MedTalkAI: Assisted Anamnesis Creation With Automatic Speech Recognition

Yanna Torres Gonçalves, João Victor B. Alves, Breno Alef Dourado Sá, Lazaro Natanael da Silva, José A. Fernandes de Macedo, Ticiana L. Coelho da Silva

[1] Insight Data Science Lab
Universidade Federal do Ceará (UFC) – Fortaleza, CE – Brazil

Abstract. *Conventional approaches to documenting patient medical histories are often time-consuming and require significant healthcare professional involvement. This paper introduces MedTalkAI, which integrates ASR models, including Whisper and Wav2Vec 2.0, to transcribe audio recordings of patient histories in Brazilian Portuguese efficiently. MedTalkAI validates, corrects, and evaluates transcriptions, facilitating the creation of a unique medical audio-text database. Additionally, MedTalkAI enhances ASR models for medical applications using language models. This approach aims to improve medical history transcription and analysis, contributing to the development of more reliable ASR models and automating the documentation process.*

1. Introduction

Collecting medical histories through traditional methods is often inefficient, consuming a significant portion of healthcare professionals' time. As highlighted by Hapvida NotreDame Intermédica[1], up to 50% of a consultation can be devoted to this task. This can lead to medical histories that are insufficiently detailed, resulting in potential inaccuracies or incomplete diagnoses.

Moreover, the inconsistency in the format, structure, and content of medical histories can hinder the comparison and sharing of information among healthcare providers. Professionals may also struggle to remember all the pertinent details provided by patients during consultations [Gür 2012]. Addressing these issues requires solutions that enhance efficiency, ensure precise details, standardize formats, and improve access to relevant information in the medical history collection process.

Automatic Speech Recognition (ASR) methods aim to convert speech signals into textual representations[Besacier et al. 2014]. Pre-trained audio encoders like Wav2Vec 2.0 [Baevski et al. 2020, Schneider et al. 2019] and Jasper [Li et al. 2019] excel in generating high-quality speech representations. However, when it comes to specific domains, they often lack specialized capabilities, consequently, a fine-tuning stage is necessary to optimize their performance. However, this process is complex and usually requires the expertise of a qualified professional. Another constraint is the need for a comprehensive dataset that includes pairs of audio and text specific to the domain and language, particularly for medical histories in Brazilian Portuguese, which is the focus of this work.

Ideally, ASR models should perform reliably across diverse domains without extensive supervised fine-tuning for each deployment scenario. Large language models

[1] `https://www.hapvida.com.br/site/`

(LLMs) such as Whisper [Radford et al. 2023] and AudioPALM [Rubenstein et al. 2023] offer a promising solution. By leveraging knowledge learned from training on multiple datasets, these models can generalize effectively to new, related datasets. This approach minimizes the need for extensive fine-tuning, enhancing their versatility and efficiency across various practical applications.

To explore relevant ASR solutions, we created a benchmark using real medical histories to test models like Whisper and others. The results revealed challenges with clinical terms in Brazilian Portuguese, highlighting the need for an audio-text database and continuous model fine-tuning. In this paper, we present MedTalkAI[2], a tool that transcribes doctors' audio reports of patient anamnesis into text. It integrates ASR models like Whisper and Wav2Vec 2.0, the latter with an n-gram language model for error correction. MedTalkAI streamlines consultations and helps build a medical audio-text database for fine-tuning ASR models with specialized vocabulary.

In the literature, there are several NLP annotation tools, such as [Li et al. 2021, da Silva et al. 2019], and commercial tools like Prodigy[3] and Tagtog[4]. They offer user-friendly interfaces, active learning, and evaluation features. However, to the best of our knowledge, no equivalent tool exists for ASR models, highlighting the novelty of our approach.

2. Background and Related Works

In this section, we provide an overview of the main ASR approaches used in this work.

Wav2Vec 2.0 bypasses the problem of training a model with huge datasets [Baevski et al. 2020]. It is trained with limited amounts of labeled data. Wav2Vec 2.0 jointly learns discrete speech units with contextualized representations. The model architecture comprises a feature encoder that receives the raw waveform as input and feeds several blocks containing a temporal convolution followed by layer normalization and a GELU activation function. The output of the feature encoder is fed to a context network that follows a Transformer architecture.

Wav2Vec 2.0 + KenLM [Sullivan et al. 2022] integrates a 4-gram language model into Wav2Vec 2.0's decoding process using KenLM [Heafield 2011], which is more efficient in memory and CPU usage than SRILM [Stolcke 2002]. This addition reduces spelling errors and improves word sequence predictions by considering language probabilities. The results align with [Baevski et al. 2020], showing that while Wav2Vec 2.0 can work without a language model, adding one enhances performance, especially with limited audio data. Additionally, the language model relies on a text-only corpus, avoiding the need for audio fine-tuning.

