Fine-Tuning de modelo ASR para Reconhecimento Automático de Fala em Domínios específicos de Baixo Recurso

Vanessa R. Nakamura[1], **Gabriel T. A. Sousa**[1]

[1] Instituto de Informática
Universidade Federal de Goiás (UFG) – Goiânia, GO – Brazil

vanessanakamura@discente.ufg.br, gabrielteixeira@discente.ufg.br

Abstract. *This study fine-tunes the Whisper model, developed by OpenAI, for automatic speech recognition (ASR) in low-resource domains. Using audio data from health sector podcasts, the audio segments were preprocessed and segmented for effective model training. The fine-tuning resulted in a final Word Error Rate (WER) of 0.4, significantly improving ASR accuracy. These results show that even with limited domain-specific data, substantial performance enhancements can be achieved, making the technology more accessible and applicable across various professional fields.*

Resumo. *Este estudo realiza o fine-tuning do modelo Whisper, desenvolvido pela OpenAI, para reconhecimento automático de fala (ASR) em domínios de baixo recurso. Utilizando dados de áudio de podcasts da área da saúde, os segmentos de áudio foram pré-processados e segmentados para um treinamento eficaz do modelo. O fine-tuning resultou em uma Taxa de Erro de Palavras (WER) final de 0,4, melhorando significativamente a precisão do ASR. Os resultados mostram que, mesmo com dados específicos limitados, é possível alcançar melhorias substanciais, tornando a tecnologia mais acessível e aplicável em diversos campos profissionais.*

1. Introdução

A tecnologia de reconhecimento automático de fala (ASR) tem avançado significativamente nas últimas décadas, sendo amplamente aplicada em diversas áreas da produção social e da vida humana, incluindo economia, militar e cultura [Graves et al. 2013, Chan 2016]. Sistemas de ASR desenvolvidos para línguas internacionais como o inglês já alcançaram capacidades de reconhecimento ao nível humano, apresentando desempenho robusto, rápido e preciso [Zhang 2020]. No entanto, entre as mais de 7000 línguas existentes globalmente, a maioria não possui recursos de treinamento suficientes, o que resulta em um desempenho insatisfatório dos sistemas de ASR para essas línguas [Miao 2022, Krishna 2021], da mesma forma, há escasses de dados nos cenários de domínio específico.

A escassez de dados transcritos para comunidades linguísticas de baixo recurso impede o treinamento substancial de grandes redes neurais, levando a um desempenho pobre e à falta de aplicações práticas.[Yadav and Sitaram 2022]. Este cenário destaca a necessidade de focar em soluções que possam melhorar o desempenho de ASR para línguas e domínios específicos com recursos limitados.

Com o advento dos grandes modelos de linguagem, como o GPT (Generative Pre-Training Transformer) e o BERT (Bidirectional Encoder Representations from Transformers), a oportunidade para o desenvolvimento de grandes modelos no domínio da fala também surgiu. Recentemente, modelos de fala multilíngues e multitarefa tornaram-se uma abordagem popular para resolver o problema do ASR de baixo recurso. O modelo Whisper, desenvolvido pela OpenAI, exemplifica essa tendência ao realizar múltiplas tarefas de processamento de fala, como ASR, tradução de fala, identificação de idioma e detecção de atividade de fala para 100 línguas simultaneamente [Radford 2023].

Embora o Whisper apresente capacidades excepcionais, ele ainda não atende plenamente os requisitos práticos para tarefas de ASR em cenários de baixo recurso, deixando um espaço considerável para melhorias. Estudos recentes têm utilizado uma pequena quantidade de dados supervisionados para realizar o fine-tuning do Whisper, visando melhorar seu desempenho em línguas-alvo específicas. Pesquisas de [Sicard 2023] e [Liu 2023] mostraram que o fine-tuning pode reduzir significativamente a taxa de erro de reconhecimento de fala.

Apesar dos avanços, ainda existem questões que necessitam de pesquisa mais aprofundada. Entre elas, destacam-se a extensão em que o fine-tuning pode melhorar o desempenho, quais partes do modelo são mais críticas para o fine-tuning e as vantagens e desvantagens dos métodos de fine-tuning eficientes em termos de parâmetros [Liu et al. 2024]. Esses aspectos são essenciais para selecionar a estratégia de fine-tuning mais adequada para casos de uso específicos e recursos disponíveis.

