Universidade Federal de São Carlos– UFSCar
Centro de Ciências Exatas e de Tecnologia– CCET
Departamento de Computação– DC
Programa de Pós-Graduação em Ciência da Computação– PPGCC

Ricardo Alexandre Gracelli

Sistema de Reconhecimento de Fala Disártrica usando Aprendizagem Autossupervisionada

Sorocaba
2025

Ricardo Alexandre Gracelli

Sistema de Reconhecimento de Fala Disártrica usando Aprendizagem Autossupervisionada

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação do Centro de Ciências Exatas e de Tecnologia da Universidade Federal de São Carlos, como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação.

Área de concentração: Visão Computacional

Orientador: Jurandy G. Almeida Jr.

Sorocaba
2025

UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Ciências Exatas e de Tecnologia
Programa de Pós-Graduação em Ciência da Computação

Folha de Aprovação

Defesa de Dissertação de Mestrado do candidato Ricardo Alexandre Gracelli, realizada em 08/09/2025.

Comissão Julgadora:

Prof. Dr. Jurandy Gomes de Almeida Junior (UFSCar)

Prof. Dr. Cesar Henrique Comin (UFSCar)

Prof. Dr. Thiago Oliveira dos Santos (UFES)

O Relatório de Defesa assinado pelos membros da Comissão Julgadora encontra-se arquivado junto ao Programa de Pós-Graduação em Ciência da Computação.

Este trabalho é dedicado a todos que, direta ou indiretamente, lutam por uma sociedade mais justa e mais inclusiva

Agradecimentos

Agradecimentos são devidos aos autores Shahamiri et al. pela valiosa clarificação fornecida em relação à sua pesquisa. Além disso, expressa-se profunda gratidão ao Professor Mark Allan Hasegawa-Johnson, da Universidade de Illinois, pela generosa disponibilização do banco de dados UA-Speech. Suas contribuições foram fundamentais para o avanço deste estudo. Parte desta pesquisa foi apoiada pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPq (processos 315220/2023-6 e 420442/2023-5).

Reconhece-se, igualmente, o apoio e a compreensão da esposa, Márcia, e dos filhos, Alec e Agatha, ao longo de toda esta jornada. Agradecimentos são estendidos aos pais, Maria e Hélio, pela constante motivação e suporte. Finalmente, manifesta-se gratidão à UFSCar e a todos os docentes envolvidos, direta ou indiretamente, neste projeto e, em especial, ao orientador, Prof. Jurandy Almeida, pela paciência, pelas sugestões e pelas críticas construtivas que foram essenciais para o desenvolvimento deste trabalho.

"A tecnologia pode ser uma extensão do nosso corpo e mente, permitindo-nos superar limitações físicas e explorar novas fronteiras da consciência."

(Miguel Angelo Laporta Nicolelis)

Resumo

Este estudo tem como objetivo desenvolver e avaliar sistemas de Reconhecimento Automático de Fala (do inglês: *Automatic Speech Recognition*) (ASR) que atendam às necessidades de indivíduos com fala disártrica — condição que compromete a clareza da comunicação e limita o uso de tecnologias assistivas baseadas em voz. Um dos principais problemas em relação ao reconhecimento de fala disártrica reside na escassez de dados rotulados e, para enfrentar esse desafio, foram examinadas duas abordagens complementares e interdependentes.

Na primeira, investigaram-se técnicas de *data augmentation* orientadas à patologia, aplicadas a dois fluxos de processamento com arquiteturas baseadas em Transformadores: *FW1* e *FW2*. Os métodos de perturbação de sinais — ruído aditivo, dilatação temporal e Oclusão Espectral (do inglês: *Spectral Occlusion*) (SO) (método adicional proposto) — foram aplicados isoladamente e em combinações às gravações da base UA-Speech. A análise exploratória das curvas de Acurácia no Reconhecimento de palavras (do inglês: *Word Recognition Accuracy)* (WRA), Taxa de Erro de Palavras (do inglês: *Word Error Rate*) (WER) e Taxa de Erro de Caracteres (do inglês: *Character Error Rate*) (CER) mostrou que ruídos somados à dilatação temporal reduzem consistentemente os erros em falantes de inteligibilidade moderada; a SO adiciona ganhos pontuais; e a união das três perturbações beneficia casos severos, embora possa degradar vozes quase normotípicas devido ao sobreaquecimento espectral.

A segunda abordagem, construída sobre os achados da primeira, utilizou um pré-treinamento supervisionado em fala normotípica da base LJSpeech. Esses modelos de referência foram então submetidos a um ciclo autossupervisionado contrastivo na partição de falantes normotípicos da base UA-Speech, em que as mesmas transformações avaliadas na Fase 1 (ruído, dilatação temporal e SO) foram reutilizadas para gerar pares positivos de treinamento contrastivo. Dessa forma, as estratégias de aumento de dados não apenas foram validadas isoladamente, mas também forneceram o alicerce para a etapa contrastiva. Avaliamos dois métodos: Estrutura Simples para Aprendizado Contrastivo de

Representações Visuais (do inglês: *Simple Framework for Contrastive Learning of Visual Representations*) (SimCLR) e Trocando Atribuições entre Visões (do inglês: *Swapping Assignments between Views*) (SwAV). Os pesos assim refinados foram então transferidos para o ajuste final nos conjuntos disártricos. Em comparação ao modelo de referência treinado apenas em LJSpeech, ambos os métodos contrastivos ampliaram os ganhos: o SimCLR mostrou maior sensibilidade a falantes severos, enquanto o SwAV manteve desempenho estável em todos os níveis de inteligibilidade, reduzindo adicionalmente a CER e a WER e elevando a WRA. Em alguns casos severos, observou-se redução superior a 20 pontos percentuais em WER.

Em síntese, a integração entre aumento de dados específico e pré-treinamento contrastivo resultou em modelos de ASR mais robustos às variações articulatórias da disartria, favorecendo a inclusão de pessoas com fala disártrica em sistemas de comunicação baseados em voz.

Palavras-chave: Disartria, Reconhecimento de Fala, Visão Computacional, Aprendizado Profundo, Aprendizado Autossupervisionado, Aprendizado Contrastivo.

Abstract

This study aims to develop and evaluate Automatic Speech Recognition (ASR) systems tailored to the needs of individuals with dysarthric speech — a condition that compromises communication clarity and limits the use of voice-based assistive technologies. One of the main challenges in dysarthric speech recognition lies in the scarcity of labeled data, and to address this issue, two complementary and interdependent approaches were examined.

The first investigated pathology-oriented *data augmentation* techniques applied to two Transformer-based processing pipelines: *FW1* and *FW2*. Signal perturbation methods — additive noise, time-stretching, and the proposed Spectral Occlusion (SO) — were applied individually and in combination to recordings from the UA-Speech corpus. Exploratory analysis of Word Recognition Accuracy (WRA), Word Error Rate (WER), and Character Error Rate (CER) curves showed that combining noise and time-stretching consistently reduced errors in speakers with moderate intelligibility; SO provided additional gains in specific cases; and the union of all three perturbations benefited severe cases, although it could degrade nearly typical voices due to spectral overheating.

The second approach, built upon the findings of the first, employed supervised pre-training on typical speech from the LJSpeech corpus. These baseline models were then subjected to a self-supervised contrastive cycle on the typical partition of UA-Speech, where the same transformations from Phase 1 (noise, time-stretching, and SO) were reused to generate positive pairs for contrastive training. Thus, augmentation strategies were not only validated in isolation but also served as the foundation for the contrastive stage. We evaluated two methods: Simple Framework for Contrastive Learning of Visual Representations (SimCLR) and Swapping Assignments between Views (SwAV). The refined weights were then transferred for fine-tuning on dysarthric datasets. Compared to the baseline trained solely on LJSpeech, both contrastive methods enhanced performance: SimCLR showed higher sensitivity to severe speakers, while SwAV maintained stable performance across all intelligibility levels, further reducing CER and WER and increasing WRA. In some severe cases, WER was reduced by more than 20 percentage

points. In summary, the integration of targeted data augmentation and contrastive pre-training resulted in ASR models more robust to the articulatory variability of dysarthria, supporting the inclusion of dysarthric speakers in voice-based communication systems.

Keywords: Dysarthria, Speech Recognition, Computer Vision, Deep learning, Self-Supervised Learning, Contrastive Learning.

Lista de ilustrações

Lista de tabelas

Lista de siglas

AE Autocodificadores (do inglês: *Autoencoders*)

ASR Reconhecimento Automático de Fala (do inglês: *Automatic Speech Recognition*)

AudioLIME Explicações Interpretáveis e Auditivas para Modelos de Áudio (do inglês: *Audio-based Local Interpretable Model-agnostic Explanations*)

Audio-XAI Inteligência Artificial Explicável Aplicada a Áudio (do inglês: *Audio Explainable Artificial Intelligence*)

AVC Acidente Vascular Cerebral

BERT Representações de Codificadores Bidirecionais de Transformadores (do inglês: *Bidirectional Encoder Representations from Transformers*)

CADSR Reconhecimento Automático de Fala Disártrica com Aprendizado Contrastivo (do inglês: *Contrastive Automatic Dysarthric Speech Recognition*)

CBMS Sistemas Médicos Baseados em Computador (do inglês: *Computer-Based Medical Systems*)

CER Taxa de Erro de Caracteres (do inglês: *Character Error Rate*)

CHiME-5 Audição Computacional em Ambientes com Múltiplas Fontes – 5ª Edição (do inglês: *Computational Hearing in Multisource Environments – 5th Edition*)

CNNs Redes Neurais Convolucionais (do inglês: *Convolutional Neural Networks*)

CTC Classificação Temporal Conexiva (do inglês: *Connectionist Temporal Classification*)

DASR Reconhecimento Automático de Fala Disártrica (do inglês: *Dysarthric Automatic Speech Recognition*)

DCT Transformada Discreta do Cosseno (do inglês: *Discrete Cosine Transform*)

DC-TTS Conversor de Texto para Fala com Redes Convolucionais Profundas (do inglês: *Deep Convolutional Text-to-Speech*)

DIRHA Ambiente Doméstico para Reconhecimento de Fala em Ambientes Reais (do inglês: *Distant-speech Interaction for Robust Home Applications*)

DistilHuBERT Modelo Compactado de Representação de Fala Autossupervisionado (do inglês: *Distilled Hidden-Unit Bidirectional Encoder Representations from Transformers*)

DSP Processador Digital de Sinais (do inglês: *Digital Signal Processor*)

DyPCL Aprendizado Contrastivo Dinâmico em Nível de Fonema (do inglês: *Dynamic Phoneme-level Contrastive Learning*)

E2E-DASR Reconhecimento Automático de Fala Disártrica de Ponta a Ponta (do inglês: *End-to-End Dysarthric Automatic Speech Recognition*)

ELA Esclerose Lateral Amiotrófica

FFT Transformada Rápida de Fourier (do inglês: *Fast Fourier Transform*)

Grad-CAM Mapeamento de Ativação por Gradiente para Classes (do inglês: *Gradient-weighted Class Activation Mapping*)

HuBERT Representações de Codificadores Bidirecionais de Unidade Oculta de Transformadores (do inglês: *Hidden-Unit Bidirectional Encoder Representations from Transformers*)

LAMB Otimizador com Escalonamento Adaptativo por Camada para Treinamento em Larga Escala (do inglês: *Layer-wise Adaptive Moments optimizer for Batch training*)

LARS Escalonamento Adaptativo com Regressão por Etapas (do inglês: *Layer-wise Adaptive Rate Scaling*)

LightHuBERT Aprendizado Leve e Configurável de Representações de Fala (do inglês: *Lightweight and Configurable Speech Representation Learning with Hidden-Unit BERT*)

LLMs Modelos de Linguagem em Grande Escala (do inglês: *Large Language Models*)

MHAT Transformador com Atenção Multicabeça (do inglês: *Multi-Head Attention Transformer*)

MFCC Coeficientes Cepstrais em Mel-frequência (do inglês: *Mel Frequency Cepstral Coefficients*)

MOS Pontuação Média de Opinião (do inglês: *Mean Opinion Score*)

PASE Codificador de Fala Agnóstico ao Problema (do inglês: *Problem Agnostic Speech Encoder*)

PASE+ Versão Aprimorada do Codificador de Fala Agnóstico ao Problema (do inglês: *Problem Agnostic Speech Encoder Plus*)

PB-DSR Reconhecimento de Fala Baseado em Protótipos (do inglês: *Prototype-Based Dysarthric Speech Recognition*)

PLN Processamento de Linguagem Natural

QRNN Rede Neural Quase Recorrente (do inglês: *Quasi-Recurrent Neural Network*)

ReLU Unidade Linear Retificada (do inglês: *Rectified Linear Unit*)

RNN-T Transdutor de Rede Neural Recorrente (do inglês: *Recurrent Neural Network Transducer*)

SALR Regularização Latente Independente de Falante (do inglês: *Speaker-Adversarial Latent Regularization*)

SCL Aprendizado Contrastivo Supervisionado (do inglês: *Supervised Contrastive Learning*)

SCNN Rede Neural Convolucional Espacial (do inglês: *Spatial Convolutional Neural Network*)

SGD Descida do Gradiente Estocástica (do inglês: *Stochastic Gradient Descent*)

SimCLR Estrutura Simples para Aprendizado Contrastivo de Representações Visuais (do inglês: *Simple Framework for Contrastive Learning of Visual Representations*)

SO Oclusão Espectral (do inglês: *Spectral Occlusion*)

SSL Aprendizado Autossupervisionado (do inglês: *Self-Supervised Learning*)

STFT Transformada de Fourier de Curto Prazo (do inglês: *Short-Time Fourier Transform*)

SwAV Trocando Atribuições entre Visões (do inglês: *Swapping Assignments between Views*)

TDNN Rede Neural com Atraso Temporal (do inglês: *Time-Delay Neural Network*)

TIMIT Corpus de Fala do Instituto de Tecnologia de Massachusetts (do inglês: *Texas Instruments/Massachusetts Institute of Technology*)

t-SNE Embutimento Estocástico de Vizinhança Distribuída t (do inglês: *t-Distributed Stochastic Neighbor Embedding*)

TTS Conversor de Texto em Fala (do inglês: *Text-to-Speech*)

UTran-DSR Reconhecimento de Fala Disártrica com Transformer Aprimorado (do inglês: *Transformer-based Dysarthric Speech Recognition with Feature Enhancement*)

VAE Autocodificadores Variacionais (do inglês: *Variational Autoencoders*)

ViTs Transformadores de Visão (do inglês: *Vision Transformers*)

VTAB Benchmark de Adaptação de Tarefas Visuais (do inglês: *Visual Task Adaptation Benchmark*)

WRA Acurácia no Reconhecimento de palavras (do inglês: *Word Recognition Accuracy)*

WER Taxa de Erro de Palavras (do inglês: *Word Error Rate*)

XLS-R Modelo Multilíngue de Representação de Fala Autossupervisionado (do inglês: *Cross-Lingual Speech Representations*)

Sumário

Capítulo 1

Introdução

A disartria é uma condição que abrange diversos distúrbios da fala decorrentes de mudanças no controle dos músculos envolvidos na produção dos fones. Essas alterações são resultados de lesões no sistema nervoso central ou periférico, ocasionando problemas na comunicação oral devido à paralisia, fraqueza ou falta de coordenação muscular na articulação das palavras (DARLEY; ARONSON; BROWN, 1969) apud (ORTIZ, 2010). É definida por um conjunto de modificações que resultam do controle muscular inexistente devido à lesão no sistema nervoso central ou periférico. Essa alteração atinge um ou mais elementos da produção de comunicação, que consiste na articulação imprecisa e/ou restrita, alteração na fonação, na prosódia, na respiração e na ressonância. Vale ressaltar que, a depender da etiologia neurológica, há diferentes quadros disártricos (FRACASSI et al., 2011).

Esta condição afeta a capacidade de falar corretamente, causando uma fala lenta e arrastada. Está associada a condições neurológicas, como encefalite, tumores cerebrais e doenças neurodegenerativas, como Doença de Parkinson, Acidente Vascular Cerebral (AVC), Esclerose Múltipla e Esclerose Lateral Amiotrófica (ELA).

As causas da disartria podem variar dependendo da região do sistema nervoso afetada. Ela pode ser classificada em diferentes tipos, incluindo disartria hipocinética, que causa rouquidão e tremores na voz, e disartria hipercinética, que causa problemas na pronúncia vocálica e soprosidade na voz (PORTALETE et al., 2019).

Face às características intrínsecas às deficiências motoras da fala, os sistemas de ASR, não apresentam bons resultados quando utilizados por pessoas com dificuldades de articulação da linguagem oral, a exemplo dos pacientes disártricos. Muitos esforços foram empregados por diversos pesquisadores, das mais diferentes áreas, em busca de uma solução eficiente para o problema de comunicação desses pacientes, figurando como o estado da

arte na Ciência da Computação, trabalhos como o *SpeechVision* (SHAHAMIRI, 2021), o E2E-DASR (ALMADHOR et al., 2023), e o *Dysarthric Speech Transformer: A Sequence-to-Sequence Dysarthric Speech Recognition System* (SHAHAMIRI; LAL; SHAH, 2023). Contudo, esses esforços esbarram na baixa oferta de bases de dados robustas para vozes disártricas, levando os pesquisadores a adotarem técnicas de aumento artificial dos dados, incluindo síntese de fala e adição de ruído aos registros sonoros.

Nesse sentido, em um artigo elaborado pelos autores e apresentado à Sistemas Médicos Baseados em Computador (do inglês: *Computer-Based Medical Systems*) (CBMS) 2024, foram realizados testes com três métodos de aumento de dados, sendo um deles, chamado de Oclusão Espectral, que mascara partes dos espectrogramas sonoros de falas normais, buscando simular as características das falas disártricas, obtendo-se resultados similares a outros métodos adotados na literatura para estender os *datasets* disárticos disponíveis.

Em outra direção, arquiteturas de Redes Neurais Profundas Autossupervisionadas como as que se utilizam de modelos como os Autocodificadores (do inglês: *Autoencoders*) (AE) e os Autocodificadores Variacionais (do inglês: *Variational Autoencoders*) (VAE) têm se mostrado ferramentas promissoras em cenários com baixa oferta de dados.

AEs são redes neurais projetadas para aprender uma representação compacta (codificação) dos dados de entrada. Eles consistem em duas partes principais: o Codificador, que transforma os dados de entrada em uma representação de menor dimensão, e o Decodificador, que reconstrói os dados originais a partir dessa representação compacta. Em cenários com baixa oferta de dados, os AEs podem ser usados para redução de dimensionalidade, simplificando os dados enquanto mantêm as características mais importantes, e para *denoising*, removendo ruídos dos dados e melhorando a qualidade dos dados disponíveis (GOODFELLOW; BENGIO; COURVILLE, 2016a).

Os VAEs são uma extensão dos AEs que introduzem uma abordagem probabilística para a codificação dos dados. Eles não apenas aprendem uma representação compacta, mas também modelam a distribuição dos dados no espaço latente, permitindo a geração de novos dados semelhantes aos dados de entrada. Os VAEs são particularmente úteis em cenários com baixa oferta de dados porque podem gerar novos exemplos de dados que seguem a mesma distribuição dos dados de treinamento, aumentando efetivamente o tamanho do conjunto de dados, e ajudam a evitar o sobreajuste, um problema comum quando se trabalha com conjuntos de dados pequenos (KINGMA; WELLING, 2014).

Representações de Codificadores Bidirecionais de Unidade Oculta de Transformadores (do inglês: *Hidden-Unit Bidirectional Encoder Representations from Transformers*) é uma arquitetura autossupervisionada utilizada para modelar fala sem a necessidade de rótulos explícitos em grandes quantidades de dados de treinamento. Ela combina aprendizado de representação com um objetivo de classificação de unidade oculta, onde os segmentos de fala são agrupados e representados por *tokens* discretos, com os quais o modelo é treinado. Esse modelo provou ser eficiente na extração de representações de fala robustas,

especialmente em cenários com recursos limitados (HSU et al., 2021a).

MelHuBERT é uma extensão do HuBERT que integra características de espectrogramas Mel para fornecer uma representação de fala ainda mais refinada, adequada para tarefas que envolvam a análise prosódica e da qualidade do som, como em sistemas de reconhecimento de fala para pacientes com disartria. Outra abordagem relevante para o aprimoramento de sistemas ASR é o aprendizado contrastivo, representado pelo *Estrutura Simples para Aprendizado Contrastivo de Representações Visuais (do inglês:* Simple Framework for Contrastive Learning of Visual Representations*)* proposto por Chen et al. (2020a). Enquanto métodos como HuBERT (HSU et al., 2021a) e MelHuBERT (HSU et al., 2021c) se destacam na extração de representações robustas com base em predição mascarada, o SimCLR utiliza uma estratégia que explora a similaridade entre diferentes transformações de um mesmo dado (amostras positivas) e a dissimilaridade em relação a outros dados (amostras negativas). Tal técnica é particularmente eficaz em cenários com baixa disponibilidade de dados rotulados, permitindo ao modelo capturar características discriminativas mesmo em condições adversas.

No contexto do reconhecimento de fala disártrica, o SimCLR pode contribuir significativamente ao facilitar a extração de padrões acústicos que diferenciam a fala normotípica da disártrica. Combinado a técnicas como o mascaramento espectral, esse modelo não apenas destaca as nuances específicas da fala disártrica, mas também reduz a interferência de ruídos e inconsistências comuns em gravações, promovendo sistemas mais precisos. Desta forma, a adoção do SimCLR reforça as estratégias já mencionadas, ampliando o potencial das abordagens autossupervisionadas para atender às necessidades desses pacientes.

Complementarmente, o método Trocando Atribuições entre Visões (do inglês: *Swapping Assignments between Views*) idealizado por Caron et al. (2020), propõe um mecanismo distinto, que evita a comparação explícita entre amostras positivas e negativas. Em vez disso, o modelo aprende a agrupar diferentes representações em torno de um conjunto compartilhado de protótipos latentes, a partir da atribuição cruzada entre diferentes vistas de um mesmo dado. Essa abordagem permite ao SwAV alcançar estabilidade semântica mais consistente, o que é especialmente relevante em cenários com alta variabilidade de produção, como na fala disártrica. A robustez promovida por essa estrutura prototípica torna o SwAV promissor para sistemas que precisam lidar com distorções articulatórias severas e padrões acústicos menos previsíveis, ampliando ainda mais o espectro de aplicação das técnicas de aprendizado autossupervisionado no ASR clínico.

1.1 Objetivos

Esta proposta visa aprimorar o reconhecimento de fala disártrica por meio de técnicas de autossupervisão, superando a limitação de dados rotulados. A fala disártrica, devido

a dificuldades motoras, compromete a precisão dos sistemas tradicionais de ASR. Dado o número reduzido de corpora específicos para fala disártrica, a autossupervisão oferece uma solução, permitindo que o modelo aprenda representações significativas a partir de dados não rotulados, utilizando tarefas auxiliares como predição ou reconstrução de características. O objetivo é melhorar a precisão e a robustez dos modelos ASR para este público, sem depender de grandes volumes de dados anotados.

