IV Workshop de Matemática, Estatística e Computação Aplicadas à Indústria

MedDialogue-Audio: Corpus Sintético de Diálogos Médicos para Avaliação de Modelos ASR

Aline Elí Gassenn[1], José F. Rodrigues-Jr[2]
ICMC-USP
Luis G. M. Andrade[3]
UNESP
Douglas Teodoro[4]
Universidade de Genebra

1 Introdução

O uso de sistemas baseados em inteligência artificial tem se expandido para diferentes setores industriais, incluindo o setor de saúde, que vem incorporando tecnologias de processamento de fala para otimizar fluxos de trabalho e reduzir custos operacionais. Entre essas tecnologias, o reconhecimento automático de fala (ASR) tem sido aplicado à automatização de registros clínicos e à integração de informações em sistemas hospitalares, com potencial para aumentar a produtividade e minimizar erros de transcrição em ambientes de alta demanda.

A adoção de soluções dessa natureza depende de modelos capazes de operar sob condições acústicas variáveis, típicas de clínicas e hospitais. Entretanto, o desenvolvimento e a validação de modelos robustos enfrentam a escassez de bases públicas de fala médica, o que limita a reprodução de experimentos e o avanço de aplicações voltadas ao setor produtivo. Restrições éticas e legais relacionadas à privacidade de dados médicos tornam a coleta de amostras reais um processo custoso e de difícil padronização.

Neste contexto, o trabalho apresenta o MedDialogue-Audio, um corpus sintético de diálogos em inglês entre médicos e pacientes, desenvolvido para apoiar pesquisas e aplicações industriais em reconhecimento de fala no domínio da saúde. O conjunto foi derivado do corpus textual MedDialog-EN e transformado em áudio por meio de um processo automatizado de síntese e adição controlada de ruído, de modo a simular diferentes condições acústicas.

[1]aline.gassenn@usp.br
[2]junio@icmc.usp.br
[3]gustavo.modelli@unesp.br
[4]douglas.teodoro@unige.ch

O objetivo é oferecer um recurso aberto e reprodutível que permita avaliar o desempenho de modelos ASR em cenários de fala médica, contribuindo para o desenvolvimento de soluções tecnológicas aplicáveis a ambientes hospitalares, clínicas e sistemas de telemedicina.

2 Materiais e Métodos

O corpus MedDialogue-Audio foi desenvolvido a partir do conjunto textual MedDialog-EN [3] [4], composto por pares de frases em inglês entre pacientes e médicos, obtidos de plataformas de teleatendimento. Cada registro contém uma descrição breve do paciente e a resposta correspondente do profissional. A escolha dessa base se deve à sua cobertura de múltiplas especialidades clínicas e à estrutura adequada para geração de amostras de fala segmentadas e independentes.

Os textos passaram por um processo de normalização linguística com o objetivo de corrigir inconsistências e preparar o material para a conversão em áudio. Foram removidos caracteres não textuais e substituídas abreviações e unidades de medida por suas formas completas. Em seguida, aplicou-se o modelo de linguagem *gpt-4o-mini* para revisão sintática e padronização das expressões clínicas. O modelo também inferiu informações básicas sobre o paciente, como gênero e faixa etária, utilizadas apenas para seleção de vozes na etapa de síntese.

A geração de fala foi realizada com o modelo Orpheus TTS [2], responsável pela conversão de texto em áudio. Foram criados dois arquivos por par de frases, correspondentes ao paciente e ao médico, totalizando 21.068 amostras originais. Para representar condições de gravação observadas em ambientes hospitalares, cada arquivo recebeu seis variações com ruído: três níveis de ruído branco (2%, 6% e 10%) e três níveis de ruído de fundo provenientes de gravações hospitalares (20%, 40% e 60%) [5] [6]. O corpus final contém 147.476 amostras de áudio.

As amostras foram nomeadas segundo o padrão `[ID]_[SPEAKER][TIPO][NÍVEL].wav`, em que o identificador representa o par de frases, `SPEAKER` indica o interlocutor, `TIPO` define o tipo de ruído (original, branco ou ambiente) e `NÍVEL` expressa sua intensidade percentual. Um arquivo de metadados complementa o conjunto, reunindo informações como duração, energia média, frequência fundamental, centróide espectral, razão harmônico-ruído e transcrição textual associada [1].

A etapa de avaliação experimental foi conduzida com três modelos de reconhecimento automático de fala (ASR): Whisper-base, Wav2Vec 2.0 e HuBERT. Os testes utilizaram uma amostra correspondente a 10% do corpus, incluindo áudios originais e com ruído. As métricas consideradas foram a Taxa de Erro de Palavras (WER), o Reconhecimento de Termos Médicos (MTRA) e o BERTScore, empregadas para medir o efeito do ruído sobre a precisão lexical e semântica das transcrições.

3 Resultados

O processo de construção do corpus MedDialogue-Audio resultou em 147.476 amostras de áudio distribuídas entre versões originais e com ruído. As medições acústicas apresentaram valores regulares entre as amostras, indicando consistência no processo de síntese. A normalização textual reduziu o número de erros ortográficos e estruturais, garantindo maior uniformidade nas transcrições.

