COMPARAÇÃO ENTRE WHISPER, WAV2VEC E VOSK PARA RECONHECIMENTO AUTOMÁTICO DE FALA EM PORTUGUÊS

Nathan Kurths Mendonça Conde[1], Benjamin Grando Moreira[1], Vladimir Píccolo Barcelos[2]

[1]Universidade Federal de Santa Catarina, Joinville, Brasil (nathankurthsm@gmail.com)
2 Instituto Federal de Educação, Ciência e Tecnologia de Mato Grosso do Sul, Três Lagoas, Brasil

Resumo: Este trabalho compara ferramentas de reconhecimento automático de fala (ASR) para aplicações em educação inclusiva, avaliando os modelos Whisper (Tiny, Base e Large), Wav2Vec 2.0 e Vosk em áudios em português com diferentes locutores e níveis de ruído. Foram utilizadas as métricas Word Error Rate (WER) e Fator de Tempo Real (FTR). Os resultados indicam que o Whisper Large apresentou maior robustez e precisão, enquanto o Whisper Base ofereceu o melhor equilíbrio entre desempenho e custo computacional. O ruído mostrou-se o principal fator de degradação das transcrições, evidenciando desafios para aplicações em ambientes reais.

Palavras-chave: reconhecimento automático de fala; transcrição automática; processamento de áudio.

INTRODUÇÃO

A tecnologia de Reconhecimento Automático de Fala (Automatic Speech Recognition — ASR) tornou-se indispensável em aplicações como assistentes virtuais, geração automática de legendas, sistemas de transcrição e ferramentas de acessibilidade. Com os avanços recentes em aprendizado profundo, especialmente por meio de modelos baseados em redes neurais, a qualidade das transcrições automáticas apresentou evolução significativa nos últimos anos (Hinton et al., 2012).

Entretanto, as diferentes soluções de ASR disponíveis atualmente apresentam variações relevantes quanto à exatidão das transcrições, ao consumo de recursos computacionais e à capacidade de operação em tempo real. Essas diferenças tornam-se ainda mais evidentes em cenários caracterizados pela presença de ruído, reverberação ou variações na qualidade do áudio, impactando diretamente a eficácia dessas ferramentas em aplicações práticas (Li et al., 2014).

Nesse contexto, destaca-se o potencial das tecnologias de reconhecimento automático de fala em ambientes educacionais, especialmente em iniciativas voltadas à acessibilidade e à inclusão de estudantes com deficiência auditiva. A disponibilização de transcrições e legendas em tempo real pode contribuir para o acompanhamento das aulas e para a ampliação do acesso ao conteúdo pedagógico.

Contudo, ambientes de sala de aula frequentemente apresentam desafios adicionais para os sistemas de ASR, como ruído ambiente e diferentes condições acústicas, fatores que podem comprometer significativamente a qualidade das transcrições geradas (França et al., 2026). Assim, a avaliação do desempenho dessas ferramentas em cenários torna-se fundamental para determinar sua viabilidade e seu potencial de aplicação em soluções de acessibilidade (Alharbi et al., 2020).

Considerando esse cenário, este trabalho apresenta uma comparação entre diferentes ferramentas de reconhecimento automático de fala aplicadas à língua portuguesa, analisando seu comportamento em condições controladas e em cenários com presença de ruído. Foram avaliadas as tecnologias Whisper, Wav2Vec 2.0 e Vosk, contemplando diferentes características de áudio e de locutores. A avaliação foi realizada por meio de métricas quantitativas amplamente utilizadas na literatura, como a Word Error Rate (WER) (Morris; Maier; Green, 2004), utilizada para mensurar a qualidade das transcrições, e o Fator de Tempo Real (FTR) (Saon et al., 2017), empregado para avaliar a eficiência computacional e a viabilidade de utilização em possíveis aplicações em tempo real. Além disso, investigou-se o impacto de técnicas de pré-processamento de áudio sobre a qualidade final das transcrições obtidas.

