

Avaliação de Ferramentas de Reconhecimento Automático de Fala Visando Aplicações em Educação Inclusiva

Evaluation of Automatic Speech Recognitions Tools for Inclusive Education Applications

Nathan Kurths Mendonça Conde
2026

Universidade Federal de Santa Catarina
Bacharelado em Ciência e Tecnologia, Joinville, Santa Catarina, Brasil
E-mail: nathankurthsm@gmail.com

Sob a orientação do Dr. Benjamin Grando Moreira
e coorientação do Me. Vladimir Píccolo Barcelos

Resumo: Apresenta-se neste artigo uma análise do desempenho de ferramentas de reconhecimento automático de fala (ASR) voltadas para a transcrição de áudios em português. O estudo surgiu a partir do crescimento do uso de soluções baseadas em inteligência artificial em atividades acadêmicas e profissionais que dependem da conversão de fala em texto. O principal objetivo foi comparar diferentes ferramentas de transcrição, observando aspectos como precisão dos resultados, tempo de processamento e comportamento em diferentes condições de áudio. Para a realização dos testes foram utilizados arquivos com variações de duração, qualidade sonora e presença de ruídos, buscando aproximar a análise de situações reais de uso. As transcrições obtidas foram comparadas com textos de referência para identificar erros, limitações e pontos positivos de cada solução avaliada. Durante a análise, foi possível perceber que a qualidade do áudio influencia diretamente no desempenho das ferramentas, principalmente em gravações longas ou com interferências externas. Os resultados indicam que as tecnologias atuais já oferecem resultados satisfatórios em muitos cenários, mas ainda apresentam dificuldades em situações mais complexas, mostrando que o reconhecimento automático de fala continua sendo uma área em constante evolução.

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

Abstract: This study analyzes the performance of automatic speech recognition (ASR) tools focused on audio transcription in Portuguese. The research emerged from the growing use of artificial intelligence-based solutions in academic and professional activities that depend on converting speech into text. The main objective was to compare different transcription tools, considering aspects such as transcription accuracy, processing time, and performance under different audio conditions. To conduct the tests, audio files with variations in duration, sound quality, and background noise were used in order to approximate real-world usage scenarios. The generated transcriptions were compared with reference texts to identify errors, limitations, and strengths of each evaluated solution. During the analysis, it was observed that audio quality directly influences the performance of the tools, especially in long recordings or those containing external interference. The results indicate that

current technologies already provide satisfactory outcomes in many scenarios; however, they still face difficulties in more complex situations, demonstrating that automatic speech recognition remains an area under continuous development.

Keywords: speech recognition; automatic transcription; audio processing.

1. INTRODUÇÃO

O reconhecimento automático de fala (Automatic Speech Recognition - ASR) tornou-se indispensável em áreas como assistentes virtuais, geração automática de legendas e ferramentas de acessibilidade. Com os avanços recentes no aprendizado profundo, especialmente por meio de modelos de redes neurais, a qualidade das transcrições automáticas evoluiu (Hinton *et al.*, 2012).

No entanto, as diferentes soluções de ASR disponíveis apresentam variações relevantes quanto à exatidão, ao consumo de recursos computacionais e à capacidade de operação em tempo real. Essas diferenças tornam-se ainda mais evidentes em cenários com presença de ruído ou variações na qualidade do áudio (Li *et al.*, 2014), impactando diretamente a eficácia dessas ferramentas em aplicações práticas.

Nesse contexto, destaca-se a aplicação dessas tecnologias em ambientes educacionais para colaborar como uma nova abordagem de qualidade requisitada para acrescentar na inclusão de alunos surdos (França *et al.*, 2026), como no acompanhamento de aulas, onde a presença de ruído ambiente, múltiplos falantes e diferentes condições acústicas pode comprometer a qualidade das transcrições. Assim, a avaliação do desempenho dos sistemas de ASR nesse tipo de cenário torna-se fundamental, especialmente considerando seu potencial para promover acessibilidade e apoio ao aprendizado (Alharbi *et al.*, 2021).

Considerando esse contexto, este artigo apresenta uma comparação entre diferentes ferramentas de reconhecimento automático de fala, seguida da análise dos resultados obtidos a partir de testes realizados tanto em ambientes controlados quanto em condições com presença de ruído. A avaliação considera métricas quantitativas, como a *Word Error Rate* (WER) (Morris; Maier; Green; 2004) e o Fator de Tempo Real (FTR) (Saon *et al.*, 2017). além de identificar o impacto de técnicas de pré-processamento de áudio na qualidade das transcrições.

Por fim, apresenta-se a viabilidade dessas ferramentas para uso em tempo real e em contextos educacionais, contribuindo para a identificação de soluções adequadas a cenários que demandam acessibilidade, eficiência e robustez.

A metodologia definida para a análise que resultou na identificação da viabilidade das ferramentas, inclui:

- **Avaliar a precisão das transcrições:** Apontar a exatidão das ferramentas Whisper, Wav2Vec 2.0 e Vosk utilizando a métrica de *Word Error Rate* (WER) (Morris; Maier; Green; 2004);
- **Desempenho em tempo real:** Calcular o FTR para compreender se o processamento dessas ferramentas consegue acompanhar a cadência de uma aula padrão sem gerar atrasos acumulados;

- **Apontar o impacto do ambiente:** Comparar o comportamento dos modelos em áudios com e sem a presença de ruído;
- **Verificar o impacto do pré-processamento:** Distinguir como técnicas de redução de ruído e filtragem de frequências influenciam a qualidade final do texto transcrito;
- **Identificar soluções para acessibilidade:** Esclarecer quais são as ferramentas mais robustas e eficientes para o auxílio de estudantes com deficiência auditiva em ambientes de sala de aula.

2. FUNDAMENTAÇÃO TEÓRICA

Visando comparar o desempenho de diferentes ferramentas de reconhecimento automático de fala, as transcrições de áudio são o ponto de partida para as fases de tratamento e análise, sendo fundamental verificar como diferentes tecnologias de identificação automática da fala podem afetar a qualidade dos resultados. Dessa forma, foram selecionadas e avaliadas três opções de ferramentas de reconhecimento de fala reconhecidas atualmente: Whisper, Wav2Vec2.0 e Vosk, apresentadas a seguir.

O Whisper é um sistema de identificação automática da fala desenvolvido pela OpenAI (Openai, 2022), com base em uma estrutura Transformer encoder-decoder (Vaswani *et al.*, 2017). O modelo passou por um treinamento com quase 680 mil horas de áudios em várias línguas e tarefas, o que garante sua eficiência em vários contextos e idiomas. Um dos maiores benefícios do Whisper é produzir texto bruto de forma direta, eliminando etapas adicionais de normalização inversa, além disso, o modelo realiza outras tarefas além da transcrição, como tradução, identificação de idioma e organização temporal (Radford *et al.*, 2022).

O modelo Whisper disponibiliza diferentes versões, projetadas para equilibrar desempenho e custo computacional. As variantes incluem tiny, base, small, medium e large, que se diferenciam principalmente pelo número de parâmetros, consumo de memória e precisão nos resultados de transcrição (Radford *et al.*, 2022).

De modo geral, modelos menores, como tiny, base e small, apresentam menor custo computacional e maior velocidade de processamento, sendo mais adequados para aplicações em tempo real ou dispositivos com recursos limitados. Por outro lado, modelos maiores, como medium, large e large-v2 oferecem maior precisão, especialmente em cenários com ruído ou múltiplos idiomas, porém exigem maior capacidade de processamento e tempo de execução. A Tabela 1 apresenta um resumo comparativo das principais características dessas versões.

Tabela 1 - Comparação entre as versões do modelo Whisper

Modelo	Parâmetros	Velocidade	Precisão
Tiny	~39M	Muito alta	Baixa
Base	~74M	Alta	Moderada
Small	~244M	Média	Boa

Medium	~769M	Baixa	Muito Boa
Large	~1550M	Muito baixa	Excelente

Fonte: Elaborado pelo Autor (2026).

O Wav2Vec 2.0, proposto por Baevski *et al.* (2020), é um sistema de reconhecimento automático da fala que emprega aprendizado auto-supervisionado para obter representações diretamente de sinais de áudio em estado bruto. Em contraste com os métodos mais comuns baseados em Hidden Markov Models (HMM) e Deep Neural Networks (DNN) supervisionadas, que precisam de muitas informações já marcadas, o Wav2Vec 2.0 consegue identificar elementos importantes em dados sem rótulos, diminuindo muito a necessidade de transcrições feitas por pessoas no processo de aprendizado (Baevski *et al.*, 2020).

O modelo Wav2Vec 2.0 envolve, primeiramente, a obtenção de características, seguida por uma rede Transformer que organiza as relações temporais no sinal sonoro. Após, pode ser refinado para funções específicas de identificação da fala. Essa estratégia se mostra eficiente em situações com pouca informação rotulada, sendo bastante aplicada em estudos e usos que pedem versatilidade e capacidade de se adequar a diferentes áreas (Baevski *et al.*, 2020).

O Vosk é um sistema de reconhecimento automático de fala de código aberto, capaz de converter áudio em texto localmente, sem a necessidade de conexão com a internet (Alphacephei, 2026). Esse sistema utiliza modelos de som compactos e funciona em diversas línguas, podendo ser usado com Python, Java, C++ e outras linguagens. O grande diferencial do Vosk, de acordo com Alphacephei (2026) é rodar sem problemas em aparelhos mais simples, como sistemas embutidos e aplicativos que não usam a internet, complementar a isso, o Vosk se caracteriza pela a rapidez, tamanho reduzido e por não depender da nuvem, sendo ideal para situações que exigem respostas rápidas e em locais com pouca estrutura (Alphacephei, 2026).

2.1. Influência do ruído

Em sistemas de reconhecimento automático de fala, a presença de ruídos pode afetar significativamente a qualidade das transcrições geradas. O ruído corresponde a qualquer sinal indesejado presente na gravação que interfira na captura da fala, podendo ter origem em fatores ambientais, equipamentos de gravação ou outras fontes sonoras. Os ruídos podem ser classificados em diferentes categorias, como ruído de fundo contínuo (ventiladores, ar-condicionado e trânsito), ruídos impulsivos (batidas e impactos), interferência de outras vozes e reverberação do ambiente. Esses fatores dificultam a identificação correta dos fonemas e palavras pelos modelos de reconhecimento de fala, aumentando a taxa de erros de transcrição. A degradação da relação sinal-ruído (Signal-to-Noise Ratio – SNR) impacta diretamente métricas de desempenho como a Word Error Rate (WER) (Radford *et al.*, 2022). Em cenários com elevado nível de ruído, os modelos tendem a apresentar aumento de substituições, inserções e omissões de palavras durante a transcrição. Por esse motivo, técnicas de pré-processamento, como filtragem de sinais, redução de

ruído e normalização do áudio, são frequentemente empregadas para melhorar a qualidade do sinal de entrada e aumentar a precisão dos sistemas de reconhecimento automático de fala. Entre os trabalhos relacionados encontrados na literatura, destaca-se o estudo de Silva (2023), que avaliou o uso de tecnologias de reconhecimento automático de fala em um processo de extração de informações de documentos médicos.