Whisper [Radford et al. 2023] is a robust ASR system designed to perform reliably without fine-tuning. Unlike models like Wav2Vec 2.0, which focus on audio representation, Whisper uses an encoder-decoder Transformer for accurate transcription. It also handles language identification and translation to English. Trained on 680,000 hours of data, including 117,000 hours in 96 languages and 125,000 hours of translations, Whis-

[2] `https://youtu.be/r-swiJUqzW4`

[3] `https://prodi.gy/`

[4] `https://docs.tagtog.com/`

per delivers high-quality results across tasks without needing dataset-specific fine-tuning, making it a versatile model for multilingual applications.

3. MedTalkAI Architecture

The architecture of MedTalkAI comprises the front end developed using Next.js, a powerful and versatile JavaScript framework. This choice allows us to create a dynamic and responsive user interface that enhances the overall user experience. On the back end, we leverage Flask, a lightweight and flexible web framework implemented in Python. Flask serves as the engine that handles data processing, communication with the front end, and integration with the AI models. The core functionality, involving the implementation of ASR models, is accomplished using Python. An overview of our system is shown in Figure 1, which has five main modules:

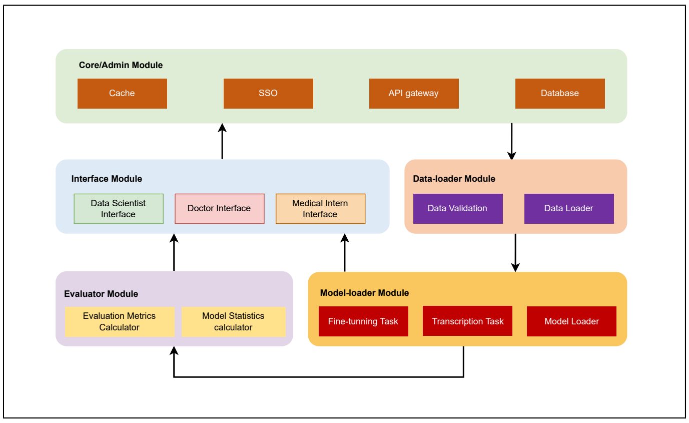


Figure 1. MedTalkAI Architecture

Core/admin module controls all data and ASR model flows, providing the gateway for other modules. Additionally, it features an administrator control panel, empowering system management and database administration. It is essential to note the three user roles within the system: doctors, data scientists, and medical interns, each with distinct access and interaction requirements, primarily handled through Single Sign-On (SSO).

Data-loader is responsible for loading data tailored to the specific requirements of each ASR model. Given that different models may have varying input data requirements, the data loader ensures that the pertinent data is correctly supplied to each model.

Model-loader The primary function of a model loader is to load different ASR models into the comparison tool. It handles the initialization and setup of model architectures, weights, and configurations.

Interface module encompasses three distinct web interfaces, each tailored to the specific roles of doctors, data scientists, and medical interns. For doctors, the interface facilitates the recording of medical histories and validating and correcting transcriptions generated by ASR models. The interface for data scientists supports choosing ASR models, evaluating transcription metrics, generating transcriptions for other models, and performing fine-tuning of ASR models. For medical interns, the interface provides tools for recording medical histories, editing pre-existing texts, and contributing to the creation of a clinical dataset by transcribing and correcting audio recordings.

Evaluator calculates a range of metrics to assess the efficacy of ASR models. Common metrics, such as Word Error Rate (WER), Cosine Similarity, and BLEU, are employed to quantify the performance of the models. The Evaluator module presents these evaluation results, providing valuable insights for the fine-tuning and optimization of ASR models. This information is instrumental in refining models and improving their overall performance. The interface associated with this module is designed for data scientists, offering a user-friendly platform to interact with and interpret the evaluation results effectively.