1.2 Hipóteses e Questões de Pesquisa

Esta dissertação investiga o uso de transformações de dados e técnicas de aprendizado autossupervisionado para aprimorar o reconhecimento automático de fala disártrica. Parte-se da hipótese de que a aplicação de estratégias de aumento de dados, tanto tradicionais quanto propostas neste estudo, pode beneficiar significativamente o desempenho de sistemas ASR voltados a esse tipo de fala. Presume-se que transformações como adição de ruído branco, dilatação temporal e, principalmente, a Oclusão Espectral proposta sejam capazes de simular de forma eficaz as características da disartria, contribuindo para a geração de representações mais informativas.

Adicionalmente, considera-se que o uso de pré-treinamento com aprendizado contrastivo, por meio das abordagens SimCLR e SwAV, resulte em melhorias substanciais nas representações latentes, elevando os índices de inteligibilidade das transcrições. Esta dissertação utiliza como base duas arquiteturas inspiradas no trabalho de Shahamiri, Lal e Shah (2023), que são aqui denominadas *FW1* (modelo com menor complexidade estrutural) e *FW2* (modelo com maior profundidade). Supõe-se que arquiteturas mais simples, como a *FW1*, sejam mais suscetíveis a limitações de generalização quando treinadas exclusivamente com dados supervisionados, sobretudo em contextos adversos como o da fala disártrica severa. No entanto, essas mesmas arquiteturas podem se beneficiar de maneira proporcionalmente mais expressiva quando submetidas a estratégias de pré-treinamento contrastivo, capazes de extrair representações mais robustas a partir de dados não rotulados. Já modelos mais complexos, como o *FW2*, com maior capacidade de representação, tendem a apresentar desempenho mais estável mesmo em cenários com alta variabilidade na entrada.

No que diz respeito às abordagens de aprendizado contrastivo, espera-se que o SimCLR, ao aplicar a maximização da similaridade entre diferentes visões da mesma entrada, seja eficaz na compressão fonêmica das representações, contribuindo para a redução dos erros de transcrição. Por outro lado, o SwAV, por utilizar um mecanismo de agrupamento em protótipos semânticos, tende a favorecer a estabilidade de representação em tarefas com variação intensa entre os exemplos — como ocorre na fala disártrica severa. Desse modo, postula-se que falantes com maior grau de comprometimento articulatório sejam os que mais se beneficiam dessas técnicas, especialmente quando combinadas a arquiteturas

adequadas e pré-treinadas com representações contrastivas.

Com base nessas premissas, formulam-se as seguintes hipóteses de pesquisa:

1. A introdução de múltiplas técnicas de aumento (ruído aditivo, alongamento temporal, Oclusão Espectral e suas combinações) nos espectrogramas de fala disártrica leva os modelos ASR a aprenderem representações acústicas mais robustas, resultando em menores CER, menores WER e em transcrições perceptivelmente mais inteligíveis do que aquelas obtidas a partir de dados não aumentados.

2. Empregar essas mesmas transformações como base para formar pares positivos no pré-treinamento contrastivo (SimCLR ou SwAV) gera *embeddings* que, após o ajuste fino (do inglês: *fine-tuning* supervisionado, aprimoram de modo consistente o desempenho do reconhecimento de fala disártrica.

1.3 Contribuições

Este trabalho propõe estratégias para a melhoria da eficácia de sistemas de reconhecimento automático de fala disártrica (DASR), tendo como objetivo central investigar se diferentes métodos de aumento de dados e a adoção de paradigmas autossupervisionados aplicados a espectrogramas podem contribuir para a construção de representações acústicas mais robustas. A etapa inicial da pesquisa será dedicada à avaliação de abordagens convencionais, como ruído aditivo e alongamento temporal, amplamente utilizadas na literatura (KO et al., 2015; PARK et al., 2019). Essas técnicas servirão de referência para validar a eficácia de perturbações mais específicas ao domínio da fala patológica.

Na etapa seguinte, será introduzida a técnica de Oclusão Espectral, cuja proposta consiste em mascarar seletivamente regiões do espectro de frequência, simulando padrões acústicos associados a distúrbios motores da fala, como os encontrados na disartria. Tal abordagem busca expandir os dados disponíveis de maneira controlada e orientada por características clínicas, oferecendo uma alternativa às técnicas genéricas de aumento de dados.

Espera-se que, ao aplicar essas transformações isoladamente e em combinação, os modelos resultantes apresentem melhorias consistentes nas métricas de avaliação (CER, WRA/WER), aproximando-se de configurações robustas mesmo em contextos de escassez de dados anotados. A validação dessas estratégias será realizada por meio da análise do desempenho de diferentes configurações de aumento, bem como da inteligibilidade perceptiva das transcrições geradas, conforme recomendado por trabalhos recentes voltados à fala disártrica (SHAHAMIRI; LAL; SHAH, 2023).

A abordagem Oclusão Espectral, proposta neste trabalho, resultou no artigo *Exploring Alternative Data Augmentation Methods in Dysarthric Automatic Speech Recognition* de Gracelli e Almeida (2024), apresentado na CBMS 2024. Essa publicação consolida os

achados da primeira etapa da pesquisa, evidenciando os benefícios da técnica no reconhecimento de fala disártrica com dados limitados.

Como desdobramento da etapa inicial, as transformações espectrais aplicadas serão reutilizadas no paradigma de aprendizado autossupervisionado, com o objetivo de formar pares positivos mais informativos e fisiologicamente plausíveis durante o pré-treinamento contrastivo. Em abordagens autossupervisionadas como o SimCLR (CHEN et al., 2020c) e o SwAV (CARON et al., 2020), a eficácia do aprendizado depende criticamente da qualidade das amostras aumentadas: idealmente, duas vistas de um mesmo dado devem preservar semântica relevante e estimular a invariância desejada nos espaços latentes.

Nesse contexto, espera-se que os aumentos inspirados em distorções reais da fala disártrica – como a Oclusão Espectral – possam induzir variações acústicas mais representativas das condições clínicas enfrentadas por usuários reais de sistemas de ASR. Em vez de depender unicamente de ruídos ou transposições artificiais genéricas, a formação dos pares positivos será guiada por perturbações que refletem limitações motoras ou articulatórias, conforme sugerido por estudos como Shahamiri, Lal e Shah (2023), que indicam ganhos significativos quando os dados de pré-treinamento se alinham às características do domínio-alvo.

A expectativa é que essa integração favoreça o aprendizado de representações mais robustas e sensíveis à variação patológica, contribuindo para a generalização dos modelos mesmo em situações de baixa inteligibilidade. Além disso, acredita-se que essa estratégia poderá reduzir o impacto do desbalanceamento entre falas normais e disártricas durante o *fine-tuning* supervisionado, fornecendo *embeddings* pré-treinados mais estáveis e informativos.

Por fim, o uso dessas transformações como base para contrastes no aprendizado autossupervisionado também poderá estimular investigações futuras sobre o design de aumentos clínico-específicos como alternativa ou complemento aos tradicionais aumentos de áudio, ampliando as possibilidades de aplicação do paradigma em contextos de fala patológica.

1.4 Organização do trabalho

O restante deste documento está estruturado da seguinte forma:

O Capítulo 2 apresenta os conceitos básicos e fundamentos necessários para compreender os modelos e técnicas abordados neste trabalho, como aprendizado autossupervisionado, redes neurais convolucionais, espectrogramas Mel e STFT, bem como os modelos SimCLR e SwAV.

O Capítulo 3 revisa os trabalhos relacionados e as abordagens existentes no estado da arte para reconhecimento de fala disártrica, discutindo métodos de aumento de dados, arquiteturas de redes neurais e bases de dados utilizadas em estudos anteriores. Além disso, o capítulo detalha técnicas de transformação de dados, incluindo a Oclusão Espec-

tral, uma técnica proposta pelo autor deste texto, em colaboração com seu orientador, e apresentada em artigo publicado na CBMS 2024.

O Capítulo 4 detalha a metodologia proposta, a qual, resumidamente, busca integrar as técnicas de aumento de dados exploradas no capítulo anterior, com as técnicas de aprendizado autossupervisionado. Ainda, é explorado o estado da arte no que tange à aplicação de métodos de autossupervisão em atividades de reconhecimento de fala.

Capítulo 2

Conceitos Básicos

Neste capítulo, apresentam-se conceitos importantes que fundamentam o desenvolvimento deste trabalho, em particular: a definição de redes convolucionais e a descrição de suas principais características e aplicações; o uso de espectrogramas como representação visual de sinais de áudio e sua relevância no contexto do processamento de fala; e, por fim, os métodos de aprendizado autossupervisionado, abrangendo aprendizado de representações e aprendizado contrastivo, com destaque para suas particularidades e contribuições no contexto da extração de características em tarefas relacionadas ao reconhecimento de fala.

2.1 Redes Neurais Convolucionais

As Redes Neurais Convolucionais (do inglês: *Convolutional Neural Networks*) (CNNs) formam uma categoria de redes profundas desenvolvidas para lidar com dados que possuem uma organização espacial ou temporal, como imagens, séries temporais e vídeos. O uso dessas arquiteturas ganhou notoriedade a partir do trabalho de Krizhevsky et al. (2012), com a proposta da AlexNet, que superou outras abordagens no desafio ImageNet e demonstrou o potencial das CNNs em tarefas visuais de larga escala (KRIZHEVSKY; SUTSKEVER; HINTON, 2012).

A Figura 1 ilustra a estrutura geral de uma rede neural convolucional profunda, individualizando seus principais blocos como as camadas convolucionais e as camadas de *pooling*

Diferente das redes densas, onde todos os neurônios se conectam a todas as ativações da camada anterior, as CNNs utilizam filtros com dimensões reduzidas que percorrem a entrada extraindo padrões locais. Essa operação, denominada convolução, realiza mul-

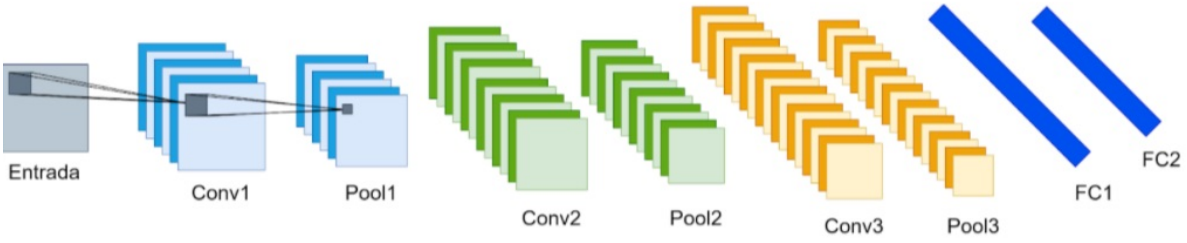


Figura 1 – Diagrama geral de uma CNN. (Fonte: Alura)

tiplicações ponto a ponto entre os valores de entrada e os pesos do filtro, gerando um novo conjunto de ativações chamado de mapa de características. Em uma dimensão, a convolução discreta entre uma entrada x e um filtro w é definida por:

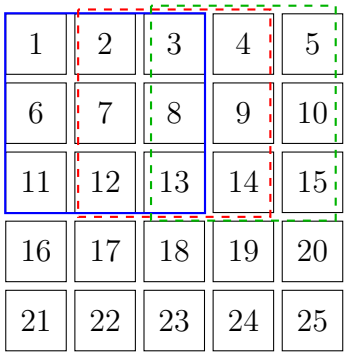
$$s(t) = (x * w)(t) = \sum_{a=-\infty}^{\infty} x(a) \cdot w(a-t) \tag{1}$$

Para o caso bidimensional, como em imagens, a operação generaliza-se para:

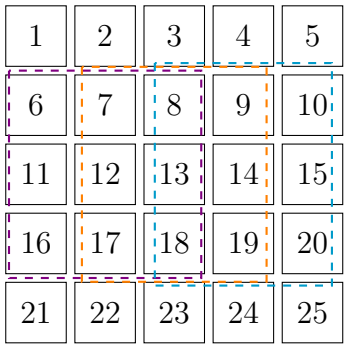
$$S(i,j) = (I * K)(i,j) = \sum_{m} \sum_{n} I(m,n) \cdot K(i-m, j-n) \tag{2}$$

onde I representa a entrada e K o filtro convolucional. O deslizamento do filtro sobre a entrada permite detectar padrões em diferentes regiões do espaço, como bordas, contornos, texturas, linhas horizontais ou verticais e outras estruturas locais características do dado de entrada. O deslocamento de cada passo do filtro é definido pelo parâmetro *stride*, que controla o quanto o kernel se move a cada iteração. Um *stride* de valor 1 indica que o filtro se move um pixel por vez, enquanto valores maiores provocam maior redução dimensional e menos sobreposição entre regiões vizinhas. Com o aumento do *stride*, a resolução espacial da saída diminui, mas a operação torna-se computacionalmente mais eficiente.

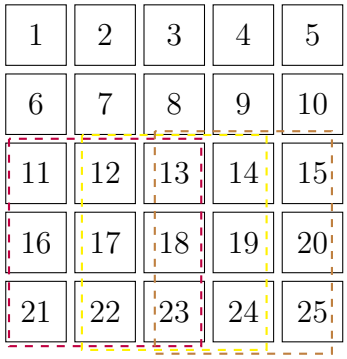
Outro parâmetro importante é o tamanho do filtro (ou janela). Filtros menores, como os de tamanho , são capazes de captar padrões locais finos e são frequentemente empilhados para aumentar a profundidade da rede. Já filtros maiores, como , capturam padrões de maior escala com uma única operação, mas também implicam em maior número de parâmetros e maior custo computacional. As figuras 2 e 3 ilustram o processo de deslocamento de um filtro (*kernel)* aplicado a fatores de *stride* 1 e 2, respectivamente.



(a) Deslocamentos iniciais do kernel 3×3 com stride 1.



(b) Deslocamentos intermediários do kernel 3×3 com stride 1.



(c) Deslocamentos finais do kernel 3×3 com stride 1

Figura 2 – Operação de convolução com kernel 3×3 e stride 1

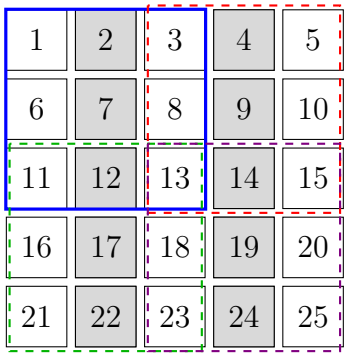


Figura 3 – Operação de convolução com kernel 3×3 e stride 2. Células não utilizadas estão em cinza. Janelas ativas são destacadas com cores distintas para melhor visualização.

As tabelas 1 e 2 mostram os resultados da aplicação de um kernel 3×3 com pesos $\begin{bmatrix} 1 & 0 & -1 \\ 1 & 0 & -1 \\ 1 & 0 & -1 \end{bmatrix}$ sobre a entrada das Figuras 2 e 3, para diferentes valores de stride:

Tabela 1 – Resultado da convolução com stride 1

-6	-6	-6
-6	-6	-6
-6	-6	-6

Tabela 2 – Resultado da convolução com stride 2

-6	-6
-6	-6

Entre os principais benefícios das CNNs estão a redução no número de parâmetros devido ao compartilhamento de pesos, a capacidade de explorar a estrutura local dos dados por meio de conexões esparsas, e a propriedade de equivariância a translações, que preserva a localização relativa dos padrões detectados (GOODFELLOW; BENGIO; COURVILLE, 2016b).

As CNNs são compostas por blocos que aplicam sequencialmente à convolução, uma função de ativação não linear e uma etapa de redução espacial, como o *pooling*. A função de ativação: Unidade Linear Retificada (do inglês: *Rectified Linear Unit*) (ReLU), definida por $\text{ReLU}(x) = \max(0, x)$, é amplamente utilizada por sua simplicidade e eficiência na propagação do gradiente. A seguir, o *pooling* reduz a dimensão dos mapas de ativação, promovendo invariância a pequenas mudanças na posição dos padrões e reduzindo a complexidade computacional. Entre os tipos mais comuns está o *max pooling*, que retém o valor máximo em uma janela local. As Figuras 4 e 5 ilustram, respectivamente a topologia da função ReLU e o processo de *max pooling* aplicado a uma matriz bidimensional.

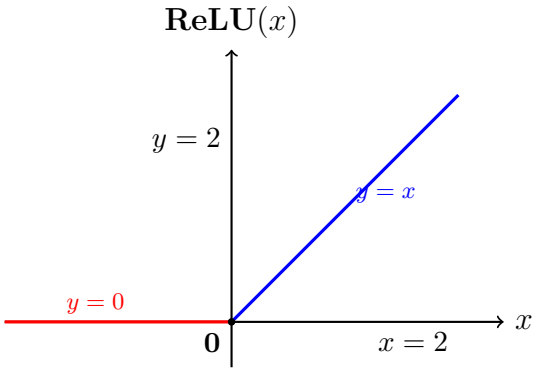


Figura 4 – Gráfico da função ReLU. Os valores negativos são anulados (vermelho) e os positivos são mantidos linearmente (azul).

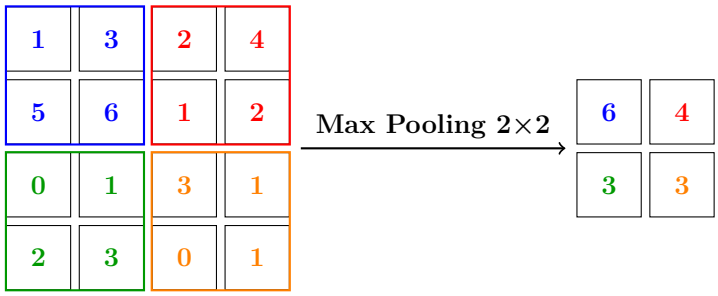
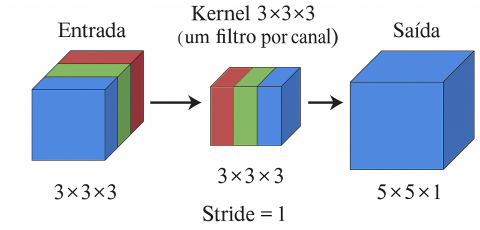


Figura 5 – Operação de Max Pooling com janela 2×2 e stride 2. Os valores máximos de cada região são destacados com cores correspondentes.

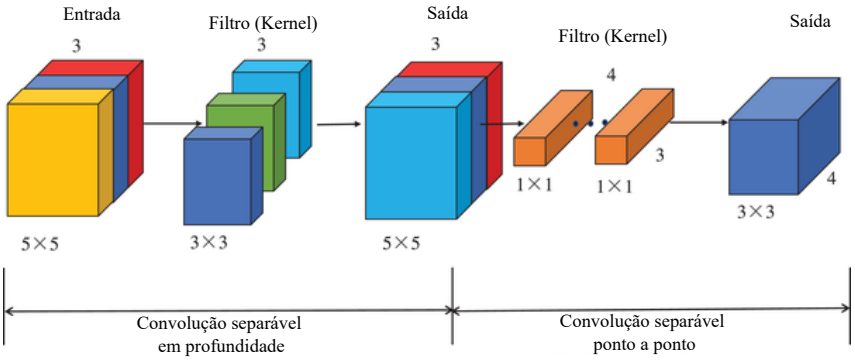
Diversas variações das camadas convolucionais surgiram nos últimos anos, como o uso de filtros 1×1 para redução de profundidade, convoluções transpostas para expansão espacial, e convoluções separáveis, como as convoluções separáveis por profundidade (do inglês: *depthwise separable convolutions*) utilizadas em arquiteturas eficientes como o MobileNet. A Figura 6 ilustra os tipos de convolução citados.

A convolução transposta é usada para aumentar a resolução espacial de um mapa de ativação. Ao contrário da convolução tradicional, que reduz a dimensão das imagens, a transposta expande a entrada, inserindo zeros entre os pixels antes da aplicação do filtro. Essa técnica é essencial em tarefas como reconstrução de imagens e segmentação semântica, especialmente em arquiteturas *autoencoder* - será abordada futuramente neste capítulo - e GANs (DUMOULIN; VISIN, 2016).

As convoluções separáveis reduzem a complexidade computacional da rede sem comprometer sua capacidade de extração de características. A convolução espacialmente separável decompõe um filtro 2D em dois filtros 1D. Já a convolução separável em profundidade, amplamente utilizada em redes eficientes como MobileNet, divide o processo em duas etapas: aplicação individual de filtros por canal (*depthwise*), seguida por uma convolução 1×1 para combinar os canais (*pointwise*) (CHOLLET, 2017).



(a) - Diagrama convolução transposta. Fonte: Os autores



(b) - Diagrama convolução separável em profundidade. Fonte: (WU; LI; ZHOU, 2022)

Figura 6 – Tipos de convoluções

A evolução das CNNs levou ao desenvolvimento de arquiteturas notáveis. A VGG-Net aprofundou as redes usando filtros pequenos (3×3), enquanto a ResNet introduziu conexões residuais que atenuam o problema do desaparecimento do gradiente em redes profundas (HE et al., 2016). A DenseNet, por sua vez, promove conexões entre todas as camadas de um bloco, favorecendo a reutilização de representações e o fluxo do gradiente (HUANG et al., 2017).

Durante o treinamento, variações na distribuição interna das ativações podem dificultar a convergência. Para lidar com esse fenômeno, Ioffe e Szegedy (2015) propuseram a normalização por lote (*Batch Normalization*), que estabiliza os dados normalizando a média e a variância dos mapas de ativação dentro de cada mini-batch. A normalização é descrita por:

$$\hat{x} = \frac{x - \mathbb{E}[x]}{\sqrt{Var[x]}} \quad , \quad y = \gamma\hat{x} + \beta \tag{3}$$

onde γ e β são parâmetros aprendíveis. Essa técnica acelera o treinamento, melhora a estabilidade do gradiente e atua como regularizador (IOFFE; SZEGEDY, 2015).

2.2 Espectrogramas STFT e MFCC

O uso de espectrogramas na análise de sinais de fala tem se mostrado uma ferramenta eficiente para a representação visual das características temporais e espectrais do sinal.

A Figura 7 ilustra os dois tipos mais comumente utilizados em análises acústicas: o espectrograma da Transformada de Fourier de Curto Prazo (do inglês: *Short-Time Fourier Transform*) (STFT) e o espectrograma de Coeficientes Cepstrais em Mel-frequência (do inglês: *Mel Frequency Cepstral Coefficients*) (MFCC).

Transformada de Fourier de Curto Prazo (STFT)

A STFT é amplamente utilizada para analisar sinais de fala ao longo do tempo. Essa técnica consiste em segmentar o sinal contínuo $x(t)$ em janelas de curta duração, normalmente de 20 a 40 milissegundos, assumindo que o sinal é aproximadamente estacionário dentro de cada janela. Para cada janela $w(t)$, aplica-se a transformada de Fourier, gerando um espectro de frequência local:

$$X(t,\omega) = \int_{-\infty}^{\infty} x(\tau) w(\tau - t) e^{-j\omega\tau} d\tau \tag{4}$$

Na prática, essa operação é discretizada e computada via Transformada Rápida de Fourier (do inglês: *Fast Fourier Transform*) (FFT), gerando o espectrograma como o módulo ao quadrado da STFT:

$$\text{Espectrograma}_{\text{STFT}}(n,k) = |X(n,\omega_k)|^2 \tag{5}$$

onde n é o índice temporal (posição da janela) e ω_k é a frequência discretizada.

