Speech Recognition Models in Assisting Medical History

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. *This paper addresses challenges highlighted by health professionals, where up to 50% of a medical consultation's time is spent on history creation. To streamline this process, we propose leveraging Automatic Speech Recognition (ASR) models to convert spoken language into text. In our study, we assess the effectiveness of pre-trained ASR models for medical history transcription in Brazilian Portuguese. By incorporating language models to enhance ASR output, we aim to improve the accuracy and semantic fidelity of transcriptions. Our results demonstrate that integrating a 5-gram model with Wav2Vec2 PT significantly reduces transcription errors, while also maintaining superior performance in capturing textual nuances and similarity.*

1. Introduction

The traditional process of collecting medical histories can be time-consuming and requires considerable time from healthcare professionals [Chiu et al. 2017]. According to Hapvida NotreDame Intermédica[1], during a medical consultation, on average, up to 50% of the total time is consumed by the creation of the medical history. Often, medical histories may not contain detailed enough information, which can lead to inaccurate or incomplete diagnoses.

Medical histories may also vary in terms of format, structure, and content, making comparison and information sharing among healthcare professionals difficult. Another possible problem is that during a consultation, for example, healthcare professionals may face difficulties in remembering all the relevant details provided by the patient. In short, it is necessary to seek solutions to the problems related to inefficiency, lack of precise details, lack of standardization, and difficulty in accessing relevant information during the process of collecting and creating medical histories.

In response to the challenges posed by the traditional process of collecting medical histories, automatic speech recognition (ASR) emerges as a promising alternative. ASR involves converting speech into text using computer programs, often employing methods like pattern recognition and artificial intelligence [Reddy 1976]. While pre-trained audio encoders like Wav2Vec2 [Baevski et al. 2020, Schneider et al. 2019] and Jasper [Li et al. 2019] have proven effective in learning high-quality speech representations, their unsupervised nature typically necessitates a fine-tuning stage for specific tasks, such as speech recognition in specialized domains like medical histories. However, fine-tuning can be complex and often requires the expertise of a qualified professional. Additionally, a comprehensive dataset containing pairs of audio and text specific to the

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

domain and language is crucial. Nevertheless, such datasets are currently lacking for medical histories in Brazilian Portuguese, rendering fine-tuning impractical.

ASR models should ideally perform reliably across various domains without the need for supervised fine-tuning for every deployment distribution. Large language models (LLMs) like Whisper [Radford et al. 2023] or AudioPALM [Rubenstein et al. 2023] present an alternative, allowing the knowledge gained during training on one or multiple datasets to be applied to different yet related datasets. To comprehensively explore alternatives for ASR in addressing our specific problem, we establish a benchmark focused on real audio and text data from medical histories. This benchmark enables us to evaluate various ASR models, including advanced language models like Whisper.

Our findings highlighted challenges in Portuguese medical transcription, including phonetic similarities ("SS" and "S") and the silent "H", leading to inaccuracies in ASR transcriptions, medical acronyms ("FC" for *frequência cardíaca* - heart rate) and measurements ("bpm" for *batimentos por minuto* - beats per minute) further complicate accurate transcription. Furthermore, while a doctor might verbally report a patient's condition as "the patient presents a heart rate of 90 beats per minute", it's common practice for physicians to document this as "HR: 90 bpm" in the medical history.

Such obstacles underscore the need for precise decoding by ASR models for Brazilian Portuguese in the clinical domain, where accurate medical history documentation is crucial for legal compliance. To address these challenges arising from medical terminology and better transcribe audio recordings of patient histories, we conduct a comparative study involving different ASR models, including Whisper and Wav2Vec2. Moreover, to balance the accuracy of clinical term transcription with the associated semantics and context, we enhance the decoding of Wav2Vec2 PT by integrating language models to correct the transcriptions.

We chose to incorporate a language model with Wav2Vec2 transcriptions because, in our benchmarks, Wav2Vec2 PT outperformed Whisper. Specifically, Wav2Vec2 PT achieved an average Word Error Rate (WER) of 0.24 and a cosine similarity of 0.88, while Whisper achieved a WER of 0.37 and a cosine similarity of 0.83. Furthermore, integrating Wav2Vec2 PT with a 5-gram model further enhanced its performance, achieving a WER of 0.17 and a cosine similarity of 0.91. Even after fine-tuning Wav2Vec2 PT, the improvement was more significant when a language model was integrated with Wav2Vec2 PT.