Silva (2023) comparou diferentes abordagens de transcrição, observando que o modelo Whisper apresentou os melhores resultados gerais de precisão, alcançando índices superiores a 93% nos melhores cenários avaliados. O autor também verificou que a presença de ruídos nos áudios influencia a estabilidade dos resultados, embora o Whisper tenha demonstrado boa robustez mesmo em condições menos favoráveis. Além disso, foi observado que soluções executadas em infraestrutura de nuvem apresentaram menor tempo de processamento quando comparadas às abordagens locais.

Os resultados apresentados por Silva (2023) reforçam o potencial dos modelos baseados em redes neurais profundas para aplicações de reconhecimento automático de fala em português. Entretanto, o estudo concentrou-se principalmente na utilização do Whisper em um cenário específico de aplicação, não contemplando uma comparação direta com modelos como Wav2Vec 2.0 e Vosk sob as mesmas condições experimentais. Dessa forma, o presente trabalho busca ampliar essa análise por meio da comparação entre diferentes tecnologias de ASR, considerando áudios com e sem ruído e utilizando métricas padronizadas de avaliação.

2.2. Trabalhos relacionados

Diversos estudos têm investigado o desempenho de sistemas de reconhecimento automático de fala em diferentes idiomas e contextos de aplicação. Rabelo (2022) avaliou o modelo Whisper Large para transcrição automática de fala em português utilizando amostras do Multilingual TEDx Corpus. Rabelo (2022) utilizou a métrica Word Error Rate (WER) para avaliar a qualidade das transcrições, obtendo um valor médio de aproximadamente 0,4. O estudo concluiu que o modelo apresentou desempenho satisfatório para a tarefa proposta, porém, destacou a necessidade de avaliações futuras envolvendo áudios com ruído e conjuntos de dados mais extensos.

Outro trabalho relevante é o de Radford *et al.* (2022), que apresentou o modelo Whisper, treinado com aproximadamente 680 mil horas de dados multilíngues coletados da internet. Os autores demonstraram que o uso de grandes volumes de dados diversificados contribui para a robustez do modelo em diferentes idiomas, sotaques e condições acústicas. Da mesma forma, Chun *et al.* (2023) trouxeram dados de WER de 0,307 do Whisper Large contra 0,249 do Riva Conformer para áudios coreanos.

A Tabela 2 apresenta uma síntese dos principais trabalhos relacionados encontrados na literatura e suas respectivas características.

Tabela 2 – Trabalhos relacionados

Trabalho	Modelo	Idioma	Ruído	Tempo Real	Métrica	Resultados
Radford et al. (2022)	Whisper	Multilíngue	Sim	Não informado	WER	Robustez a sotaques, ruídos e diferentes idiomas
Chun et al. (2023)	Whisper Large e Riva Conformer	Coreano	Sim	Não informado	WER	Whisper: WER = 0,307; Riva Conformer: WER = 0,249
Rabelo (2022)	Whisper Large	Português	Não	Não	WER	WER médio ~ 0,40
Silva(2023)	Whisper e outros	Português	Sim	Não informado	Precisão e WER	Precisão superior 93%
Autor	Whisper, Vosk, Wav2Vec	Português	Sim	Sim (FTR)	WER , Acurácia e FTR	Comparação entre três tecnologias mostrados ao final do trabalho

Fonte: Elaborado pelo Autor (2026).

3. METODOLOGIA

Os testes foram conduzidos com o objetivo de avaliar o desempenho de diferentes ferramentas de reconhecimento automático de fala em termos de precisão e eficiência computacional. Foram analisadas as seguintes soluções: Whisper (nas versões tiny, base e large), Wav2Vec 2.0 e Vosk.

Cada ferramenta foi testada em um conjunto de áudios previamente definidos, sendo um áudio do autor, um áudio sintético gerado por uma voz masculina e outro áudio sintético gerado com uma voz feminina, ambos os áudios sintéticos foram gerados a partir da ElevenLabs¹. Cada um dos três áudios possuem uma versão com e sem ruído.

O conjunto de testes foi composto por seis arquivos de áudio distintos, classificados conforme o tipo de voz e presença de ruído:

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

¹ <https://elevenlabs.io/>

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

O Texto 1 foi extraído de uma notícia sobre o esporte boxe disponível no portal ge.globo², enquanto os Textos 2 e 3 foram gerados com auxílio da ferramenta Gemini³, da Google, visando diversificar os contextos linguísticos utilizados nos testes, todos eles se encontram no Apêndice A. Antes da transcrição, todos os sinais de áudio passaram por um processo de pré-processamento padronizado, composto pelas seguintes etapas:

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

Os áudios foram padronizados para a taxa citada por questão de compatibilidade, porque os modelos utilizados são treinados em 16 kHz. Para avaliar a robustez dos modelos em condições adversas, foram geradas versões ruidosas dos áudios originais. O ruído foi produzido artificialmente por meio da geração de ruído branco gaussiano utilizando a biblioteca NumPy. Esse sinal foi então combinado ao áudio original com fator de intensidade de 0,05, produzindo arquivos contendo fala degradada por interferência sonora.

Após a adição do ruído, os sinais foram limitados ao intervalo de amplitude válido para evitar saturação e armazenados em formato PCM de 16 bits com taxa de amostragem de 16 kHz. Embora o ruído branco não represente fielmente todos os cenários encontrados em ambientes reais, sua utilização permite avaliar de forma controlada a sensibilidade dos modelos à degradação do sinal de fala.

Esse processo foi aplicado de forma consistente em todas as ferramentas avaliadas, garantindo coerência e padrão entre os resultados. Para cada ferramenta e cada tipo de áudio foram realizadas 20 execuções independentes, totalizando múltiplas amostras para análise estatística.

Durante cada execução, foram registrados:

- a) Tempo total de processamento da transcrição;
- b) Texto gerado pela ferramenta.
- c) Taxa de erro de palavras.

Os resultados foram armazenados em arquivos no formato CSV para posterior análise, seguindo as colunas condizente com a Tabela 3.

² <https://ge.globo.com/combate/noticia/2026/03/07/hebert-conceicao-defende-cinturao-latino-americano-do-boxe-contra-johan-gonzalez-em-abril.ghtml>

³ <https://gemini.google.com/app>

Tabela 3 – Colunas dos resultados armazenados em CSV

Nome da coluna	Descrição
modelo	Qual tecnologia abordada nesse trabalho
audio	Arquivo com o áudio dos textos
tipo	Humana ou sintética, com ou sem ruído
run	Ciclo 1 a 20 de cada áudio e tipo
tempo_seg	Tempo de transcrição em segundos
rtf	Métrica para eficiência computacional (FTR)
wer	Métrica para taxa de erro
acuracia	Métrica para taxa de acerto

Fonte: Elaborado pelo Autor (2026).

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

O WER foi utilizada para medir a diferença entre o texto transcrito e o texto de referência. Essa métrica considera inserções, deleções e substituições de palavras (Morris; Maier; Green, 2004). Equação 1 (WER):

$$\text{WER} = (\text{Substituições} + \text{Deleções} + \text{Inserções}) / \text{Número de palavras do texto} \quad (1)$$

A acurácia foi calculada com base na Equação 2:

$$\text{Acurácia} = 1 - \text{WER} \quad (2)$$

O FTR foi utilizado para avaliar a eficiência computacional, sendo definido como a razão entre o tempo de processamento e a duração do áudio (Saon *et al.*, 2017) (Equação 3):

$$\text{FTR} = \text{Tempo de processamento} / \text{Duração do áudio} \quad (3)$$

Valores de FTR menores ou igual que 1 indicam processamento em tempo real, valores acima de 1 indicam processamento mais lento que o tempo real. Além disso, o FTR foi calculado considerando todo o processo de transcrição, no numerador estamos incluindo pré-processamento do áudio e inferência do modelo.

Por fim, houve auxílio de ferramentas de Inteligência Artificial (ChatGPT da OpenAI) para o tratamento e análise dos dados resultantes dos experimentos deste trabalho.

3.1. Ambiente experimental

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. O sistema utilizado possui arquitetura de 64 bits e armazenamento em unidade SSD com capacidade aproximada de 477 GB.

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

4. ANÁLISE DOS RESULTADOS

As análises foram realizadas considerando métricas de acurácia, *Word Error Rate* (WER) e Fator de Tempo Real (FTR). Os dados obtidos foram organizados e analisados com auxílio da biblioteca Pandas, enquanto os gráficos comparativos foram gerados utilizando Matplotlib e Seaborn.

4.1. Resultados do Texto 1 - Boxe (Apêndice A.1)

Os resultados apresentados nesta seção referem-se ao Texto 1, descrito no Apêndice A, composto por um conteúdo relacionado ao contexto esportivo de boxe. Foram avaliados diferentes modelos de reconhecimento automático de fala em cenários com e sem ruído, utilizando gravações humanas e vozes sintéticas masculinas e femininas.

4.1.1. Whisper

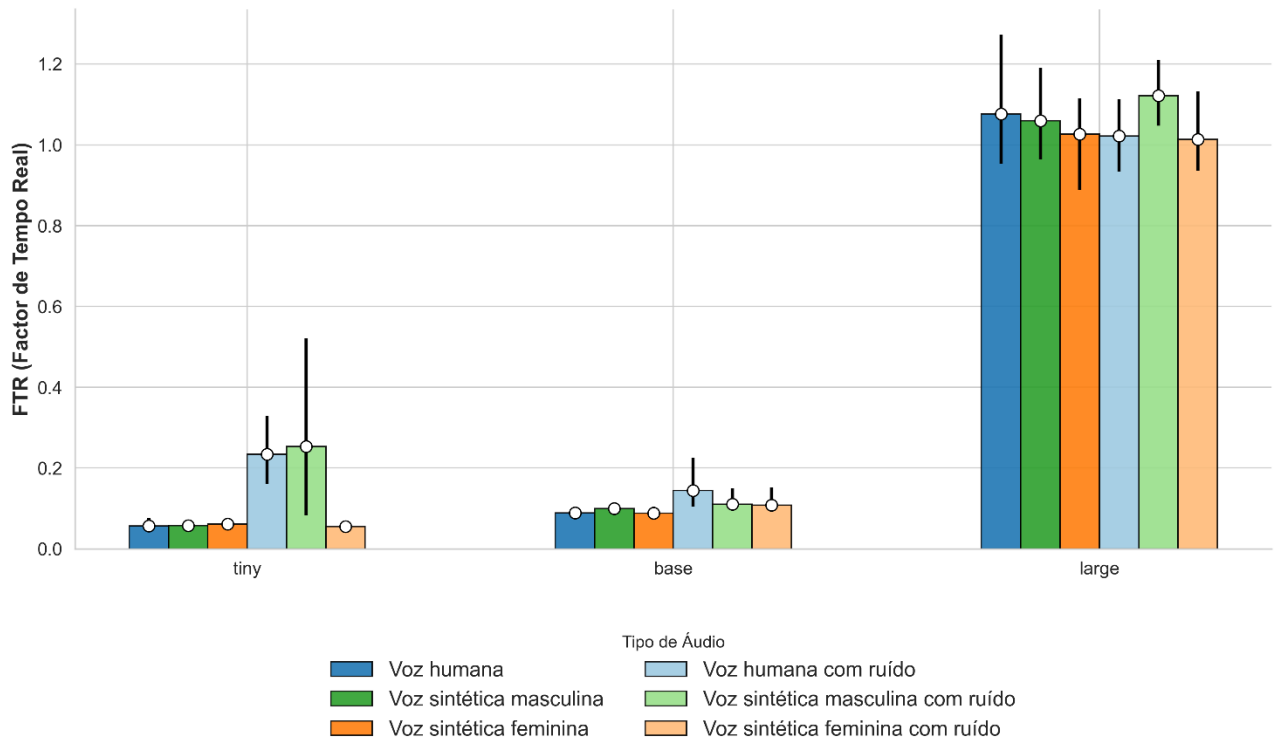
Entre os modelos avaliados visualiza-se nas Figuras 1 e 2 que o Whisper Tiny apresentou o menor custo computacional, ou seja, registrando baixos valores de FTR. Em vozes sem ruído, os resultados mostraram desempenho satisfatório para aplicações em tempo real. Contudo, em cenários ruidosos, observou-se aumento expressivo do WER, principalmente nas vozes masculinas e humanas com ruído, nos quais o modelo apresentou elevada degradação na qualidade das transcrições, se aproximando no valor total 1 da escala.