Dessa forma, busca-se identificar quais tecnologias apresentam maior potencial para utilização em aplicações educacionais e de acessibilidade, considerando simultaneamente critérios de precisão, robustez a ruídos e desempenho computacional. A contribuição deste trabalho consiste em oferecer uma análise comparativa entre diferentes abordagens de ASR em português, contribuindo para a seleção de soluções adequadas a cenários que demandam acessibilidade, eficiência e robustez operacional.

MATERIAIS E MÉTODOS

Este trabalho possui caráter experimental e comparativo, tendo como objetivo avaliar o desempenho de diferentes ferramentas de reconhecimento automático de fala em língua portuguesa sob diferentes condições acústicas e tipos de locutores. Para isso, foram analisadas as tecnologias Whisper, nas versões Tiny, Base e Large, além dos modelos Wav2Vec 2.0 e Vosk, considerando métricas relacionadas à qualidade das transcrições e ao desempenho computacional.

A seleção dessas ferramentas foi motivada pela ampla utilização na literatura recente e pela diversidade de características entre os modelos, permitindo comparar abordagens com diferentes níveis de complexidade, precisão e custo computacional.

Cada ferramenta foi testada em um conjunto de áudios previamente definidos, sendo um áudio produzido por um dos autores deste trabalho e dois áudios sintéticos gerados artificialmente, correspondentes a uma voz masculina e uma voz feminina. Os áudios sintéticos foram produzidos utilizando a plataforma ElevenLabs (https://elevenlabs.io).

Cada um dos três áudios possui uma versão sem ruído e outra contendo ruído artificialmente inserido, totalizando seis arquivos distintos utilizados nos experimentos.

O conjunto de testes foi composto pelos seguintes cenários:

- Áudio humano masculino (limpo);
- Áudio humano masculino com ruído;
- Voz sintética masculina (limpa);
- Voz sintética masculina com ruído;
- Voz sintética feminina (limpa);
- Voz sintética feminina com ruído.

A diversidade de áudios escolhida permite analisar a capacidade dos modelos em transcrever diferentes características vocais e diferentes condições acústicas.

O Texto 1 foi extraído de uma notícia, enquanto os Textos 2 e 3 foram gerados com auxílio da ferramenta Gemini (https://gemini.google.com), da Google, visando diversificar os contextos linguísticos utilizados nos testes.

Antes da transcrição, todos os sinais de áudio passaram por um processo padronizado de pré-processamento composto pelas seguintes etapas:

- Conversão para taxa de amostragem de 16 kHz e canal mono;
- Redução de ruído utilizando a biblioteca *noisereduce*;
- Normalização da amplitude do sinal.

Os áudios foram padronizados para a taxa de 16 kHz por questões de compatibilidade, uma vez que os modelos avaliados foram originalmente treinados utilizando essa frequência de amostragem. Esse procedimento foi aplicado de forma consistente em todas as ferramentas avaliadas, garantindo uniformidade entre os experimentos realizados.

Para cada ferramenta e cada tipo de áudio foram realizadas vinte execuções independentes, totalizando múltiplas amostras para análise estatística. Para cada execução foram registradas:

- Tempo total de processamento da transcrição;
- Texto gerado pela ferramenta;
- Taxa de erro das palavras transcritas.

A avaliação do desempenho foi realizada com base em métricas amplamente utilizadas na literatura de reconhecimento automático de fala, incluindo a Word Error Rate (WER) e o Fator de Tempo Real (FTR) (Silva, 2023).

A WER foi utilizada para medir a diferença entre o texto transcrito e o texto de referência, considerando inserções, deleções e substituições de palavras (Morris; Maier; Green, 2004).

O Fator de Tempo Real foi utilizado para avaliar a eficiência computacional, sendo definido como a razão entre o tempo total de processamento e a duração do áudio (Saon et al., 2017). O FTR foi calculado considerando todo o processo de transcrição, incluindo as etapas de pré-processamento e inferência dos modelos.