The rest of this article is organized as follows. Section 2 presents the main related works. Section 3 explains the data, evaluation metrics, and methods used. Section 4 discusses our experimental results. Finally, Section 5 summarizes this work and proposes future directions.

2. Related Work

This section provides an overview of key studies related to ours. Some ASR approaches in the medical domain have been explored, such as the survey by [Lee et al. 2023], which assessed a machine learning-based SR system's efficacy in reducing nursing documentation workload within a psychiatry ward. Conducted at Cheng Hsin General Hospital in Taiwan, nurses evaluated the SR system's documentation time and error rate compared to keyboard entry. Results revealed that the SR system processed 30,112 words in 32,456 seconds, achieving a recognition accuracy improvement from 87.06% to 95.07% across four sessions. However, despite the improvements, the system still produced errors, indicating that further refinement is necessary to ensure reliability in clinical settings.

In [Paats et al. 2018], an analysis of different language models' impact on an Estonian ASR system's clinical performance was conducted. Initially using a Gaussian Mixture Model (GMM) acoustic model, the system transitioned to a Deep Neural Network (DNN) acoustic model. The fine-tuning process involved adapting the acoustic model with in-domain data and the language model with spoken data. Testing with 11 radiologists dictating 219 reports in a clinical environment showed performance improvement, with the average WER decreasing from 18.4% to 5.8%. While demonstrating acceptable accuracy, these studies underscore the importance of adapting the model to the specific domain. They show that although improvements are achieved, there are still errors and a need for further enhancements in ASR transcription.

Researchers are collectively concerned about improving ASR model transcription accuracy due to the potential for clinical harm arising from speech recognition inaccuracies, as evidenced in prior studies like [Chiu et al. 2017, Kar et al. 2021, Sullivan et al. 2022], among others. While fine-tuning is important, researchers are exploring other strategies, such as the use of language models to enhance ASR model output. The same idea that we follow in this paper.

[Chiu et al. 2017] explores two methods for constructing speech recognition models: one uses recurrent neural networks with connectionist temporal classification (CTC), while the other employs Listen Attend and Spell (LAS) models. The CTC system trains an acoustic model with CTC loss, using context-dependent phoneme outputs, n-gram language models, and pronunciation dictionaries, and decodes using a finite-state transducer (FST) decoder. Both unidirectional and bidirectional CTC models were trained. LAS models consist of an encoder, attention mechanism, and decoder. The CTC models achieved a WER of 20.1%, while the LAS models achieved 18.3%.

[Kar et al. 2021] developed a system to extract medical information from audio recordings in critical medical scenarios. The authors utilized a multi-style training approach and noise reduction techniques, integrating specific medical terms into the ASR system for enhanced accuracy. The transcription texts were processed with MetaMap, a tool identifying clinical concepts in UMLS ontologies, which notably improved accuracy, particularly with medical terms. Results showcased a substantial reduction in WER of up to 52.27% compared to the baseline model.

Another relevant study in this area is [Sullivan et al. 2022], which integrates a 4-gram language model into the decoding process of Wav2Vec2. This integration aids in reducing spelling errors and improbable word sequences, thereby increasing the likelihood of accurately predicting words commonly found in the language. Moreover, an advantage of incorporating a language model into decoding is the ability to calculate probabilities from a text-only corpus, eliminating the need for audio data as in a fine-tuning approach.

The present study follows the strategies outlined in the aforementioned studies. First, based on the work of [Paats et al. 2018], the focus will be on adapting the model to the medical domain and the specific language of Brazilian Portuguese. In line with [Kar et al. 2021], while creating an ontology of medical terms may seem appealing, it presents significant challenges, such as development complexity and the difficulty of finding Portuguese-language databases for clinical domain. Therefore, we experiment with language models to enhance decoding, as suggested by [Chiu et al. 2017] and [Sullivan et al. 2022]. Furthermore, we compare different language models to assess which are most effective for Portuguese. This approach was chosen due to the wide availability of these pre-trained models and their flexibility.