O modelo Whisper Base demonstrou melhora significativa em relação ao Tiny, apresentando maior estabilidade e melhores índices de acurácia. Em diversos cenários sem ruído, os resultados superaram 80% de acurácia, mantendo também tempos de processamento evidentemente baixos quando comparados aos modelos maiores.

Já o Whisper Large apresentou os melhores resultados de transcrição entre os modelos Whisper. O modelo obteve menores taxas de TEP e maior robustez em ambientes ruidosos, alcançando acurácias superiores a 90% em diferentes cenários. Entretanto, o aumento da precisão veio acompanhado de elevado custo computacional, limitando sua utilização em aplicações de baixa latência.

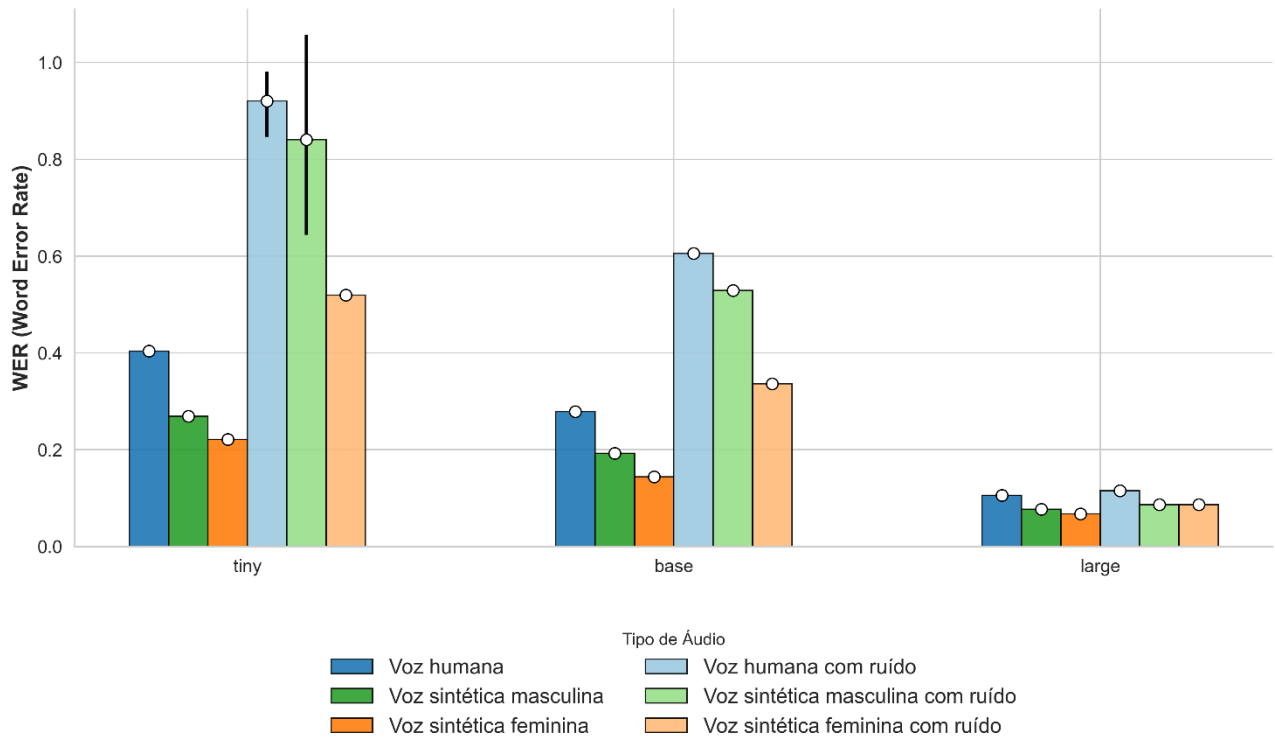
Figura 1 – FTR Whisper no texto 1

Comparação dos Modelos Whisper - FTR



Fonte: Elaborado pelo Autor (2026).

Figura 2 - WER Whisper no texto 1
 Comparação dos Modelos Whisper - WER



Fonte: Elaborado pelo Autor (2026).

4.1.2. Wav2Vec

O modelo Wav2Vec 2.0 apresentou nas figuras 3 e 4 um desempenho intermediário, combinando boa qualidade de transcrição em vozes sem ruído, e péssima transcrição em vozes com ruído, inclusive todas as repetições avaliadas tiveram exatamente o mesmo valor de WER. Além disso, para o FTR, os valores foram muito baixos, no máximo se aproximaram no valor de 0,25 da escala, bem distante do limite, esclarecendo que em questão de processamento o Wav2Vec 2.0 é bastante eficiente em áudios com ou sem ruído.

Figura 3 - FTR Wav2Vec no texto 1

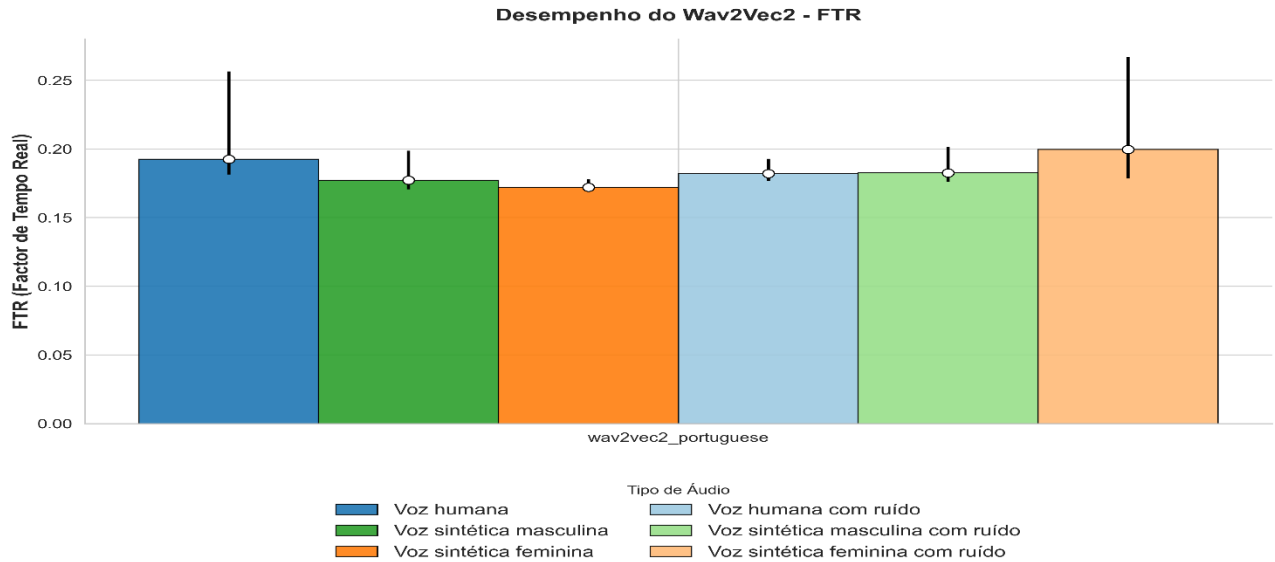
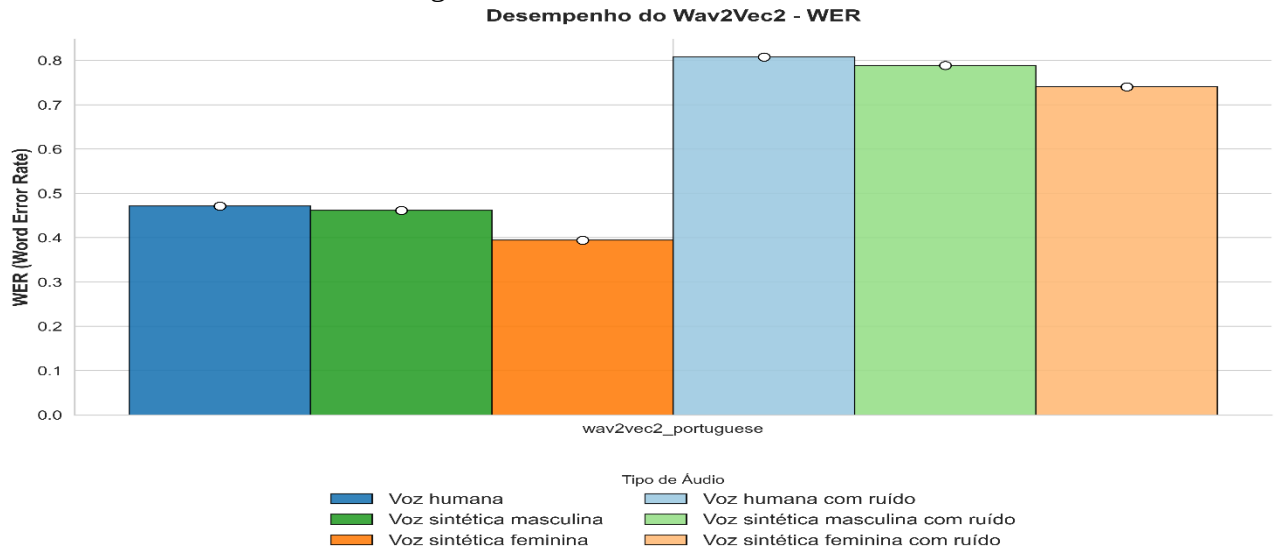


