Geração de Corpus Sintético Sociolinguístico para Avaliação de Reconhecimento de Fala no Contexto Clínico Brasileiro

Ana Carla Sergina N. de Lima[1], Clauirton A. Siebra[2]

[1]Departamento de Informática – Universidade Federal da Paraíba (UFPB)
João Pessoa – PB – Brasil

ana.sergina@academico.ufpb.br, clauirton@ci.ufpb.br

Abstract. *Electronic health record documentation consumes a significant portion of clinical care time, motivating the use of Automatic Speech Recognition (ASR) systems for consultation transcription. In Brazil, the advancement of such solutions is limited by the scarcity of public datasets containing real medical interactions, due to ethical and legal restrictions. This work proposes a pipeline for generating a synthetic corpus of teleconsultations in Brazilian Portuguese with cultural and linguistic variability. A total of 120 dialogues were generated using a language model and converted into audio through neural text-to-speech synthesis. The results indicate that the controlled insertion of regionalisms enables the analysis of limitations in general-purpose models, thereby motivating the development of specialized clinical transcription systems tailored to the regional differences of the Brazilian context.*

Resumo. *A documentação em prontuários eletrônicos consome tempo significativo do atendimento clínico, motivando o uso de sistemas de Reconhecimento Automático de Fala (ASR) para transcrição de consultas. No Brasil, o avanço dessas soluções é limitado pela escassez de bases públicas de interações médicas reais, devido a restrições éticas e legais. Este trabalho propõe um pipeline para geração de um corpus sintético de teleconsultas em português brasileiro com variabilidade cultural e linguística. Foram gerados 120 diálogos por meio de modelo de linguagem e convertidos em áudio com síntese neural de voz. Os resultados indicam que a inserção controlada de regionalismos permite analisar limitações de modelos generalistas, o que motiva o desenvolvimento de sistemas de transcrição clínica especializados para as diferenças regionais do contexto brasileiro.*

1. Introdução

A crescente adoção de prontuários eletrônicos ampliou a digitalização da documentação clínica, mas também intensificou a carga administrativa sobre os profissionais de saúde (Tran et al., 2023). Estudos observacionais indicam que médicos podem dedicar entre 30% e 40% do tempo da consulta exclusivamente ao preenchimento e à atualização de

registros clínicos (Sinsky et al., 2016). Em alguns contextos, essa carga administrativa pode chegar a representar até duas horas de documentação para cada hora de atendimento efetivo (Sinsky et al., 2016). No Brasil, a digitalização dos registros ampliou as demandas de preenchimento obrigatório e codificação padronizada, o que pode comprometer a fluidez da comunicação médico-paciente, aumentar a sobrecarga cognitiva do profissional e impactar a qualidade do atendimento (Catapan et al., 2020).

Para mitigar esse cenário, sistemas de Reconhecimento Automático de Fala (Automatic Speech Recognition – ASR) têm sido utilizados para automatizar a transcrição de consultas e apoiar soluções conhecidas como digital scribes (Tran et al., 2023). Este trabalho se insere nesse contexto, propondo uma abordagem baseada em transcrição automática de teleconsultas como estratégia de suporte a sistemas médicos de automação de prontuários, com aplicação direta no projeto BLIND REVIEW.

Entretanto, o desenvolvimento e a validação de soluções como o BLIND REVIEW são severamente limitados pela ausência de bases públicas de áudio com interações médicas reais. A coleta desse tipo de dado envolve custos elevados e enfrenta restrições éticas e legais rigorosas, especialmente no que tange à privacidade e ao sigilo das informações de saúde estabelecidos pela Lei Geral de Proteção de Dados (LGPD) e discutidos amplamente na literatura internacional de geração de dados (Chen et al., 2021).