3. Data and Methods

This section outlines our data source and methodology for addressing the research questions guiding our experiments. First, we employ data preprocessing, comprising two main phases: audio standardization and text processing. Subsequently, we select several ASR models and provide them with audio inputs. Each ASR model generates text outputs corresponding to the given audio. Additionally, we introduce a novel step aimed at enhancing the decoding of ASR models by incorporating language models to mitigate syntactic errors, given the critical importance of accurate medical histories. Finally, we evaluate the text outputs of the ASR models using various evaluation metrics, including WER, Cosine Similarity, and BLEU. All these steps are illustrated in Figure 1.

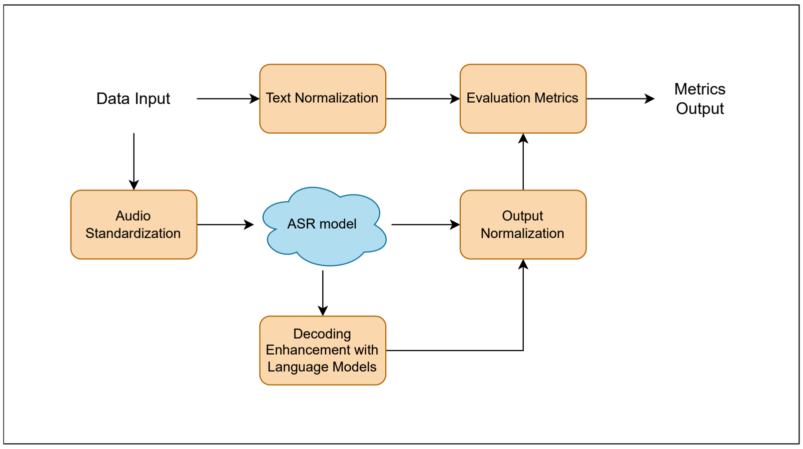

Figure 1. Methodology Overview

3.1. Data source and preprocessing

We establish a benchmark by recording 224 audio samples derived from authentic medical history texts sourced from the private database of Hapvida NotreDame Intermédica. This initiative addresses the lack of publicly accessible databases specifically tailored to Brazilian Portuguese in the medical domain. The audio pairs are meticulously curated and recorded by the authors. In the preprocessing stage, we standardize the audio files to 16kHz frequency and conduct text preprocessing. This involves converting all text to lowercase, removing accents, converting certain punctuations, such as ':' to 'colon', numbers to their written-out form and eliminating characters not present in the Wav2Vec2 vocabulary. These steps ensure uniformity and consistency throughout the dataset.

3.2. ASR models

In this paper, we use three ASR models: Wav2Vec2, Whisper, and HuBERT [Hsu et al. 2021]. In what follows, we provide an overview of such ASR models.

Wav2Vec2 addresses the challenge of training models with large datasets by utilizing limited labeled data. It employs a unique approach by jointly learning discrete speech units alongside contextualized representations. The model architecture consists of a feature encoder that takes raw waveform input and passes it through blocks containing temporal convolution, layer normalization, and a GELU activation function. The output is then fed into a context network following a Transformer architecture.

HuBERT [Hsu et al. 2021] is a self-supervised BERT model [Devlin et al. 2018] designed for audio inputs. To handle variable-length sound units and multiple sound units per input, the authors introduce hidden units (Hu), which are clusters assigned to input audio segments using k-means. These hidden units are then mapped to embedding vectors and used as targets to train the BERT model. After pre-training to assign contextual representations to audio inputs, the model is fine-tuned for ASR using a softmax layer. According to [Hsu et al. 2021], HuBERT performs comparably to Wav2Vec2 across all evaluated fine-tuning subsets when pre-trained on Librispeech (960h) and Libri-light (60,000h).

Whisper [Radford et al. 2023] is a robust speech processing system designed to overcome the limitations of unsupervised pre-trained audio encoders like Wav2Vec2. While these encoders excel at learning speech representations from raw audio, they lack a decoder of comparable performance, requiring fine-tuning for specific tasks like ASR. Whisper introduces an encoder-decoder Transformer architecture [Vaswani et al. 2017], utilizing sequence-to-sequence models for transcript prediction without extensive standardization. It also performs tasks such as language identification and translation to English. Trained with 680,000 hours of data across 96 languages, Whisper transfers well to other datasets without requiring dataset-specific fine-tuning, making it a versatile multi-language multitask model.