A resolução temporal da STFT depende do tamanho da janela $w(t)$: janelas curtas oferecem boa resolução temporal, mas menor resolução em frequência; já janelas longas proporcionam maior resolução espectral, mas perdem detalhes temporais (ALLEN; RABINER, 1977). Essa limitação é conhecida como o princípio de incerteza de Gabor: não é possível simultaneamente obter alta resolução temporal (σ_t) e alta resolução espectral (σ_f) para o mesmo sinal, pois

$$\sigma_t \, \sigma_f \;\geq\; \tfrac{1}{4\pi}$$

(GABOR, 1946; COHEN, 1995). Janelas mais curtas melhoram a localização temporal, mas alargam as bandas de frequência; janelas mais longas produzem o efeito oposto.

Coeficientes Cepstrais na Escala Mel (MFCC)

Os MFCC derivam do espectrograma STFT, mas introduzem transformações inspiradas na percepção auditiva humana. O processo de cálculo do MFCC envolve as seguintes etapas matemáticas:

1. Aplicação da STFT: gera-se o espectro de magnitude $|X(n,\omega)|$ para cada janela.

2. Mapeamento para a Escala Mel: o espectro é passado por um banco de filtros triangulares sobrepostos definidos na escala Mel, que é dada por:

$$m(f) = 2595 \cdot \log_{10}\left(1 + \frac{f}{700}\right) \tag{6}$$

   Cada filtro $H_m(k)$ soma as energias das bandas correspondentes, aproximando a sensibilidade auditiva do ouvido humano.

3. Logaritmo da Energia: transforma-se cada energia filtrada E_m com a função logarítmica:

$$\log E_m = \log\left(\sum_k |X(n,k)|^2 H_m(k)\right) \tag{7}$$

4. Transformada Discreta do Cosseno (do inglês: *Discrete Cosine Transform*) (DCT): é aplicada à sequência de log-energias para obter os coeficientes cepstrais:

$$c_n = \sum_{m=1}^{M} \log E_m \cdot \cos\left[\frac{\pi n(m-0.5)}{M}\right] \tag{8}$$

   onde M é o número de filtros Mel e c_n são os coeficientes MFCC.

Esses coeficientes descrevem o envelope espectral do sinal, condensando suas características acústicas mais relevantes em poucas dimensões, o que facilita o uso em sistemas de reconhecimento automático de fala (DAVIS; MERMELSTEIN, 1980).

Apesar das vantagens perceptuais da escala Mel, a transformação reduz a resolução em altas frequências, o que pode limitar a detecção de detalhes sutis do sinal. Contudo, essa perda é geralmente compensada pelos modelos de aprendizado profundo, que aprendem representações discriminativas diretamente a partir dos espectrogramas (GOODFELLOW; BENGIO; COURVILLE, 2016b). A escolha entre STFT e MFCC depende do objetivo da tarefa: enquanto a STFT oferece uma representação mais completa do conteúdo de frequência, o MFCC fornece uma codificação mais compacta e robusta, adequada especialmente para tarefas como reconhecimento de fala.

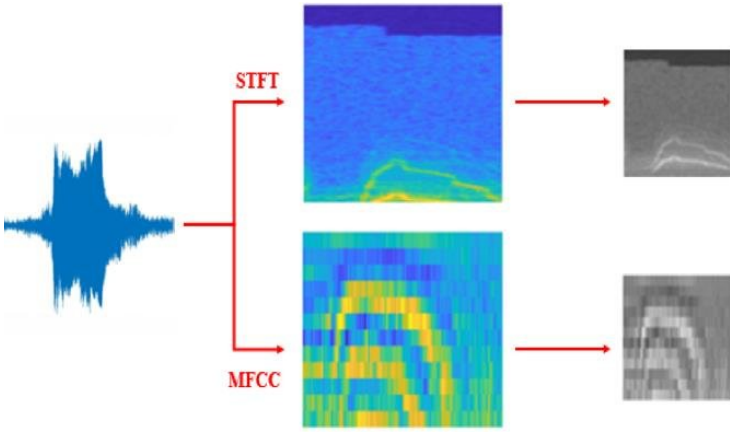


Figura 7 – Comparativo entre espectrogramas MFCC e STFT. Fonte: (SAAD; AHMED; ELARABY, 2024)

2.3 Aprendizado Autossupervisionado

O Aprendizado Autossupervisionado (do inglês: *Self-Supervised Learning*) (SSL) em redes neurais é uma técnica avançada que permite que modelos de aprendizado de máquina aprendam a partir de dados não rotulados, gerando seus próprios rótulos a partir das informações disponíveis. Esta abordagem é particularmente útil em cenários onde a rotulação manual de dados é impraticável devido ao volume ou custo.

O SSL envolve a criação de tarefas auxiliares que permitem ao modelo gerar rótulos a partir dos dados não rotulados. Por exemplo, em Processamento de Linguagem Natural (PLN), um modelo pode ser treinado para prever a próxima palavra em uma sequência de texto, utilizando grandes volumes de dados textuais não rotulados (DEVLIN et al., 2019b).

O modelo é treinado utilizando esses rótulos gerados automaticamente. Durante o treinamento, a rede neural ajusta seus pesos e biases para minimizar a diferença entre suas predições e os rótulos gerados. Este processo é iterativo e contínuo, permitindo que o modelo aprenda representações cada vez mais precisas dos dados (LECUN; BENGIO; HINTON, 2015).

Após o treinamento inicial, o modelo é avaliado utilizando um conjunto de dados de validação. Com base nos resultados, ajustes são feitos para melhorar a precisão e a robustez do modelo. Este ciclo de treinamento e avaliação continua até que o modelo atinja um desempenho satisfatório (GOODFELLOW; BENGIO; COURVILLE, 2016b).

2.3.1 Aprendizado de Representações

O aprendizado de representações é uma abordagem central no aprendizado de máquina que busca transformar dados brutos, como imagens, áudios ou textos, em representações mais úteis para tarefas específicas, como classificação ou reconhecimento. Essa abordagem reduz a dependência de características manuais, permitindo que o modelo extraia padrões relevantes diretamente dos dados (LECUN; BENGIO; HINTON, 2015). AEs e VAEs são métodos de aprendizado autossupervisionado de representações que aprendem a codificar dados em um espaço latente informativo. São amplamente utilizados em tarefas como redução de ruído, compressão de dados, geração de amostras e detecção de anomalias, sendo os VAEs especialmente indicados quando se deseja modelar distribuições probabilísticas e realizar geração controlada de dados. Em processamento de fala, modelos como o Representações de Codificadores Bidirecionais de Unidade Oculta de Transformadores (do inglês: *Hidden-Unit Bidirectional Encoder Representations from Transformers*) (HuBERT) destacam-se por aprender representações de áudio bruto, facilitando o uso em tarefas como reconhecimento automático de fala, mesmo com dados rotulados limitados (HSU et al., 2021b).

Autocodificadores(AEs)

AEs são redes neurais projetadas para aprender uma representação compacta dos dados de entrada. Sua estrutura típica envolve um codificador que condensa as informações e um decodificador que tenta reconstruir os dados a partir dessa codificação reduzida.

Estrutura e Funcionamento Codificador (*Encoder*): O codificador é uma rede neural que mapeia os dados de entrada ($\mathbf{x}$) para uma representação latente ($\mathbf{z}$). Isso é feito através de uma série de camadas de neurônios que reduzem progressivamente a dimensionalidade dos dados.

$$\mathbf{z} = f(\mathbf{x}) \tag{9}$$

Decodificador (*Decoder*): O decodificador é uma rede neural que mapeia a representação latente ($\mathbf{z}$) de volta para os dados de entrada ($\mathbf{x}$). Isso é feito através de uma série de camadas de neurônios que aumentam progressivamente a dimensionalidade dos dados.

$$\mathbf{x}' = g(\mathbf{z}) \tag{10}$$

A figura 8 ilustra a arquitetura básica de um autocodificador:

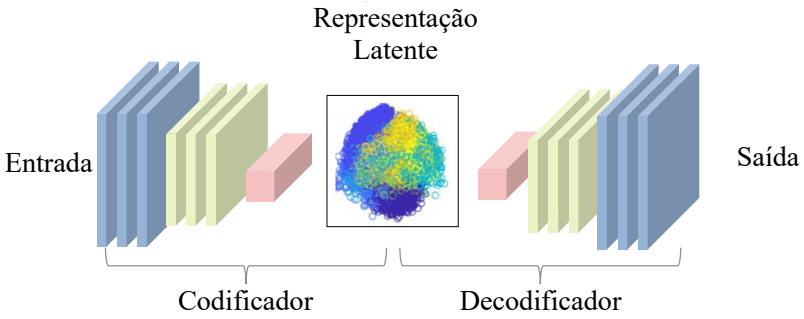


Figura 8 – Arquitetura de um autocodificador. Fonte: MathWorks (2024)

Matemática dos codificadores: O objetivo do treinamento de um codificador é minimizar a diferença entre os dados de entrada ($\mathbf{x}$) e a reconstrução ($\mathbf{x}'$). Isso é feito minimizando uma função de perda, geralmente utiliza-se a perda quadrática média (MSE):

$$\mathcal{L}(\mathbf{x}, \mathbf{x}') = |\mathbf{x} - \mathbf{x}'|^2 \tag{11}$$

Autocodificadores Variacionais (VAEs)

Os VAEs são uma extensão dos AEs que introduzem uma abordagem probabilística para a codificação dos dados. Eles não apenas aprendem uma representação compacta, mas também modelam a distribuição dos dados no espaço latente.

Estrutura e Funcionamento do Codificador (*Encoder*): O codificador mapeia os dados de entrada ($\mathbf{x}$) para uma distribuição probabilística no espaço latente, geralmente assumindo uma distribuição normal.

$$q(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z}; \mu(\mathbf{x}), \sigma^2(\mathbf{x})) \tag{12}$$

Decodificador (*Decoder*): O decodificador mapeia amostras da distribuição latente de volta para os dados de entrada.

$$p(\mathbf{x}|\mathbf{z}) \tag{13}$$

A figura 9 ilustra a arquitetura básica de um VAE:

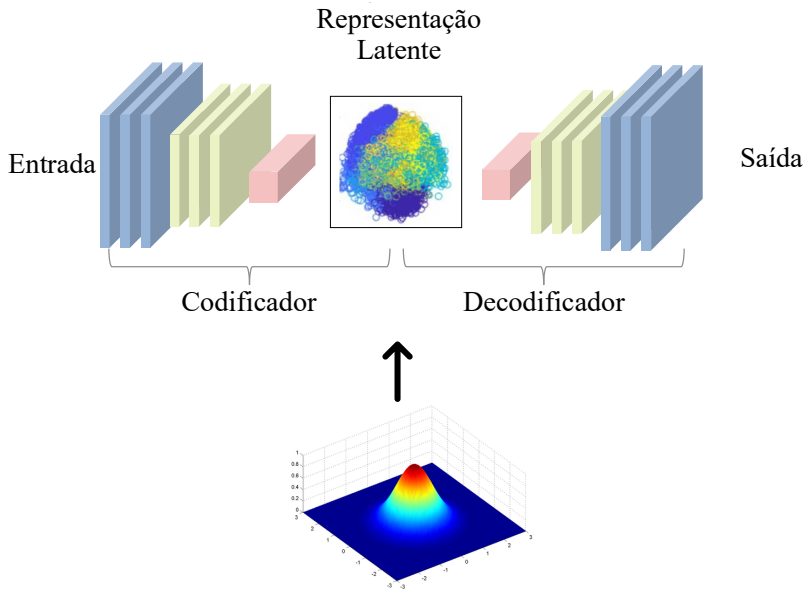


Figura 9 – Arquitetura de um autocodificador variacional. Fonte: MathWorks (2024 - Modificado)

Matemática dos VAEs: Os VAEs utilizam a inferência variacional para aproximar a distribuição posterior ($p(\mathbf{z}|\mathbf{x})$) com uma distribuição ($q(\mathbf{z}|\mathbf{x})$). A função de perda dos VAEs é composta por dois termos:

Erro de Reconstrução: Mede a diferença entre os dados de entrada ($\mathbf{x}$) e a reconstrução ($\mathbf{x}'$).

$$\mathcal{L}_{\text{rec}}(\mathbf{x}, \mathbf{x}') = -\mathbb{E}_{q(\mathbf{z}|\mathbf{x})}[\log p(\mathbf{x}|\mathbf{z})] \tag{14}$$

Divergência KL: Mede a diferença entre a distribuição aproximada ($q(\mathbf{z}|\mathbf{x})$) e a distribuição prior ($p(\mathbf{z})$).

$$\mathcal{L}_{\text{KL}}(q(\mathbf{z}|\mathbf{x})|p(\mathbf{z})) = \text{KL}(q(\mathbf{z}|\mathbf{x})|p(\mathbf{z})) \tag{15}$$

A função de perda total é a soma desses dois termos:

$$\mathcal{L}(\mathbf{x}, \mathbf{x}') = \mathcal{L}_{\text{rec}}(\mathbf{x}, \mathbf{x}') + \mathcal{L}_{\text{KL}}(q(\mathbf{z}|\mathbf{x})|p(\mathbf{z})) \tag{16}$$

HuBERT

O HuBERT é um modelo de aprendizado de representações de fala autossupervisionado que utiliza uma abordagem de predição mascarada de unidades ocultas (HSU et al., 2021a). Seu funcionamento baseia-se em duas etapas principais: a extração de pseudo-rótulos e o aprendizado de representações de fala, empregando diferentes funções de perda, como a Entropia Cruzada e a Classificação Temporal Conexiva (do inglês: *Connectionist Temporal Classification*) (CTC)

Extração de Pseudo-rótulos

Na primeira etapa, o modelo utiliza STFT para converter o sinal de áudio em um espectrograma. A STFT é matematicamente representado pela seguinte equação:

$$X(t, f) = \sum_{n=-\infty}^{\infty} x[n] \cdot w[n - t] \cdot e^{-j2\pi fn} \tag{17}$$

em que $x[n]$ é o sinal de entrada, $w[n - t]$ é a janela de tempo aplicada ao sinal, e $e^{-j2\pi fn}$ é a exponencial complexa usada na Transformada de Fourier. Isso resulta em uma representação tempo-frequência do sinal de fala. Em seguida, um algoritmo de agrupamento como o *k-means* é aplicado aos coeficientes espectrais, gerando pseudo-rótulos que serão utilizados como alvos no treinamento do modelo (HSU et al., 2021a).

Predição Mascarada e Entropia Cruzada

Na segunda etapa, o HuBERT é treinado para prever as partes mascaradas do espectrograma utilizando uma abordagem de predição mascarada, semelhante ao modelo Representações de Codificadores Bidirecionais de Transformadores (do inglês: *Bidirectional Encoder Representations from Transformers*) (BERT) (DEVLIN et al., 2019a). O modelo mascara aleatoriamente uma fração do espectrograma e tenta prever as unidades ocultas associadas a essas partes mascaradas. A função de perda utilizada neste caso é a Entropia Cruzada, definida como:

$$L_{\text{Entropia Cruzada}} = -\sum_{i} y_i \log(\hat{y}_i) \tag{18}$$

em que y_i são os pseudo-rótulos verdadeiros, e $\hat{y}_i$ são as previsões do modelo. Essa função mede a discrepância entre as previsões do modelo e os pseudo-rótulos, com o objetivo de minimizar a perda e melhorar a acurácia do modelo (HSU et al., 2021a).

Perda CTC

Além da entropia cruzada, o HuBERT e outros modelos de reconhecimento de fala frequentemente empregam a CTC para lidar com a natureza sequencial e não alinhada dos sinais de áudio. Essa abordagem é particularmente eficaz em tarefas em que não há correspondência explícita entre as entradas (espectrogramas) e as saídas (transcrições ou rótulos).

A CTC mapeia probabilidades de sequência para rótulos mesmo quando as durações das entradas e saídas divergem. Seu objetivo é maximizar a probabilidade da transcrição correta, considerando todas as possíveis correspondências entre as previsões do modelo e a sequência-alvo (GRAVES et al., 2006a).

Matematicamente, é definida como:

$$L_{\text{CTC}} = -\log P(C|X) \tag{19}$$

em que $P(C|X)$ representa a probabilidade da sequência de caracteres C dada a entrada X. O algoritmo utiliza uma estratégia de eliminação de rótulos repetidos e a introdução de um *token* especial (*blank*) para lidar com variações temporais nas previsões (GRAVES et al., 2006a).

A combinação entre entropia cruzada (para predição mascarada) e CTC (para alinhamento sequencial) permite que o HuBERT capture padrões acústicos e linguísticos em múltiplos níveis de granularidade — desde características espectrais até representações de alto nível da fala. Esta última, em particular, é fundamental para o reconhecimento automático de fala, pois possibilita o aprendizado a partir de sequências de diferentes comprimentos, sem a necessidade de alinhamento explícito (BAEVSKI et al., 2020).

Decodificação *greedy*

Na etapa de decodificação de um sistema de reconhecimento de fala, o objetivo é converter, a cada instante temporal, o vetor de probabilidades estimado pela rede neural em uma sequência final de símbolos. A forma mais simples de realizar essa conversão é a chamada decodificação *greedy* — ou *best-path*. O algoritmo escolhe, em cada passo t, o símbolo de maior probabilidade $\arg\max_k P(k \mid \mathbf{h}_t)$ e avança imediatamente para o próximo passo sem manter hipóteses alternativas. Em modelos baseados em CTC, essa seleção é feita sobre cada quadro; ao final, removem-se os rótulos *blank* e colapsam-se repetições para obter a transcrição final (GRAVES et al., 2006b). Em arquiteturas autorregressivas — como o Listen, Attend and Spell (CHAN et al., 2016a) ou o Transdutor de Rede Neural Recorrente (do inglês: *Recurrent Neural Network Transducer*) (RNN-T) (GRAVES, 2012) — o símbolo escolhido retroalimenta o decodificador, e o processo termina quando é emitido o marcador de fim de sentença.

O apelo da abordagem *greedy* reside no seu custo computacional linear no comprimento da sequência e no tamanho do vocabulário, além de exigir memória constante. Essas características permitem sua implementação vetorizada em GPU ou mesmo em Processador Digital de Sinais (do inglês: *Digital Signal Processor*) (DSP) embarcado, reduzindo a latência a poucos dezenas de milissegundos em dispositivos móveis (HANNUN et al., 2014). Por outro lado, como a estratégia não mantém hipóteses alternativas, decisões locais equivocadas não podem ser revistas, o que torna o método sensível a pequenas oscilações de probabilidade e ruídos. Comparações empíricas mostram que a diferença em WER entre a decodificação *greedy* e a busca em feixe (*beam search*) com feixes estreitos (largura 8 a 10) pode variar de cinco a quinze pontos percentuais, dependendo do grau de ruído e da complexidade do vocabulário (KIM; HORI; WATANABE, 2017; WATANABE et al., 2017). Assim, a decodificação *greedy* é frequentemente a opção preferida em aplicações que priorizam latência mínima e execução totalmente local — como interfaces de comando por voz —, enquanto técnicas baseadas em busca em feixe são reservadas a cenários que visam máxima precisão de transcrição, mesmo com maior custo computacional.

2.3.2 Aprendizado Contrastivo

O aprendizado contrastivo tem se estabelecido como uma abordagem central no aprendizado de representações autossupervisionadas, oferecendo um mecanismo eficiente para lidar com dados não rotulados. Fundamentado em princípios de similaridade semântica, o método busca mapear dados de entrada para um espaço latente onde pares positivos (amostras semanticamente relacionadas) são aproximados, enquanto pares negativos (amostras não relacionadas) são afastados (OORD; LI; VINYALS, 2018). Esse processo é mediado por funções de perda específicas, como a InfoNCE, que maximiza a similaridade entre pares positivos em relação aos negativos dentro de um mesmo minibatch. Essa estrutura tem demonstrado resultados superiores em tarefas *downstream*, reduzindo a dependência de rótulos extensivos, especialmente em contextos onde a anotação de dados é limitada ou onerosa.

Na prática, o aprendizado contrastivo é frequentemente implementado em pipelines que utilizam redes neurais profundas para codificação das representações. Modelos como SimCLR destacaram-se ao integrar técnicas de aumento de dados (e.g., transformações geométricas e alterações no domínio do espectro) para gerar pares contrastivos robustos, explorando a variabilidade intrínseca dos dados (CHEN et al., 2020c). Além disso, o aprendizado contrastivo foi estendido para domínios além de imagens, com aplicações notáveis em sinais de áudio. Por exemplo, o wav2vec e suas variantes demonstraram que é possível aprender representações úteis para reconhecimento de fala sem supervisão explícita, utilizando objetivos contrastivos para distinguir segmentos temporais de áudio com base em seu contexto (SCHNEIDER et al., 2019).

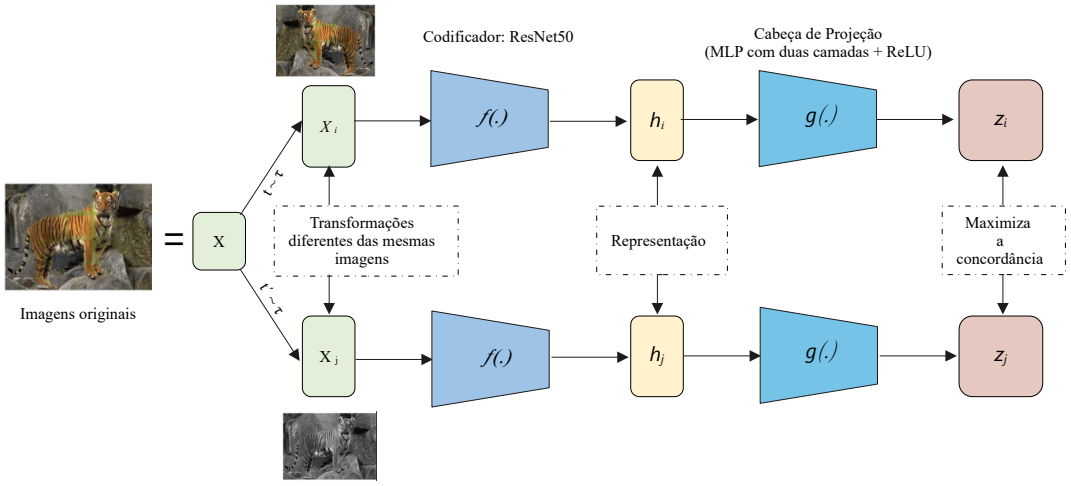


Figura 10 – Diagrama SimCLR. FONTES: Diagrama adaptado de: Chen et al. (2020c) - Imagens adaptadas de: https://medium.com/data-science/a-framework-for-contrastive-self-supervised-learning-and-designing-a-new-approach-3caab5d29619

SimCLR

O SimCLR, é um modelo de aprendizado autossupervisionado baseado em aprendizado contrastivo que visa melhorar as representações visuais sem a necessidade de rótulos supervisionados. O modelo utiliza uma rede ResNet como *backbone*, e a projeção é realizada em um espaço latente reduzido por uma camada adicional de projeção. O objetivo é maximizar a similaridade entre diferentes transformações de uma mesma imagem (âncoras positivas) e minimizar a similaridade entre imagens diferentes (âncoras negativas) (CHEN et al., 2020c). A figura 10 ilustra a estrutura da técnica apresentada no *paper* original.