Neste contexto, o presente trabalho busca explorar estratégias de fine-tuning do modelo Whisper para domínios específicos, como a área da saúde, onde a aquisição de dados anotados é particularmente desafiadora. A viabilidade do fine-tuning para melhorar o desempenho de ASR em domínios de baixo recurso será discutida, destacando a necessidade de uma seleção criteriosa de estratégias com base em casos de uso específicos.

Propõe-se a utilização de dados públicos, como áudios de podcasts de saúde que já possuem transcrições, para realizar o fine-tuning do modelo Whisper. Esta abordagem visa demonstrar que, mesmo com dados limitados e específicos, é possível alcançar melhorias significativas no desempenho de ASR, tornando a tecnologia mais acessível e aplicável a diferentes domínios profissionais.

2. Métodos

2.1. Pré-processamento dos dados

Para a construção do dataset utilizado neste estudo, foi coletado um episódio do podcast ”EndoDirect - Endocrinologia e Metabologia”, especificamente o episódio intitulado ”ED 85 - Lipedema: o que sabemos até agora”. Este episódio contém aproximadamente uma hora de conversa entre múltiplos locutores falantes da língua portuguesa, sendo um recurso valioso para a criação de um dataset anotado para treinamento de modelos de reconhecimento de fala.

O áudio foi segmentado em intervalos de 30 segundos, resultando em 132 segmentos distintos. Este processo de segmentação foi realizado utilizando a biblioteca pydub, que permitiu a manipulação e exportação dos segmentos em formato WAV. Esta

abordagem facilitou o manuseio dos dados e a subsequente aplicação dos modelos de reconhecimento de fala .

O modelo selecionado para a transcrição dos segmentos de áudio foi o whisper-tiny, desenvolvido pela OpenAI. Este modelo é composto por 4 camadas, 384 de largura, 6 cabeças de atenção e aproximadamente 39 milhões de parâmetros. A escolha deste modelo deve-se à sua capacidade de realizar tarefas de transcrição com alta eficiência e baixa exigência computacional, o que é ideal para cenários com recursos limitados (Figura 1) [Radford 2023].

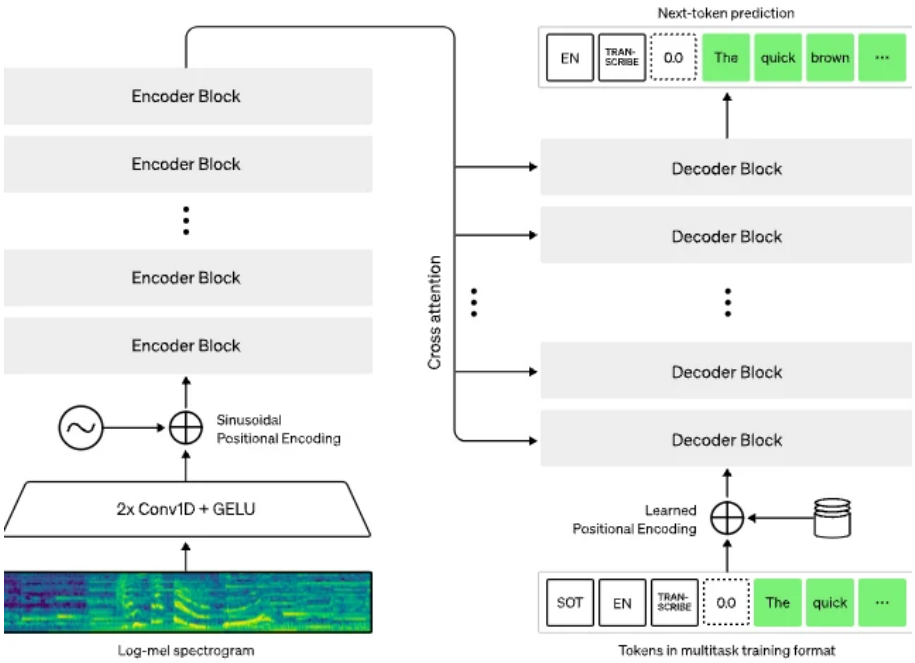

Figure 1. Arquitetura Whisper Vanilla. (Fonte: https://openai.com/index/whisper/. Acesso: 10/07/2024)

