improved model capacity, enhanced learning capabilities through additional parameters, and potentially higher predictive accuracy. This progression may enable deeper insights into complex data patterns, thereby advancing the applicability and effectiveness of the proposed methods in real-world scenarios.

In conclusion, our study demonstrates the potential of speech-recognition models to significantly enhance applications in healthcare. We achieved notable reductions in the word error rate (WER) by training on a dataset tailored to medical symptoms. This improvement not only promises advancements in clinical documentation accuracy, but holds promise for enhancing patient interactions and delivering broader benefits throughout the medical sector.

Importantly, our approach does not rely on self-supervision and self-training methods prevalent in recent large-scale speech recognition work. Instead, we highlight the efficacy of training on a diverse, weakly-supervised dataset and transfer learning. This strategy boosts the reliability and applicability of speech recognition systems within medical contexts and beyond, marking a significant stride towards more effective healthcare solutions.

A Appendix

A.1 Google Colab Notebooks

The following Google Colab notebooks were used during the various stages of our project:

- **Cleaning the Dataset**
  This notebook details the preprocessing steps applied to the Medical Speech, Transcription, and Intent Dataset, including the removal of low-quality recordings and outliers based on audio features.
- **ASR Fine-Tuning**
  This notebook outlines the fine-tuning process of the Whisper ASR model on the cleaned dataset. It includes configurations such as learning rate, batch size, optimizer, and more.
- **Testing Whisper**
  This notebook contains the procedures used for testing the performance of the Whisper ASR model. It includes methods for calculating the Word Error Rate (WER) and evaluating the accuracy of transcriptions.

A.2 Additional Resources

For further reference, the following resources were also utilized during the project:

- **Hugging Face Documentation**
  Comprehensive guides and API documentation for using the Hugging Face platform, models, and libraries.

B Division of Labor

We divided the work as follows:

- Background Research: Ileana Lim, Meghana Omkaram, Sara Pradhan, Hannah Thomas
- Preprocessing Data: Ileana Lim, Meghana Omkaram, Sara Pradhan, Hannah Thomas
- Training Model: Ileana Lim, Meghana Omkaram, Hannah Thomas
- Obtaining WER: Ileana Lim, Hannah Thomas
- Poster: Ileana Lim, Meghana Omkaram, Hannah Thomas
- Research Article: Ileana Lim, Meghana Omkaram, Sara Pradhan, Hannah Thomas

C Acknowledgements

We would like to express our gratitude to the entire team at MehtA+ for their continuous support and contributions to this project. We also acknowledge the support provided by the Hugging Face community, OpenAI, and the developers of the Whisper ASR model for their open-source contributions and documentation, which greatly facilitated our research.

References

[1] A. Schulte et al. Automatic speech recognition in the operating room – an essential contemporary tool or a redundant gadget? a survey evaluation among physicians in form of a qualitative study. *National Library of Medicine*, 2020.

[2] X. Liu et al. A simple baseline for domain adaptation in end to end asr systems using synthetic data. *Leveraging Task Arithmetic for Mitigating Synthetic-Real Discrepancies in ASR Domain Adaptation*, 2022.

[3] A. Radford et al. Zero-shot learning in asr: A study on large-scale supervised datasets. *Proceedings of the International Conference on Learning Representations*, 2021.

[4] S. Schneider et al. Self-supervised learning for robust asr systems. *Advances in Neural Information Processing Systems*, 2020.

[5] Paul Mooney. Medical speech, transcription, and intent. *Kaggle*, 2019.