3.3. Fine-tuning Wav2Vec2 PT

Due to the absence of a dedicated Brazilian Portuguese medical corpus, we engaged three medical science students (two females and one male) to record a total of 657 audio clips containing real patient anamneses. This effort aimed to enhance transcription accuracy by fine-tuning Wav2Vec2 PT, particularly in capturing the nuances of the Brazilian Portuguese language and medical terminology. We applied the same pre-processing steps outlined in Section 3.1 to this dataset.

During fine-tuning, the model's parameters were adjusted to optimize performance. This included setting a learning rate of 0.0003, a train batch size of 2, an eval batch size of 1, a seed of 42, gradient accumulation steps of 8, and a total train batch size of 16. The training process employed a linear learning rate scheduler with a warmup of 500 steps and lasted for 40 epochs. This process resulted in a loss of 0.7948 and a WER of 0.7625 on the evaluation set.

3.4. Decoding Enhancement

We notice that ASR decoding can introduce errors, and given the critical importance of precise texts, correction becomes necessary. We explore two approaches, both leveraging language models. Language models capture the fundamental structure and dependencies of language. These models serve as tools for comprehending and producing natural language text. Our investigation encompasses two strategies: one integrating an ASR and an **n-gram model**, and another utilizing a large language model (LLM) known as **Mistral** [Jiang et al. 2023]. There are several LLMs we might consider here to enhance the decoding. However, Mistral-7B outperforms LLaMA [Touvron et al. 2023] and is released under the Apache 2.0 license, contributing to its appeal in licensing and performance.

N-gram. An n-gram language model employs statistical methods to forecast the subsequent word in a sequence of text, drawing from preceding words. The "n" in "n-gram" signifies the number of consecutive words considered within the conditional probability. For instance, a unigram entails assessing the prior probability of a single word, denoted as $P(p_i)$. Meanwhile, a bigram entails considering two words $P(p_i|p_{i-1})$, and a trigram involves considering the two preceding words $P(p_i|p_{i-1}, p_{i-2})$, and so forth. Expanding further, an n-gram model is denoted by $P(p_i|p_{i-1}, ..., p_{i-n})$. Conceptually, a bigram language model analyzes pairs of adjacent words, while a trigram model examines groups of three words.

The n-gram models calculate the likelihood of word or character occurrences based on preceding context, utilizing extensive textual datasets for training. With nearly 5,000 medical records from Hapvida NotreDame Intermédica, we utilize these texts in our experiments to train the n-gram models. While integrating Wav2Vec2 + n-gram has been previously investigated by [Sullivan et al. 2022], it hasn't been explored in the medical domain. In our experiments, we apply this strategy by leveraging KenLM[2], departing from decoding audio without a language model and enabling the processor to directly receive the model's output logits. This approach, rooted in the decoding process with a language model, enables the processor to consider the probabilities of potential output characters at each time step, thus rectifying any character errors made by the ASR model.

Mistral. [Jiang et al. 2023] stands as an open-source language model focused on achieving high performance without requiring substantial hardware investments. This emphasis on efficiency while delivering superior performance underscores its design. Mistral-7B incorporates grouped-query attention (GQA) to optimize inference speed and reduce memory demands during decoding, allowing for larger batch sizes and increased throughput, crucial for real-time applications. Additionally, Mistral employs sliding window attention to handle longer sequences more efficiently, addressing a common limitation in LLMs. These attention mechanisms enhance the performance and efficiency of Mistral-7B, operating on the Transformer architecture.

Mistral-7B introduces Sliding Window Attention, Rolling Buffer Cache, and Pre-fill and Chunking mechanisms. Sliding Window Attention utilizes stacked transformer layers, while the Rolling Buffer Cache rotates the buffer to accommodate fixed attention span sizes. Pre-fill and Chunking leverage available prompts, enabling pre-filling of the cache with known prompt data during sequence generation. Additionally, the authors present a fine-tuned model, Mistral-7B-Instruct, tailored for chat-based inference. In this study, we improve ASR decoding through Mistral-Instruct, a large language model where we provide instructions to correct sentences syntactically. The prompt strategy used is

[2] `https://github.com/kpu/kenlm`

called one-shot learning.