Figura 4 - WER Wav2Vec no texto 1

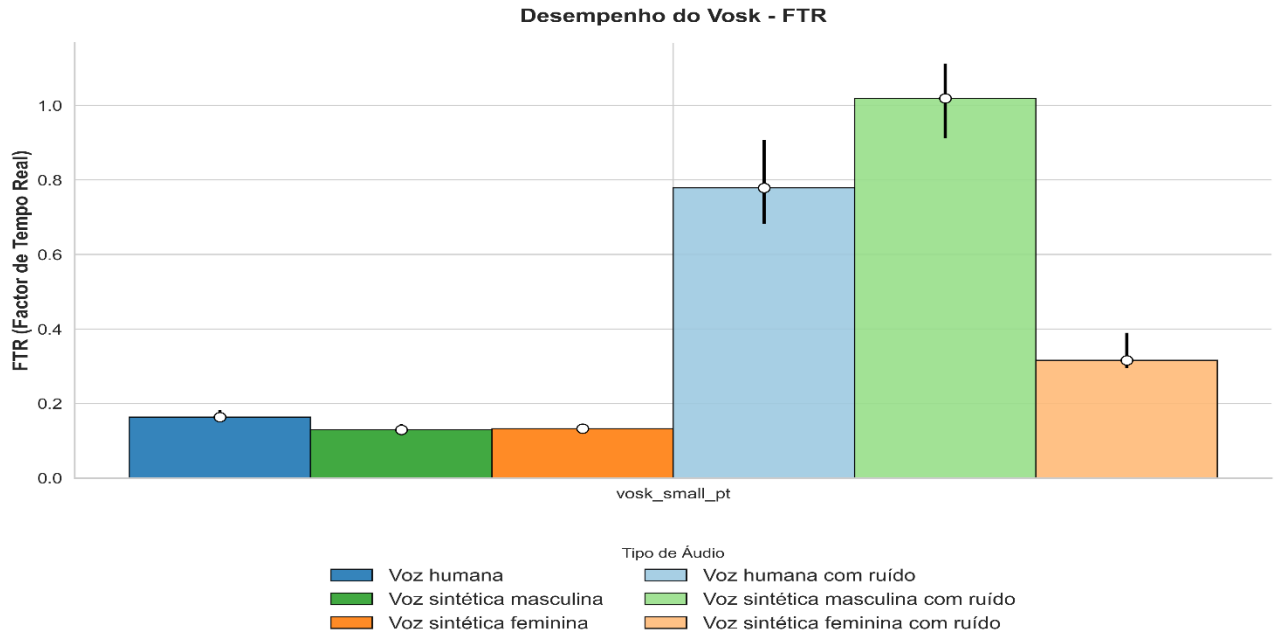


4.1.3. Vosk

O Vosk destacou-se nas figuras 5 e 6 pelo baixo tempo de processamento e excelente desempenho computacional em áudios sem ruído, da mesma maneira apresentou valores de WER decentes se comparado com os demais algoritmos discutidos até então, se aproximando ao valor de 0,4. Já os áudios com ruído sofreram bastante em processamento, onde a voz sintética masculina ultrapassou o valor 1, até então sendo o pior processamento do texto 1, e junto a essa observação o seu WER foi elevado, ultrapassando 0,8 junto com a voz humana ruidosa, que também não teve um processamento desejável. Por fim, é interessante observar que a voz feminina sintética ruidosa

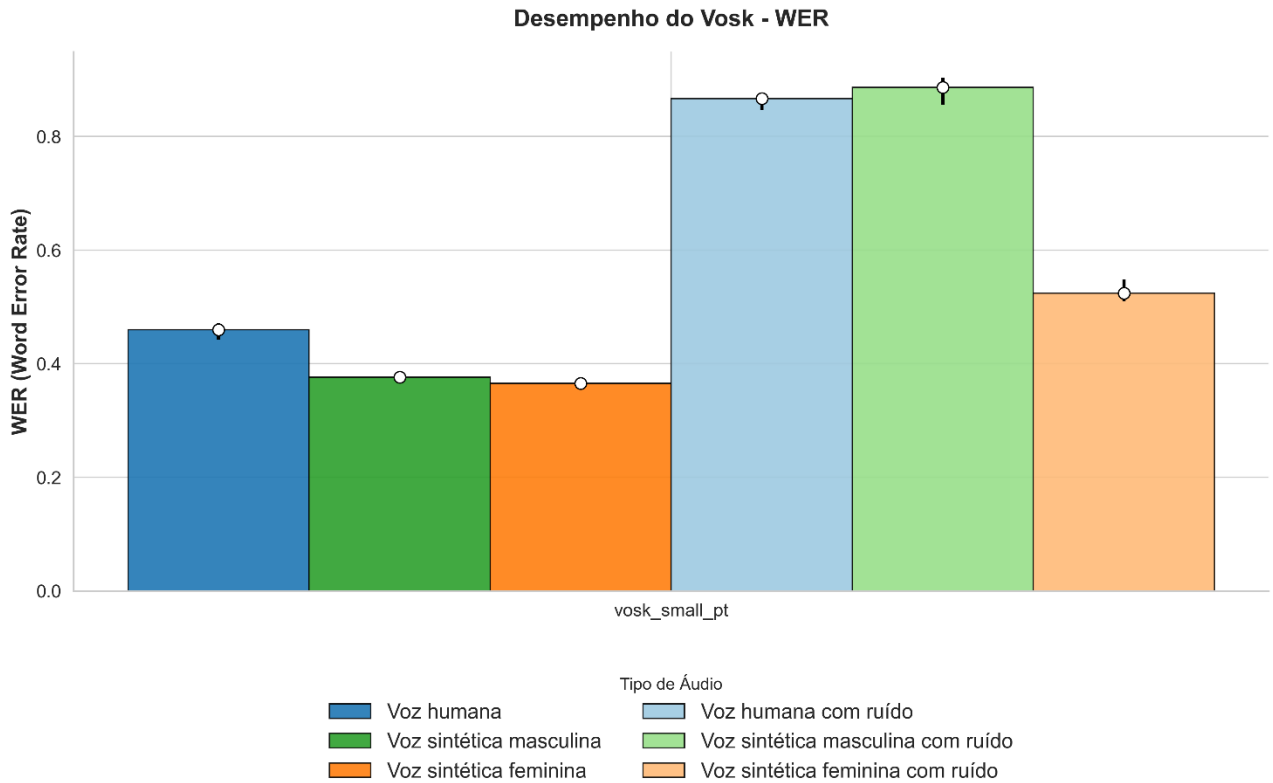
revelou métricas páreas com os áudios sem ruído, se tornando um ponto interessante das avaliações processadas no algoritmo Vosk.

Figura 5 - FTR do Vosk no texto 1



Fonte: Elaborado pelo Autor (2026).

Figura 6 - WER do Vosk no texto 1



Fonte: Elaborado pelo Autor (2026).

4.2. Resultado do Texto 2 – Aula de biologia (Apêndice A.2)

Serão discutidos nesse tópico as métricas geradas do Texto 2, que traz um texto discorrendo sobre uma aula de biologia em um cenário de educação básica, como no ensino fundamental brasileiro, se aproximando de uma aplicação educacional desejada.

4.2.1. Whisper

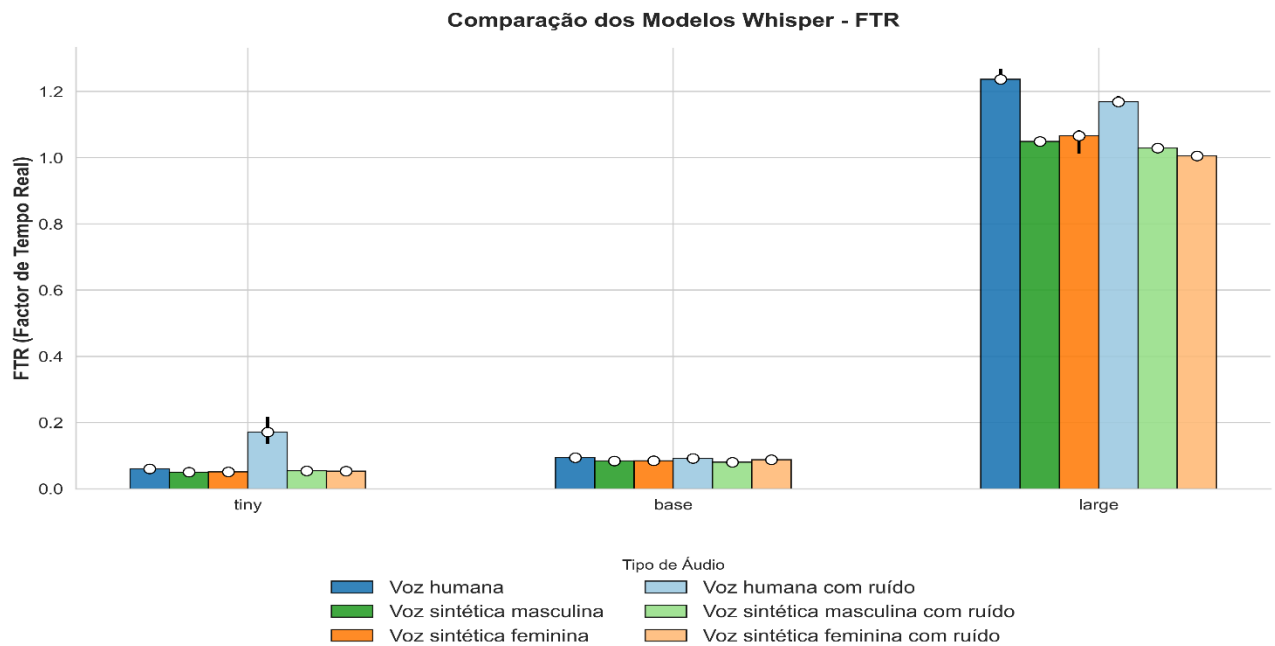
Discutindo sobre as figuras 7 e 8, em primeiro lugar, o modelo Whisper Tiny traz valores consistentes e desejáveis de processamento computacional, através de valores de FTR baixos, quase alcançando 0,2 na escala. Já sobre a qualidade da sua transcrição temos uma variação impactante e muito individual de cada voz, sendo impreciso discutir como vozes sem ou com ruído, já que a voz humana ruidosa lidera como o de pior qualidade, alcançando quase a escala de 1. Fora essa anomalia, a voz sintética feminina e masculina ruidosas competem a mesma qualidade de transcrição em valores regulares, enquanto os demais áudios possuem WER baixos se tornando experimentos positivos.

Em sequência, o modelo Whisper Base traz um valor muito controlado de FTR, aproximadamente 0,1 para todos os áudios, bastante desejável. Analisando o gráfico do WER vemos uma variância proporcional ao Modelo Tiny, se comportando de maneira semelhante quanto à

hierarquia de valores, contudo em escalas menores, sendo que a voz humana ruidosa ultrapassou por poucos valores o valor 0,6 e os demais áudios transcritos se comportam como discutido no modelo Tiny.