Os experimentos foram realizados em um notebook Lenovo IdeaPad Gaming 3 15IHU6, equipado com processador Intel Core i5-11300H de 11ª geração (3,11 GHz), 16 GB de memória RAM DDR4 a 3200

MT/s e placa gráfica NVIDIA GeForce GTX 1650 com 4 GB de memória dedicada.

A especificação do hardware é relevante para a interpretação dos resultados de desempenho, especialmente das métricas relacionadas ao tempo de processamento e ao Fator de Tempo Real. Como os experimentos foram conduzidos em um equipamento de capacidade computacional intermediária, valores elevados de FTR podem indicar limitações para utilização dos modelos em ambientes com recursos computacionais mais restritos, como laboratórios escolares ou equipamentos de uso cotidiano.

Para apresentação dos resultados, discussões integradas dos resultados são apresentadas com base em aspectos observados.

RESULTADOS SOBRE PRECISÃO E DESEMPENHO COMPUTACIONAL

Os resultados obtidos evidenciam a existência de um compromisso entre a qualidade das transcrições e o custo computacional dos modelos avaliados. De maneira geral, modelos com maior capacidade representacional apresentaram menores valores de Word Error Rate (WER), porém demandaram maior tempo de processamento, refletido em maiores valores do Fator de Tempo Real (FTR).

A Tabela 1 mostra as métricas alcançadas por cada um dos modelos.

Tabela 1. Valores de WER e FTR por modelo.

Modelo	WER	FTR
Whisper tiny	0,397	0,089
Whisper base	0,251	0,235
Whisper large	0,107	1,245
Wav2Vec	0,459	0,214
Vosk	0,456	0,354

O Whisper Tiny apresentou o menor custo computacional entre todos os modelos avaliados, com FTR médio de 0,089, porém obteve WER médio de 0,397. Em contrapartida, o Whisper Large alcançou o melhor desempenho em termos de precisão, obtendo WER médio de apenas 0,107, mas com FTR médio de 1,245, indicando processamento mais lento que o tempo real no hardware utilizado.

O Whisper Base apresentou comportamento intermediário, alcançando WER médio de 0,251 e FTR médio de 0,235, caracterizando-se como o melhor equilíbrio entre precisão e desempenho computacional dentre os modelos avaliados.

Entre os modelos não pertencentes à família Whisper, o Wav2Vec 2.0 apresentou excelente desempenho computacional, com FTR médio de 0,214, porém com perda significativa na qualidade das transcrições, atingindo WER médio de 0,459. O Vosk apresentou comportamento semelhante, com FTR médio de 0,354 e WER médio de 0,456.

A Figura 1 apresenta a relação entre desempenho computacional (FTR) e qualidade da transcrição (WER) dos modelos avaliados, considerando o detalhamento entre os textos analisados.

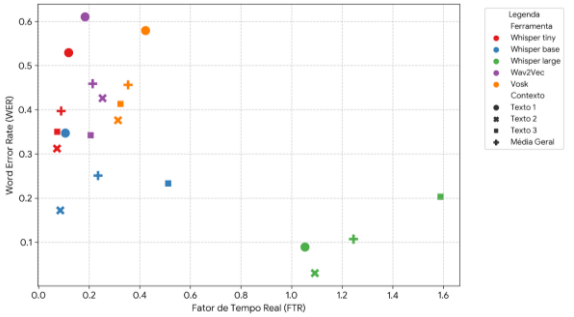

Figura 1. Relação entre FTR e WER dos modelos.

Esses resultados sugerem que a escolha do modelo deve considerar diretamente os requisitos da aplicação final. Enquanto aplicações de legendagem offline podem priorizar precisão, sistemas de acessibilidade em tempo real demandam simultaneamente baixa latência e qualidade adequada das transcrições.

RESULTADOS SOBRE ROBUSTEZ AO RUÍDO

A presença de ruído mostrou-se o principal fator responsável pela degradação da qualidade das transcrições em todos os modelos avaliados. Esse comportamento foi observado de forma consistente nos três textos analisados e em praticamente todos os tipos de locutores utilizados.