A preparação dos dados envolveu o carregamento dos segmentos de áudio, sua normalização e a geração das transcrições. Utilizou-se a biblioteca transformers da Hugging Face, que provê ferramentas avançadas para o processamento de linguagem natural e modelos de reconhecimento de fala. O modelo foi carregado e preparado para inferência, sendo movido para um dispositivo GPU disponível, otimizando assim o tempo de processamento.

Cada segmento de áudio foi processado individualmente para gerar as transcrições. Este processo envolveu a normalização do tensor de áudio [Zhao and Zhang 2022] e a utilização do modelo Whisper Large para gerar as transcrições. Este conjunto de dados foi dividido em subconjuntos de treino e teste, utilizando uma divisão de 80% para treino e 20% para teste. Esta divisão visa garantir que o modelo possa ser treinado de maneira robusta, permitindo a avaliação de seu desempenho em dados não vistos anteriormente.

A preparação do conjunto de dados para treinamento envolveu a extração de carac-

terísticas de áudio (Figura 2) e a tokenização das transcrições para gerar os identificadores de entrada para o modelo. Este processamento foi realizado utilizando as funcionalidades integradas da biblioteca transformers, garantindo a compatibilidade dos dados com o modelo Whisper.

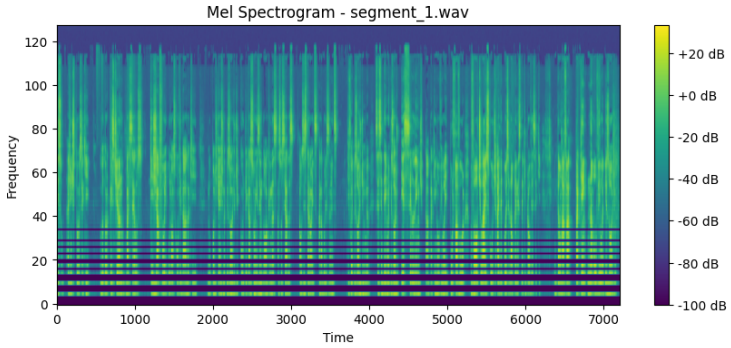

Figure 2. Mel Espectrograma de um segmento do dataset.)

2.2. Fine-Tuning do Modelo

Para a etapa de treinamento do modelo, foram definidos parâmetros específicos, incluindo estratégias de parada antecipada para evitar overfitting e otimizar o desempenho do modelo. A métrica de avaliação escolhida foi a Taxa de Erro de Palavras (WER), uma métrica amplamente utilizada para avaliar a precisão de sistemas de reconhecimento de fala (GRAVES et al., 2013; CHAN et al., 2016).

Os parâmetros de treinamento do modelo foram cuidadosamente selecionados para otimizar o desempenho e garantir a eficiência do processo de treinamento. O tamanho do batch foi definido como 32, permitindo um equilíbrio entre a utilização de memória e a estabilidade do gradiente durante o treinamento. A taxa de aprendizado foi ajustada para 1e-4, um valor que tem se mostrado eficaz em evitar tanto a subadaptação quanto a superadaptação do modelo . Além disso, foram configurados 5 passos de aquecimento (warmup_steps), o que ajuda a estabilizar o treinamento inicial, especialmente importante em modelos de aprendizado profundo. A estratégia de avaliação foi configurada para ocorrer a cada 2 passos (eval_steps), permitindo uma monitorização frequente do desempenho e ajustes oportunos. Para prevenir o overfitting, foi implementada uma estratégia de parada antecipada (early_stopping_callback) com paciência de 5 avaliações consecutivas sem melhoria significativa. O valor da taxa de decaimento do peso (weight_decay) foi definido em 0.01, ajudando a regularizar o modelo e evitar a superadaptação. A acumulação de gradientes foi configurada para 1, e a máxima norma do gradiente foi limitada a 0.1 (max_grad_norm), medidas que auxiliam na manutenção da estabilidade do treinamento e na eficiência computacional [Chan 2016] Estas configurações foram escolhidas com base em práticas recomendadas na literatura e ajustadas para otimizar o desempenho do modelo Whisper-tiny no contexto de transcrição de áudio em português [Alharbi et al. 2021].