To ensure consistently high-quality responses, Mistral-7B-Instruct is utilized within a well-defined prompt structure. The prompt is provided as follows, where "row[sentence]" is the medical report transcribed from an ASR model that should be corrected. We also define rules for how responses should be written by the model, amongst them ensuring corrections are limited to text enclosed within $<$ and $>$ delimiters, focusing on correcting only spelling and syntactic errors without sentence restructuring, and not providing explanations for the corrections made. This strategy not only enhances the accuracy of the responses but also ensures they adhere to the desired standards.

```
### Prompt:
{"role": "user", "content": Please correct the following sentence in Portuguese:
<queixa sididor em bevê há hum mês com ocorrências exporádicas corrimento com
odor e prorido>;
Delimit the correction using the characters <>;
Do not rephrase the sentence; Correct only the
spelling and syntactic errors; Don’t rewrite
the sentence. There is no need to explain your choices.}
{"role": "assistant", "content": "<queixa-se de dor em bv há um mês com
ocorrências esporádicas corriemnto com odor e prurido>"}
{"role": "user", "content": Now, correct the following sentence in Portuguese:
{row[sentence]};
Delimit the correction using the characters <>;
Do not rephrase the sentence; Correct only the
spelling and syntactic errors; Don’t rewrite
the sentence. There is no need to explain your choices."}
### End
```

3.5. Evaluation Metrics

A range of metrics are utilized to evaluate the effectiveness of ASR models, including Word Error Rate (WER), Cosine Similarity, and BLEU. This data is crucial for refining models and enhancing their performance.

The word error rate (WER) is the most used evaluation metric for ASR systems. The percentage of incorrect words gives the WER of a transcription concerning the number of input words. The incorrect words were erroneously inserted, replaced, or deleted by the system transcription. WER is defined as in Equation 1.

$$WER = \frac{I + R + D}{H + R + D} \tag{1}$$

where I is the number of inserted words, R is the number of replaced words, D is the number of deleted words, and H is the number of hits. Despite its popularity, WER is limited to accuracy at the word level.

Different from WER, BLEU [Papineni et al. 2002] can evaluate whether the transcription maintains the context and organization of the sentence. BLEU was originally proposed for neural machine translation and it claims to be highly correlated with human assessment. BLEU is based on the precision of n-grams, which compares the n-grams of reference text T^* with the n-grams of its transcription T. Let $NG(n, t)$ be the set of n-grams of text t, the n-gram precision P_n between texts T^* and T is given by Equation 2.

$$P_n = \frac{|NG(n, T^*) \cap NG(n, T)|}{NG(n, T)} \tag{2}$$

BLEU is calculated as the geometric mean (all weights equal to 1/4) of P_n, for $n = 1, 2, 3, 4$ multiplied by a factor that penalizes transcriptions shorter than the referenced text. The $bleu_{penalty}$ factor is 1 if $|T| > |T^*|$ and $e^{(1-|T^*|/|T|)}$, otherwise. BLEU is defined in Equation 3.

$$BLEU = \sqrt[4]{P_1 P_2 P_3 P_4} \times bleu_{penalty} \tag{3}$$

The Cosine Similarity allows for determining how close the two sentences are in a defined vector space. For the Cosine Similarity, we can use Word Embedding or Sentence Embedding vectors [Li et al. 2020] (in the experiments, we use the Universal Sentence Encoder). The Cosine Similarity is defined as in Equation 4, where A and B are vectors of attributes; A_i and B_i are components of vector A and B, respectively; $||\mathbf{A}||$ is the Euclidean norm of vector A. Similarly, $||\mathbf{B}||$ is the Euclidean norm of vector B.

$$\cos(\mathbf{A}, \mathbf{B}) = \frac{\mathbf{AB}}{||\mathbf{A}||||\mathbf{B}||} = \frac{\sum_{i=1}^{n} \mathbf{A}_i \mathbf{B}_i}{\sqrt{\sum_{i=1}^{n} (\mathbf{A}_i)^2} \sqrt{\sum_{i=1}^{n} (\mathbf{B}_i)^2}} \tag{4}$$

4. Experimental Results

To guide this section, we have formulated the following key research questions. In what follows, we study each research question separately.