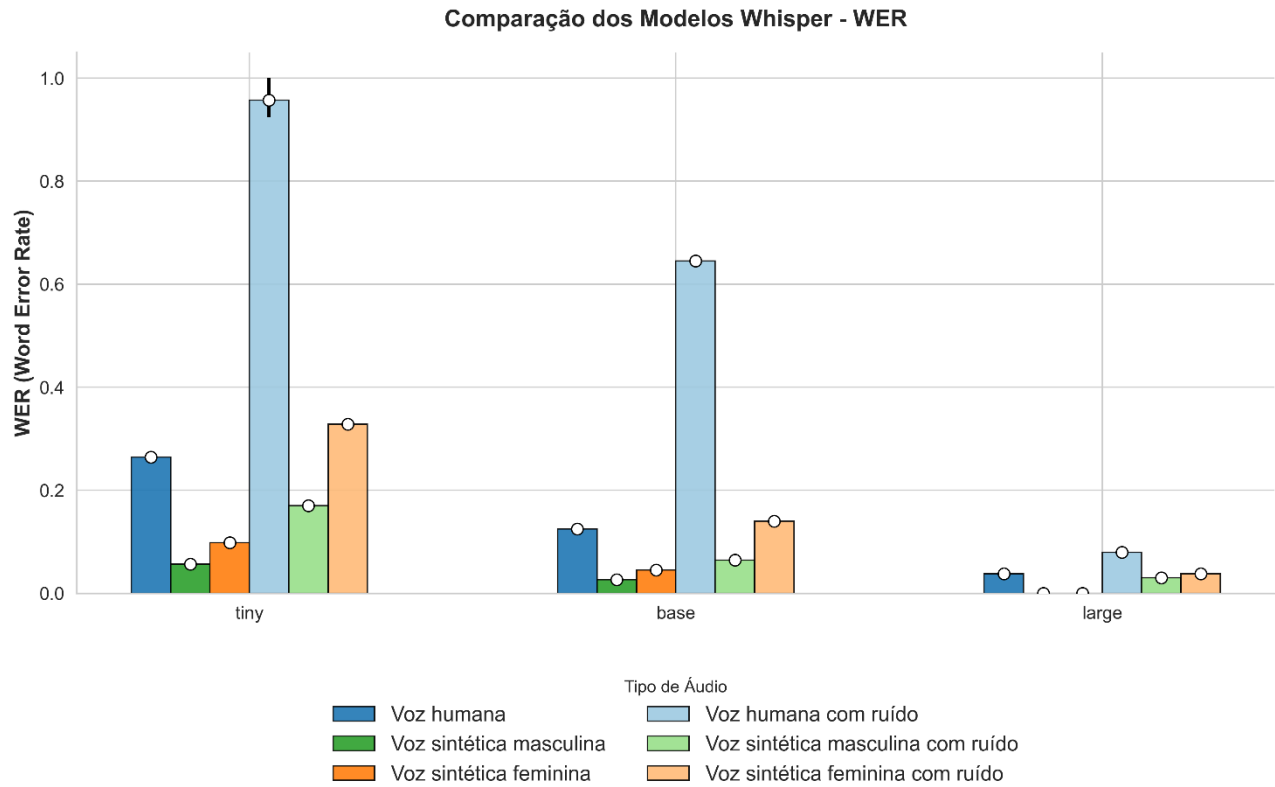
O modelo Whisper Large traz um cenário em destaque, com valores de FTR grandes estando em uma faixa de 1,0 até 1,2, sendo um processamento custoso. Em contrapartida, temos valores de WER ótimos, onde as vozes sintéticas masculina e feminina sem ruído estão no valor de aproximadamente 0, e as demais vozes estão entre o intervalo de 0 a 0,1, revelando uma transcrição excelente.

Figura 7 - FTR Whisper no texto 2



Fonte: Elaborado pelo Autor (2026).

Figura 8 - WER Whisper no texto 2

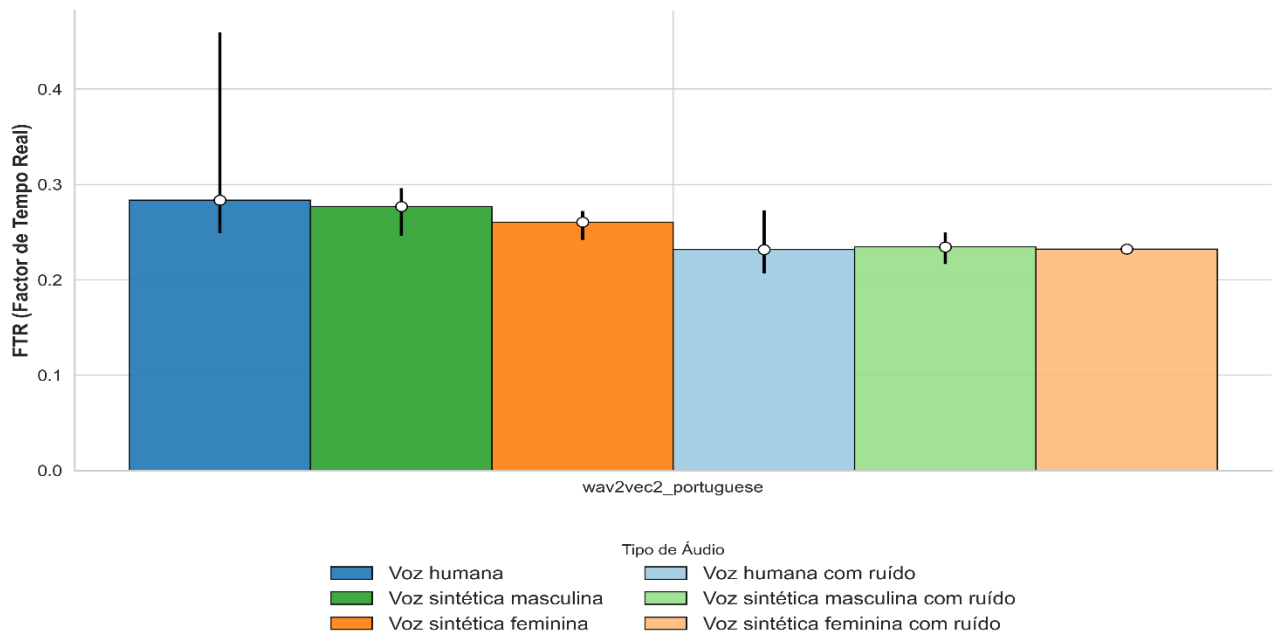


Fonte: Elaborado pelo Autor (2026).

4.2.2. Wav2Vec

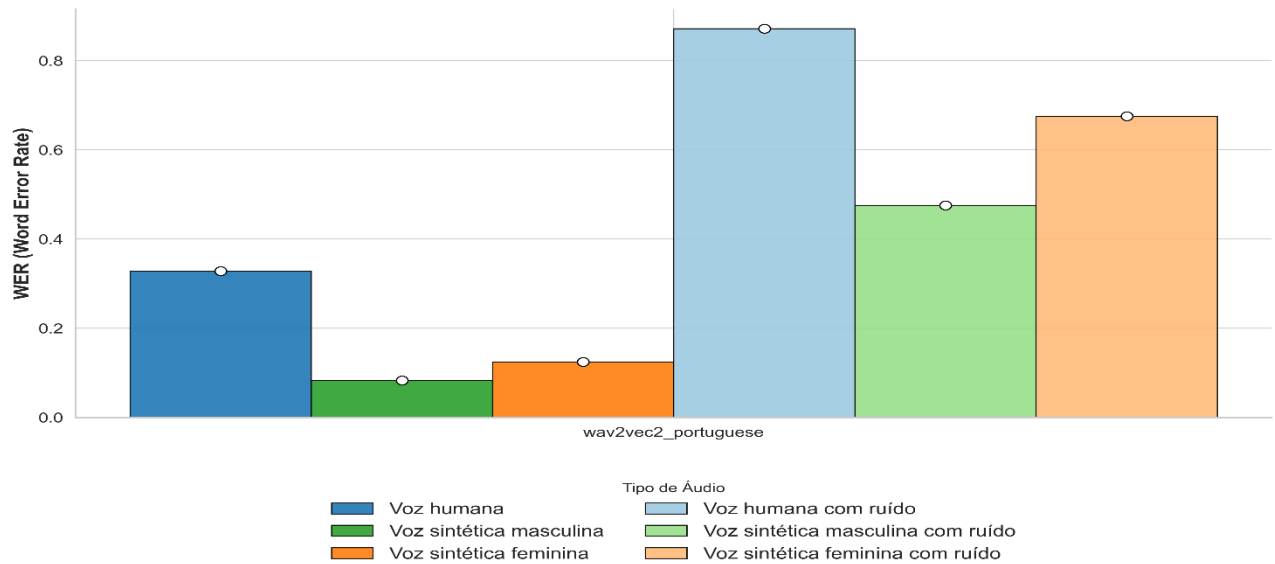
O modelo Wav2Vec 2.0 nas figuras 9 e 10 apresentou desempenho satisfatório nas vozes sem ruídos, principalmente nas amostras geradas por vozes sintéticas masculina e feminina, que obtiveram os menores valores de WER, indicando maior precisão nas transcrições. Já a voz humana apresentou um WER intermediário, demonstrando maior dificuldade do modelo em reconhecer variações naturais da fala humana em comparação às vozes sintetizadas. Em contrapartida, os cenários com ruído apresentaram degradação significativa na qualidade das transcrições, especialmente na voz humana ruidosa, que atingiu os maiores valores de WER entre todas as amostras avaliadas, evidenciando alta sensibilidade do modelo à presença de ruídos externos. Em relação ao FTR, o modelo manteve valores relativamente baixos em todos os cenários analisados, variando aproximadamente entre 0,23 e 0,28. Isso demonstra que o Wav2Vec 2.0 possui boa eficiência computacional e baixo custo de processamento, independentemente da presença de ruído ou do tipo de locutor utilizado. Além disso, observou-se pequena variação entre os valores mínimo e máximo de FTR, indicando estabilidade no tempo de processamento durante as execuções realizadas.

Figura 9 - FTR Wav2Vec no texto 2
Desempenho do Wav2Vec2 - FTR



Fonte: Elaborado pelo Autor (2026).

Figura 10 - WER Wav2Vec no texto 2
Desempenho do Wav2Vec2 - WER



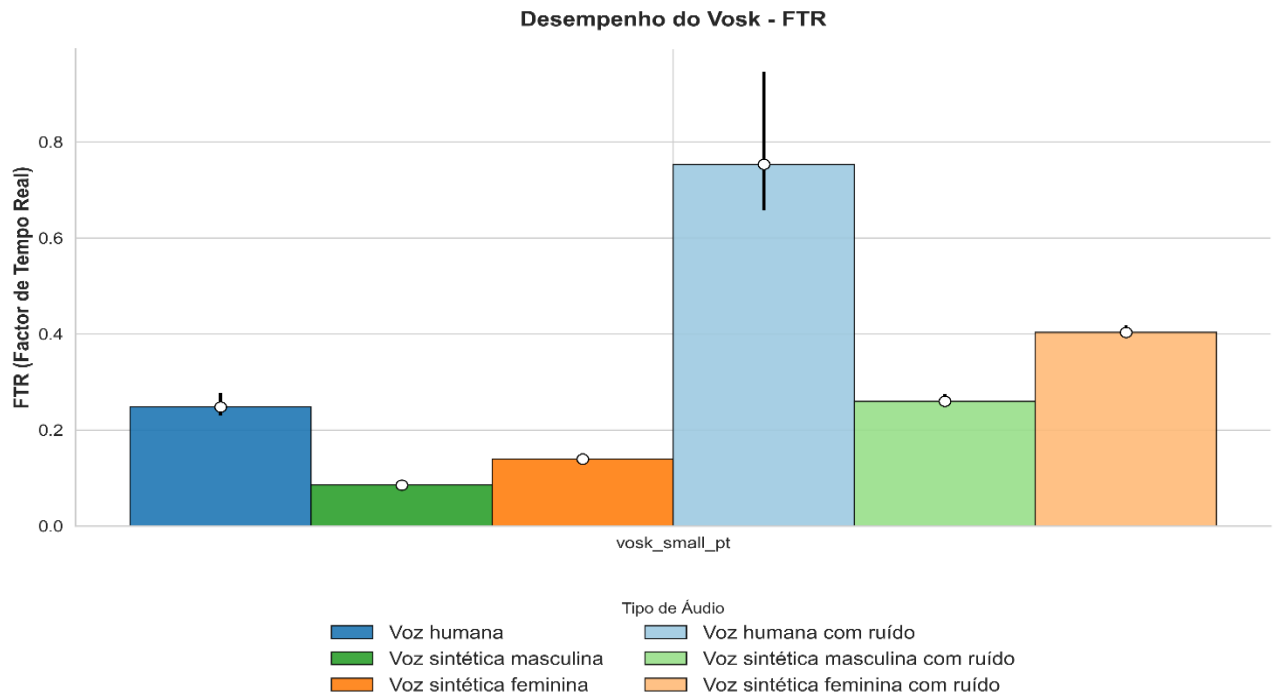
Fonte: Elaborado pelo Autor (2026).