A função de perda usada no SimCLR é a InfoNCE (*Noise Contrastive Estimation*), que pode ser expressa pela seguinte equação:

$$\ell_{i,j} = -\log \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} \mathbf{1}_{[k \neq i]} \exp(\text{sim}(z_i, z_k)/\tau)} \tag{20}$$

em que $\text{sim}(z_i, z_j)$ é a similaridade coseno entre os vetores de projeção z_i e z_j, e τ é um parâmetro de temperatura que controla a escala da distribuição de similaridades. O denominador da equação soma as similaridades de todas as outras amostras z_k, exceto i, maximizando a distância das âncoras negativas (CHEN et al., 2020c).

A arquitetura e a função de perda propostas permitem treinar modelos de forma eficiente utilizando aprendizado autossupervisionado, sem a necessidade de rótulos.

SwAV

O SwAV, é uma abordagem de aprendizado autossupervisionado que busca gerar representações discriminativas sem recorrer ao contraste explícito com amostras negativas. Em vez disso, o modelo realiza atribuições cruzadas entre diferentes transformações da mesma entrada, associando cada uma delas a um conjunto compartilhado de protótipos que representam regiões do espaço latente (CARON et al., 2020).

Durante o treinamento, o modelo aplica diferentes aumentações v_1, v_2, ..., v_n a uma amostra x e extrai seus *embeddings* $z_1 = f(v_1), z_2 = f(v_2), \ldots$, onde $f(\cdot)$ é o codificador. Cada *embedding* z_i é então normalizado para a esfera unitária e projetado em um espaço latente, sendo comparado com um conjunto fixo de K protótipos $c_1, c_2, \ldots, c_K$, geralmente armazenados como colunas de uma matriz $C \in \mathbb{R}^{d \times K}$.

A atribuição suave de um vetor z a esses protótipos é definida como:

$$p^{(k)} = \frac{\exp\left(\frac{z^\top c_k}{\tau}\right)}{\sum_{k'=1}^{K} \exp\left(\frac{z^\top c_{k'}}{\tau}\right)} \tag{21}$$

onde τ é um parâmetro de temperatura que controla a suavidade da distribuição.

No entanto, para garantir que as atribuições sejam balanceadas e evitar colapsos triviais (como todos os *embeddings* sendo atribuídos ao mesmo protótipo), o SwAV utiliza a normalização de *Sinkhorn-Knopp*, que projeta a matriz de similaridade $\mathbf{C}^\top \mathbf{Z}$ em uma distribuição quase uniforme:

$$Q^* = \mathrm{Diag}(u) \cdot \exp\left(\frac{\mathbf{C}^\top \mathbf{Z}}{\varepsilon}\right) \cdot \mathrm{Diag}(v) \tag{22}$$

onde u e v são vetores de normalização iterativamente ajustados, e ε é o parâmetro de regularização da entropia.

A função de perda do SwAV, chamada de *swapped prediction loss*, busca prever a atribuição de uma *view* a partir da representação da outra:

$$\mathcal{L}_{\mathrm{SwAV}} = \ell(z_t, q_s) + \ell(z_s, q_t) \tag{23}$$

onde $\ell(z, q)$ é uma perda de entropia cruzada entre o vetor q (a atribuição-alvo) e as probabilidades previstas para z:

$$\ell(z, q) = -\sum_{k=1}^{K} q^{(k)} \log p^{(k)} \tag{24}$$

Ao comparar diferentes versões da mesma amostra, o SwAV assegura que o modelo seja capaz de reconhecer padrões similares mesmo quando o espectro do sinal é alterado por técnicas como ruído, dilatação ou oclusão. Isso favorece o aprendizado de representações coerentes e robustas, algo fundamental quando se trabalha com sinais de fala sujeitos a distorções, como ocorre na disartria.

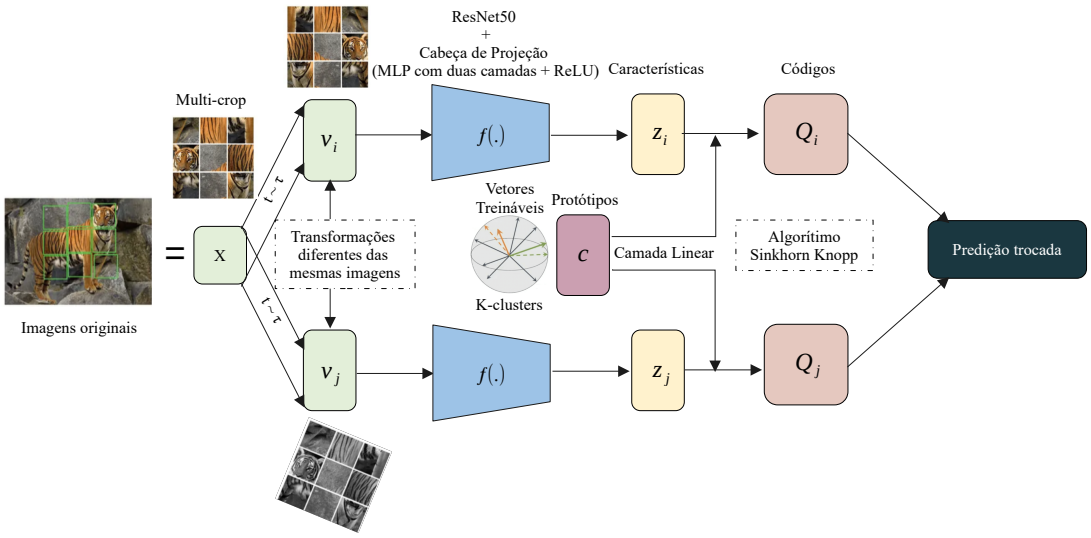


Figura 11 – Diagrama do SwAV. FONTES: Diagrama adaptado de: Caron et al. (2020) - Imagens: https://theaisummer.com/swav/ e https://medium.com/data-science/a-framework-for-contrastive-self-supervised-learning-and-designing-a-new-approach-3caab5d29619

A figura 11 ilustra o modelo proposto por Caron et al. (2020).

2.3.3 Métricas de Avaliação: CER, WER e WRA

No contexto do ASR, a avaliação do desempenho dos modelos é geralmente realizada por meio de métricas que quantificam os erros na transcrição em relação ao conteúdo de referência. As três métricas mais comuns nesse domínio são a CER, a WER e o WRA.

A métrica WER (25) é a mais amplamente utilizada na literatura para avaliar sistemas de ASR (GOLDWATER; JURAFSKY; MANNING, 2011). Ela calcula a proporção de palavras incorretas na saída do sistema, considerando inserções (I), deleções (D) e substituições (S), de acordo com a fórmula:

$$WER = \frac{S_{word} + D_{word} + I_{word}}{N_{word}} \tag{25}$$

em que N é o número total de palavras na transcrição de referência. Para avaliações de granularidade mais fina, a métrica CER é aplicada, especialmente em sistemas com vocabulário reduzido ou em línguas morfologicamente ricas. O CER (26) é computado de maneira análoga ao WER, mas com base em caracteres em vez de palavras (CHAN et al., 2016b).

$$CER = \frac{S_{char} + D_{char} + I_{char}}{N_{char}} \tag{26}$$

Já a WRA (27) é uma métrica complementar que representa diretamente a acurácia no reconhecimento de palavras, sendo definida como:

$$WRA = 1 - WER \tag{27}$$

Capítulo 3

Aumento de Dados para Fala Disártrica

Neste Capítulo, serão apresentados os desenvolvimentos relacionados à primeira etapa deste trabalho, que envolve a investigação das metodologias utilizadas no estado da arte do reconhecimento de fala disártrica, bem como as técnicas de aumento de dados e os resultados obtidos após as aplicações dessas técnicas. Em particular foram utilizados dois pipelines com arquiteturas diferenciadas e aplicados diferentes processos de aumento de dados, resultando em artigo publicado na conferência CBMS 2024 (GRACELLI; ALMEIDA, 2024).

3.1 Introdução

Dadas as características intrínsecas dos comprometimentos motores da fala, os sistemas de ASR não apresentam bom desempenho quando utilizados por indivíduos com dificuldades na articulação da linguagem oral, como pacientes disártricos. Muitos pesquisadores de várias áreas têm se esforçado para encontrar uma solução eficiente para os problemas de comunicação enfrentados por esses pacientes. No campo da Ciência da Computação, trabalhos de ponta incluem SpeechVision (SHAHAMIRI, 2021), E2E-DASR (ALMADHOR et al., 2023) e Dysarthric Speech Transformer (SHAHAMIRI; LAL; SHAH, 2023).

No entanto, esses esforços enfrentam desafios devido à disponibilidade limitada de bancos de dados robustos de voz disártrica. Para contornar essa limitação, os pesquisadores têm recorrido a técnicas como síntese de fala e adição de ruído às gravações de áudio. Essas soluções apresentam desafios como a demanda de tempo para treinar um sistema

de aprendizado profundo para síntese de fala e a adição de anomalias ao áudio original, que então precisaria passar pelo processo de conversão de gravações de áudio em imagens espectrais.

Portanto, meios de aumentar os dados durante o processo de carregamento e estruturação de conjuntos de dados no curso do treinamento de redes neurais podem se tornar uma alternativa viável às adotadas em outras pesquisas relatadas.

Motivados por tais observações, o objetivo deste estudo foi, portanto, explorar a eficácia dos métodos de aumento de dados no contexto dos sistemas de reconhecimento de fala disártrica. Especificamente, focamos na avaliação de duas técnicas distintas: a primeira, utilizando adição de ruído e dilatação temporal; e a segunda, denominada Oclusão Espectral (do inglês: *Spectral Occlusion*), que foi a abordagem proposta nesse artigo. Esta análise se concentra em observar o potencial dessas técnicas para preservar ou melhorar a precisão dos sistemas de Reconhecimento Automático de Fala Disártrica (do inglês: *Dysarthric Automatic Speech Recognition*) (DASR).

3.2 Trabalhos Relacionados

3.2.1 Mecanismos de Atenção e *Transformers*

Nos últimos anos, os mecanismos de atenção em aprendizado profundo ganharam atenção significativa devido à sua aplicação bem-sucedida em várias tarefas, como tradução automática (BAHDANAU; CHO; BENGIO, 2015), reconhecimento de fala (CHOROWSKI et al., 2015) e legendagem de imagens (XU et al., 2015). Na classificação de imagens, os mecanismos de atenção têm sido utilizados para melhorar o desempenho das **CNN!**s ao permitir que a rede se concentre seletivamente em regiões informativas de uma imagem.

Trabalhos iniciais de Mnih et al. (2014) introduziram um modelo de atenção recorrente para classificação de imagens, permitindo atenção sequencial a diferentes regiões da imagem para capturar características informativas. No entanto, essa abordagem era computacionalmente cara e intensiva em parâmetros. Avanços recentes propuseram arquiteturas de CNNs com mecanismos de atenção para classificação de imagens.Wang et al. (2017) introduziram uma rede de atenção residual empregando módulos de atenção para focar seletivamente em regiões informativas. Da mesma forma, Hu et al. (2018) propuseram uma rede de compressão e excitação utilizando atenção por canal para ponderar a importância de diferentes canais de características.

Além disso, mecanismos de atenção espacial têm sido explorados para classificação de imagens. Fu, Zheng e Mei (2017) introduziram uma rede de vislumbre usando atenção espacial para focar seletivamente em diferentes regiões da imagem. Da mesma forma, Woo et al. (2018) propuseram um módulo de atenção em bloco convolucional utilizando mecanismos de atenção espacial e por canal para melhorar o desempenho das CNNs. Essas

abordagens destacam o potencial dos mecanismos de atenção em melhorar o desempenho das CNNs na classificação de imagens.

Transformadores de Visão (do inglês: *Vision Transformers*) (ViTs) e mecanismos de atenção são conceitos intimamente relacionados usados em aprendizado de máquina para ajudar redes neurais a focar em partes relevantes da entrada. ViTs são uma técnica de aprendizado profundo que emprega mecanismos de atenção para permitir que a rede se concentre em *patches* específicos da imagem.

Dosovitskiy et al. (2021) introduziram arquiteturas ViTs que aplicam diretamente a arquitetura dos transformadores a sequências de *patches* de imagem, alcançando resultados excepcionais em tarefas de classificação com recursos computacionais reduzidos necessários para treinamento. Os ViTs superam redes convolucionais de última geração quando pré-treinados em grandes conjuntos de dados e transferidos para vários benchmarks de reconhecimento de imagens, como ImageNet, CIFAR-100 e Benchmark de Adaptação de Tarefas Visuais (do inglês: *Visual Task Adaptation Benchmark*) (VTAB).

3.2.2 Reconhecimento Automático de Fala para Disartria

O sistema *SpeechVision* (SHAHAMIRI, 2021) é um modelo de DASR projetado especificamente para reconhecer palavras isoladas por meio de espectrogramas visuais. Para contornar a escassez de dados disártricos, foram aplicadas técnicas de aumento de dados baseadas em manipulações de largura, corte e zoom dos espectrogramas. Além disso, espectrogramas sintéticos foram gerados com o auxílio do modelo Conversor de Texto para Fala com Redes Convolucionais Profundas (do inglês: *Deep Convolutional Text-to-Speech*) (DC-TTS), resultando em melhorias significativas após análise do Pontuação Média de Opinião (do inglês: *Mean Opinion Score*) (MOS).

No trabalho de Shahamiri, Lal e Shah (2023), *Transformers* de Atenção foram empregados para classificar palavras disártricas com maior precisão. Utilizou-se um pipeline de aprendizado por transferência em duas fases, combinando pré-treinamento com fala saudável e ajuste fino com fala disártrica. Técnicas de aumento de dados de áudio foram incorporadas para expandir o conjunto de treinamento, com ganhos de até 23% na precisão para 73% dos participantes do banco de dados UA-Speech.

Outro avanço importante foi o sistema Reconhecimento Automático de Fala Disártrica de Ponta a Ponta (do inglês: *End-to-End Dysarthric Automatic Speech Recognition*) (E2E-DASR) (ALMADHOR et al., 2023), que utiliza uma arquitetura híbrida com Rede Neural Convolucional Espacial (do inglês: *Spatial Convolutional Neural Network*) (SCNN) e Transformador com Atenção Multicabeça (do inglês: *Multi-Head Attention Transformer*) (MHAT). Essa combinação permite a extração de características visuais e temporais dos fonemas disártricos, superando limitações observadas em modelos tradicionais. O estudo também emprega aprendizado por transferência com dados sintéticos gerados por um Conversor de Texto em Fala (do inglês: *Text-to-Speech*) (TTS) baseado em WaveNet,

resultando em melhorias de até 20,72% na acurácia para falantes com baixa inteligibilidade.

Recentemente, Vinotha et al. (2024) propuseram a integração do modelo SepFormer com uma Rede de Atenção Hierárquica, dentro de um esquema de aprendizado por transferência em múltiplas etapas. Essa arquitetura alcançou acurácia de até 84,07% em palavras isoladas, demonstrando eficácia em cenários com diferentes graus de severidade da disartria.

No mesmo contexto, Irshad et al. (2024) apresentaram o modelo Reconhecimento de Fala Disártrica com Transformer Aprimorado (do inglês: *Transformer-based Dysarthric Speech Recognition with Feature Enhancement*), baseado em *Transformer* com módulos de aprimoramento de características (*feature enhancement*). Avaliado nos bancos de dados UA-Speech e TORGO, o modelo obteve 97,75% de acurácia, sendo considerado um dos melhores desempenhos recentes em DASR.

Outra contribuição notável é de Hsieh e Wu (2024), que aplicaram aprendizado por currículo combinado com *embeddings* articulatórios para reorganizar os dados de treinamento com base na inteligibilidade. A estratégia demonstrou ganhos de até 11,37% na acurácia comparada às abordagens tradicionais.

Por fim, Wang et al. (2024) propuseram um método de adaptação baseado em protótipos, com o objetivo de melhorar o desempenho do ASR em falantes disártricos não vistos anteriormente. A proposta elimina a necessidade de ajustes manuais para novos usuários e utiliza *Transformers* supervisionados para gerar *embeddings* de características robustas e personalizadas.

3.2.3 Bancos de Dados Utilizados

Os experimentos foram realizados com dois bancos de dados de fala: o UA-Speech e o LJSpeech, escolhidos por sua complementaridade em termos de variabilidade, condições de gravação e tipos de fala.

UA-Speech

O banco de dados UA-Speech (KIM; HASEGAWA-JOHNSON; PERLMAN, 2008) é amplamente utilizado em estudos de reconhecimento automático de fala disártrica. Ele contém gravações de 15 pacientes com Paralisia Cerebral (fala disártrica) e 13 falantes de controle (fala típica). Cada participante produziu um total de 721 enunciados, abrangendo palavras isoladas, dígitos e expressões curtas. As gravações foram realizadas em ambiente controlado, com taxa de amostragem de 16 kHz e resolução de 16 bits. A duração média por falante é de aproximadamente 1.887 segundos. Os falantes são identificados por um prefixo que indica o gênero (M para masculino e F para feminino), seguido por um número que corresponde ao seu identificador único (ex.: M08, F05).

Um aspecto central do banco de dados é a anotação de inteligibilidade dos falantes disártricos, classificada em quatro níveis: *muito baixo*, *baixo*, *médio* e *alto*. Essa organização possibilita análises comparativas entre diferentes severidades da disartria, permitindo avaliar o desempenho dos modelos de reconhecimento de forma mais detalhada. Adicionalmente, os dados de cada paciente estão organizados em três blocos de gravação independentes (denominados B1, B2 e B3), o que favorece estratégias de particionamento para treino, validação e teste em experimentos de *machine learning*.

Além da organização em falantes normotípicos e disártricos, a base UA-Speech apresenta fenômenos específicos associados às deficiências articulatórias de pacientes com Paralisia Cerebral. Do ponto de vista perceptual, observam-se alongamentos excessivos de fonemas, reduções na taxa de fala, omissões de segmentos e prosódia monotônica, resultando em inteligibilidade reduzida mesmo em enunciados curtos (BALL et al., 2011; YORKSTON et al., 2010). Esses aspectos perceptuais estão diretamente relacionados a limitações no controle motor fino dos articuladores, como língua, lábios e mandíbula.

Sob a ótica acústica, tais dificuldades se refletem em padrões espectrográficos característicos, já bem documentados em estudos prévios (KIM; HASEGAWA-JOHNSON; PERLMAN, 2008; SCHUSTER; SCHULLER; WENINGER, 2019). Em falantes com inteligibilidade muito baixa, observam-se o alongamento dos formantes vocálicos, frequentemente acompanhados de transições mais lentas e pouco definidas; a presença de zonas de silêncio ou *gaps* espectrais decorrentes de falhas de coarticulação; variações abruptas na intensidade de segmentos consonantais, associadas a explosões fracas ou ausentes em oclusivas; e instabilidades periódicas na frequência fundamental (F0), que refletem irregularidades no controle da fonte glótica. Já em falantes classificados com inteligibilidade média ou alta, embora as deficiências permaneçam perceptíveis, os padrões espectrográficos apresentam maior preservação das transições formânticas e uma relação sinal-ruído mais favorável, o que possibilita um reconhecimento automático mais eficaz.

LJSpeech

O banco de dados LJSpeech (ITO; JOHNSON, 2017) é composto por aproximadamente 24 horas de fala contínua em inglês, gravada por uma única falante feminina norte-americana. Ele contém cerca de 13.100 enunciados extraídos de livros de domínio público, com locução clara e consistente. Diferentemente do UA-Speech, que é organizado em falantes e níveis de inteligibilidade, o LJSpeech não apresenta variabilidade de locutor, mas fornece uma base extensa de fala típica conectada.

Esse corpus foi utilizado como dado auxiliar para o pré-treinamento das redes neurais, fornecendo um ponto de partida robusto antes da adaptação aos dados mais restritos e heterogêneos da base UA-Speech. Tal estratégia é consistente com abordagens de *transfer learning*, em que grandes corpora de fala normotípica são empregados para inicializar modelos, posteriormente refinados em bases clínicas menores e mais específicas.

3.2.4 Métodos

Foram utilizados dois *frameworks* de *Speech-Transformer* para testes de aumento de dados. O primeiro, denominado *FW1*, é um sistema baseado em *Transformers* desenvolvido para o banco de dados LJSpeech e ajustado conforme os procedimentos de Shahamiri, Lal e Shah (2023). Consiste em uma rede inicial com três camadas convolucionais 1D seguidas por camadas ReLU, quatro codificadores empilhados, uma camada decodificadora e transformador de duas cabeças.

O segundo *framework*, *FW2*, também baseado em *Transformers*, foi desenvolvido originalmente para o mandarim e modificado de acordo com uma arquitetura superior relatada por Shahamiri, Lal e Shah (2023). Inclui cinco codificadores empilhados, três camadas decodificadoras e transformador de duas cabeças, com camadas separáveis por profundidade substituindo as camadas densas nos codificadores. Ambos os *frameworks* utilizaram função de perda de entropia cruzada, otimizador Adam e tamanho de lote de 64 amostras.

A figura 12 ilustra as arquiteturas *FW1* e *FW2*, respectivamente e a figura 13 ilustra de forma generalista o fluxo de trabalho adotado.

Os métodos adotados para aumento de dados incluíram a inserção de ruído branco com um fator de 10 (tf_1), de acordo com a metodologia aplicada por Shahamiri, Lal e Shah (2023) e a dilatação temporal com um fator de 0,9, simulando fala mais alongada (tf_2). Referimos a essa técnica de aumento como "AUG" ($tf_1 + tf_2$). Essas técnicas foram implementadas no grupo controle do banco de dados UA-Speech.

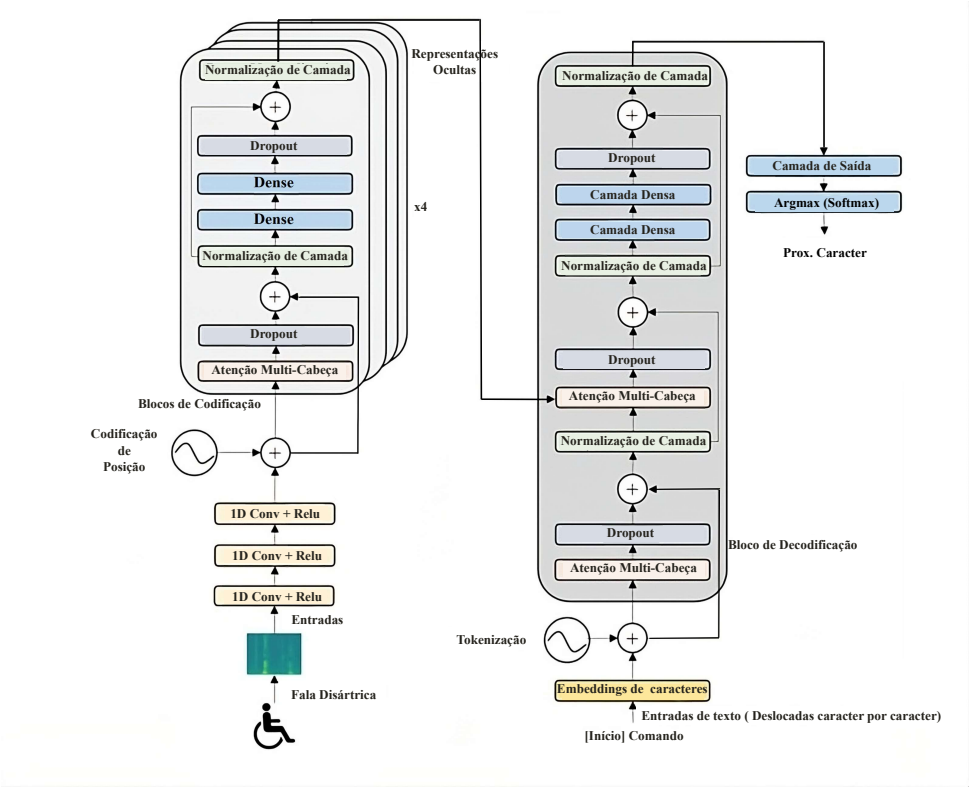
A título de ilustração, a Tabela 3 ilustra de forma prática o funcionamento do processo de decodificação. Nas primeiras épocas de treinamento, as hipóteses geradas pelo *beam search*[1] tendem a apresentar sequências desconexas ou truncadas, evidenciando a elevada incerteza do modelo nesse estágio inicial. À medida que o treinamento avança, tais hipóteses tornam-se mais estáveis e passam a convergir gradualmente para a palavra correta, demonstrando a eficácia do mecanismo de refinamento proporcionado pelo feixe de busca. No caso da decodificação *greedy*, o procedimento segue a mesma lógica de predição passo a passo, porém sem a geração de múltiplas hipóteses alternativas, limitando-se a escolher sempre o símbolo de maior probabilidade em cada iteração.