1. **RQ1.** What are some descriptive statistics from the training data?
2. **RQ2.** Which are the pre-trained ASR models most effective for assisting with medical history?
3. **RQ3.** What are the most effective decoding strategies that leverage language models to enhance the accuracy and reliability of ASR transcriptions?

4.1. Study on the results of RQ1

To investigate the impact of utilizing automated transcription tools on cognitive load and efficiency for healthcare professionals, we prefer to analyze this research question using the dataset employed during fine-tuning. We engaged three medical science students to record a total of 657 audio clips containing real patient anamneses. Analysis of these recordings revealed notable statistics regarding duration: the average duration was 94.69 seconds, with the shortest clip lasting 4.18 seconds and the longest stretching to 346.50 seconds, resulting in a standard deviation of 46.15 seconds. According to Hapvida NotreDame Intermédica, medical appointments usually last around 15 minutes, with nearly half of this time consumed by the manual entry of medical text into systems (between 7 to 8 minutes).

With an average recording duration of 94.69 seconds, considerably shorter than a typical medical appointment, there is ample opportunity for verification and potential re-recording of transcriptions using automated tools such as ASR models. This has the potential to alleviate the cognitive burden and improve efficiency for healthcare professionals. Moreover, it suggests that more time remains available for interaction between doctors and patients. Thus, this highlights a crucial avenue for research to assess the practical implications and effectiveness of integrating automated transcription tools into medical workflows.

4.2. Study on the results of RQ2

In this study, we selected the following pre-trained ASR models designed to support Brazilian Portuguese language: Wav2Vec2 PT[3], HuBERT[4] and Whisper[5]. Our benchmark includes 224 audio recordings (differing from those utilized in fine-tuning) that were transcribed and evaluated using the pre-trained models. Utilizing the widely used WER metric for this task, HuBERT achieved a WER of 0.66, Whisper achieved a WER of 0.37, and Wav2Vec2 PT achieved a WER of 0.24. Wav2Vec2 PT also demonstrated superior performance in capturing textual nuances and similarity, scoring a BLEU of 0.55 and cosine of 0.88, compared to HuBERT's BLEU of 0.12 and cosine of 0.68, and Whisper's BLEU of 0.43 and cosine of 0.83.

Table 1. Examples of medical histories generated by the ASR pre-trained models.

Model	Original	Transcription	WER	BLEU SCORE (n-gram)	Cosine Similarity
Wav2Vec2 PT	PT: paciente fez ultrassonografia que acusou nódulo tireoidiano fez nova ultrassonografia com surgimento de um novo nódulo EN: patient underwent ultrasound which accused of thyroid nodule performed another ultrasound with emergence of a new nodule	paciente fez ultrasonografia que acusou o nódulo tireoidiano fez nova ultrasonografia com o surgimento de um novo nódulo	0.31	0.32	0.92
Whisper		paciente fez ultra sonografia que acusou o nóluo tiréoediano fez nova ultra sonografia com o surgimento de um novo nóluo	0.5	3.88e-78	0.91
HUBERT		paciente fez utra so nografia que acusou nodlotiraoe de ano fez nova utração nografia com sulgimento de um novo nódulo	0.56	0.18	0.88
Wav2Vec2 PT	PT: nega dor sialose halitose pigarro ou outras queixas EN: denies pain, sialosis, halitosis, throat clearing or other complaints.	nega dor cialose alitose pigarro ou outras queixas	0.25	0.41	0.93
Whisper		mega doce se arose alitosse pigarro ou outras queixas	0.62	0.29	0.60
HuBERT		meca dor se alos alitose pigarro ou outras queixhas	0.62	4.34e-78	0.78

We performed the Wilcoxon signed-rank test [Wilcoxon 1992] since it is a non-parametric test. From the statistical point of view, the test is safer since it does not assume normal distributions. Our null hypothesis (H_O) states that the models perform equally well for WER results. So, we conducted the statistical test to compare the performance of the Whisper, HuBERT, and Wav2Vec2 PT models. The Wilcoxon tests revealed highly significant differences between all pairs of models. Specifically, the ρ-values were extremely low for each comparison: between Whisper and HuBERT (ρ = 2.072e-29), between Whisper and Wav2Vec2 PT (ρ = 9.566e-15), and between HuBERT and Wav2Vec2 PT (ρ = 1.359e-37). These results indicate strong evidence to reject the null hypothesis and conclude that the models' performances are statistically different. From such a comparison of the three pre-trained ASR models, Wav2Vec2 PT outperformed the other two models.