4.2.3. Vosk

O Vosk traz dados consistentes entre o FTR e o WER nas figuras 11 e 12, onde a voz humana com ruído revela valores maiores nas duas métricas e o seu FTR se aproxima de 0,8 enquanto o seu

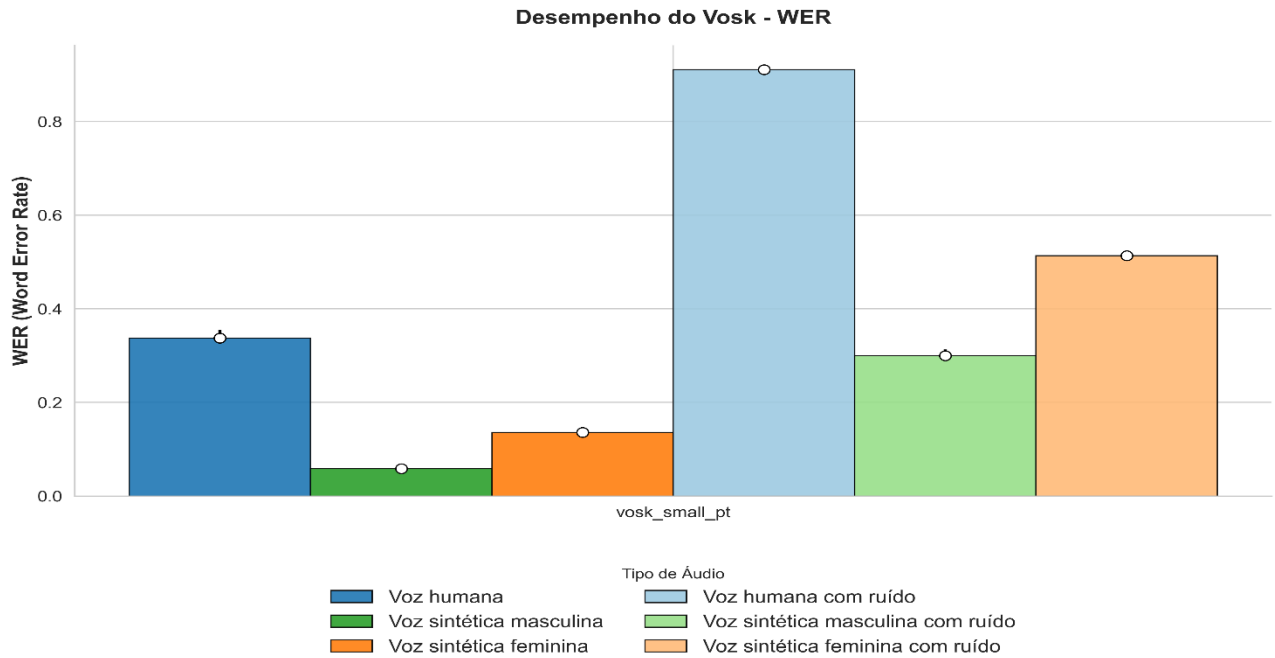
WER ultrapassa esse valor evidentemente, assim evidenciando que essa é a voz com o pior desempenho e transcrição. As demais vozes estão com valores regulares, e como comentado, consistentes, reproduzindo aproximadamente os seus números nas métricas, sendo as vozes sem ruído com os melhores valores, e os com ruído demonstram queda de qualidade e desempenho de modo minimalista.

Figura 11 - FTR Vosk para o texto 2



Fonte: Elaborado pelo Autor (2026).

Figura 12 - WER Vosk para o texto 2



Fonte: Elaborado pelo Autor (2026).

4.3. Resultado do Texto 3 – Aula de potenciação (Apêndice A.3)

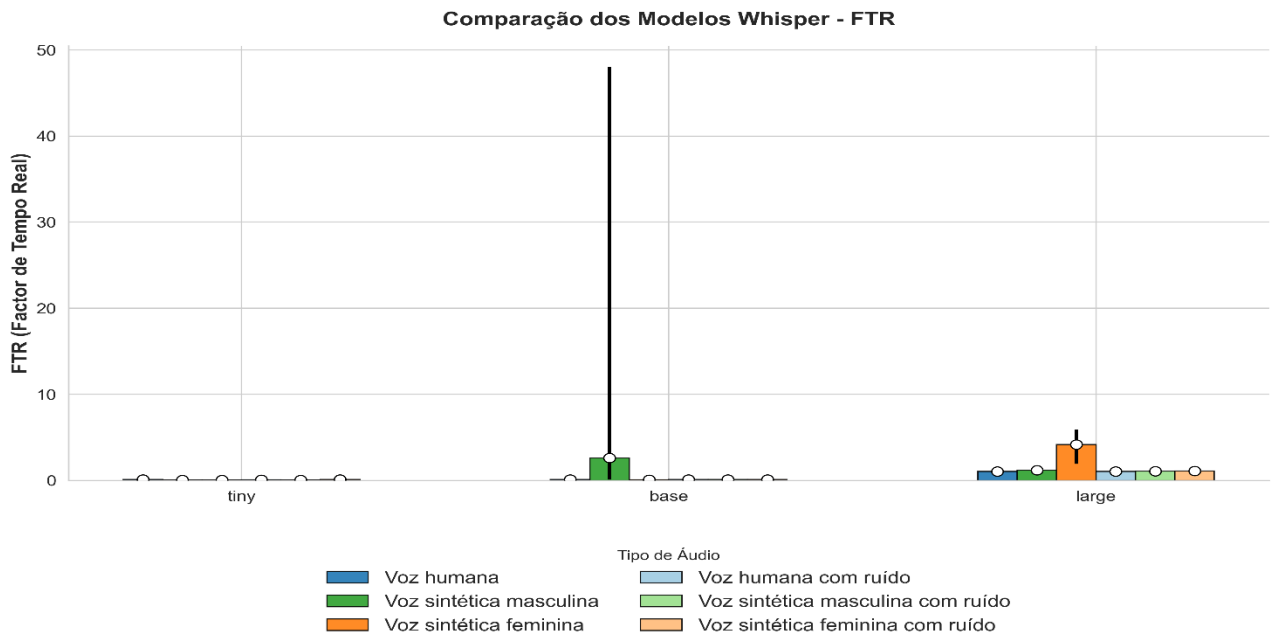
4.3.1. Whisper

Observando as figuras 13 e 14, o modelo Whisper Tiny apresentou o menor custo computacional entre as versões avaliadas, mantendo valores de FTR extremamente baixos em praticamente todos os cenários analisados. Isso demonstra alta velocidade de processamento e grande eficiência computacional. Entretanto, em relação ao WER, o modelo apresentou desempenho limitado, principalmente em áudios com ruído, nos quais ocorreram os maiores índices de erro. O pior resultado foi observado na voz humana ruidosa, enquanto os melhores desempenhos ocorreram nas vozes sintéticas masculina e feminina sem ruído. Dessa forma, embora o Whisper Tiny seja rápido, sua precisão é significativamente afetada pela presença de ruídos.

Discutindo sobre o Whisper Base, apresentou desempenho intermediário, equilibrando eficiência computacional e qualidade de transcrição. Os valores de FTR permaneceram baixos na maioria dos testes, apesar de uma ocorrência isolada apresentar um valor extremamente elevado, indicando possível instabilidade durante uma das execuções. Em relação ao WER, o modelo demonstrou melhora significativa quando comparado ao Tiny, principalmente em vozes sem ruído. O melhor resultado foi obtido na voz sintética feminina ruidosa, enquanto a voz humana ruidosa continuou sendo o cenário mais desafiador. Assim, o Whisper Base mostrou maior robustez e estabilidade nas transcrições em comparação à versão Tiny.

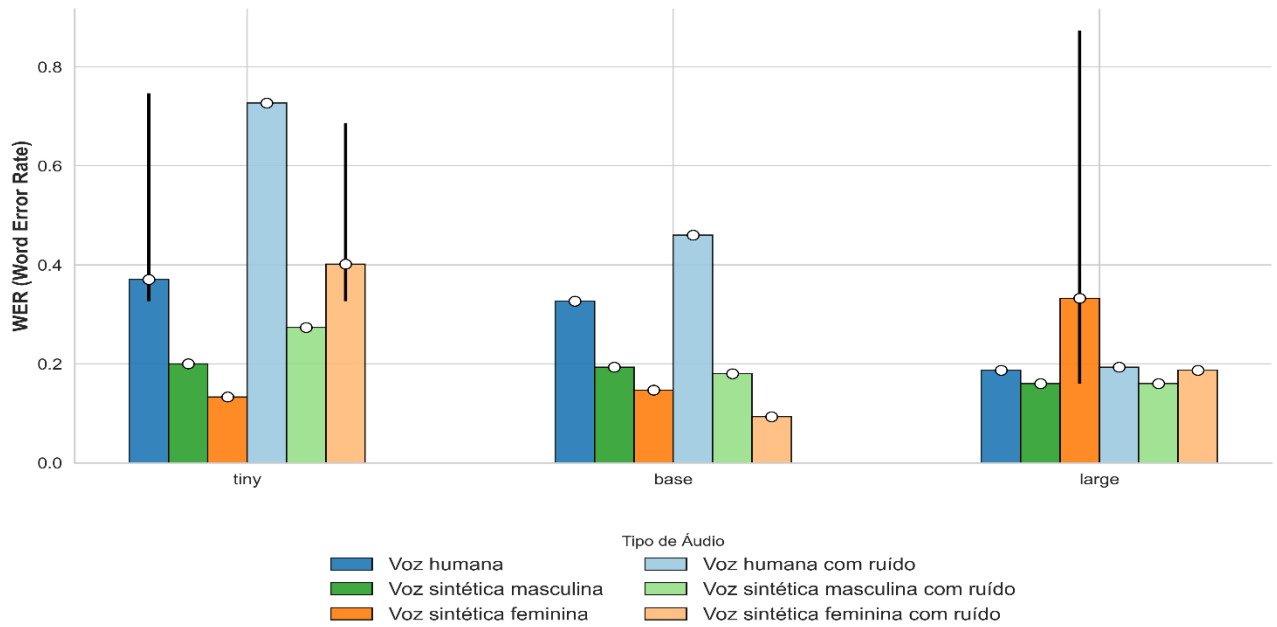
Já o Whisper Large apresentou os maiores valores de FTR entre os três modelos avaliados, evidenciando maior custo computacional e maior tempo de processamento. Apesar disso, o modelo apresentou os menores valores médios de WER em grande parte dos cenários analisados, demonstrando maior precisão nas transcrições. Os resultados indicam que o Whisper Large foi o modelo mais robusto diante de diferentes tipos de áudio, incluindo cenários com ruído. Entretanto, também foram observadas variações elevadas em algumas execuções, especialmente na voz sintética feminina, indicando certa inconsistência no processamento em situações específicas. Ainda assim, o modelo Large apresentou o melhor equilíbrio entre robustez e precisão geral das transcrições.