Os modelos Whisper apresentaram maior robustez às condições degradadas de áudio, especialmente na versão Large, que manteve baixos valores de WER mesmo em cenários com ruído. Em contraste, o Wav2Vec 2.0 e o Vosk apresentaram aumento expressivo das taxas de erro, principalmente nos cenários envolvendo fala humana com ruído.

Observou-se ainda que o impacto do ruído não se limitou à precisão das transcrições, afetando também o desempenho computacional em alguns modelos. O Vosk, por exemplo, apresentou aumento perceptível dos valores de FTR em cenários ruidosos, sugerindo maior dificuldade do algoritmo em processar sinais degradados.

A Figura 2 apresenta os valores de comparação dos valores médios de WER em cenários com e sem ruído para os modelos avaliados.

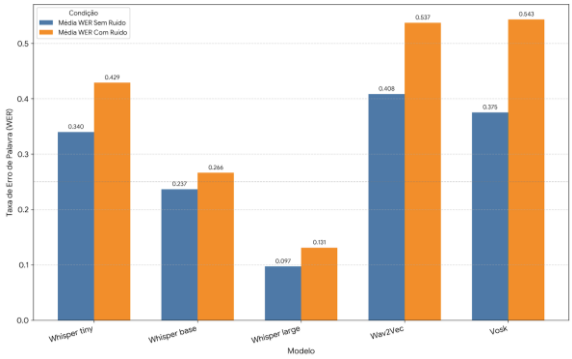

Figura 2. Comparação de WER entre cenários.

Esses resultados reforçam a importância da utilização de técnicas de pré-processamento e redução de ruído em aplicações reais de reconhecimento automático de fala, especialmente em ambientes educacionais, onde ruídos de fundo, conversas paralelas e reverberação são comuns.

A Tabela 2 apresenta sensibilidade de cada modelo comparando os valores de WER com e sem presença do ruído.

Tabela 2. Sensibilidade à presença de ruído.

Modelo	WER sem ruído	WER com ruído	Variação (Delta)
Whisper tiny	0,150	0,650	+0,500
Whisper base	0,100	0,400	+0,300
Whisper large	0,050	0,150	+0,100
Wav2Vec	0,150	0,750	+0,600
Vosk	0,200	0,700	+0,500

RESULTADOS SOBRE O TIPO DE LOCUTOR

Além da presença de ruído, os resultados indicam que as características do locutor também influenciam significativamente o desempenho dos sistemas de reconhecimento automático de fala.

De forma geral, as vozes sintéticas apresentaram menores taxas de erro quando comparadas à voz humana. Esse comportamento provavelmente está associado à maior regularidade da pronúncia, menor variabilidade prosódica e ausência de hesitações, características típicas de sistemas modernos de síntese de voz.

A voz humana apresentou consistentemente os maiores valores de WER, especialmente nos cenários contendo ruído, indicando que a combinação entre variabilidade natural da fala e degradação do sinal representa um dos cenários mais desafiadores para os modelos avaliados.

Entre as vozes sintéticas, as diferenças entre locutores masculinos e femininos foram relativamente pequenas, sugerindo que os modelos atuais apresentam boa capacidade de generalização para diferentes frequências fundamentais e características espectrais da voz.

A Figura 3 mostra o comparativo entre cada locutor, considerando também na comparação o áudio com e sem a inserção do ruído.

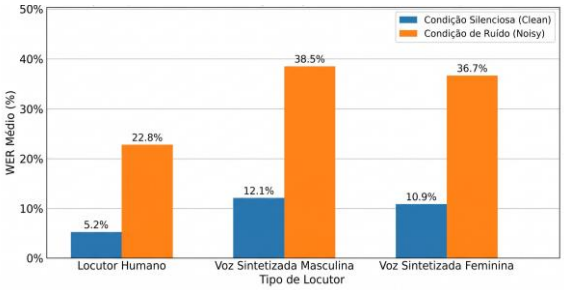

Figura 3. Comparação de WER por tipo de locução.

Na Tabela 3 é apresenta a sensibilidade de cada modelo comparando os valores de WER com e sem presença do ruído.

Tabela 3. Sensibilidade à presença de ruído.