Through our evaluation, we identified challenges encountered by established ASR models such as Wav2Vec2 PT and Whisper, particularly when confronted with medical

[3] https://huggingface.co/jonatasgrosman/wav2vec2-xls-r-1b-portuguese

[4] https://huggingface.co/jonatasgrosman/exp_w2v2t_pt_hubert_s807

[5] https://huggingface.co/openai/whisper-large-v3

terms such as “thyroid nodule (nódulo tireoidiano)”, “sialosis (sialose)”, and other terms outlined in Table 1. Additionally, we noticed common errors in Brazilian Portuguese transcriptions, stemming from phonetic similarities such as “SS” and “S” (ultrassonografia), “S” and “C” (sialose) as well as the presence of silent “H” (halitose). Our analysis further underscores the complexities associated with accurately transcribing medical terminology, thereby shedding light on the limitations of widely used ASR models in capturing the nuances of healthcare-related vocabulary.

To address such errors, we have opted to fine-tune the Wav2Vec2 PT using additional pairs of audio and text in Brazilian Portuguese derived from real medical histories. We refer to this fine-tuned model as Wav2Vec2 PT Fine-tuned[6]. The fine-tuning process has been previously described in Section 3.3, our primary objective was to reduce grammatical errors.

Despite the intended objective, the fine-tuned model's performance worsened unexpectedly, resulting in more inaccurate transcriptions of medical terms, with a WER of 0.66. The pre-trained Wav2Vec2 PT model was initially trained on the CommonVoice and LibriSpeech datasets. This degradation may have been amplified by regional differences in the audio data used for fine-tuning, hindering rather than aiding the model compared to its previous training data. Additionally, only a limited number of text-audio pairs were utilized in fine-tuning. It's crucial to acknowledge the regional context of this research, as it's part of a larger R&D project intended for use by local healthcare professionals. Though the benchmark recordings were made by regional researchers, the specific region is undisclosed due to blind submission. Given the critical importance of medical history, improving the decoding accuracy of our best ASR model (Wav2Vec2 PT) according to the benchmark remains imperative.

4.3. Study on the results of RQ3

We examined two approaches employing language models to handle corrections in ASR transcriptions. The first method consisted of training an n-gram with almost 5,000 medical reports collected from Hapvida NotreDame Intermédica. We varied n using the values $\{3,5,7\}$. The implementation used an integrated Wav2Vec2 PT and an n-gram model via KenLM. Even with variations in n, the models exhibit similar performance, as evidenced by the Wilcoxon tests, which achieved $\rho > 0.05$ (ρ around 0.4), all indicating statistical equivalence. Due to space constraints, we opted to outline the findings of Wav2Vec2 PT+5-gram[7].

The second method uses Wav2Vec2 PT+Mistral-7B-Instruct[8], prompting the model to correct the spelling and syntactic errors. Using the benchmark, the Wav2Vec2 PT+5-gram outperformed the other models, achieving an average WER of 0.17, and Wav2Vec2 PT+Mistral-7B-Instruct achieved an average WER of 0.48. The Wilcoxon test results demonstrate statistical differences between Wav2Vec2 PT+5-gram and Wav2Vec2 PT+Mistral-7B-Instruct (ρ = 8.18e-28), besides Wav2Vec2 PT and Wav2Vec2 PT+5-gram (ρ = 2.17e-13). Wav2Vec2 PT+5-gram achieved an average BLEU score of 0.65, and an average cosine similarity of 0.91.

[6] https://huggingface.co/medtalkai/wav2vec2-xls-r-1b-medical-domain02

[7] https://huggingface.co/medtalkai/wav2vec_kenlm5

[8] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2

Wav2Vec2 PT+Mistral-7B-Instruct exhibited a lower performance than the others due to its tendency to slightly alter the wording while preserving the intended meaning, as reported in Table 2. Unlike others, Wav2Vec2 PT+5-gram excels at fixing transcription errors. Even with language models, it's still a challenge to deal with medical acronyms, like 'fc' (*frequência cardíaca*) and 'bv' (*baixo ventre*), and measurements, such as 'bpm' (*batimentos por minuto*) as highlighted in Table 2. Note that the punctuation ":" was replaced with a colon (dois pontos, in Portuguese), and numbers are written out in full.