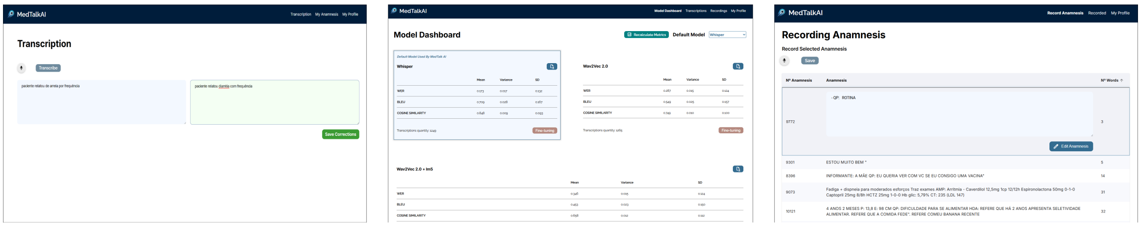
4. System Demonstration and Operation

The MedTalkAI interface serves as the access point for users to: (i) record anamneses; (ii) choose an ASR model to transcribe the recording; (iii) generate the transcription of the recorded anamnesis using an ASR model, allowing the user to validate and correct the result as needed; (iv) evaluate the metrics of the resulting transcriptions; (v) generate transcriptions using other available models; (vi) fine-tune ASR models based on the corrected transcriptions; (vii) include pre-existing medical histories; and (viii) edit the pre-existing medical histories.

The tool has three main types of users: doctor, data scientist, and medical intern. The doctor can perform tasks (i) and (iii), the data scientist can perform tasks (ii), (iv), (v), (vi), and (vii), and the medical intern can perform tasks (i) and (viii).

Scenario I: Doctor. In the MedTalkAI user interface, when a doctor logs in, they are presented with the option to either review previous transcriptions or record a new anamnesis as shown in Figure 2(a). If the user chooses to record a new one, after completion, the interface enables the doctor to assess the audio quality, ensuring there are no interruptions or cuts in the recording. If any deficiencies are detected, the doctor can choose to re-record the session for a more accurate output. Notably, the doctor is not required to manually choose an ASR model, as MedTalkAI automatically configures the model with the best statistics, simplifying this process for the user.

Upon submission of the recording, MedTalkAI automatically initiates the transcription pipeline. Once the transcription is complete, the transcribed output is displayed, allowing the doctor to review and make any necessary modifications. This approach streamlines the workflow for healthcare professionals, enabling them to focus on patient care rather than dealing with technical intricacies or the documentation of anamnesis.



Transcription

paciente relatou de arreia por frequência

paciente relatou diarréia com frequência

Model Dashboard

Default Model Used By MedTalk AI

Whisper

	Mean	Variance	SD
WER	0.173	0.017	0.132
BLEU	0.709	0.028	0.167
COSINE SIMILARITY	0.848	0.009	0.093

Transcriptions quantity: 1249

Wav2Vec 2.0

	Mean	Variance	SD
WER	0.287	0.015	0.124
BLEU	0.549	0.025	0.157
COSINE SIMILARITY	0.749	0.010	0.100

Transcriptions quantity: 1265

Wav2Vec 2.0 + lm5

	Mean	Variance	SD
WER	0.346	0.015	0.124
BLEU	0.453	0.023	0.150
COSINE SIMILARITY	0.658	0.012	0.112

Recording Anamnesis

Record Selected Anamnesis

Nº Anamnesis	Anamnesis	Nº Words ↑
9772	- QP: ROTINA	3
9301	ESTOU MUITO BEM "	5
8396	INFORMANTE: A MÃE QP: EU QUERIA VER COM VC SE EU CONSIGO UMA VACINA"	14
9073	Fadiga + dispneia para moderados esforços Traz exames AMP: Arritmia - Caverdilol 12,5mg 1cp 12/12h Espironolactona 50mg 0-1-0 Captopril 25mg 8/8h HCTZ 25mg 1-0-0 Hb glic: 5,79% CT: 235 (LDL 147)	31
10121	4 ANOS 2 MESES P: 13,8 E: 98 CM QP: DIFICULDADE PARA SE ALIMENTAR HDA: REFERE QUE HÁ 2 ANOS APRESENTA SELETIVIDADE ALIMENTAR. REFERE QUE A COMIDA FEDE". REFERE COMEU BANANA RECENTE	32

Figure 2. (a) Doctor, (b) Data Scientist, and (c) Medical Intern Visualization

Scenario II: Data Scientist. MedTalkAI provides data scientists with a model dashboard that displays all model statistics and offers the option to switch the default model to another one. Additionally, it allows for recalculating the metrics of the models if a new one is introduced or if the implementation is changed (see Figure 2(b)). Furthermore, the data scientist can access all transcriptions generated from a recording and compare them individually with the results of other models. They can also initiate the benchmark pipeline to generate transcriptions for any missing models, ensuring all models have an equal number of transcriptions.

Finally, the data scientist can upload new medical histories into the application to be recorded by medical interns and initiate the pipeline to generate transcriptions for all models based on the recordings made by the medical interns.

Scenario III: Medical Intern. After the medical intern logs into MedTalkAI, they are presented with all the medical histories that have not yet been recorded. If they encounter any grammatical mistakes in the existing text, they can edit it to ensure good quality data (see Figure 2(c)). Afterward, they can record the medical history and also review all previously recorded ones. The recordings made by the medical interns help create a clinical dataset that can be used to train and/or fine-tune models.