3.2.5 Nossa Abordagem

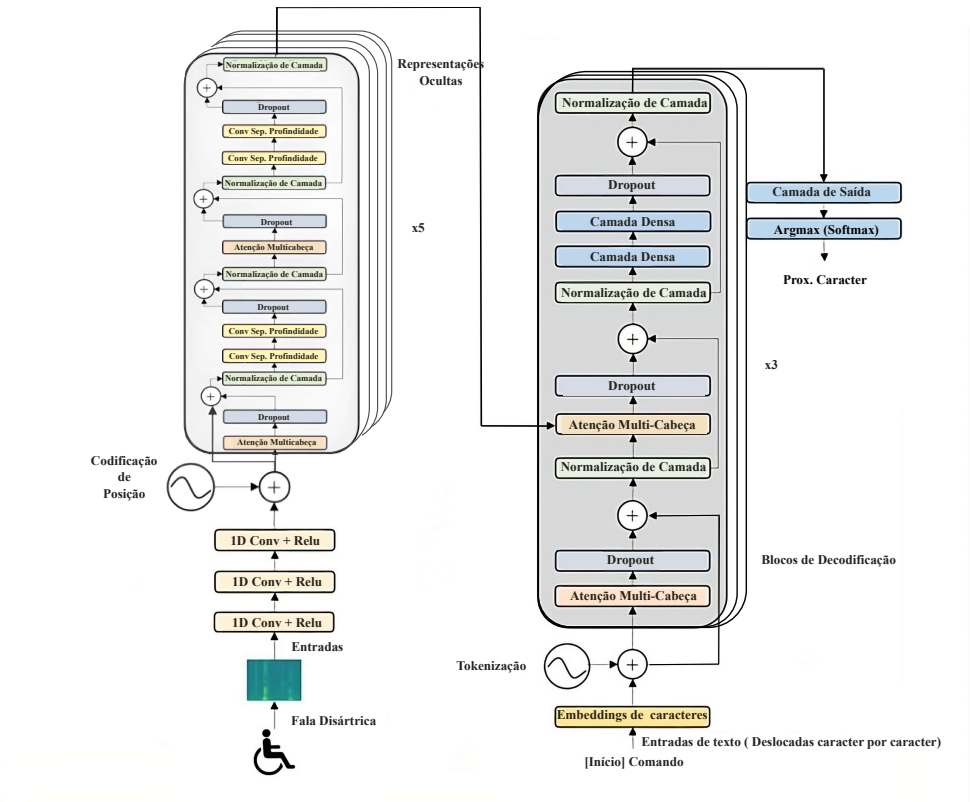
Visão Geral

Inserção de ruído, dilatação temporal e síntese de fala são técnicas frequentemente citadas na literatura como métodos comuns para aumentar dados de áudio (VACHHANI; BHAT; KOPPARAPU, 2018; WEI et al., 2020; SHAHAMIRI, 2021; ABAYOMI-ALLI

[1] O *beam search* é uma estratégia de decodificação que mantém várias hipóteses candidatas em paralelo, refinando-as progressivamente até selecionar a mais provável. Em contraste, o método *greedy decoding* escolhe apenas o símbolo de maior probabilidade a cada passo, sem explorar alternativas.



(a) Arquitetura FW1



(b) Arquitetura FW2

Figura 12 – Diagramas de Blocos das arquiteturas *Speech-Transformer*. (Adaptado de Shahamiri, Lal e Shah (2023))

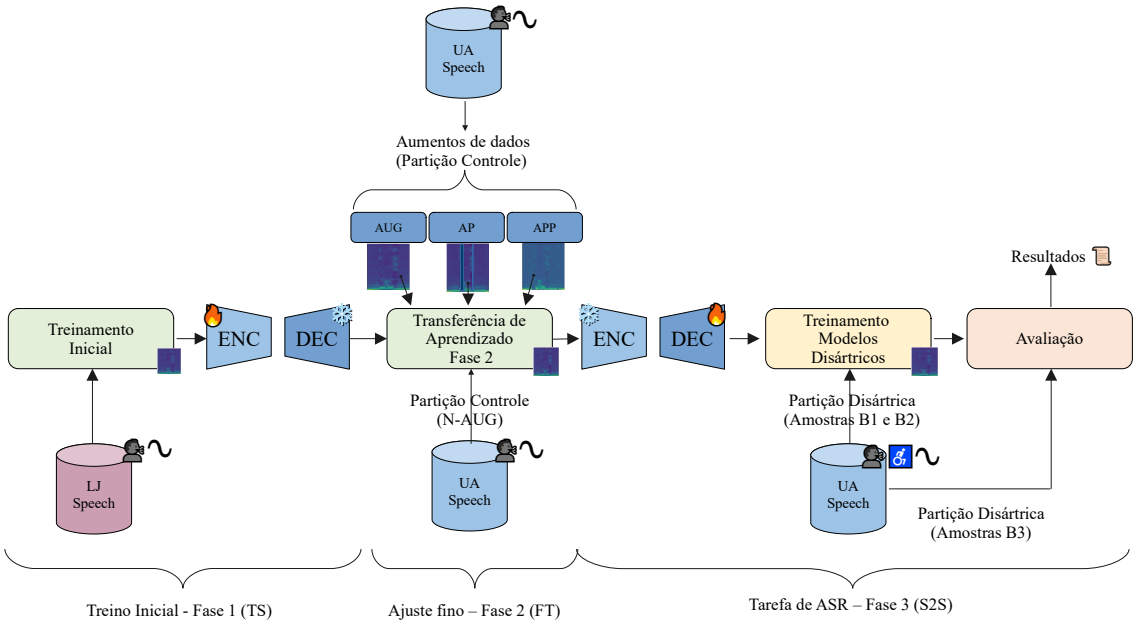


Figura 13 – Visão geral da metodologia supervisionada

Tabela 3 – Exemplos de decodificação em diferentes estágios de treinamento. São exibidas as hipóteses geradas pelo *beam search*, a saída final escolhida e a saída correspondente em *greedy decoding*.

Época	**Hipóteses do beam search (parciais)**	**Saída final (beam)**	**Greedy**	**Groundtruth**
Inicial (época 1)	“bseinerfdSe...”; “beiner...”; “bsen...”	“bsen<eos>”	“bs...” (sequência truncada)	“from”
Avançada (época 100)	“d” → “di” → “did”; “di” → “die”; “did<eos>”	“did<eos>”	“did” (sem hipóteses alternativas)	“did”

et al., 2022). A inserção de ruído é uma técnica prevalente devido à sua simplicidade e eficácia. Ao adicionar ruído aleatório ao sinal de áudio original, os modelos podem aprender a focar nas características essenciais do sinal de áudio e ignorar o ruído, melhorando assim sua robustez em ambientes ruidosos do mundo real. A dilatação temporal, outro método comumente usado, consiste em alterar a velocidade do sinal de áudio sem mudar seu tom. Este método é particularmente útil em tarefas como reconhecimento de fala, onde a mesma palavra pode ser falada em diferentes velocidades por diferentes falantes. A síntese de fala é um método mais complexo, mas altamente eficaz. Envolve a geração artificial de sinais de fala, o que pode aumentar significativamente a quantidade de dados de treinamento disponíveis, especialmente em cenários onde os dados de fala reais são limitados.

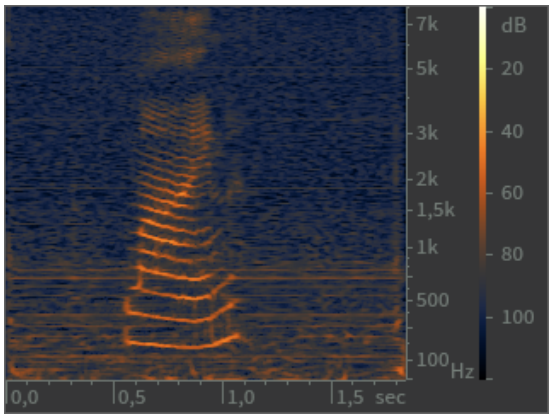
Essas técnicas não são apenas recorrentes na literatura, mas também se mostraram eficazes em várias tarefas de aprendizado de máquina relacionadas ao áudio. Elas são frequentemente usadas em combinação para criar um conjunto de dados de treinamento

diversificado e robusto, levando a melhorias significativas no desempenho do modelo. No entanto, é importante notar que a escolha das técnicas de aumento de dados pode depender da tarefa específica e das características dos dados disponíveis. Portanto, embora esses métodos sejam comuns, eles não são os únicos, e novas técnicas estão constantemente sendo desenvolvidas e exploradas neste campo em constante evolução.

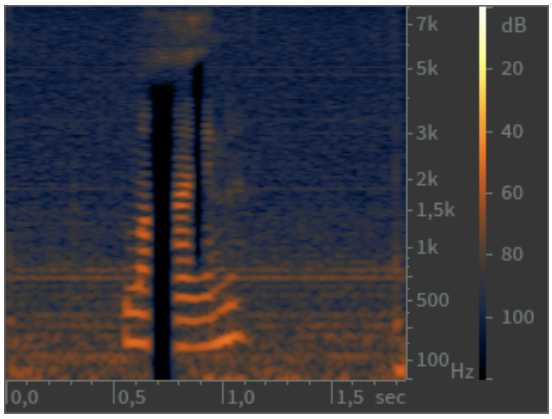
Dessa forma, dada a especificidade da fala disártrica, outras técnicas podem melhorar o desempenho dos sistemas de reconhecimento de fala, e uma possível técnica que desenvolvemos é apresentada a seguir.

Proposta

Para o método proposto, denominado SO (tf_3) e referido neste trabalho como "AP", foi adotado um sistema de mascaramento de seções aleatórias do espectro de fala para simular deficiências durante a articulação de certos fonemas. Além disso, o método proposto foi combinado aos demais métodos de transformação de dados, resultando na configuração denominada "APP". Esses métodos foram aplicados aos dados de controle do banco de dados UA-Speech. A Figura 14 ilustra um exemplo antes e depois das transformações propostas.



(a) Espectrograma nativo



(b) Espectrograma transformado

Figura 14 – Espectrograma antes e depois das transformações propostas.

A adoção desse método de transformação é justificada pelo estudo das dinâmicas espectrais da fala disártrica em comparação com a fala saudável. Na fala disártrica, podem ser observadas prolongações de alguns fonemas e lacunas espectrais causadas pela dificuldade de articulação perfeita durante a ligação dos átomos sonoros. Além disso, em alguns casos, surgem padrões de ondas que caracterizam instabilidades na emissão da fala, isto é, zonas de intermitência na pressão sonora do áudio, associadas a deficiências no controle dos órgãos fonatórios, frequentemente presentes em pacientes disártricos, como mostrado na Figura 15.

Formulação do método

Para viabilizar a aplicação controlada de oclusões em regiões acusticamente salientes, partimos da representação tempo–frequência do sinal e da respectiva distribuição de energia. Denotamos por $x[n]$ o sinal discreto e por $S(k, m)$ sua STFT, calculada com janela $w[\cdot]$ e avanço H, adotando k como índice espectral e m como índice temporal. A partir do módulo ao quadrado, obtemos o mapa de energia $E(k, m)$, que serve de base para selecionar, de forma estatisticamente informada, as áreas a serem ocluídas e fundamenta a formalização do método.

Seja o sinal discreto $x[n]$ e sua STFT:

$$S(k, m) = \sum_{n=0}^{N-1} x[n]\, w[n - mH]\, e^{-j2\pi kn/N},$$

onde $w[\cdot]$ é a janela de análise, H é o *hop size*, k é o índice de frequência e m o índice temporal. A energia espectral é

$$E(k, m) = |S(k, m)|^2.$$

Define-se a energia total como $\mathcal{E}_{\text{tot}} = \sum_{k,m} E(k, m)$. Para um parâmetro $\rho \in (0, 1]$, escolhe-se um limiar τ tal que o conjunto

$$\Omega_\tau = \{(k, m) : E(k, m) \geq \tau\} \quad \text{satisfaça} \quad \sum_{(k,m)\in\Omega_\tau} E(k, m) \geq \rho\, \mathcal{E}_{\text{tot}}.$$

O trecho de maior energia $\mathcal{B}$ é definido como o retângulo mínimo que envolve Ω_τ, com dimensões $H_\mathcal{B} \times W_\mathcal{B}$.

Sobre $\mathcal{B}$, são aplicadas até duas janelas de oclusão ($M \in \{1, 2\}$), cada uma com dimensões (h_i, w_i) amostradas uniformemente como

$$h_i \sim \mathcal{U}(1, \lfloor 0.2\, H_\mathcal{B} \rfloor), \quad w_i \sim \mathcal{U}(1, \lfloor 0.2\, W_\mathcal{B} \rfloor).$$

A posição (k_i, m_i) do canto superior esquerdo de cada janela é escolhida aleatoriamente dentro de $\mathcal{B}$, com probabilidade proporcional à energia local:

$$\mathbb{P}((k, m)) = \frac{E(k, m)\, \mathbf{1}_{(k,m)\in\mathcal{B}}}{\sum_{(u,v)\in\mathcal{B}} E(u, v)}.$$

Por fim, os coeficientes de S são zerados nos retângulos sorteados:

$$\tilde{S}(k,m) = \begin{cases} 0, & \text{se } (k,m) \in \bigcup_{i=1}^{M} R_i, \\ S(k,m), & \text{caso contrário.} \end{cases}$$

Pseudo-código da transformação

Com base na formulação anterior, a Tabela 4 apresenta a instância procedimental usada nos experimentos, detalhando entradas, saídas e passos do algoritmo. O procedimento recebe $x[n]$, os parâmetros da STFT e três hiperparâmetros interpretáveis: $\rho \in (0, 1]$, que controla a fração mínima da energia total considerada ao definir o limiar τ; α, que limita as dimensões de cada janela de oclusão como fração do retângulo de maior energia; e $M_{\max}$, que fixa o número máximo de janelas aplicadas. A amostragem das posições é ponderada por $E(k,m)$ para concentrar oclusões em regiões informativas, enquanto o ajuste de contorno assegura que os retângulos permaneçam contidos em $\mathcal{B}$.

Tabela 4 – Pseudocódigo da Oclusão Espectral (SO)

Entrada: $x[n]$, $params_{\text{STFT}}$, $\rho \in (0, 1]$, α, $M_{\max}$
Saída: $\tilde{S}$

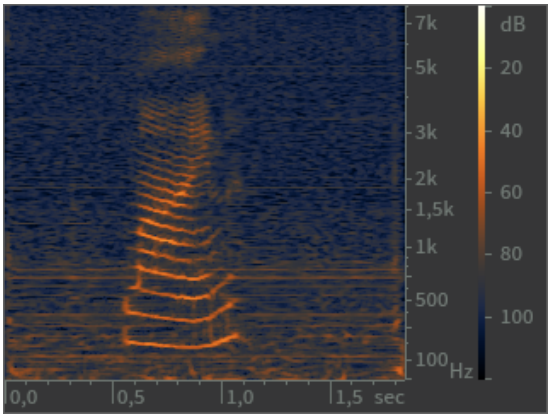
1. $S \leftarrow \text{STFT}(x; params_{\text{STFT}})$
2. $E \leftarrow |S|^2$
3. Escolher τ tal que $\sum_{E \geq \tau} E \geq \rho \sum E$
4. $\Omega_\tau \leftarrow \{(k,m) : E[k,m] \geq \tau\}$
5. $B \leftarrow$ bounding_box(Ω_τ); altura H_B, largura W_B
6. $M \leftarrow \text{randint}(1, M_{\max})$
7. Para $i = 1..M$:
   $h_i \leftarrow \text{randint}(1, \lfloor \alpha H_B \rfloor)$, $\quad w_i \leftarrow \text{randint}(1, \lfloor \alpha W_B \rfloor)$
   $(k_0, m_0) \leftarrow$ amostragem_ponderada($E|_B$); ajustar para caber em B
   $S[k_0 : k_0 + h_i,\ m_0 : m_0 + w_i] \leftarrow 0$
8. $\tilde{S} \leftarrow S$; **return** $\tilde{S}$

Fonte: Elaborado pelo autor (2025).

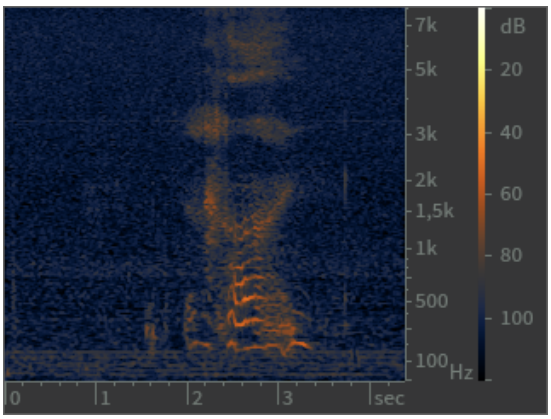
Disponibilidade do código e dados

Com vistas à reprodutibilidade científica, todo o código desenvolvido neste trabalho, bem como instruções de execução e organização dos dados, está disponível em repositório público no GitHub[2]. Esse repositório contém as implementações correspondentes às etapas metodológicas descritas neste capítulo, abrangendo desde o pré-processamento até as estratégias de aumento de dados e treinamento supervisionado.

[2] <https://ragracelli.github.io/CONTRASTIVE_DASR/>



(a) grupo de controle

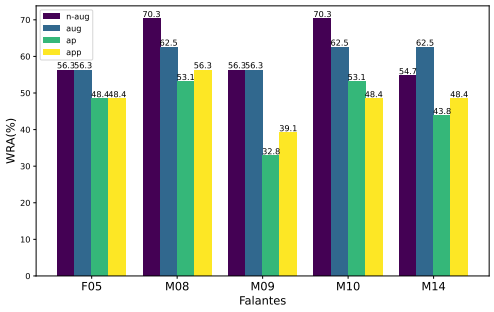


(b) grupo disártrico

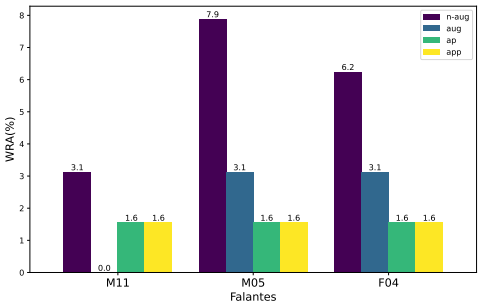
Figura 15 – Amostra espectral da palavra “Line” reproduzida por diferentes grupos.

3.3 Experimentos e Discussão dos Resultados

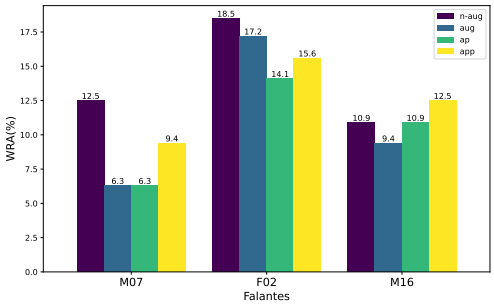
Os experimentos foram inicialmente conduzidos dividindo o conjunto de dados LJSpeech em uma proporção de 99:1 e treinados por 200 épocas. Os melhores resultados, determinados com base na menor perda de validação, foram salvos e usados como pesos pré-treinados para a próxima etapa. A etapa subsequente envolveu o ajuste fino na partição de controle do conjunto de dados UA-Speech, que foi dividido em conjuntos de treinamento, validação e teste para validação cruzada. A proporção foi definida em 80:20, após separação dos dados marcados com a chave B3 para os testes. Durante esta etapa, os decodificadores foram congelados e o treinamento persistiu por 100 épocas. Após a conclusão da fase de treinamento e decodificação, os conjuntos de teste alcançaram uma taxa de reconhecimento de palavras de 95%. Após a conclusão desta etapa, os melhores pesos também foram salvos e usados nas fases de teste subsequentes aplicadas aos conjuntos de dados disártricos. Os dados foram divididos em conjuntos de treinamento, validação e teste, como é comumente feito em estudos semelhantes. Os arquivos nomeados B1 e B2 foram usados para compor os conjuntos de treinamento e validação. Por outro lado, os arquivos identificados como B3 foram usados para formar o conjunto de teste. Esta metodologia de divisão de dados, associada ao banco de dados UA-Speech, é semelhante



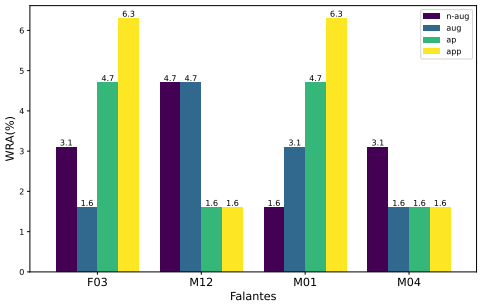
(a) - WRA - Alta Intelig. - FW1



(b) - WRA - Moderada Intelig. - FW1

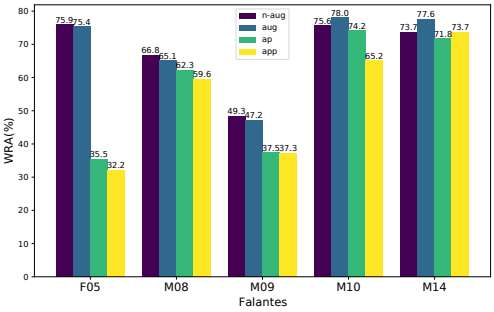


(c) - WRA - Baixa Intelig. - FW1

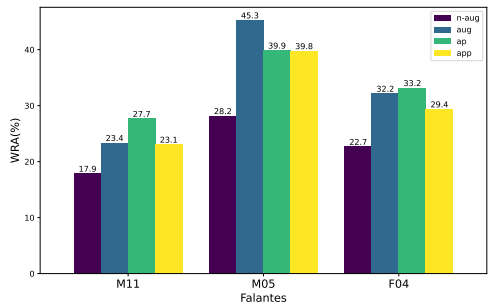


(d) - WRA - Muito Baixa Intelig. - FW1

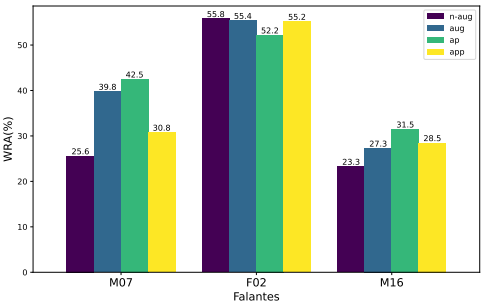
Figura 16 – Comparativos em WRA dos métodos adotados no modelo FW1. N-AUG: (tf_0); AUG: $(tf_1 + tf_2)$; AP: (tf_3); APP: $(tf_1 + tf_2 + tf_3)$.