Original Sentence	Transcription
queixa-se de dor em **bv** há um mês com ocorrências esporádicas corriemnto com odor e prurido	*Wav2Vec2 PT*: queixa-se de dor em **bvê** a um mês com ocorrências hesporádicas corrimento com odor e prurido *Wav2Vec2 PT+5-gram*: dor em **bv** a um mês com ocorrências esporádicas corrimento com odor e prurido *Wav2Vec2 PT+Mistral-7B-Instruct*: queixa-se de dor em côito a um mês com ocasiões esporádicas correndo com odores e prurido
fc dois pontos cento e quatro **bpm**	*Wav2Vec2 PT*: **efc** dois pontos cento e quatro **bê p eme** *Wav2Vec2 PT+5-gram*: **fc** dois pontos cento e quatro **b p em** *Wav2Vec2 PT+Mistral-7B-Instruct*: **efc** dois pontos cento e quatro **beém**

Table 2. Examples of transcriptions from the language models combined with Wav2Vec2 PT

5. Conclusion and Future Works

This paper has addressed the critical need for accurate and efficient transcription of medical histories through the development and evaluation of an ASR tool. Given the absence of a dedicated Portuguese medical history database, we constructed a comprehensive benchmark consisting of 224 pairs of audio and text sourced from authentic medical history documents. Our evaluation of established ASR models, including Wav2Vec2 and Whisper, revealed notable challenges, particularly in accurately transcribing specific medical terms. The intricate nature of healthcare-related vocabulary and the nuances of the Brazilian Portuguese language highlight the limitations inherent in widely used ASR models. Furthermore, we have shown that incorporating language models such as n-gram can significantly enhance the transcription accuracy of Wav2Vec2 PT, offering promising avenues for further improvement in medical transcription technology.

In future research, we aim to fine-tune Mistral-7B-Instruct for the medical domain to enhance transcription accuracy. We will also investigate advanced language models such as GPT to further improve ASR model transcription accuracy. Additionally, we plan to explore the use of knowledge graphs and evaluate approaches like constructing controlled vocabularies to compare their performance against n-gram models and/or in combination with them.

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.

Chiu, C.-C., Tripathi, A., Chou, K., Co, C., Jaitly, N., Jaunzeikare, D., Kannan, A., Nguyen, P., Sak, H., Sankar, A., et al. (2017). Speech recognition for medical conversations. *arXiv preprint arXiv:1711.07274*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., and Mohamed, A. (2021). Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM TASLP*, 29:3451–3460.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. d. l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al. (2023). Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Kar, S., Mishra, P., Lin, J., Woo, M.-J., Deas, N., Linduff, C., Niu, S., Yang, Y., McClendon, J., Smith, D. H., et al. (2021). Systematic evaluation and enhancement of speech recognition in operational medical environments. In *IJCNN*, pages 1–8.

Lee, T.-Y., Li, C.-C., Chou, K.-R., Chung, M.-H., Hsiao, S.-T., Guo, S.-L., Hung, L.-Y., and Wu, H.-T. (2023). Machine learning-based speech recognition system for nursing documentation–a pilot study. *IJMI*, 178:105213.

Li, B., Zhou, H., He, J., Wang, M., Yang, Y., and Li, L. (2020). On the sentence embeddings from pre-trained language models. In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, *Proceedings of the EMNLP*, pages 9119–9130.

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.

Paats, A., Alumäe, T., Meister, E., and Fridolin, I. (2018). Retrospective analysis of clinical performance of an estonian speech recognition system for radiology: effects of different acoustic and language models. *JDI*, 31(5):615–621.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th ACL*, page 311–318, USA. Association for Computational Linguistics.

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.

Reddy, D. R. (1976). Speech recognition by machine: A review. *Proceedings of the IEEE*, 64(4):501–531.

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.

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.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *NIPS*, pages 6000–6010.

Wilcoxon, F. (1992). Individual comparisons by ranking methods. In Kotz, S. and Johnson, N. L., editors, *Breakthroughs in Statistics: Methodology and Distribution*, pages 196–202. Springer New York, New York, NY.