Experimental Results. Our benchmark consists of 397 pairs of audio and text from real patient anamneses, recorded by three medical students via MedTalkAI. The average duration of these recordings was 64.42 seconds, with the shortest at 4.38 seconds and the longest at 148.01 seconds. Medical appointments typically last around 15 minutes, with approximately half of this time (7-8 minutes) spent on manually entering medical text into systems. This highlights the potential time savings that automated transcription can offer.

Challenges in transcribing Portuguese medical terms include phonetic similarities, silent letters, and medical acronyms, which lead to inaccuracies in ASR transcriptions. While a doctor might say "the patient presents a heart rate of 90 beats per minute," they would typically document it as "HR: 90 bpm." In our benchmarks, Whisper outperformed Wav2Vec2 PT, achieving a Word Error Rate (WER) of 0.16 and a cosine similarity of 0.95, compared to Wav2Vec2's WER of 0.24 and cosine similarity of 0.93. Integrating Wav2Vec2 PT with a 5-gram model worsened its performance, resulting in a WER of 0.30 and a cosine similarity of 0.92, even though initial tests showed improvements in decoding.

5. Conclusion and Future Works

This paper addresses the need for accurate and efficient transcription of medical histories by developing and evaluating an ASR tool. In the absence of a dedicated Portuguese medical history database, we constructed a benchmark of 397 audio-text pairs from real medical histories and evaluated ASR models like Wav2Vec 2.0 and Whisper, identifying challenges in transcribing specific medical terms. Our findings highlight the necessity of fine-tuning ASR models for Portuguese medical contexts. Introducing MedTalkAI, we provide a tool that enables healthcare professionals to efficiently transcribe and correct audio recordings, contributing to a unique medical audio-text database and allowing data scientists to fine-tune ASR models. Future works will focus on refining and expanding this tool by conducting tests with doctors to evaluate its functionality and improve its impact on healthcare documentation and diagnosis.

Acknowledge. This research was conducted as part of the CEREIA project and was supported by multiple institutions and partners. We gratefully acknowledge the contributions from Hapvida NotreDame Intermédica, the Federal University of Ceará (UFC), and the São Paulo Research Foundation (FAPESP). This research was supported by grant 2020/09706-7, São Paulo Research Foundation (FAPESP). We extend our gratitude to all the institutions, partners, and funders involved in this project.

References

Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. In *NeurIPS*, pages 12449–12460.

Besacier, L., Barnard, E., Karpov, A., and Schultz, T. (2014). Automatic speech recognition for under-resourced languages: A survey. *Speech Communication*, 56:85–100.

da Silva, T. L. C., Magalhães, R. P., de Macêdo, J. A., Araújo, D., Araújo, N., de Melo, V. T., Olímpio, P., Rego, P. A., and Neto, A. V. L. (2019). Improving named entity recognition using deep learning with human in the loop. In *EDBT*, pages 594–597.

Gür, B. (2012). *Improving speech recognition accuracy for clinical conversations*. PhD thesis, Massachusetts Institute of Technology.

Heafield, K. (2011). Kenlm: Faster and smaller language model queries. In *Proceedings of the sixth workshop on statistical machine translation*, pages 187–197.

Li, J., Lavrukhin, V., Ginsburg, B., Leary, R., Kuchaiev, O., Cohen, J. M., Nguyen, H., and Gadde, R. T. (2019). Jasper: An End-to-End Convolutional Neural Acoustic Model. In *Proc. Interspeech 2019*, pages 71–75. ISCA.

Li, Y., Yu, B., Quangang, L., and Liu, T. (2021). Fitannotator: A flexible and intelligent text annotation system. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, pages 35–41.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. In *ICML*, pages 28492–28518.

Rubenstein, P. K., Asawaroengchai, C., Nguyen, D. D., Bapna, A., Borsos, Z., Quitry, F. d. C., Chen, P., Badawy, D. E., Han, W., Kharitonov, E., et al. (2023). Audiopalm: A large language model that can speak and listen. *arXiv preprint arXiv:2306.12925*.

Schneider, S., Baevski, A., Collobert, R., and Auli, M. (2019). wav2vec: Unsupervised pre-training for speech recognition. In *Interspeech 2019*, pages 3465–3469.

Stolcke, A. (2002). Srilm-an extensible language modeling toolkit. In *Seventh international conference on spoken language processing*.

Sullivan, P., Shibano, T., and Abdul-Mageed, M. (2022). Improving automatic speech recognition for non-native english with transfer learning and language model decoding. In *AANLSP*, pages 21–44.