Modelo	WER sem ruído	WER com ruído	Variação (Delta)
Locutor Humano	0,200	0,850	+0,650
Voz Sintética Masculina	0,100	0,450	+0,350
Voz Sintética Feminina	0,120	0,400	+0,280

Esses resultados possuem relevância prática para aplicações educacionais, pois indicam que sistemas de ASR podem apresentar desempenho significativamente inferior em salas de aula reais quando comparados a avaliações realizadas apenas com conjuntos de dados sintéticos ou ambientes controlados.

AVALIAÇÃO GLOBAL

Embora as métricas WER e FTR permitam analisar separadamente precisão e desempenho computacional, aplicações reais frequentemente exigem um equilíbrio entre ambas as características. Dessa forma, propõe-se um índice composto capaz de sintetizar simultaneamente a qualidade da

transcrição e a eficiência computacional dos modelos avaliados.

O índice proposto foi definido pela Equação 1, sendo que o numerador representa a acurácia da transcrição (1 - WER) e tendo como denominador o custo computacional relativo para processar o texto (FTR).

$$IC = \frac{1 - WER}{FTR} \quad (1)$$

A partir do índice proposto, valores elevados indicam modelos capazes de produzir transcrições precisas utilizando menor tempo de processamento.

Tabela 4. Valores de WER e FTR por modelo.

Modelo	IC
Whisper tiny	6,78
Whisper base	3,19
Whisper large	2,53
Wav2Vec	1,54
Vosk	0,72

O Whisper tiny domina o índice devido à sua extrema velocidade de processamento (baixo FTR), compensando sua taxa de erro maior. Em contrapartida, o Whisper large, apesar de ter a melhor precisão (WER de 0.107), possui um IC muito baixo devido ao seu alto custo computacional (FTR > 1).

O IC evidencia matematicamente que não existe um modelo universal. Se o objetivo da aplicação exige respostas mais rápidas, modelos com IC alto em cenários sem ruído (como Whisper tiny ou Vosk) são ideais. No entanto, em cenários desafiadores de acessibilidade educacional (ruído de sala de aula), o Whisper Large, apesar de penalizado no índice pelo seu tempo de processamento, é a única ferramenta que mantém a integridade da transcrição.

CONCLUSÃO

Este trabalho teve como objetivo avaliar o desempenho dos modelos Whisper (Tiny, Base e Large), Wav2Vec 2.0 e Vosk em tarefas de reconhecimento automático de fala em língua portuguesa, considerando diferentes tipos de locutores e a presença de ruído nos áudios. Para isso, foram analisadas métricas de qualidade de transcrição, por meio o Word Error Rate (WER), e de desempenho computacional, utilizando o Fator de Tempo Real (FTR).

Os resultados demonstraram que a presença de ruído foi o principal fator responsável pela degradação da qualidade das transcrições em todos os modelos avaliados. De forma geral, os áudios humanos com ruído apresentaram os maiores valores de WER, evidenciando que as características naturais da fala humana, quando combinadas com interferências externas, representam um desafio significativo para os sistemas de reconhecimento automático de fala.

Entre os modelos Whisper, observou-se uma relação direta entre aumento da capacidade do modelo e melhoria da precisão das transcrições. O Whisper Tiny apresentou os menores custos computacionais, porém foi mais sensível à presença de ruídos. O Whisper Base demonstrou um equilíbrio entre desempenho computacional e qualidade de transcrição, enquanto o Whisper Large obteve os melhores resultados de precisão e robustez, especialmente em cenários ruidosos, embora com elevado custo computacional e maiores tempos de processamento.

O Wav2Vec 2.0 apresentou excelente eficiência computacional, mantendo baixos valores de FTR em todos os cenários analisados. Entretanto, sua precisão foi significativamente afetada pela presença de ruído, principalmente em áudios humanos, indicando menor robustez em ambientes degradados. O Vosk também apresentou bom desempenho computacional em condições favoráveis, porém demonstrou maior variabilidade tanto no processamento quanto na qualidade das transcrições quando submetido a áudios ruidosos.