Embora Modelos de Linguagem de Grande Escala (LLMs), como o ChatGPT, e ferramentas de síntese de voz, como o Azure Speech, possibilitem a geração de dados sintéticos para contornar essas barreiras (Chen, Lu e Wong, 2019), esses recursos usualmente apresentam uma limitação crítica: a baixa variabilidade fonética e linguística. Dados sintéticos gerados de forma genérica falham na simulação de construções linguísticas regionais, entonações naturais e expressões coloquiais, que são elementos estruturais comuns e fundamentais nas interações reais entre médico e paciente no vasto contexto cultural brasileiro (Da Silva, Freitas e Souza, 2019). A necessidade de datasets localizados tem sido comprovada por iniciativas recentes em outros idiomas, que demonstram que modelos treinados em bases genéricas perdem precisão significativa quando aplicados ao contexto clínico regional (Nguyen et al., 2024).

2. Trabalhos Relacionados

Esta seção apresenta os principais avanços relacionados ao uso de Reconhecimento Automático de Fala (ASR) na saúde, à geração de dados sintéticos em contextos clínicos e às limitações associadas à variabilidade linguística e cultural. A análise desses eixos permite evidenciar a lacuna científica que motiva o presente trabalho.

2.1. Reconhecimento Automático de Fala na Documentação Clínica

O uso de sistemas de ASR para apoio à documentação médica tem crescido significativamente, especialmente com a popularização de soluções conhecidas como

digital scribes. Estudos recentes avaliaram o desempenho de sistemas comerciais e modelos de uso geral na transcrição de consultas clínicas, destacando ganhos de produtividade e redução no tempo de registro manual.

Pesquisas como as de Lybarger et al. (2023) analisaram o impacto do ASR na edição de notas clínicas e compararam modelos generalistas com modelos ajustados para o domínio médico. Trabalhos mais recentes, como o de Tran et al., (2023), compararam arquiteturas especializadas para conversas médico-paciente, evidenciando que modelos treinados especificamente para o contexto clínico tendem a apresentar melhor desempenho em termos de Word Error Rate (WER).

Outros estudos, como o de Peivandi et al. (2022), avaliaram erros em registros de enfermagem produzidos por tecnologias de reconhecimento de fala online e offline, reforçando que a precisão ainda varia conforme o vocabulário técnico utilizado. Já investigações mais recentes em ambientes de emergência, conduzidas por Luo et al. (2025), demonstraram diferenças significativas de eficácia entre quatro motores comerciais de ASR quando aplicados a diálogos clínicos reais, evidenciando a necessidade constante de avaliação comparativa. Entretanto, observa-se que a maior parte desses avanços concentra-se em bases e modelos treinados predominantemente em língua inglesa, havendo um número reduzido de estudos voltados ao português brasileiro (PT-BR) ou a cenários multilíngues.

2.2. Geração de Dados Sintéticos e Restrições de Privacidade

A avaliação de sistemas de ASR em saúde enfrenta barreiras relacionadas à privacidade e confidencialidade de dados clínicos. Regulamentações como a LGPD no Brasil e a HIPAA nos Estados Unidos impõem restrições severas ao compartilhamento de gravações reais de consultas médicas, limitando a disponibilidade de datasets públicos (Chen et al., 2021).

Diante desse cenário, pesquisadores têm explorado o uso de dados sintéticos como alternativa para experimentação e benchmarking. Trabalhos recentes utilizaram modelos de linguagem de grande porte (LLMs), como GPT-3 e GPT-4, para gerar roteiros clínicos simulados (Chen, Lu e Wong, 2019). Paralelamente, sistemas de síntese de voz (Text-to-Speech – TTS) vêm sendo empregados para converter textos em áudios artificiais, preservando a estrutura conversacional sem expor informações sensíveis.

Embora tais abordagens tenham demonstrado viabilidade técnica, muitas bases sintéticas existentes apresentam limitações quanto à diversidade linguística, à naturalidade conversacional e à representatividade cultural. Em vários casos, os diálogos gerados mantêm uma estrutura excessivamente formal ou padronizada, não refletindo adequadamente a espontaneidade da fala em consultas reais.

2.3. A Lacuna da Variabilidade Cultural e Dialetal

Estudos recentes indicam que o desempenho de sistemas de ASR pode variar significativamente conforme sotaques regionais, variações dialetais e uso de expressões coloquiais (Nguyen et al., 2024). Trabalhos como o de Zolnoori et al., (2024), que investigaram a equidade em sistemas de reconhecimento de fala, demonstraram disparidades no desempenho do ASR ao transcrever a comunicação verbal entre pacientes e enfermeiros, associadas a características demográficas e sociolinguísticas dos falantes.