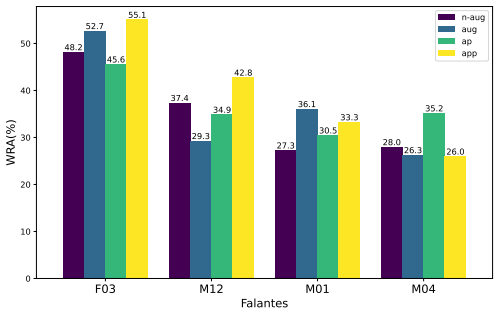
(a) - WRA - Alta Intelig. - FW2



(b) - WRA - Moderada Intelig. - FW2



(c) - WRA - Baixa Intelig. - FW2



(d) - WRA - Muito Baixa Intelig. - FW2

Figura 17 – Comparativos em WRA dos métodos adotados no modelo FW2. N-AUG: (tf_0); AUG: $(tf_1 + tf_2)$; AP: (tf_3); APP: $(tf_1 + tf_2 + tf_3)$.

às adotadas e discutidas em Almadhor et al. (2023) e Shahamiri, Lal e Shah (2023).

O treinamento utilizou o otimizador Adam, com parâmetros $\beta_1 = 0{,}8$ e $\beta_2 = 0{,}9$, associado a uma política de taxa de aprendizado dinâmica. A taxa inicial foi definida como 1×10^{-5}, com aquecimento linear até 1×10^{-3} ao longo das primeiras 15 épocas. Após essa fase, foi aplicado um decaimento linear até retornar a 1×10^{-5} ao longo de 85 épocas. Essa configuração favoreceu a estabilidade na fase inicial do treinamento e evitou sobreajuste nas fases finais, conforme já demonstrado em arquiteturas Transformer para ASR.

Para cada falante disártrico, os conjuntos de dados foram testados aplicando o pré-treino da base normotípica em seu estado natural (tf_0), depois com ruído de fundo e alongamento temporal adicionados ($tf_1 + tf_2$), seguido pela aplicação da transformação proposta (tf_3), e finalmente adicionando ruído de fundo e alongamento temporal ao SO ($tf_1 + tf_2 + tf_3$). Ademais, conforme proposto por Shahamiri, Lal e Shah (2023), foram estabelecidas rotinas de congelamento direcionadas a blocos específicos da arquitetura, conforme apresentado na Tabela 5.

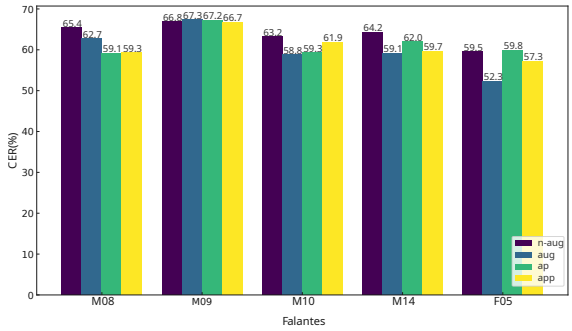
Tabela 5 – Rotina de congelamento adotada para as arquiteturas FW1 e FW2 conforme o nível de inteligibilidade dos falantes disártricos

Nível de Inteligibilidade	**Falantes Disártricos**	**FW1 (sem aumento de dados)**	**FW1 (com aumento de dados)**	**FW2 (com aumento de dados)**
Muito Baixo	M04, F03, M12, M01	Decodificador	Decodificador	Último decodificador
Baixo	M07, F02, M16	Último componente feed-forward do codificador	Último componente feed-forward do codificador	Último decodificador
Moderado	M05, M11, F04	Componente feed-forward do decodificador	Decodificador	Componente feed-forward do último decodificador
Alto	M09, M14, M10, M08, F05	Componente feed-forward do decodificador	Segunda camada densa do decodificador	Componente feed-forward do último decodificador

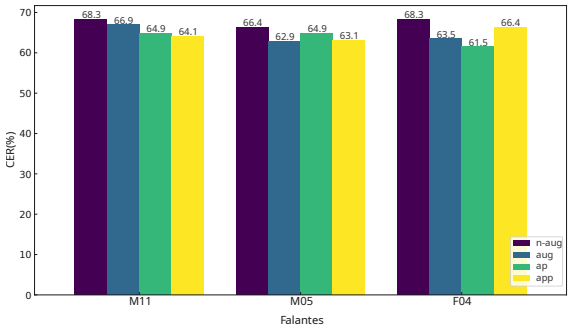
Os resultados em WRAdos testes fornecidos pela arquitetura *FW1* são ilustrados na Figura 16: (a), (b), (c) e (d). Os resultados alcançados através da arquitetura *FW2*, por outro lado, são representados na Figura 17: (a), (b), (c) e (d), também expressos em termos de WRA.

Ainda, os resultados foram aferidos em termos de CER e WER cujos valores são, respectivamente, apresentados nas Figuras 18 a 21, também subdivididas em termos de nível de inteligibilidade.

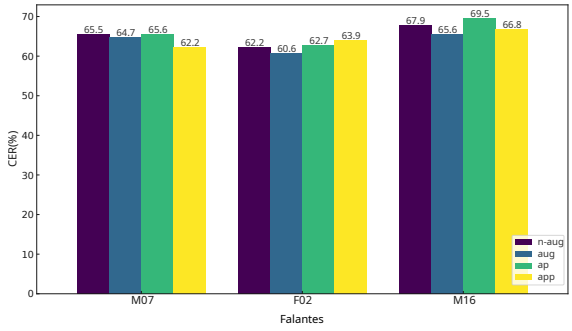
Exploratoriamente, os testes em WRA demonstraram que, para a fala disártrica com uma maior inteligibilidade, como F05, a ampliação de dados, independentemente do método empregado, prejudicou o desempenho do reconhecimento. Por outro lado, à medida que a inteligibilidade da fala em teste diminuía, a eficácia dos métodos de ampliação de dados testados apresentava melhora para alguns falantes.



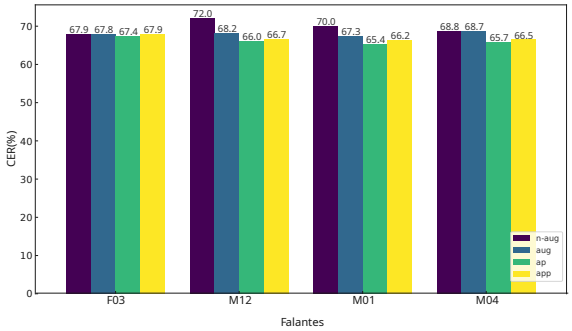
(a) - CER - Alta Intelig. - FW1



(b) - CER - Moderada Intelig. - FW1

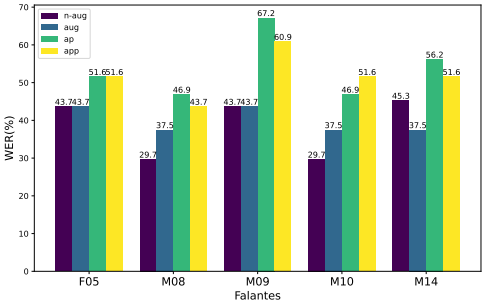


(c) - CER - Baixa Intelig. - FW1

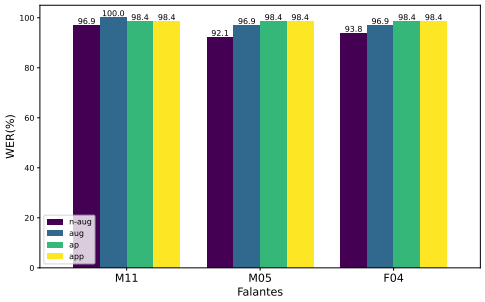


(d) - CER - Muito Baixa Intelig. - FW1

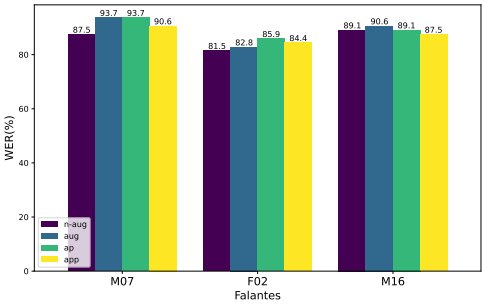
Figura 18 – Comparativos em CER, dos métodos adotados no modelo FW1. N-AUG: (tf_0); AUG: $(tf_1 + tf_2)$; AP: (tf_3); APP: $(tf_1 + tf_2 + tf_3)$.



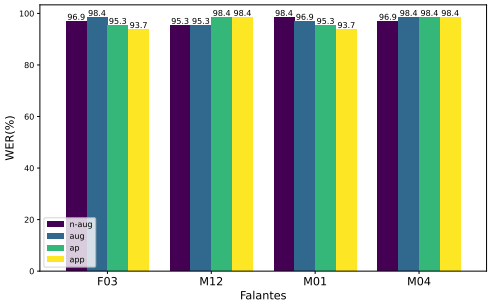
(a) - WER - Alta Intelig. - FW1



(b) - WER - Moderada Intelig. - FW1

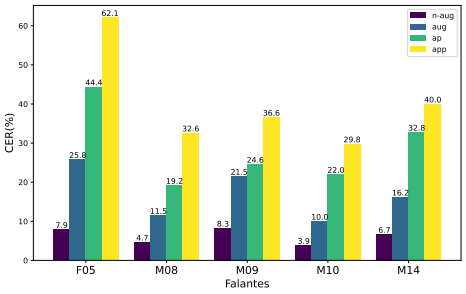


(c) - WER - Baixa Intelig. - FW1

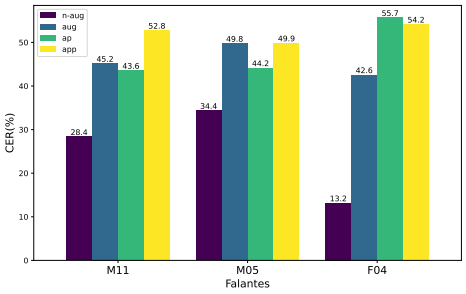


(d) - WER - Muito Baixa Intelig. - FW1

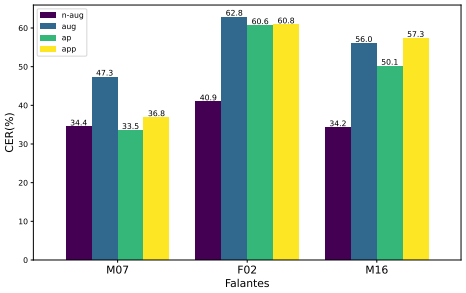
Figura 19 – Comparativos em WER, dos métodos adotados no modelo FW1. N-AUG: (tf_0); AUG: $(tf_1 + tf_2)$; AP: (tf_3); APP: $(tf_1 + tf_2 + tf_3)$.



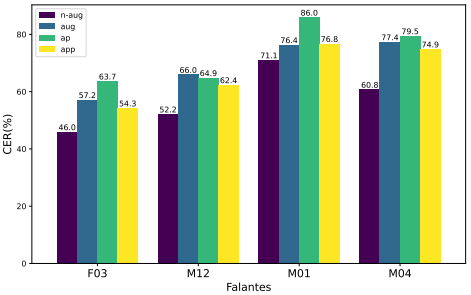
(a) - CER - Alta Intelig. - FW2



(b) - CER - Moderada Intelig. - FW2

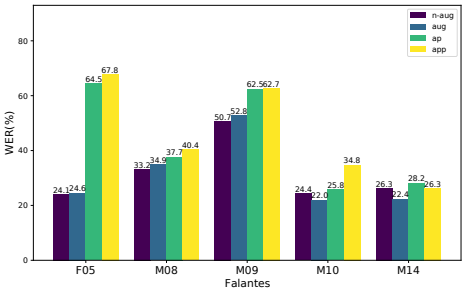


(c) - CER - Baixa Intelig. - FW2

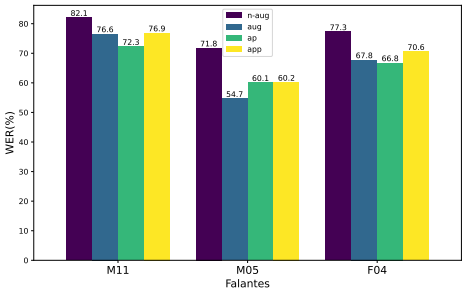


(d) - CER - Muito Baixa Intelig. - FW2

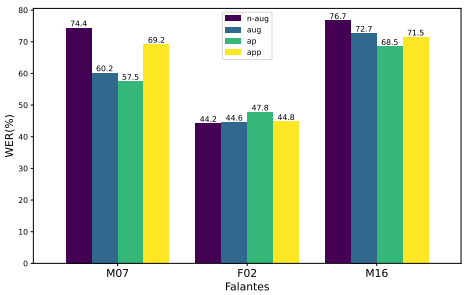
Figura 20 – Comparativos em CER, dos métodos adotados no modelo FW2. N-AUG: (tf_0); AUG: $(tf_1 + tf_2)$; AP: (tf_3); APP: $(tf_1 + tf_2 + tf_3)$.



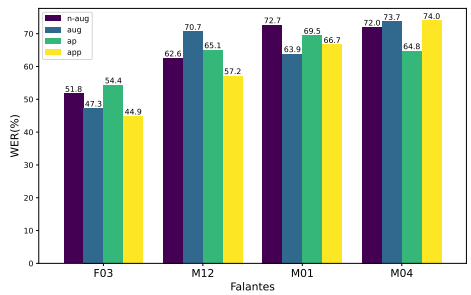
(a) - WER - Alta Intelig. - FW2



(b) - WER Moderada Intelig. - FW2



(c) - WER - Baixa Intelig. - FW2



(d) - WER - Muito Baixa Intelig. - FW2

Figura 21 – Comparativos em WER, dos métodos adotados no modelo FW2. N-AUG: (tf_0); AUG: $(tf_1 + tf_2)$; AP: (tf_3); APP: $(tf_1 + tf_2 + tf_3)$.

Embora os gráficos de WRA tenham evidenciado, para alguns pacientes, ganhos relevantes com as técnicas de aumento de dados, é fundamental considerar também as métricas de erro CER e WER para uma avaliação mais precisa da inteligibilidade automática. Diferentemente do WRA, que mede acertos, WER e CER quantificam os erros de transcrição, oferecendo uma visão mais crítica do desempenho do sistema.

Enquanto a arquitetura *FW1* apresentou ganhos modestos ou mesmo deterioração de desempenho com o uso de "APP", excetuando-se os falantes M16, F03 e M01, a *FW2* se destacou com reduções relevantes em WER, especialmente quando a técnica combinada foi aplicada a falantes com fala mais comprometida.

No reconhecimento fonêmico, medido pelo CER, a arquitetura *FW1* apresentou melhor desempenho, com reduções consistentes nos índices de erro em alguns falantes após a aplicação das técnicas de aumento. Em contraste, a arquitetura *FW2* mostrou-se sensível às transformações, resultando em aumento generalizado do CER, especialmente nas falas com maior inteligibilidade. A exceção foi o falante M07, cuja performance com dados aumentados foi comparável ao cenário sem aumento ("N-AUG"). Esses resultados indicam que, na *FW2*, as transformações aplicadas comprometeram a precisão fonêmica, possivelmente ao introduzir variações que dificultam a correspondência entre entrada e saída. Ainda assim, entre as técnicas testadas, a combinação completa ("APP") mostrou-se a mais eficaz para falas severamente comprometidas, ao passo que, para falas mais próximas do padrão normotípico, o método "AUG" isolado se mostrou mais adequado. Logo, a aplicação das técnicas de aumento propostas sofrem impactos negativos dependendo da profundidade da arquitetura utilizada.

Cabe destacar que a métrica CER se mostrou particularmente sensível em falas disártricas com palavras mais longas. Como a base de dados utilizada é composta por palavras isoladas, aquelas com maior número de caracteres tendem a amplificar os efeitos das distorções fonêmicas características da disartria severa. Além disso, enquanto o WER considera a palavra como unidade de acerto ou erro, o CER penaliza proporcionalmente cada caractere incorreto, o que resulta em valores mais elevados mesmo em casos com erros mínimos de articulação. Essa diferença estrutural entre as métricas ajuda a explicar por que, em alguns falantes, observou-se um WRA razoável acompanhado de CERs significativamente altos.

Resumo Estatístico dos Resultados

Embora os gráficos apresentados neste capítulo detalhem o comportamento individual de cada falante e técnica de aumento, é relevante sintetizar tendências globais. De forma agregada, observou-se que aproximadamente dois terços dos falantes obtiveram ganhos em inteligibilidade (WRA) com pelo menos uma das estratégias de transformação avaliadas. No modelo *FW1*, de arquitetura mais simples, a técnica combinada ("APP") mostrou-se a mais abrangente, sendo a melhor escolha para cerca de 40% dos falantes, seguida pela

Oclusão Espectral ("AP"), vencedora isolada em 20% dos casos. No modelo *FW2*, mais profundo, a Oclusão Espectral destacou-se em 27% dos falantes, enquanto a combinação "APP" beneficiou 20%.

Os maiores ganhos absolutos foram concentrados nos grupos de inteligibilidade baixa e muito baixa, com reduções de WER superiores a 20 pontos percentuais em alguns pacientes. Em contrapartida, falantes quase normotípicos apresentaram melhorias discretas ou mesmo degradação, sugerindo que a eficácia das transformações depende fortemente do grau de comprometimento articulatório. Esses achados corroboram a hipótese de que métodos de aumento de dados são especialmente vantajosos em cenários severos de disartria, enquanto sua aplicação em vozes de alta inteligibilidade deve ser mais cautelosa.

Para consolidar de forma quantitativa os resultados discutidos, a Tabela 6 apresenta um resumo agregado da proporção de falantes que obtiveram melhorias em cada técnica de aumento de dados, considerando separadamente as arquiteturas *FW1* e *FW2*. Esse panorama evidencia tendências gerais: a SO destacou-se em falantes de baixa inteligibilidade, a combinação de transformações ("APP") mostrou maior abrangência em *FW1*, e, por outro lado, falantes quase normotípicos apresentaram ganhos discretos ou mesmo degradação.

Tabela 6 – Resumo agregado dos melhores desempenhos por técnica de aumento de dados, considerando o percentual de falantes beneficiados em cada arquitetura.

Técnica	**FW1**	**FW2**	**Tendência Geral**
Ruído + Dilatação (AUG)	15%	15%	Ganhos moderados, sobretudo em falantes de inteligibilidade média.
Oclusão Espectral (AP)	20%	27%	Destaque em falantes de baixa inteligibilidade.
Combinação (APP)	40%	20%	Melhor abrangência em FW1, efeito estável em FW2.
Nenhuma melhoria	25%	20%	Sobreaquecimento em falantes quase normotípicos.

3.4 Considerações finais

Neste capítulo, foram conduzidos experimentos com diferentes estratégias de transformação de dados aplicadas à fala disártrica, com o objetivo de simular deficiências articulatórias de maneira controlada e avaliar seu impacto sobre modelos ASR. As análises permitiram responder às questões levantadas na introdução deste trabalho, especialmente no que se refere à eficácia dos métodos tradicionais e da técnica proposta de SO no reconhecimento automático de fala disártrica.

Os resultados obtidos confirmam a hipótese de que métodos tradicionais de aumento de dados, como a adição de ruído branco e a dilatação temporal, contribuem positivamente para o desempenho dos modelos, especialmente em falantes classificados com inteligibilidade baixa e muito baixa. Além disso, foi possível verificar que a técnica de SO ("AP"), desenvolvida neste estudo, oferece vantagens adicionais ao simular omissões e distorções fonêmicas presentes na fala disártrica severa. Dessa forma, a hipótese inicial

sobre a eficácia dos métodos de aumento de dados para a melhoria do reconhecimento de fala disártrica encontram respaldo nos achados experimentais.

Também se confirmou que a combinação dessas técnicas de transformação ("APP") pode potencializar os ganhos do modelo em determinados contextos. No entanto, observou-se que tal combinação deve ser aplicada com cautela: enquanto é eficaz em falas altamente comprometidas, ela pode causar perda de desempenho em falantes com fala mais próxima do padrão normotípico. Assim, responde-se também à uma questão natural e implícita da pesquisa, indicando que a combinação de transformações não é universalmente benéfica e que sua eficácia depende do grau de inteligibilidade da fala original. Nas falas de alta inteligibilidade, a aplicação das transformações "APP" introduz distorções fonéticas que reduzem a representatividade dos dados em relação ao padrão original, resultando em perda de desempenho do modelo devido à ampliação desnecessária da variabilidade.

Entre as arquiteturas testadas, de modo geral, a *FW2* apresentou o melhor desempenho absoluto em todas as métricas avaliadas (WRA, WER e CER), confirmando seu potencial em contextos com maior volume e diversidade de dados. A arquitetura *FW1*, embora mais simples, também se beneficiou das transformações, mas os ganhos foram menos consistentes e mais dependentes do tipo de falante. Isso reforça que arquiteturas profundas tendem a absorver melhor as variações introduzidas, embora a escolha da técnica de transformação continue sendo um fator decisivo.

Esses resultados sustentam a hipótese de que não basta adicionar dados ao treinamento; é necessário garantir que esses dados sejam informativos e representem adequadamente as variações reais da fala disártrica. As evidências indicam que o impacto das transformações depende diretamente do perfil fonético dos falantes, da severidade da disartria e da compatibilidade entre os dados transformados e a arquitetura do modelo utilizado. A análise conjunta das métricas WRA/WER e CER revelou que as maiores reduções de erro foram observadas em alguns casos severos, como M16. No entanto, outros falantes com inteligibilidade muito baixa, como F03 e M04, mantiveram taxas de erro elevadas, indicando que o benefício das técnicas não é uniforme entre todos os perfis severos. As particularidades fonéticas desses falantes podem ter limitado a eficácia das transformações, indicando que a severidade não é o único fator determinante para o sucesso das técnicas.

Adicionalmente, os resultados indicam que a métrica CER pode apresentar maior sensibilidade a distorções fonêmicas em comparação ao WRA/WER, sobretudo em contextos com palavras mais longas. Como as bases empregadas são compostas por palavras isoladas, erros na articulação de fonemas — mesmo que pontuais — impactam diretamente a taxa de caracteres corretos, elevando o CER mesmo quando a palavra é reconhecida parcialmente. Isso sugere que a interpretação dos resultados deve considerar o comprimento médio das palavras testadas e a natureza fonética dos erros disártricos, principalmente em falantes com severidade acentuada.

A principal vantagem das técnicas aqui propostas reside na sua viabilidade prática. A implementação direta das transformações no código-fonte elimina a necessidade de infraestrutura especializada, como novos pacientes, microfones ou ambientes de gravação controlados. Tal abordagem se mostra especialmente útil em contextos clínicos ou de pesquisa com restrições orçamentárias, nos quais a coleta de novos dados não é viável.