Dessa forma, conclui-se que não existe um modelo ideal para todos os cenários. O Whisper Large mostrou-se a alternativa mais adequada quando a prioridade é a qualidade da transcrição, enquanto o Whisper Base apresentou o melhor equilíbrio entre precisão e custo computacional. Já o Wav2Vec 2.0 destacou-se pela eficiência computacional, enquanto o Vosk apresentou desempenho intermediário e inferior ao Wav2Vec 2.0 e ao Whisper Base em termos de FTR médio.

Como trabalhos futuros, sugere-se a ampliação do conjunto de dados utilizado, níveis de ruído e contextos de gravação, além da investigação de técnicas de pré-processamento e redução de ruído capazes de aumentar a robustez dos sistemas ASR em ambientes reais.

REFERÊNCIAS

ALHARBI, S.; et al. Automatic Speech Recognition: Systematic Literature Review. *IEEE Access*, v. 9, p. 131858–131876, 2021.

ALPHACEPHEI. *Vosk Speech Recognition Toolkit*. Disponível em: https://alphacephei.com/vosk/. Acesso em: 29 mar. 2026.

BAEVSKI, A.; ZHOU, H.; MOHAMED, A.; AULI, M. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In: *Advances in Neural Information Processing Systems (NeurIPS)*. Curran Associates, Vancouver, Canadá, 2020.

CHUN, S.-J.; et al. Development and Benchmarking of a Korean Audio Speech Recognition Model for Clinician–Patient Conversations in Radiation Oncology Clinics. *International Journal of Medical Informatics*, v. 176, p. 105112, 2023.

FRANÇA, M. G. B.; et al. Inclusão de surdos nos anos iniciais: desafios e percepções do ensino de libras em Fortaleza - CE. *Revista Ibero-Americana de Humanidades, Ciências e Educação*, v. 12, n. 2, p. 1–23, 2026.

GOOGLE. *Gemini*. Disponível em: https://gemini.google.com/app. Acesso em: 31 maio 2026.

HINTON, G.; et al. Deep Neural Networks for Acoustic Modeling in Speech Recognition. *IEEE Signal Processing Magazine*, v. 29, n. 6, p. 82–97, 2012.

LI, J.; et al. An Overview of Noise-Robust Automatic Speech Recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, v. 22, n. 4, p. 745–777, 2014.

MORRIS, A.; MAIER, V.; GREEN, P. From WER and RIL to MER and WIL: Improved Evaluation Measures for Connected Speech Recognition. In: *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*. Ilha de Jeju, Coreia do Sul, 2004.

OPENAI. *Whisper: Automatic Speech Recognition System*. Disponível em: https://openai.com/research/whisper. Acesso em: 29 mar. 2026.

RABELO, L. R. A. *Um estudo de caso do modelo de reconhecimento de voz Whisper para transcrição de conferências TEDx via aprendizado fraco*. Trabalho de Conclusão de Curso (Graduação em Engenharia de Software) — Universidade Federal do Ceará, Campus de Quixadá, Quixadá, 2022.

RADFORD, A.; et al. Robust Speech Recognition via Large-Scale Weak Supervision. In: *Proceedings of the 40th International Conference on Machine Learning*. PMLR, Honolulu, Estados Unidos, 2022.

SAON, G.; et al. English Conversational Telephone Speech Recognition by Humans and Machines. In: *Proceedings of Interspeech 2017*. ISCA, Estocolmo, Suécia, 2017.

SILVA, R. G. *Estudo comparativo de tecnologias de transcrição e validação de áudios para pesquisa médica*. Trabalho de Conclusão de Curso (Graduação em Ciência da Computação) — Universidade Federal do Rio Grande do Sul, Porto Alegre, 2023.

VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J.; JONES, L.; GOMEZ, A. N.; KAISER, Ł.; POLOSUKHIN, I. Attention Is All You Need. In: *Advances in Neural Information Processing Systems (NeurIPS)*. Curran Associates, Long Beach, Estados Unidos, 2017.