No contexto clínico, essa limitação torna-se particularmente relevante, uma vez que pacientes utilizam descrições informais de sintomas, regionalismos e vocabulário não padronizado. Modelos treinados predominantemente em conjuntos formais tendem a apresentar maior taxa de erro quando expostos a variações culturais e dialetais.

Apesar dos avanços recentes, grande parte das bases utilizadas para avaliação de sistemas de ASR em saúde não incorpora explicitamente variabilidade regional ou sociolinguística. Tal limitação evidencia uma lacuna na literatura: a necessidade de conjuntos de dados que representem não apenas terminologia médica técnica, mas também diversidade linguística compatível com o contexto cultural brasileiro. A Tabela 1 sintetiza o posicionamento do presente estudo em relação aos principais trabalhos relacionados, destacando o diferencial proposto.

Tabela 1. Comparação entre trabalhos relacionados e a proposta deste estudo

Estudo	Aplicação Clínica	Uso de Base Sintética	Avaliação Comparativa de Modelos	Foco em Português Brasileiro	Considera Variabilidade Sociolinguística
Tran et al. (2023)	Sim	Não	Sim	Não	Não
Peivandi et al. (2022)	Sim	Não	Sim	Não	Não
Zolnoori et al. (2024)	Sim	Não	Sim	Não	Parcial (equidade demográfica)
Luo et al. (2025)	Sim	Não	Sim	Não	Não
Este trabalho	Sim	Sim	Sim	Sim	Sim

3. Metodologia

Esta seção descreve a metodologia adotada para o desenvolvimento e a validação do corpus sintético de teleconsultas médicas. O pipeline proposto foi estruturado em três fases principais: (1) geração dos roteiros clínicos textuais com inserção intencional de

variabilidade cultural; (2) síntese de voz baseada em modelos neurais (Text-to-Speech); e (3) validação do corpus por meio de avaliação com sistemas de Reconhecimento Automático de Fala (ASR).

3.1. Geração de Roteiros Clínicos (Simulação com LLM)

A geração primária dos roteiros textuais foi realizada por meio da API do ChatGPT. Para garantir a verossimilhança clínica e conversacional, os prompts fornecidos ao modelo foram estruturados de acordo com diferentes especialidades médicas. Em cada requisição, estabeleceu-se a obrigatoriedade da inclusão de elementos essenciais da anamnese: queixa principal, histórico clínico, sintomas associados, hipóteses diagnósticas e conduta terapêutica.

Foram selecionadas oito especialidades clínicas com alta prevalência em contextos de teleconsulta, especialmente no âmbito do Sistema Único de Saúde (SUS) e da atenção primária. A escolha incluiu especialidades como Clínica Geral, Pediatria, Ginecologia, Psiquiatria, Cardiologia, Dermatologia, Infectologia e Ortopedia/Reumatologia. Essa seleção contempla tanto especialidades de demanda ampla e contínua (como Clínica Geral) quanto áreas com linguagens e sintomas específicos, o que possibilita avaliar a performance dos modelos de transcrição frente a diferentes complexidades léxicas e temáticas.

Decidiu-se gerar um total de 120 teleconsultas simuladas, distribuídas de forma balanceada entre as especialidades (com variações entre 10 e 20 por área), garantindo uma amostra estatisticamente relevante para testes de transcrição. Não existe um padrão único na literatura para o número de amostras de diálogo necessárias em tarefas ASR específicas; entretanto, trabalhos recentes de benchmark de reconhecimento de fala em contextos de saúde definem conjuntos de teste com dezenas de conversas simuladas para avaliação robusta de modelos de transcrição em diálogos médicos. Por exemplo, o dataset Afrispeech-Dialog contém 50 conversas simuladas para avaliar sistemas ASR em contextos clínicos (Sanni et al., 2025). Ademais, estudos clássicos em ASR mostram a importância de conjuntos representativos e balanceados para avaliação e validação de desempenho de modelos, mesmo quando o tamanho total varia de acordo com as necessidades do domínio e os recursos disponíveis (Chen et al. 2021).