Os resultados mostraram aumento da taxa de erro de palavras com o crescimento do nível de ruído. O modelo Whisper-base apresentou menor variação de desempenho entre as condições testadas, enquanto o Wav2Vec 2.0 foi o mais sensível à degradação do sinal. O HuBERT apresentou resultados intermediários. A Figura 1 apresenta a variação da Taxa de Erro de Palavras (WER) sob diferentes níveis de ruído ambiente hospitalar.

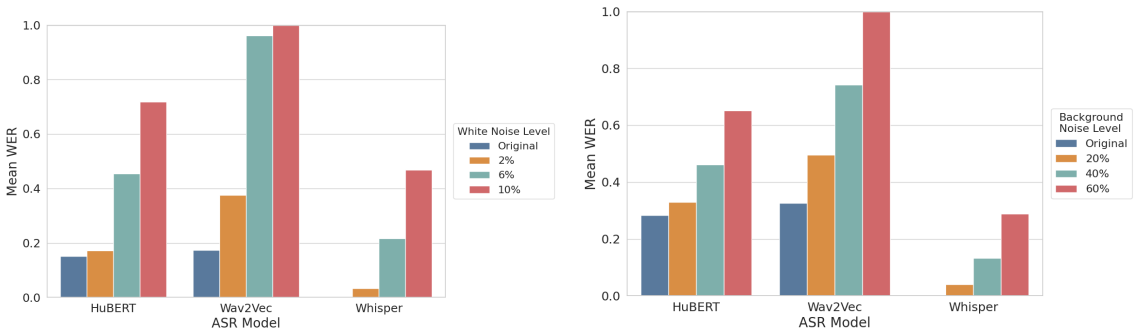

Figura 1: Taxa de erro de palavras (WER) sob condições de ruído branco (à esquerda) e ruído de fundo (à direita). Observa-se que, para o modelo Whisper, o valor de WER no áudio original foi tão baixo que não é visualmente perceptível na escala do gráfico adotada, resultando na aparente ausência da barra correspondente.

Os experimentos indicam que o corpus pode ser utilizado em análises de desempenho de modelos ASR sob condições acústicas variáveis. A estrutura do conjunto permite avaliar, de modo controlado, os efeitos do ruído sobre a transcrição de dados clínicos e apoiar o desenvolvimento de sistemas voltados à automação de registros em saúde.

4 Conclusão

Este trabalho apresentou o MedDialogue-Audio, um corpus sintético de diálogos médico-paciente em inglês desenvolvido para experimentos com modelos de reconhecimento automático de fala em condições acústicas variáveis. O conjunto foi derivado do MedDialog-EN e gerado por meio de um processo automatizado que combinou normalização linguística, síntese de fala e adição controlada de ruídos. O resultado compreende 147.476 amostras de áudio acompanhadas de metadados e transcrições padronizadas.

Os experimentos realizados com os modelos Whisper-base, Wav2Vec 2.0 e HuBERT mostraram que o desempenho de sistemas ASR varia conforme o nível e o tipo de ruído, o que reforça a necessidade de bases que representem cenários reais de operação. O corpus permite avaliar o comportamento de modelos de transcrição em contextos semelhantes aos encontrados em aplicações industriais, como registros automatizados de atendimento, integração de sistemas hospitalares e monitoramento de comunicação em processos produtivos.

O MedDialogue-Audio está disponível publicamente em: `https://huggingface.co/datasets/aline-gassenn/MedDialog-Audio`.

Agradecimentos

Este estudo foi parcialmente financiado pela Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP – processos 2024/04761-0, 2019/07665-4, 2024/13328-9), pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq 304805/2025-4), pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES – processo 001) e pelo Fundo Nacional de Ciência da Suíça (SNSF).

Referências

[1] A. E. Gassenn, L. G. M. Andrade, D. Teodoro e J. F. Rodrigues-Jr. *Medical Dialogue Audio Transcription: Dataset and Benchmarking of ASR Models*. In: Dataset Showcase Workshop (DSW), 7., 2025, Fortaleza, CE. Anais [...]. Porto Alegre: Sociedade Brasileira de Computação, 2025, pp. 71–82. DOI: 10.5753/dsw.2025.248010.

[2] CanopyAI. *Orpheus-TTS: Towards Human-Sounding Speech*. GitHub, 2025. Disponível em: `https://github.com/canopyai/Orpheus-TTS`.

[3] C. Tang, H. Zhang, T. Loakman, C. Lin e F. Guerin. *Terminology-Aware Medical Dialogue Generation*. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2023, pp. 1–5.

[4] G. Zeng, W. Yang, Z. Ju, Y. Yang, S. Wang, R. Zhang, M. Zhou, J. Zeng, X. Dong, R. Zhang, H. Fang, P. Zhu, S. Chen e P. Xie. *MedDialog: Large-scale Medical Dialogue Datasets*. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 9241–9250. DOI: 10.18653/v1/2020.emnlp-main.743.

[5] S. N. Ali e S. B. Shuvo. *Hospital Ambient Noise Dataset*. Kaggle, 2021. DOI: 10.34740/KAGGLE/DSV/2173743. Disponível em: `https://www.kaggle.com/dsv/2173743`.

[6] S. N. Ali, S. B. Shuvo, M. I. S. Al-Manzo, A. Hasan e T. Hasan. *An End-to-End Deep Learning Framework for Real-Time Denoising of Heart Sounds for Cardiac Disease Detection in Unseen Noise*. IEEE Access, vol. 11, pp. 87887–87901, 2023. DOI: 10.1109/ACCESS.2023.3292551.