Por fim, os resultados deste capítulo apontam para direções promissoras de pesquisa futura. A exploração de diferentes tamanhos de janelas de oclusão espectral, métodos alternativos de preenchimento (como valores médios ou ruído sintético), e outras métricas de distorção perceptual pode aprimorar ainda mais a eficácia dessas transformações — especialmente em falas com alta inteligibilidade, onde o uso inadequado de aumento de dados pode levar à degradação do desempenho.

Capítulo 4

Métodos Autossupervisionados de ASR para Fala Disártrica

4.1 Introdução e definição do problema.

Nesta seção, será apresentada a proposta da pesquisa, que tem como objetivo principal utilizar os tipos de dados obtidos a partir das transformações descritas no Capítulo 3, os quais servirão como base para o treinamento de uma rede neural autossupervisionada, batizada de Reconhecimento Automático de Fala Disártrica com Aprendizado Contrastivo (do inglês: *Contrastive Automatic Dysarthric Speech Recognition*) (CADSR). A finalidade desse treinamento é reconhecer a fala disártrica, buscando alcançar uma boa acurácia no processo.

A pesquisa se concentra em explorar técnicas avançadas de aprendizado de máquina para melhorar a qualidade da fala de pessoas com disartria. Ao aplicar os dados transformados no treinamento da rede neural, espera-se desenvolver um modelo capaz de ajustar e corrigir as imperfeições na fala disártrica, buscando resultados comparáveis aos obtidos no estado da arte.

A motivação para esta abordagem surge de uma limitação amplamente reconhecida na literatura: a escassez de bases de dados disártricos, tanto em volume quanto em diversidade de falantes e condições fonéticas. Segundo Hu, Phadnis e Shahamiri (2023), Leung, Li e Li (2024) e Jin et al. (2022) essa limitação compromete significativamente a capacidade dos sistemas de ASR de generalizar para diferentes perfis de disartria. Diante desse cenário, optou-se por iniciar a pesquisa com o estudo apresentado no Capítulo 3, focado em técnicas de aumento de dados para simular distorções características da fala disártrica. A aplicação controlada de transformações como adição de ruído, dilatação

temporal e SO buscou enriquecer a base original e avaliar os efeitos dessas simulações na performance dos modelos ASR supervisionados.

Com base nos resultados desse primeiro estudo, que indicaram que transformações bem direcionadas podem beneficiar substancialmente o desempenho dos modelos — especialmente em falantes com disartria severa —, delineou-se a segunda parte da pesquisa. Nesta etapa, investiga-se o uso de aprendizado autossupervisionado, por meio de técnicas contrastivas, como forma de lidar com a limitação de dados rotulados. A proposta é utilizar uma base normotípica abundante e aplicar sobre ela as mesmas transformações desenvolvidas para simular fala disártrica, permitindo ao modelo contrastar representações típicas e atípicas da fala.

Essa estratégia estabelece uma conexão direta entre os dois blocos experimentais da pesquisa: o primeiro, voltado à simulação e avaliação supervisionada da fala disártrica; o segundo, centrado na exploração de representações latentes aprendidas a partir de contrastes entre fala típica e fala artificialmente alterada. O objetivo final é proporcionar uma ferramenta eficaz que possa ser utilizada em terapias fonoaudiológicas e sistemas de reconhecimento de fala, melhorando a qualidade de vida dos indivíduos com disartria e facilitando sua integração social.

4.2 Trabalhos relacionados

O artigo de Ravanelli et al. (2020) apresenta uma abordagem inovadora para o reconhecimento de fala, utilizando um modelo de SSL chamado Versão Aprimorada do Codificador de Fala Agnóstico ao Problema (do inglês: *Problem Agnostic Speech Encoder Plus*) (PASE+). O PASE+ é uma versão aprimorada do Codificador de Fala Agnóstico ao Problema (do inglês: *Problem Agnostic Speech Encoder*) (PASE) original, que combina um codificador convolucional com múltiplas redes neurais, conhecidas como trabalhadores, que resolvem tarefas autossupervisionadas. Essa arquitetura é projetada para capturar informações relevantes da fala, como a impressão de voz do falante e fonemas, enquanto se mantém robusta em ambientes ruidosos e reverberantes. O modelo foi treinado em um conjunto de dados diversificado e demonstrou desempenho superior em comparação com versões anteriores e características acústicas tradicionais, evidenciando sua capacidade de aprender representações transferíveis em condições acústicas desafiadoras.

Além disso, o PASE+ incorpora um módulo de distorção de fala online que contamina os sinais de entrada com uma variedade de distúrbios aleatórios, melhorando a robustez do modelo. O uso de Rede Neural Quase Recorrente (do inglês: *Quasi-Recurrent Neural Network*) (QRNN) permite que o PASE+ aprenda dependências de longo prazo de forma eficiente, enquanto as conexões de atalho melhoram o fluxo de gradiente e a aprendizagem de representações. Os resultados obtidos em conjuntos de dados como Corpus de Fala do Instituto de Tecnologia de Massachusetts (do inglês: *Texas Instruments/Massachusetts*

Institute of Technology) (TIMIT), Ambiente Doméstico para Reconhecimento de Fala em Ambientes Reais (do inglês: *Distant-speech Interaction for Robust Home Applications*) (DIRHA) e Audição Computacional em Ambientes com Múltiplas Fontes – 5ª Edição (do inglês: *Computational Hearing in Multisource Environments – 5th Edition*) (CHiME-5) demonstram que o PASE+ não apenas supera seu predecessor, mas também se destaca em condições acústicas altamente desafiadoras, mostrando o potencial das técnicas de SSL no desenvolvimento de sistemas de reconhecimento de fala mais robustos.

Hu et al. (2024) investigaram o uso de representações SSL, como Wav2Vec 2.0 e HuBERT, integradas a modelos ASR baseados em Rede Neural com Atraso Temporal (do inglês: *Time-Delay Neural Network*) (TDNN) e *Conformer*, demonstrando reduções significativas em WER e CER ao combinar *embeddings* adaptados ao domínio com *front-ends* acústicos tradicionais.

Cadet et al. (2024) destacaram a importância da classificação da disartria para o diagnóstico e tratamento eficazes. No entanto, a obtenção de amostras de fala disártrica foi desafiadora devido ao número limitado de falantes disponíveis em conjuntos de dados. Este estudo investigou o uso de SSL em diferentes avaliações disártricas, propondo uma ferramenta para avaliar representações extraídas de modelos SSL em cenários de ruído variados. As abordagens de avaliação de inteligibilidade poderiam ser baseadas em referências ou não, sendo que as primeiras utilizaram dados de fala saudável para determinar características de fala inteligível.

A pesquisa de Cadet et al. (2024) revelou que ASRs treinados apenas com fala saudável apresentaram desempenho inferior em fala disártrica, especialmente à medida que a severidade da disartria aumentava. Os falantes foram categorizados com base na severidade da disartria, e foram extraídas características acústicas e representações de modelos SSL como HuBERT e wav2vec2. As tarefas de classificação incluíram a classificação de disartria, palavras e inteligibilidade, utilizando modelos de classificação como Regressão Logística e Perceptron de Múltiplas Camadas. Os resultados mostraram que modelos treinados com representações autossupervisionadas superaram aqueles baseados em características acústicas em todas as tarefas de avaliação disártrica.

Técnicas de *data augmentation* também têm desempenhado um papel importante. Soleymanpour et al. (2023) propuseram a síntese de fala disártrica utilizando modelos TTS neurais multi-falantes com coeficientes de severidade e modelagem de pausas. Os experimentos com o conjunto TORGO mostraram que a inclusão de fala sintética no treinamento melhora a acurácia do reconhecimento, reforçando a necessidade de estratégias de aumento de dados no campo.

Abordagens baseadas em aprendizado contrastivo têm ganhado destaque recentemente. Keshvari et al. (2024) aplicaram perda contrastiva em *embeddings* de locutor extraídos do Modelo Multilíngue de Representação de Fala Autossupervisionado (do inglês: *Cross-Lingual Speech Representations*) (XLS-R), com o objetivo de melhorar a inteligibi-

lidade preservando a identidade do falante. A técnica demonstrou reduções significativas em WER para amostras de fala disártrica moderada e severa.

Stumpf et al. (2024) abordaram o problema da classificação de severidade da disartria de forma independente do falante, utilizando *embeddings* do Wav2Vec 2.0 e um classificador baseado em *Transformer*. A proposta chamada Regularização Latente Independente de Falante (do inglês: *Speaker-Adversarial Latent Regularization*) obteve 70,48% de acurácia, um novo benchmark para a tarefa.

Wang et al. (2024) propuseram o método Reconhecimento de Fala Baseado em Protótipos (do inglês: *Prototype-Based Dysarthric Speech Recognition*), combinando *embeddings* do HuBERT com aprendizado Aprendizado Contrastivo Supervisionado (do inglês: *Supervised Contrastive Learning*) (SCL) para criar protótipos por palavra. O método reduziu o WER em 15,59% na média, com melhora adicional de 1,21% ao incorporar SCL, destacando-se como uma abordagem eficiente para o reconhecimento de novos padrões disártricos.

Por fim, Lee et al. (2025) propuseram o método Aprendizado Contrastivo Dinâmico em Nível de Fonema (do inglês: *Dynamic Phoneme-level Contrastive Learning*) (DyPCL), que introduz discriminação em nível fonêmico e currículo dinâmico com amostras negativas organizadas por similaridade fonética. Avaliado na base UA-Speech, o método alcançou uma redução relativa de 22,10% no WER, evidenciando os benefícios do contraste fonêmico.

4.3 Metodologia

Para a presente proposta, os codificadores descritos na Seção 3 (*FW1* e *FW2*) foram ajustados para realizar um pré-treinamento com base no paradigma de aprendizado autossupervisionado, utilizando os métodos contrastivos SimCLR – (CHEN et al., 2020b) e SwAV (CARON et al., 2020). Esses métodos são projetados para extrair representações discriminativas a partir de dados não rotulados, por meio da comparação entre diferentes transformações de uma mesma entrada.

Os codificadores utilizados neste pré-treinamento foram inicializados com os pesos previamente obtidos durante o treinamento na base LJSpeech, conforme detalhado na Seção 3. Esse procedimento visa aproveitar representações acústicas já aprendidas a partir de uma base ampla e diversa, proporcionando uma base sólida para a adaptação contrastiva aos dados disártricos.

O SimCLR utiliza pares positivos - obtidos a partir de diferentes transformações de uma mesma amostra de um espectrograma - e negativos - obtidos de outras amostras do lote - para otimizar uma função de perda contrastiva baseada em similaridade no espaço latente. O objetivo é aproximar as representações de pares positivos e distanciar as de pares negativos, promovendo a generalização do codificador.

Já o SwAV propõe uma abordagem alternativa, combinando aprendizado contrastivo com técnicas de *clustering* online. Em vez de comparar diretamente os pares de amostras, SwAV associa cada representação a protótipos aprendidos e realiza o alinhamento entre atribuições obtidas de diferentes vistas, buscando consistência entre as diferentes transformações de uma mesma amostra.

Durante o pré-processamento, os dados de entrada foram submetidos a transformações específicas, incluindo oclusão espectral, adição de ruído e dilatação temporal. Essas transformações visam simular variações acústicas e temporais na fala, permitindo que os codificadores aprendam a distinguir características invariantes dos sinais de fala. Instâncias não modificadas da fala foram mantidas como referência, compondo os pares positivos utilizados na formação das representações contrastivas.

As Figuras 22 e 23 apresentam uma visão geral do processo metodológico adotado para o modelos baseados no SimCLR e SwAV. A principal distinção em relação à abordagem descrita na Seção 3 está na Fase 2 (CL), onde o treinamento convencional com os dados da partição controle e suas transformações foi substituído por um processo de aprendizado contrastivo. Nessa nova abordagem, as transformações aplicadas buscaram realçar semelhanças entre espectrogramas de fala normotípica e versões modificadas que simulam características comuns em pacientes com disartria. As representações geradas durante essa fase foram, então, utilizadas na Fase 3 ($S2S$), em que se realizou o ajuste fino com dados reais de fala disártrica, individualmente para cada falante.

O procedimento de pré-treinamento contrastivo foi realizado por 100 épocas, com tamanho de lote igual a 32 e temperatura fixada em 0,5 para a função de perda contrastiva. Tais parâmetros foram escolhidos com base em estudos prévios que indicam sua efetividade na estabilidade do treinamento e na qualidade das representações aprendidas. Os codificadores *FW1 e FW2*, conforme definidos na Seção 3, foram utilizados integralmente como arquitetura de base neste processo.

Tanto o codificador quanto o decodificador utilizados neste trabalho são baseados exclusivamente em arquiteturas do tipo *Transformer*, sendo as entradas da rede compostas por espectrogramas STFT extraídos dos sinais de fala. Essa escolha visa garantir uma modelagem sequencial eficaz e uma captura robusta de dependências temporais e espectrais ao longo da fala.

Para avaliar a eficácia dos modelos propostos, foram medidas as taxas de reconhecimento da fala presente na base de dados UA-Speech para diferentes níveis de severidade, utilizando métricas como WER, CER e WRA. Essas métricas possibilitam mensurar tanto a precisão no nível lexical quanto fonêmico, permitindo uma análise abrangente da inteligibilidade das saídas.

Após o treinamento, os modelos foram testados com amostras reais de fala disártrica não vistas previamente. O objetivo foi verificar se as saídas geradas eram compatíveis com os rótulos esperados, evidenciando a capacidade dos modelos em aprender e processar as

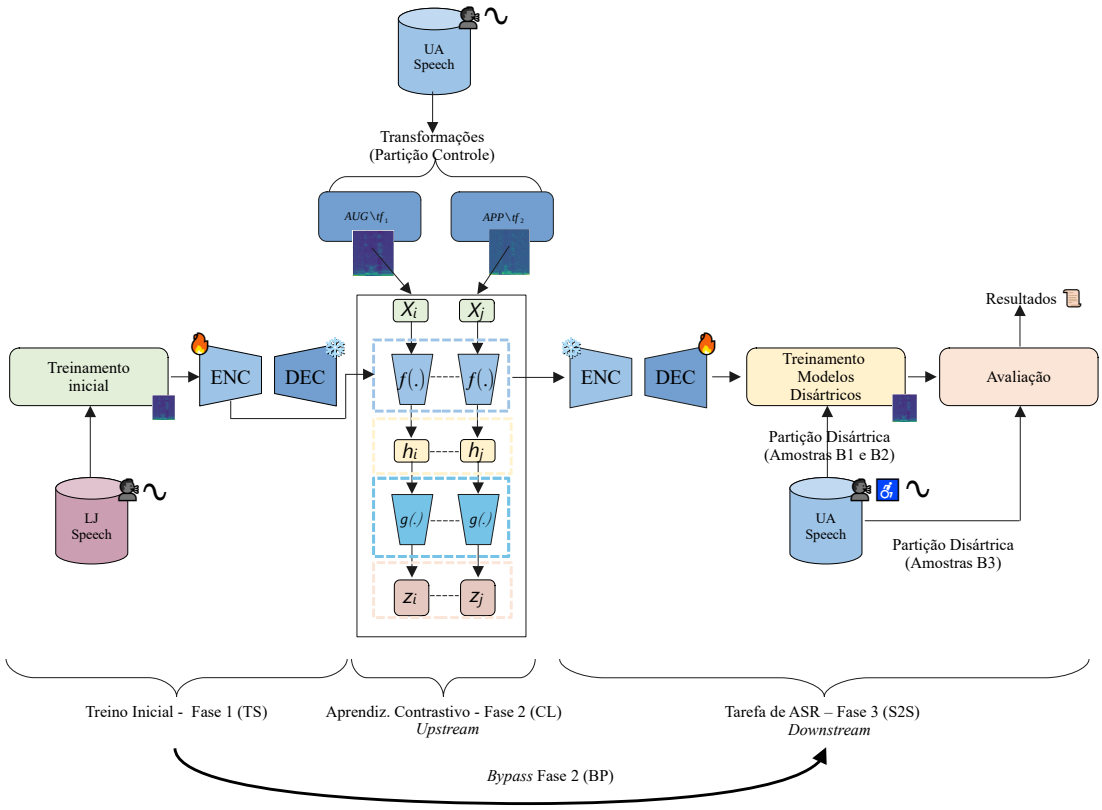


Figura 22 – Visão geral do fluxo de trabalho adotado para a técnica SimCLR. FONTE: Os Autores.

nuances da fala com comprometimento motor.

Os resultados foram analisados com base nas mesmas métricas iniciais (WER, CER e WRA), possibilitando uma comparação direta do desempenho antes e depois da aplicação do pré-treinamento contrastivo.

4.3.1 Arquiteturas dos Modelos

Aprendizado Contrastivo (CL) - *Upstream*

O codificador utilizado neste módulo foi projetado para extrair representações discriminativas a partir de espectrogramas STFT gerados a partir de sinais de fala. As entradas foram submetidas a transformações que incluem inserção de ruído branco (tf_1), dilatação temporal (tf_2) e oclusão espectral (tf_3), com o objetivo de gerar diferentes vistas da mesma amostra. Essas técnicas de *data augmentation* foram aplicadas exclusivamente sobre a partição controle da base UA-Speech, composta por amostras de fala normotípica.

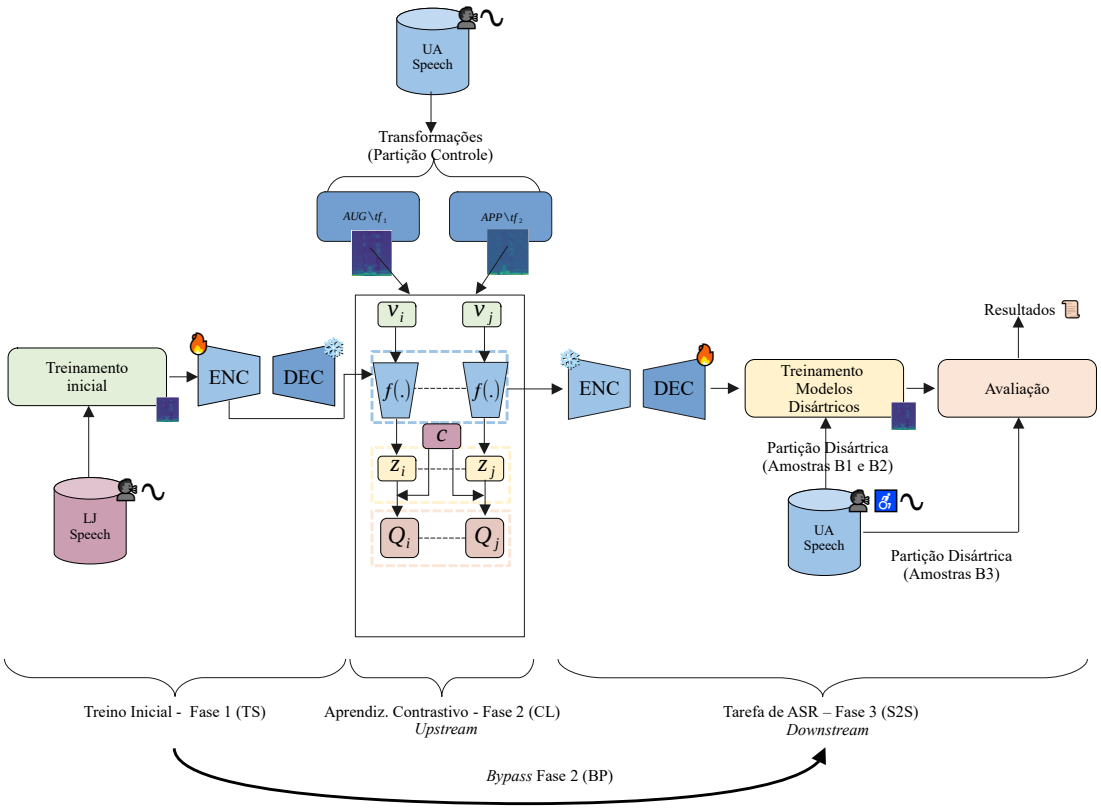


Figura 23 – Visão geral do fluxo de trabalho adotado para a técnica SwAV. FONTE: Os Autores.

Essa partição foi gerada em 80% para treinamento e 20% para validação, e utilizada no pré-treinamento contrastivo do codificador.

Após as etapas de *augmentation*, os espectrogramas são normalizados, cortados em frequência e preenchidos com zeros para padronizar sua dimensão. A arquitetura é composta inicialmente por um bloco convolucional com três camadas 1D sequenciais, cada uma utilizando filtros com ativação ReLU e *stride* reduzido, com o intuito de comprimir a representação temporal e capturar padrões locais relevantes na estrutura espectral. Em seguida, o sinal processado é encaminhado a uma pilha de camadas *Transformer*, nas quais são aplicados mecanismos de atenção multicabeça com 2 cabeças de atenção, normalização por camadas e sub-redes do tipo *feedforward*. Essas camadas têm como função principal modelar dependências temporais de longo alcance e aprimorar a expressividade da representação extraída.

Ao final do codificador, a saída é agregada ao longo do tempo por meio de média global, sendo posteriormente projetada em um espaço latente de 128 dimensões por uma cabeça contrastiva. No caso do SimCLR, essa cabeça segue o mesmo padrão adotado no restante da arquitetura: no modelo *FW1*, é composta por duas camadas densas com ativação

ReLU e normalização L2 na saída; já no modelo *FW2*, utiliza *separable convolutions* com ativação intermediária e projeção final normalizada. A perda contrastiva é calculada com base na similaridade cosseno entre pares positivos e negativos, com temperatura fixada em 0,5.

No caso do SwAV, a cabeça de projeção é seguida por uma camada de protótipos, com 3.000 unidades, que projeta os *embeddings* normalizados em um conjunto fixo de vetores de referência aprendidos. Diferente da abordagem SimCLR, que realiza comparações explícitas entre pares de amostras, o SwAV associa os *embeddings* a protótipos por meio de um processo de agrupamento supervisionado indiretamente. Para isso, é utilizado o algoritmo de normalização de *Sinkhorn* com 3 iterações e regularização por entropia com valor $\epsilon = 0{,}05$. Essa operação converte os *scores* dos protótipos em distribuições de probabilidade que indicam a associação de cada amostra a um protótipo.

A função de perda do SwAV busca maximizar a consistência entre as atribuições obtidas de diferentes visões de uma mesma amostra. A distribuição obtida de uma vista é utilizada como alvo para a predição da outra, caracterizando um alinhamento cruzado que dispensa a comparação explícita com amostras negativas. Esse mecanismo possibilita a extração de representações discriminativas de forma eficiente, mesmo em grandes volumes de dados.

O pré-treinamento contrastivo do codificador foi realizado por 100 épocas, com tamanho de lote igual a 32, utilizando exclusivamente a base de fala normotípica da UA-Speech com as técnicas de aumento de dados já descritas. Para otimização, foram testados diferentes algoritmos, incluindo Escalonamento Adaptativo com Regressão por Etapas (do inglês: *Layer-wise Adaptive Rate Scaling*) (LARS), Otimizador com Escalonamento Adaptativo por Camada para Treinamento em Larga Escala (do inglês: *Layer-wise Adaptive Moments optimizer for Batch training*) (LAMB) e Descida do Gradiente Estocástica (do inglês: *Stochastic Gradient Descent*) (SGD) com *momentum*. Neste trabalho, optou-se pelo uso do otimizador SGD, com taxa de aprendizado inicial de 1×10^{-3}, ajustada por uma estratégia combinada de *warmup* nas primeiras épocas e agendamento com Reinicializações Cíclicas por Decaimento do Cosseno (do inglês: *Cosine Annealing with warm Restarts*).