Para tornar a base mais representativa da realidade brasileira, cada teleconsulta contará com um paciente com perfil aleatório e controlado, variando atributos como gênero, idade, grau de escolaridade, estado emocional e região geográfica (com foco em sotaques e expressões típicas). Essa diversidade é fundamental para testar a robustez dos modelos de transcrição frente à variabilidade fonética e semântica da língua portuguesa falada no Brasil, um ponto recorrente de limitação nos sistemas de ASR existentes, conforme apontado na literatura revisada. Ou seja, para além da estruturação clínica, o diferencial da metodologia concentrou-se nas instruções de controle sociolinguístico. Os prompts incluíam comandos explícitos orientando o LLM a: (i) variar o nível de formalidade na fala entre os interlocutores; (ii) incluir expressões coloquiais; (iii)

simular diferentes perfis socioeconômicos de pacientes; e (iv) incorporar regionalismos inerentes às diversas regiões geográficas do português brasileiro. Um exemplo de diretriz genérica fornecida à API segue a estrutura abaixo:

> *"Gere um diálogo de teleconsulta entre médico e paciente da especialidade X, incluindo sintomas realistas, histórico clínico e orientação médica. Utilize linguagem natural, incluindo expressões coloquiais típicas da região Y."*

Dessa forma, obteve-se um conjunto textual que reflete de maneira mais fidedigna a espontaneidade observada na prática telemédica no Brasil.

3.2. Síntese de Voz (Azure Speech Service)

Com os roteiros textuais consolidados, a etapa de conversão para áudio foi executada utilizando as vozes neurais gratuitas disponibilizadas pelo serviço Azure Speech (Text-to-Speech). Foram selecionados locutores configurados para o idioma português brasileiro (pt-BR), garantindo a alternância sistemática entre vozes masculinas e femininas para representar adequadamente a dicotomia médico-paciente.

Cabe destacar uma particularidade técnica desta etapa metodológica: as vozes selecionadas pertencem ao catálogo padrão de vozes neurais do Azure. Uma vez que o serviço gratuito não disponibiliza recursos de configuração explícita de sotaques regionais nativos, a simulação dos regionalismos baseou-se quase exclusivamente no conteúdo léxico dos diálogos gerados na etapa anterior. Assim, a identidade cultural do falante foi assegurada pela sintaxe informal, escolha de vocabulário e gírias presentes no roteiro textual, provando-se um método viável e de baixo custo computacional para a simulação sociolinguística pretendida.

3.3. Protocolo de Avaliação (Validação com ASR)

A etapa final do pipeline consistiu na avaliação preliminar da base sintética, objetivando comprovar sua aplicabilidade no treinamento e aferição de soluções do tipo digital scribe. Para o processo de transcrição automatizada do corpus recém-gerado, foi utilizado o modelo Whisper (versão base).

Desenvolvido a partir de uma arquitetura baseada em Transformers, o Whisper foi escolhido por ser um modelo de referência, treinado em múltiplos idiomas, incluindo o português, com elevada robustez na captação de falas ruidosas ou coloquiais. A aplicação deste modelo permitiu extrair transcrições das consultas simuladas que, posteriormente, puderam ser confrontadas com os roteiros originais (textos de referência). O processamento dos áudios e a quantificação dos erros foram implementados por meio de scripts em Python no ambiente Google Colab, utilizando como métricas avaliativas a Taxa de Erro de Palavra (Word Error Rate - WER) e a Taxa de Erro de Caractere (Character Error Rate - CER).

4. Resultados

Nesta seção, apresentam-se os resultados obtidos a partir da transcrição automatizada do corpus sintético composto por 120 teleconsultas simuladas, avaliando a capacidade de um modelo generalista de ASR (Whisper base) em lidar com vocabulário clínico associado a marcadores de regionalismo e informalidade.

4.1. Avaliação Quantitativa: WER e CER

A Tabela 2 apresenta os resultados globais obtidos.