Figura 13 - FTR Whisper no texto 3



Fonte: Elaborado pelo Autor (2026).

Figura 14 - WER Whisper no texto 3
 Comparação dos Modelos Whisper - WER



Fonte: Elaborado pelo Autor (2026).

4.3.2. Wav2Vec

O modelo Wav2Vec 2.0 nas figuras 15 e 16 apresentou bom desempenho em áudios sem ruído, principalmente nas vozes sintéticas masculina e feminina, que obtiveram os menores valores de WER entre todos os cenários avaliados. A voz humana sem ruído apresentou um valor intermediário de WER, indicando maior dificuldade do modelo em reconhecer características naturais da fala humana em comparação às vozes sintetizadas. Já nos cenários com ruído, houve degradação significativa na qualidade das transcrições, especialmente na voz humana ruidosa, que apresentou o maior valor de WER, evidenciando elevada sensibilidade do modelo à presença de ruídos externos. Apesar disso, os áudios ruidosos gerados sinteticamente apresentaram resultados consideravelmente melhores que a voz humana ruidosa, indicando maior estabilidade em vozes sintéticas mesmo em ambientes degradados. Em relação ao FTR, os resultados permaneceram baixos em todos os cenários analisados, variando aproximadamente entre 0,18 e 0,22, demonstrando alta eficiência computacional e rapidez no processamento das transcrições. Observou-se também baixa variação entre os valores mínimos e máximos, indicando boa estabilidade do modelo durante as execuções. Dessa forma, o Wav2Vec 2.0 mostrou-se bastante eficiente em termos de processamento, porém com perda significativa de precisão em cenários com maior presença de ruído, principalmente em gravações humanas.

Figura 15 - FTR Wav2Vec no texto 3

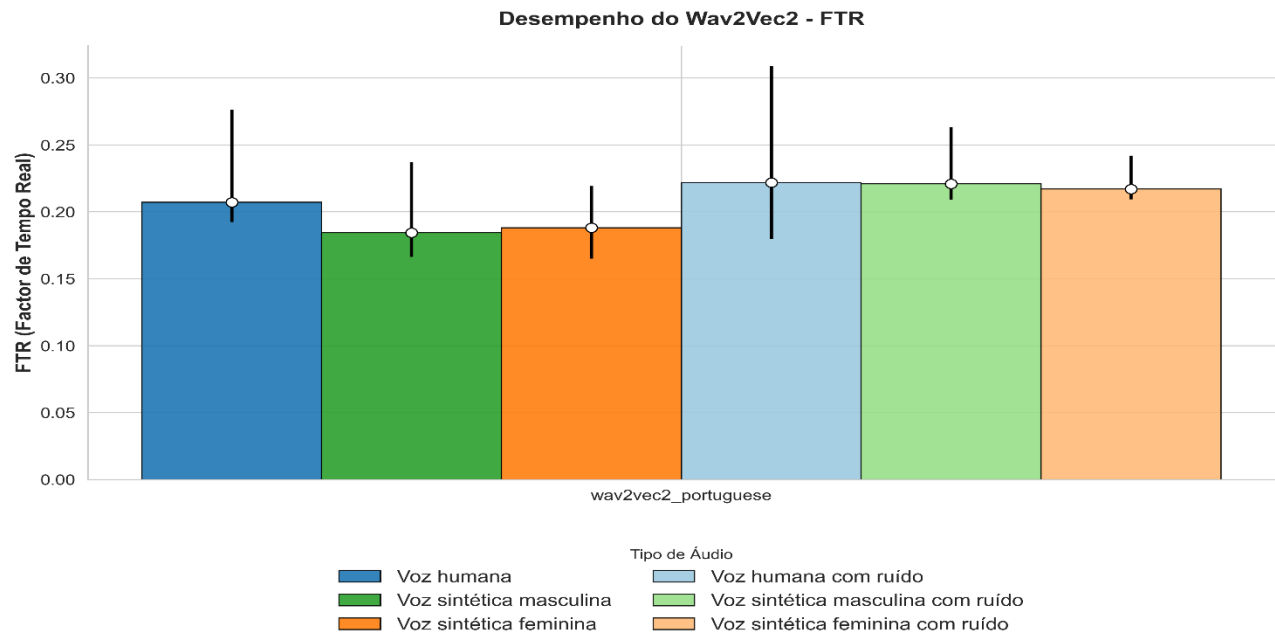
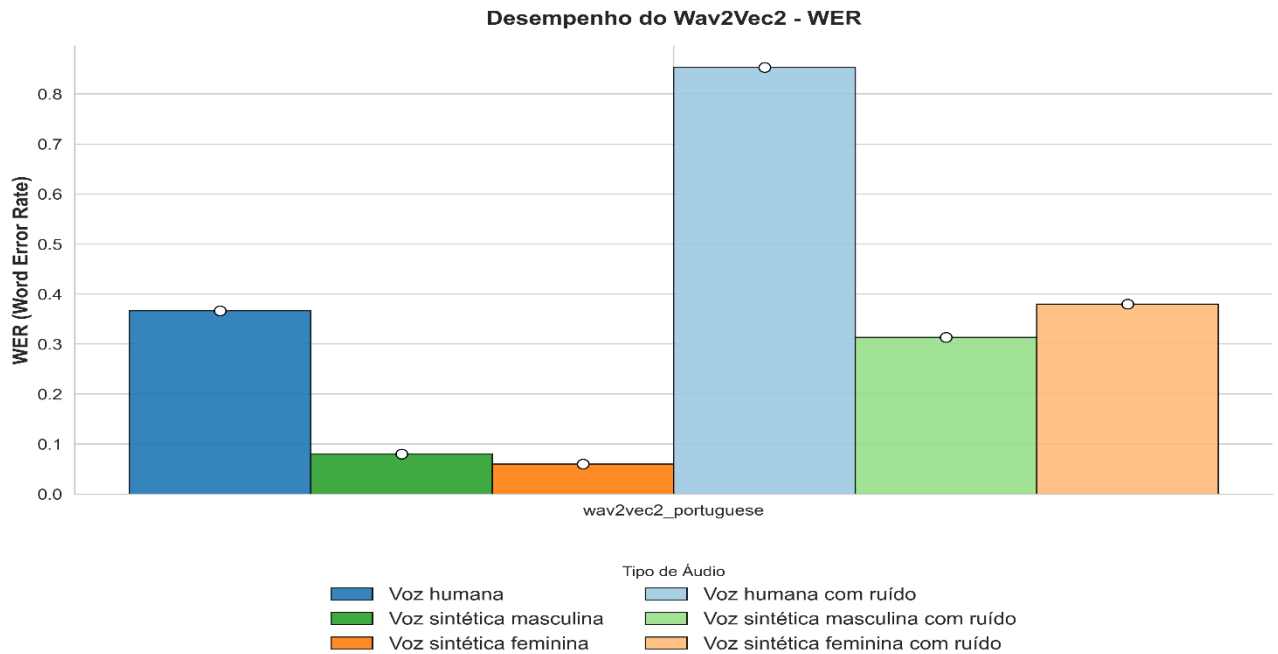


Figura 16 – WER Wav2Vec no texto 3



Fonte: Elaborado pelo Autor (2026).

4.3.3. Vosk

O modelo Vosk apresentou desempenho bastante variável nas figuras 17 e 18, dependendo do tipo de áudio analisado. Em cenários sem ruído, o melhor resultado foi obtido na voz sintética masculina, que apresentou o menor valor de WER entre todas as amostras avaliadas, indicando alta precisão na transcrição. A voz sintética feminina também apresenta desempenho satisfatório, enquanto a voz humana obteve um valor intermediário de WER, demonstrando maior dificuldade do modelo em lidar com características naturais da fala humana. Já nos cenários com ruído, houve degradação significativa na qualidade das transcrições, principalmente na voz humana ruidosa, que atingiu o maior valor de WER observado, evidenciando elevada sensibilidade do modelo à presença de ruídos externos. As vozes sintéticas ruidosas apresentaram desempenho moderado, porém ainda inferiores aos cenários sem ruído.

Em relação ao FTR, o modelo apresentou comportamento menos estável em comparação aos demais modelos analisados. Os menores valores ocorreram nas vozes sem ruído, especialmente na voz sintética masculina, demonstrando processamento rápido e eficiente nesses cenários. Entretanto, as vozes com ruído apresentaram aumento considerável no FTR, principalmente na voz humana ruidosa, que atingiu os maiores valores médios de processamento. Isso indica que o Vosk demanda maior esforço computacional em condições degradadas de áudio. Além disso, observou-se maior variação entre os valores mínimos e máximos em alguns cenários, sugerindo menor estabilidade durante as execuções.