3. Resultados

Os resultados do treinamento do modelo Whisper-tiny são apresentados a seguir, destacando o desempenho obtido após 50 passos de treinamento, correspondentes a aproximadamente 12.5 épocas. O treinamento foi interrompido devido à implementação do

parâmetro de parada antecipada (early_stopping), que interrompeu o processo após cinco iterações consecutivas sem melhoria significativa na métrica de desempenho.

A Figura 3 ilustra a evolução da loss de treinamento e validação, bem como da Taxa de Erro de Palavras (WER) ao longo dos passos de treinamento. Observa-se que o WER apresentou uma rápida diminuição nos estágios iniciais do treinamento, estabilizando-se posteriormente. A análise do gráfico revela que o WER atingiu um valor máximo de 0.6 e um mínimo de 0.4, indicando uma melhora substancial na precisão do modelo ao longo do processo de ajuste fino.

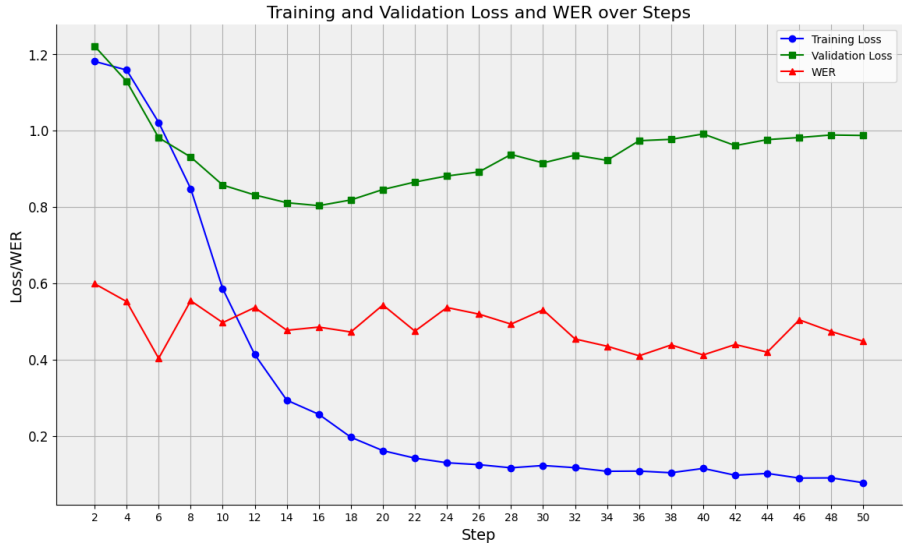

Figure 3. Training and Validation Loss and WER over Steps

Os resultados comparativos entre o modelo Whisper-tiny na sua configuração inicial (baseline) e após o fine-tuning são evidenciados na Figura 4. Este gráfico destaca a diferença significativa na Taxa de Erro de Palavras (WER) entre os dois estágios do modelo. Observa-se que o modelo ajustado (fine-tuned) apresentou uma redução considerável no WER em comparação ao modelo baseline, demonstrando a eficácia do processo de fine-tuning. Enquanto o WER do modelo baseline se manteve elevado, o modelo fine-tuned atingiu um desempenho significativamente melhor, confirmando a importância do ajuste fino para melhorar a precisão de modelos de reconhecimento de fala em domínios específicos.

4. Conclusão

O presente estudo demonstrou que o fine-tuning do modelo Whisper pode resultar em melhorias substanciais na precisão do reconhecimento automático de fala em domínios de baixo recurso. Utilizando dados de áudio de podcasts de saúde, verificou-se que o ajuste fino do modelo não apenas reduziu a taxa de erro de palavras (WER), mas também aumentou a robustez e a eficiência do modelo em ambientes com recursos limitados. Esses resultados indicam que, mesmo com uma quantidade limitada de dados específicos, é possível aprimorar significativamente a performance de modelos de ASR, contribuindo para a disseminação e aplicação dessa tecnologia em áreas profissionais diversificadas.