Tabela 2. Desempenho do modelo Whisper-base na base sintética

Métrica	Valor Médio (%)	Desvio Padrão (%)	Descrição da Métrica
WER	10,88%	2,52%	Taxa de Erro por Palavra
CER	2,93%	0,77%	Taxa de Erro por Caractere

Os resultados indicam que o modelo apresentou desempenho consistente ao longo das 120 consultas simuladas. O valor médio de WER obtido (10,88%) aproxima-se da faixa reportada na literatura para sistemas de ASR aplicados a conversas médico-paciente, como observado por Tran et al. (2023), que reportam WER entre 8,8% e 10,5% em contexto clínico. Ressalta-se que, no presente estudo, foi utilizado um modelo generalista sem ajuste fino específico para terminologia médica, o que reforça a competitividade do desempenho observado no cenário experimental considerado.

O desvio padrão relativamente baixo (2,52%) sugere estabilidade no comportamento do modelo frente à diversidade temática e linguística introduzida na base sintética, indicando ausência de variações extremas de desempenho entre especialidades médicas ou perfis sociolinguísticos simulados.

4.2. Análise Qualitativa e Impacto da Variabilidade Linguística

A análise qualitativa das transcrições revelou padrões recorrentes de substituição lexical e normalização semântica, especialmente em trechos contendo expressões coloquiais, descrições informais de sintomas e termos médicos menos frequentes. Observou-se, por exemplo, a substituição de expressões como “tô com uma dorzinha chata no peito” por “estou com dor no peito”, evidenciando tendência do modelo à formalização da linguagem. Em outros casos, termos regionais como “tontura danada” foram transcritos como “muita tontura”, indicando normalização semântica para vocabulário mais padronizado.

Também foram identificadas substituições envolvendo terminologia médica menos frequente. Em determinados trechos, “formigamento na perna” foi transcrito como “dor na perna”, alterando parcialmente a especificidade clínica do sintoma relatado. Esse tipo de erro sugere que o modelo tende a aproximar termos menos frequentes de palavras semanticamente mais comuns no treinamento.

Apesar dessas ocorrências, a baixa CER (2,93%) indica boa preservação estrutural do conteúdo textual, com erros concentrados predominantemente no nível de palavra completa, e não em fragmentação ortográfica ou perda de caracteres. Esses achados corroboram a premissa central deste estudo: a inserção controlada de variabilidade linguística em bases sintéticas permite revelar limitações e padrões de erro de modelos generalistas de ASR no contexto clínico brasileiro, fornecendo subsídios para futuros processos de ajuste fino (fine-tuning) e especialização.

5. Discussão

Os resultados obtidos evidenciam uma diferença relevante entre o desempenho de modelos de ASR avaliados em benchmarks públicos e em cenários clínicos com maior variabilidade linguística. Enquanto modelos ajustados especificamente para o português brasileiro podem alcançar WER inferiores a 8% em conjuntos de dados como o Common Voice, compostos majoritariamente por fala estruturada e acusticamente controlada, o desempenho observado no presente estudo (WER = 10,88%) reflete maior complexidade linguística e contextual.

A Tabela 3 evidencia que a diferença de desempenho não deve ser interpretada apenas como limitação do modelo, mas como reflexo da complexidade linguística introduzida no corpus sintético. Elementos como regionalismos, informalidade lexical e variações semânticas ampliam a dificuldade de transcrição automática, aproximando o cenário experimental de condições mais realistas de aplicação clínica.

Tabela 3. Comparação entre benchmark público e cenário clínico simulado

Cenário	Tipo de Dados	Modelo	WER (%)	Características
Benchmark público (Common Voice)	Leitura estruturada, dados limpos	Modelos fine-tuned PT-BR	4–8%	Baixa variabilidade, ambiente controlado
Cenário clínico simulado (este trabalho)	Teleconsulta com regionalismos e informalidade	Whisper-base	10,88%	Linguagem espontânea e diversidade sociolinguística

Essa constatação reforça a necessidade de corpora especializados e estratégias de ajuste fino direcionadas ao domínio clínico brasileiro.