Figura 17 - FTR Vosk no texto 3

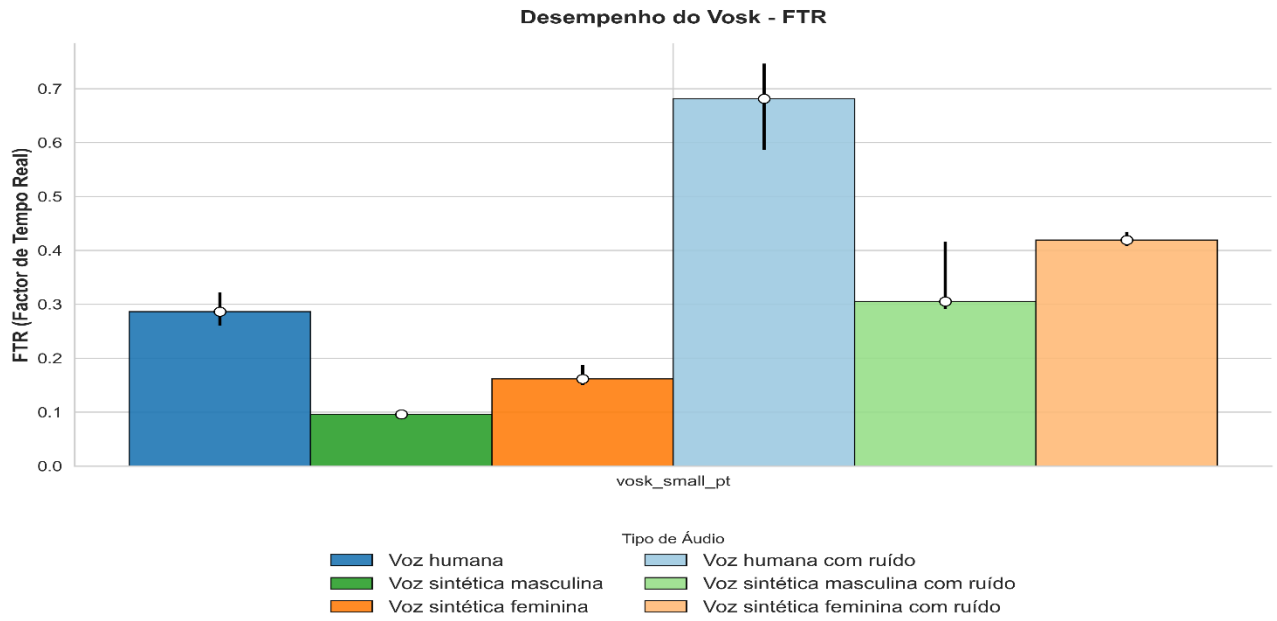
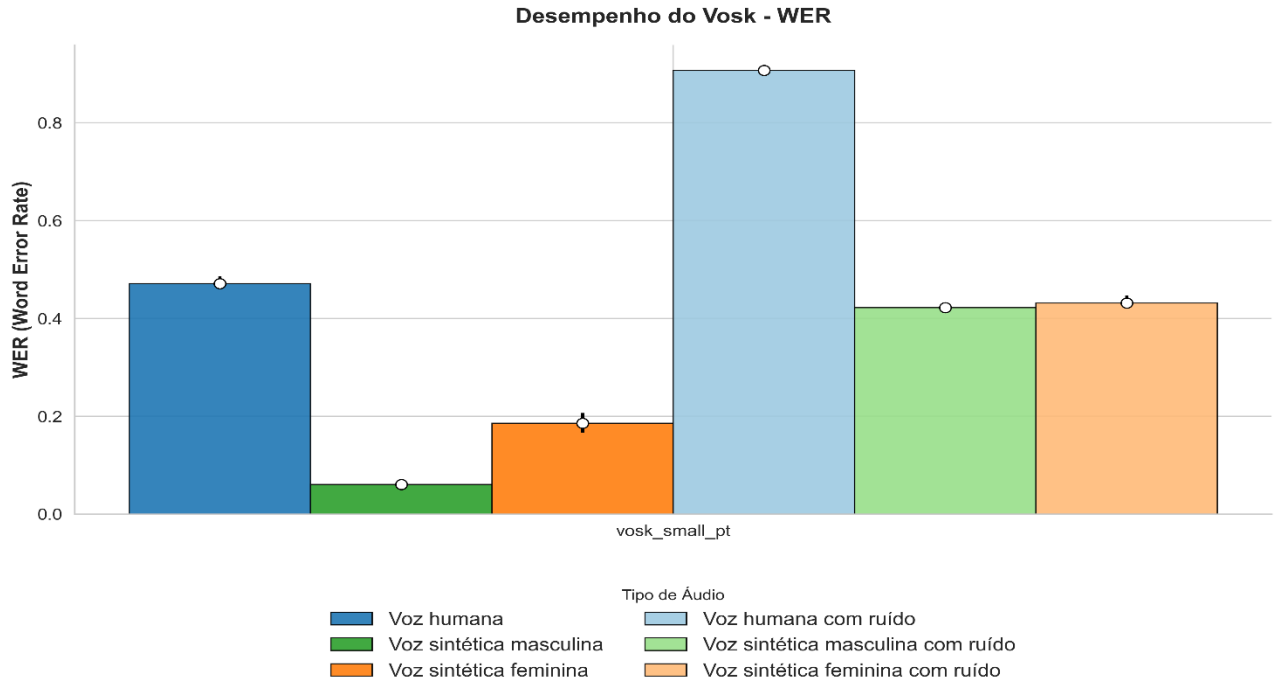


Figura 18 - WER Vosk no texto 3



4.4. Síntese das análises de resultado

Para interpretar com mais clareza os valores das métricas discutidas nesta seção, podemos consultar a Tabela 4 e 5, onde trabalhamos com os números resultantes das médias vistas nas Figuras anteriores.

Tabela 4 – Média dos valores de FTR por modelo e texto

Ferramenta	FTR Texto 1	FTR Texto 2	FTR Texto 3	Média Geral
Whisper tiny	0,119	0,073	0,074	0,089
Whisper base	0,106	0,086	0,512	0,235
Whisper large	1,053	1,092	1,589	1,245
Wav2Vec	0,184	0,253	0,206	0,214
Vosk	0,423	0,314	0,324	0,354

Fonte: Elaborado pelo Autor (2026).

Tabela 5 – Média dos valores de WER por modelo e texto

Ferramenta	WER Texto 1	WER Texto 2	WER Texto 3	Média Geral
Whisper tiny	0,529	0,312	0,350	0,397
Whisper base	0,347	0,172	0,233	0,251
Whisper large	0,089	0,030	0,203	0,107
Wav2Vec	0,610	0,426	0,342	0,459
Vosk	0,579	0,376	0,413	0,456

Fonte: Elaborado pelo Autor (2026).

Uma vez analisado as Tabelas 4 e 5, é imprescindível a percepção de que o pré-processamento abordado e a origem dos textos aplicados nos modelos apresentam limitações. O ruído branco gaussiano e a geração dos textos especificados neste trabalho representam um ambiente controlado, vinculando os valores das métricas adquiridas a apenas esta reprodutibilidade característica.

5. 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, incluindo diferentes sotaques, 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. DOI: 10.1109/ACCESS.2021.3112535

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

BAEVSKI, A. *et al.* wav2vec 2.0: A framework for self-supervised learning of speech representations. Disponível em: <https://arxiv.org/abs/2006.11477>. Acesso em: 24 maio 2026.

CHUN S.-J. *et al.* Development and benchmarking of a Korean audio speech recognition model for Clinician-Patient conversations in radiation oncology clinics. *Int J Med Inform*, 2023. Aug;176:105112. doi: 10.1016/j.ijmedinf.2023.105112. Epub 2023 Jun 1. PMID: 37276615.

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 Ilha de Jeju, Coreia do Sul, 4 a 8 de outubro de 2004. Disponível em: https://www.isca-archive.org/interspeech_2004/morris04_interspeech.html. Acesso em: 13 maio 2026.

OPENAI. *Whisper: Automatic Speech Recognition System*. 2022. 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*. 2022. Trabalho de conclusão de curso (Graduação em Engenharia de Software) – Campus de Quixadá, Universidade Federal do Ceará, Quixadá, 2022.

RADFORD, A. *et al.* Robust Speech Recognition via Large-Scale Weak Supervision. 2022. Disponível em: <https://arxiv.org/abs/2212.04356>. Acesso em: 10 maio 2026.

REDAÇÃO DO GE. *Hebert Conceição defende cinturão latino-americano do boxe contra Johan Gonzalez em abril*. ge.globo, 2026. Disponível em: <https://ge.globo.com/combate/noticia/2026/03/07/hebert-conceicao-defende-cinturao-latino-americano-do-boxe-contra-johan-gonzalez-em-abril.ghtml>. Acesso em: 31 maio 2026.

SAON, G. *et al.* English conversational telephone speech recognition by humans and machines. *Interspeech*, 2017. Disponível em: https://www.isca-archive.org/interspeech_2017/saon17_interspeech.html. Acesso em: 23 de maio de 2026.

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

VASWANI, A. *et al.* Attention is all you need. In: *Proceedings ADVANCES IN NEURAL INFORMATION PROCESSING*. Long Beach, Califórnia, EUA. 4 a 9 de dezembro de 2017. Disponível em: arXiv:1706.03762v7 [cs.CL] 2 Aug 2023. Acesso em: 13 maio 2026.

APÊNDICE A

Este apêndice apresenta os textos utilizados como base para a avaliação dos sistemas de reconhecimento automático de fala. Os conteúdos foram selecionados de forma a representar linguagem natural, com presença de nomes próprios, números e termos específicos, permitindo uma análise mais realista do desempenho dos modelos.

A.1 Texto 1

Conceição iniciou sua transição para o boxe profissional em 2022 e desde então soma 10 vitórias - cinco delas por nocaute. Na última luta, em dezembro do ano passado, derrotou o mexicano Elias Espadas, na primeira defesa do cinturão latino-americano da Organização Mundial de Boxe (WBO). Hebert é também campeão brasileiro. Adversário de Conceição em abril, o venezuelano Gonzalez se destaca pela experiência. Aos 34 anos, tem mais de 40 lutas no boxe pro-fissional: são 36 vitórias (33 por nocaute) e cinco derrotas. No fim de 2025, encarou o campeão mundial dos pesos-médios, o americano Erislandy Lara, e acabou superado por decisão unânime dos juízes.

A.2 Texto 2

Bom dia, turma! Todo mundo bem hoje? Hoje nós vamos conversar sobre uns seres vivos muito especiais que estão por toda parte: no jardim da escola, na calçada de casa e até dentro de alguns vasos na nossa sala. Vamos falar sobre as plantas! Assim como nós, as plantas são seres vivos, o que significa que elas nascem, crescem, respiram e precisam de energia para viver. Mas elas têm um "superpoder" que nós não temos: elas conseguem fabricar o próprio alimento usando apenas a luz do sol! Para que tudo isso funcione, a maioria das plantas é dividida em partes com tarefas bem importantes. Tudo começa lá embaixo, na raiz, que fica escondidinha debaixo da terra segurando a planta firme no chão e bebendo água como se fosse um canudinho. Essa água sobe pelo caule, que é o corpo da planta e serve para sustentar os galhos. Lá no alto temos as folhas, que funcionam como a "cozinha" da planta, pois é nelas que a luz do sol é transformada em comida. Algumas plantas também dão flores coloridas e frutos gostosos, onde ficam guardadas as sementes que farão nascer novas plantinhas no futuro. Para uma sementinha crescer saudável, ela precisa de quatro amigos inseparáveis: a terra, a água, a luz do sol e o ar. Sem as plantas, nossa vida seria muito difícil, pois elas limpam o ar que a gente respira, nos dão sombra fresca e garantem a nossa comida, como o feijão e as frutas do lanche. Agora, que tal a gente fazer um desenho bem bonito da planta que você mais gosta?

A.3 Texto 3

Bom dia, pessoal! Hoje nós vamos começar um conceito novo, que é a potenciação. Alguém lembra o que é uma potência? No fundo, ela é só uma forma simplificada de escrever uma

multiplicação de fatores iguais. Sabe quando a gente tem aquela conta enorme, tipo dois vezes dois vezes dois vezes dois? Então, em vez de escrever tudo isso, a gente usa a potência. Olha só esse exemplo: dois elevado a quatro é igual a dezesseis. Nesse caso, o número dois é a base, que é o número que vai se repetir. E o número quatro, lá em cima, é o expoente. É ele quem diz quantas vezes a base vai se multiplicar por ela mesma. Então, dois elevado a quatro não é dois vezes quatro, hein? É o dois multiplicado por ele mesmo quatro vezes: dois vezes dois é quatro, vezes dois é oito, vezes dois, dezesseis. Conseguiram entender?