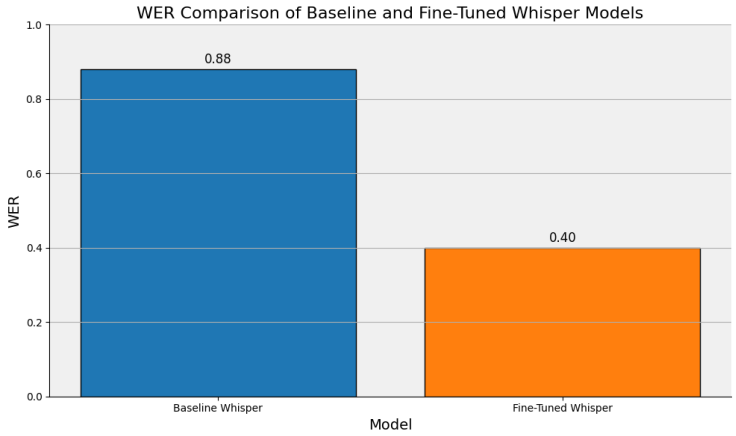

Figure 4. WER for Whisper-tiny Baseline and Fine-tuned Models

Apesar dos avanços alcançados, é necessário continuar investigando as limitações e potencialidades do fine-tuning em modelos de ASR. Futuras pesquisas devem explorar diferentes estratégias de ajuste fino, a fim de identificar os métodos mais eficazes para otimizar o desempenho em diversos cenários de aplicação. Adicionalmente, a integração de técnicas avançadas de processamento de linguagem natural (NLP) pode oferecer novas perspectivas para o desenvolvimento de sistemas de reconhecimento de fala mais precisos e adaptáveis. Em suma, o fine-tuning do Whisper representa uma abordagem promissora para melhorar o ASR em domínios de baixo recurso, promovendo a inclusão e acessibilidade tecnológica em áreas críticas como a saúde.

5. Referências

References

Alharbi, S., Alrazgan, M., Alrashed, A., Alnomasi, T., Almojel, R., Alharbi, R., Alharbi, S., Alturki, S., Alshehri, F., and Almojil, M. (2021). Automatic speech recognition: Systematic literature review. *IEEE Access*, 9:131858–131876.

Chan, W. e. a. (2016). Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. *In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4960–4964.

Graves, A., Mohamed, A.-r., and Hinton, G. (2013). Speech recognition with deep recurrent neural networks. *In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6645–6649.

Krishna, D. (2021). Multilingual speech recognition for low-resource indian languages using multi-task conformer. *CoRR*, abs/2109.03969.

Liu, W. e. a. (2023). Sparsely shared lora on whisper for child speech recognition. *ArXiv*, abs/2309.11756.

Liu, Y., Yang, X., and Qu, D. (2024). Exploration of whisper fine-tuning strategies for low-resource asr. *EURASIP Journal on Audio, Speech, and Music Processing*, 2024(29). Open Access. Licensed under a Creative Commons Attribution 4.0 International License.

Miao, L. e. a. (2022). Multilingual transformer language model for speech recognition in low-resource languages. *In: 2022 Ninth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 1–5.

Radford, A. e. a. (2023). Robust speech recognition via large-scale weak supervision. *In: Proceedings of the 40th International Conference on Machine Learning*, pages 28492–28518.

Sicard, C. e. a. (2023). Spaiche: Extending state-of-the-art asr models to swiss german dialects. *ArXiv*, abs/2304.11075.

Yadav, H. and Sitaram, S. (2022). A survey of multilingual models for automatic speech recognition. *CoRR*, abs/2202.12576.

Zhang, Q. e. a. (2020). Transformer transducer: A streamable speech recognition model with transformer encoders and rnn-t loss. *In: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7829–7833.

Zhao, J. and Zhang, W. (2022). Improving automatic speech recognition performance for low-resource languages with self-supervised models. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1227–1241.