6. Conclusão

Este trabalho apresentou a construção e validação experimental de um corpus sintético de teleconsultas médicas em português brasileiro, estruturado para incorporar variabilidade regional, informalidade lexical e diversidade temática. Diferentemente de benchmarks públicos compostos majoritariamente por fala estruturada e acusticamente

controlada, a base proposta buscou simular características linguísticas mais próximas do contexto clínico brasileiro.

A avaliação com o modelo Whisper-base resultou em WER médio de 10,88% e CER de 2,93%, demonstrando desempenho consistente ao longo das 120 consultas simuladas. No entanto, quando comparado a benchmarks públicos de fala limpa, nos quais modelos ajustados para o português brasileiro podem alcançar WER inferiores a 8%, observa-se uma degradação mensurável de desempenho no cenário com maior variabilidade sociolinguística. Essa diferença evidencia que avaliações baseadas exclusivamente em conjuntos de dados controlados podem não refletir integralmente os desafios presentes em interações clínicas reais.

Os achados reforçam que modelos generalistas, embora robustos em tarefas padronizadas, demandam estratégias de adaptação ao domínio e ajuste fino direcionado à terminologia médica e à diversidade regional. Nesse contexto, a principal contribuição deste estudo não reside apenas na geração de dados sintéticos, mas na demonstração experimental de que a inserção controlada de regionalismos e informalidade revela limitações estruturais dos sistemas atuais de ASR.

Como trabalhos futuros, propõe-se a ampliação do corpus com inclusão de ruído ambiente, sobreposição de fala e maior diversidade de especialidades, bem como a realização de experimentos de fine-tuning supervisionado para avaliar ganhos quantitativos decorrentes da especialização do modelo ao domínio clínico brasileiro.

Referências

Catapan, A. et al. (2020). Teleconsultation: Doctor–Patient Relationship. *Revista Brasileira de Educação Médica*.

Chen, R. J. et al. (2021). Synthetic data in machine learning for medicine and healthcare. *Nature Biomedical Engineering*, 5, 493–497.

Chen, R.; Lu, Y.; Wong, S. (2019). Generating Synthetic Electronic Health Records Using Generative Adversarial Networks. *Journal of the American Medical Informatics Association*, 26(8), 774–785. https://doi.org/10.1093/jamia/ocz094

Da Silva, D. F.; Freitas, E. P.; Souza, J. M. (2019). Automatic Speech Recognition for Brazilian Portuguese: A Survey of Approaches and Resources. *Speech Communication*, 114, 121–154.

Luo, X.; Zhou, L.; Adelgais, K.; Zhang, Z. (2025). Assessing the Effectiveness of Automatic Speech Recognition Technology in Emergency Medicine Settings: A Comparative Study of Four AI-powered Engines. *Journal of Healthcare Informatics Research*, 9(3), 494–512.

Lybarger, K. et al. (2023). Automatic Transcription and Structuring of Clinical Conversations. *Journal of the American Medical Informatics Association*.

Nguyen, D. P. et al. (2024). VietMedASR: A Vietnamese Medical Speech Recognition Corpus. In *Proceedings of Interspeech 2024*.

Peivandi, S.; Ahmadian, L.; Farokhzadian, J.; Jahani, Y. (2022). Evaluation and comparison of errors on nursing notes created by online and offline speech recognition technology and handwritten: an interventional study. *BMC Medical Informatics and Decision Making*, 22(1), 96.

Sanni, M. et al. (2025). Afrispeech-dialog: a benchmark dataset for spontaneous english conversations in healthcare and beyond. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (pp. 8399-8417).

Sinsky, C. et al. (2016). Allocation of Physician Time in Ambulatory Practice: A Time and Motion Study in 4 Specialties. *Annals of Internal Medicine*, 165(11), 753–760.

Tran, B. D.; Mangu, R.; Tai-Seale, M.; Lafata, J. E.; Zheng, K. (2023). Automatic speech recognition performance for digital scribes: a performance comparison between general purpose and specialized models tuned for patient-clinician conversations. *AMIA Annual Symposium Proceedings*, 1072–1080.

Zolnoori, M. et al. (2024). Decoding disparities: evaluating automatic speech recognition system performance in transcribing Black and White patient verbal communication with nurses in home healthcare. *JAMIA Open*, 7(4), ooae130.