Os pesos do codificador são salvos ao longo do treinamento e utilizados posteriormente nas etapas de avaliação e transferência para tarefas *downstream*, permitindo verificar a efetividade das representações extraídas para o reconhecimento de fala disártrica.

ASR - Seq2Seq ($S2S$) - *Downstream*

O modelo de ASR adotado nesta fase do pipeline segue uma arquitetura do tipo codificador-decodificador baseada em *Transformer*, adaptada para processar espectrogramas STFT e gerar transcrições no formato de sequência de caracteres. O modelo trata o

reconhecimento de fala como um problema de transdução sequencial, mapeando representações acústicas para representações textuais por meio de atenção contextual e codificação profunda.

Na etapa de codificação, os espectrogramas são inicialmente processados por uma série de três camadas convolucionais 1D com função de ativação ReLU e *strides* reduzidos, responsáveis por reduzir a resolução temporal e extrair padrões acústicos locais. Em seguida, os sinais transformados são repassados para uma pilha de camadas codificadoras, onde cada camada aplica atenção multicabeça seguida de uma sub-rede do tipo *feedforward*.

Duas variantes foram consideradas nesta etapa. No modelo *FW1*, a sub-rede *feedforward* das camadas do codificador é composta por camadas totalmente conectadas (*dense*), conforme a proposta original do *Transformer*. Já no modelo *FW2*, essa sub-rede é substituída por uma estrutura com *separable convolutions*, utilizando convoluções do tipo *depthwise* seguidas por *pointwise*, com o objetivo de reduzir o número de parâmetros e melhorar a modelagem de padrões locais na dimensão temporal.

Durante a fase de treinamento supervisionado (Fase 3 - $S2S$), os pesos aprendidos na etapa de pré-treinamento contrastivo (Fase 2 - CL) são carregados no codificador, permitindo que o modelo se beneficie das representações previamente aprendidas a partir de dados não rotulados. Essa estratégia visa acelerar a convergência, melhorar a generalização e preservar informações fonéticas relevantes adquiridas durante o pré-treinamento.

O treinamento na Fase 3 ($S2S$) foi realizado utilizando a partição disártrica da UA-Speech. A base foi dividida em três subconjuntos: treino, validação e teste. A separação entre treino e validação foi feita com um particionamento estratificado na proporção de 80% para o treinamento e 20% para a validação. O conjunto de teste foi composto exclusivamente por amostras com prefixo B3, previamente separadas do restante da base. O modelo foi treinado por 100 épocas, com tamanho de lote igual a 64.

A base de transcrição foi convertida para sequências de caracteres, com vocabulário composto por 34 classes, incluindo letras, espaço, pontuações e símbolos especiais de início e fim de sequência. No decodificador, as entradas são compostas por sequências de caracteres-alvo deslocadas temporalmente para a direita, incorporadas por uma camada que combina *embeddings* de caracteres e *embeddings* posicionais. As camadas do decodificador aplicam atenção causal sobre a sequência-alvo e atenção cruzada sobre a saída do codificador, além de uma sub-rede *feedforward* semelhante à do codificador, também configurável com camadas densas.

A arquitetura permite o controle de treinabilidade das camadas do codificador e do decodificador de forma seletiva. Em particular, o *feedforward* da última camada do codificador pode ser congelado ou mantido treinável durante o ajuste fino com base disártrica, viabilizando a reutilização eficaz dos pesos obtidos na Fase 2 , de acordo com os objetivos experimentais.

A inferência é realizada de forma autorregressiva, por meio de decodificação *greedy*.

A função de perda utilizada durante o treinamento supervisionado é a entropia cruzada categórica, com suavização de rótulo (*label smoothing*) e aplicação de máscara sobre os *tokens* inválidos. A taxa de aprendizado é ajustada dinamicamente por meio de uma política que combina aquecimento inicial (*warmup*) com decaimento linear ao longo das épocas.

A avaliação do modelo foi conduzida com base em métricas amplamente utilizadas na literatura de ASR, incluindo WER, CER e WRA, com validação contínua durante o treinamento e testes finais sobre amostras reais de fala disártrica. Essas métricas permitiram uma análise detalhada da inteligibilidade das saídas geradas pelo modelo ao longo do processo de ajuste fino.

4.3.2 Experimentos e Discussão dos Resultados

Foram conduzidos experimentos com dois métodos distintos de pré-treinamento contrastivo: o SimCLR e o SwAV, empregados em ambas arquiteturas analisadas no Capítulo 3. As avaliações foram realizadas individualmente por falante, utilizando como principais métricas já descritas. A análise dos resultados abrange diferentes configurações experimentais, incluindo comparações entre o modelo *FW1* com e sem pré-treinamento SimCLR, o modelo *FW2* também nas condições com e sem pré-treinamento SimCLR, bem como a aplicação do SwAV na arquitetura *FW2*. Além disso, foram realizadas comparações cruzadas entre os modelos e entre os métodos de pré-treinamento, permitindo avaliar o impacto relativo de cada abordagem sobre a capacidade de reconhecimento de fala disártrica.

Avaliação Geral dos Resultados por WRA, WER e CER

Os resultados revelaram que todos os modelos com *upstream* contrastivo (CL) apresentaram melhorias substanciais em relação às suas versões sem pré-treinamento contrastivo (BP). O modelo *FW1* com SimCLR foi o que obteve os maiores ganhos médios em WRA, com destaque para falantes como M16, F04, F02 e M07, cujos aumentos absolutos superaram 200%. Embora o WRA e o WER evidenciem melhorias expressivas, a análise da CER também confirma avanços relevantes: mesmo em casos em que a redução no WER foi moderada, observou-se diminuição consistente nos erros de caracteres, indicando que o modelo passou a gerar palavras mais próximas da grafia correta. A Tabela 7 apresenta os valores médios de todas as métricas para os modelos *FW1* e *FW2*, considerando as diferentes estratégias de pré-treinamento contrastivo. Observa-se que todos os modelos com aplicação da fase *CL* apresentaram reduções consistentes nas taxas de erro e ganhos relevantes em inteligibilidade. O modelo *FW1* com SimCLR, apesar de sua simplicidade estrutural, obteve o maior ganho absoluto em WRA (+23,9 pontos percentuais), elevando sua taxa média de reconhecimento de palavras de apenas 1,2% para 25,1%. Já

Tabela 7 – Resumo comparativo das métricas WER, CER e WRA por modelo

(a) WER e CER

Modelo	WER_{BP}	WER_{CL}	ΔWER	CER_{BP}	CER_{CL}	ΔCER
FW1 (SimCLR)	0.9875	0.7486	-0.2389	0.8942	0.7351	-0.1591
FW1 (SwAV)	0.9875	0.7664	-0.2211	0.8942	0.7700	-0.1242
FW2 (SimCLR)	0.9122	0.7240	-0.2042	0.8901	0.7189	-0.1712
FW2 (SwAV)	0.9122	0.7205	-0.2097	0.8901	0.7657	-0.1244

(b) WRA

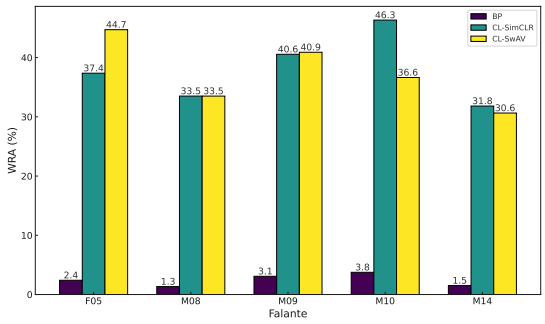
Modelo	WRA_{BP}	WRA_{CL}	ΔWRA
FW1 (SimCLR)	0.0125	0.2514	+0.2389
FW1 (SwAV)	0.0125	0.2431	+0.2306
FW2 (SimCLR)	0.0878	0.2760	+0.1882
FW2 (SwAV)	0.0878	0.2795	+0.1917

o *FW2* com SimCLR e *FW2* com SwAV alcançaram resultados finais bastante similares em WRA, com médias de 27,6% e 27,9% respectivamente. A ligeira vantagem do SwAV na WRA final sugere que, embora o SimCLR apresente reduções mais expressivas em CER, o SwAV pode oferecer maior estabilidade semântica, especialmente em contextos de maior inteligibilidade.

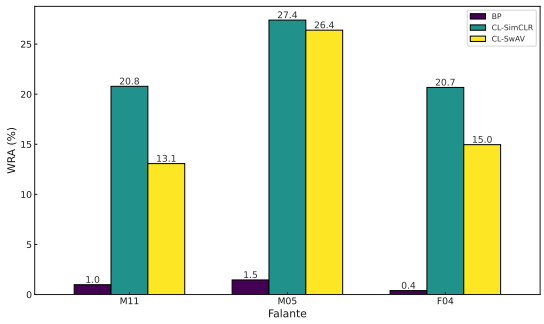
A Tabela 12 (Apêndice B) apresenta os valores absolutos de WRA e os percentuais de melhoria obtidos para cada falante. No caso do modelo *FW2*, o SimCLR também promoveu avanços relevantes, com aumento médio de +18,8 pp em WRA, reduções expressivas de até -35,9 pp em WER e quedas proporcionais na CER, especialmente entre falantes com maior severidade de comprometimento articulatório. Já a aplicação do SwAV ao *FW2* também se mostrou eficaz, atingindo desempenho superior ao SimCLR em vários falantes (como F03, F05, M05 e M09) e apresentando uma média final de WRA ligeiramente superior (+19,2 pp), além de manter uma CER mais baixa na maioria dos casos.

A Tabela 13 (apêndice B) resume detalhadamente os valores de CER e WER com e sem *upstream* contrastivo para todos os falantes. Em termos médios, o modelo *FW1* com SimCLR apresentou as maiores reduções: -15,9 pontos percentuais em CER e -23,9 pp em WER. O modelo *FW2* com SimCLR teve um ganho médio semelhante em WER (-20,4 pp), mas com uma redução mais modesta em CER (-17,2 pp). O SwAV aplicado ao *FW2* apresentou ΔWER comparável ao SimCLR (-20,9 pp), mas com ganho médio inferior em CER (-12,4 pp), sugerindo que o SwAV pode promover maior robustez semântica em contextos de fala menos comprometida, mesmo com menor redução fonética.

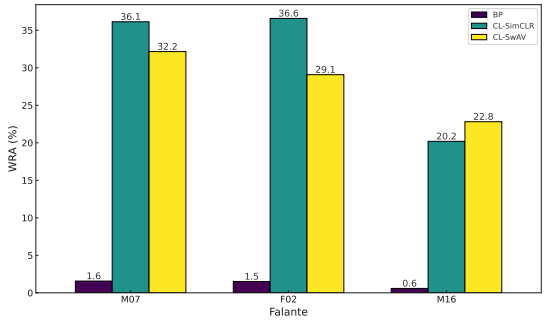
As Figuras 24 e 27 ilustram visualmente os resultados de WRA. A Figura 24 mostra os ganhos obtidos com o uso do SimCLR na arquitetura *FW1*, enquanto a Figura 27 exibe os mesmos dados para a arquitetura *FW2*. As métricas CER e WER aplicadas às



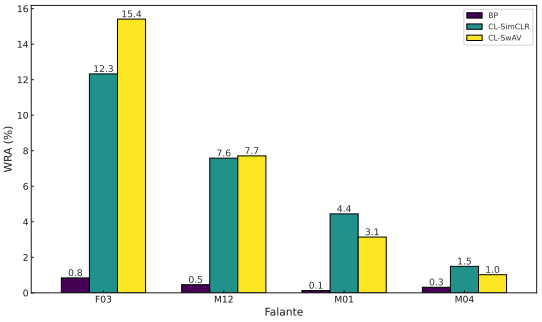
(a) - WRA - Alta Intelig. - FW1



(b) - WRA - Moderada Intelig. - FW1



(c) - WRA - Baixa Intelig. - FW1



(d) - WRA - Muito Baixa Intelig. - FW1

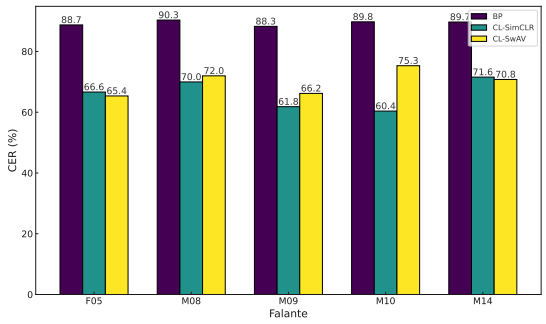
Figura 24 – Comparativos em WRA, dos métodos adotados no modelo FW1. *CL* vs. *BP*.

arquiteturas *FW1* e *FW2* estão ilustradas, respectivamente, nas Figuras 25/26 e 28/29.

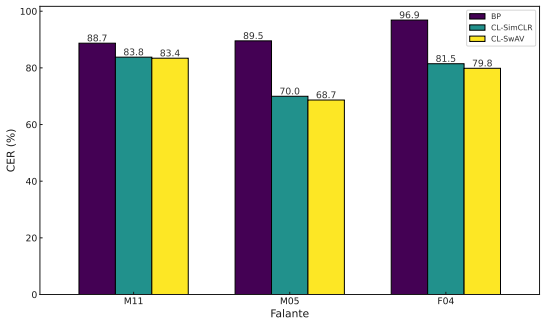
Uma observação importante, principalmente nos resultados dos falantes M04 e M11 em *FW2*, diz respeito à divergência entre as métricas fonêmica (CER) e lexical (WRA/WER). Embora o pré-treinamento contrastivo não tenha produzido os melhores resultados de CER para esses falantes, o desempenho em WRA/WER foi relativamente satisfatório. Isso sugere que o modelo, apesar de errar fonemas individuais, conseguiu preservar parte da estrutura global das palavras.

Essa diferença pode ser explicada considerando que os falantes M04 e M11 apresentam os níveis de inteligibilidade mais comprometidos dentro dos seus respectivos grupos, sendo plausível que palavras curtas sejam mais facilmente pronunciadas e reconhecidas, tanto pelo menor esforço articulatório quanto pela menor chance de erro cumulativo. Palavras mais longas, por outro lado, exigem maior coordenação motora e estão mais sujeitas a distorções acústicas. Como resultado, o modelo pode atingir alta precisão fonêmica e lexical em palavras curtas, mesmo em falas severamente disártricas, enquanto palavras longas sofrem mais com a degradação fonológica.

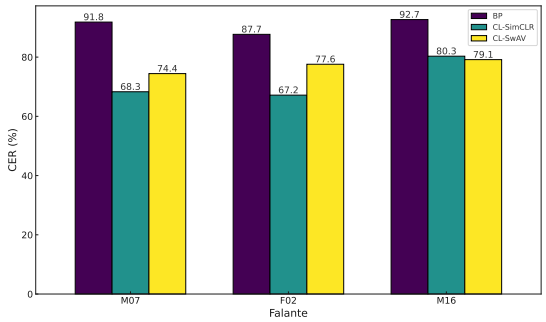
Esses achados ressaltam a importância de considerar a natureza da métrica utilizada na avaliação e reforçam que melhorias no espaço de representação global (via aprendizado contrastivo) podem não se traduzir automaticamente em ganhos fonêmicos, especialmente quando o decodificador é sensível a erros de caractere como o *greedy* utilizado no constexto



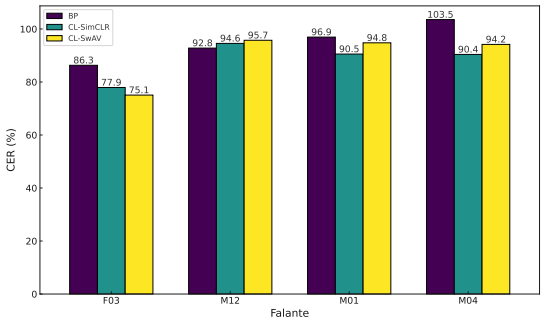
(a) - CER - Alta Intelig. - FW1



(b) - CER - Moderada Intelig. - FW1

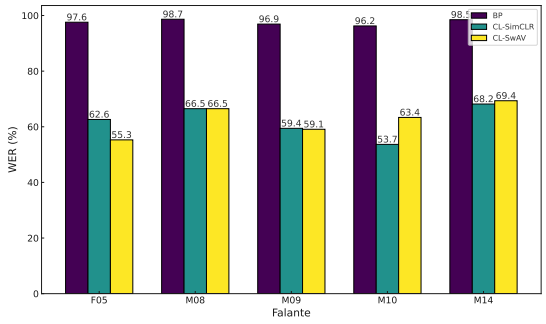


(c) - CER - Baixa Intelig. - FW1

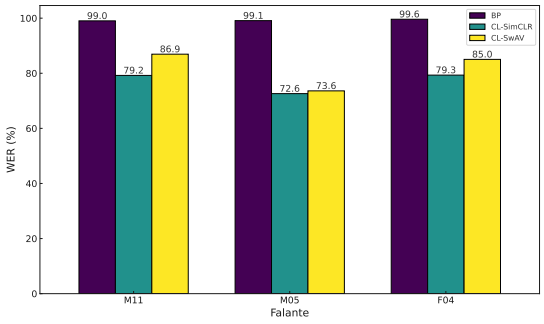


(d) - CER - Muito Baixa Intelig. - FW1

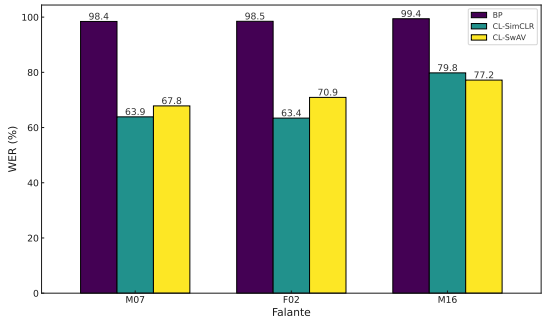
Figura 25 – Comparativos em CER, dos métodos adotados no modelo FW1. CL vs. BP.



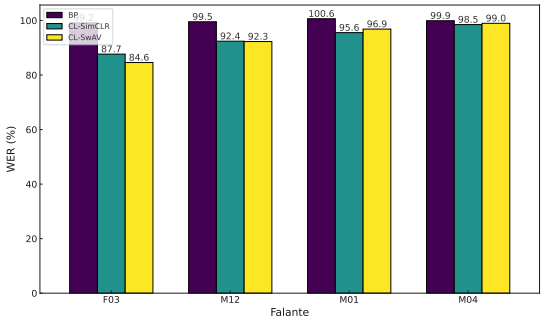
(a) - WER - Alta Intelig. - FW1



(b) - WER - Moderada Intelig. - FW1

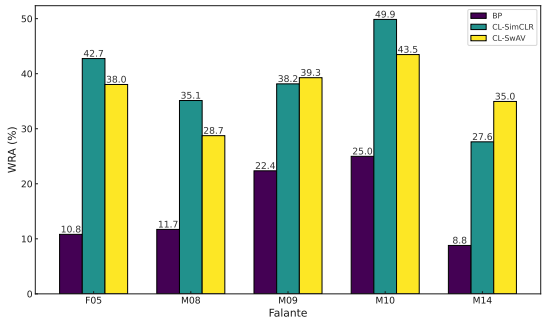


(c) - WER - Baixa Intelig. - FW1

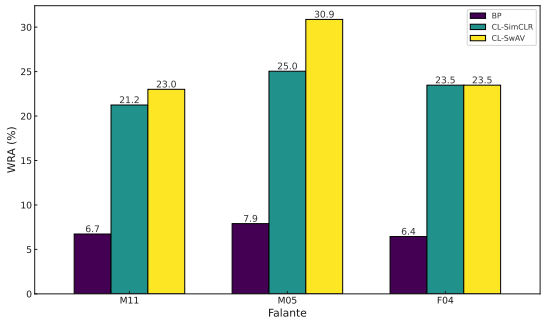


(d) - WER - Muito Baixa Intelig. - FW1

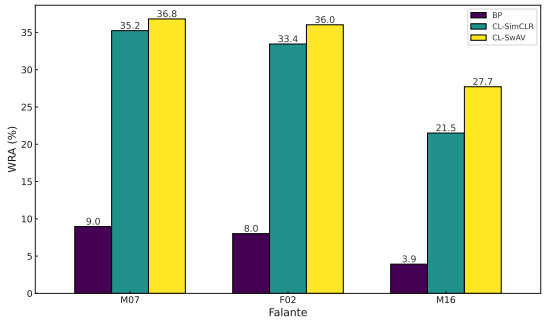
Figura 26 – Comparativos em WER, dos métodos adotados no modelo FW1. CL vs. BP.



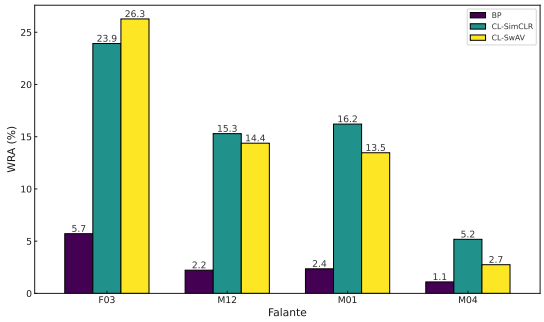
(a) - WRA - Alta Intelig. - FW2



(b) - WRA - Moderada Intelig. - FW2

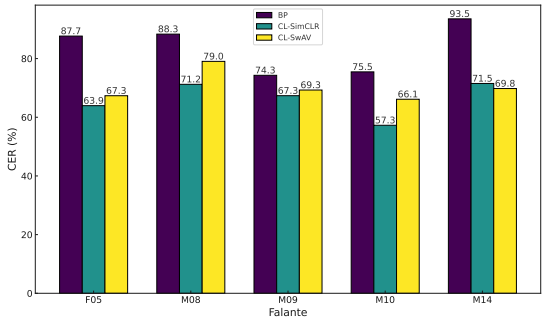


(c) - WRA - Baixa Intelig. - FW2

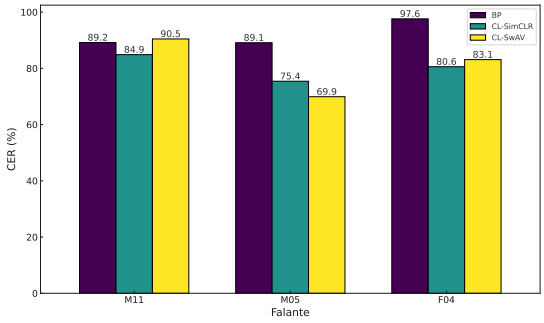


(d) - WRA - Muito Baixa Intelig. - FW2

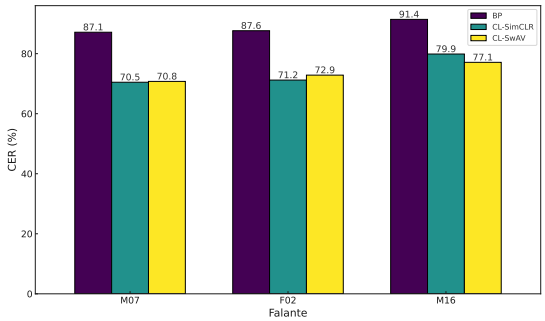
Figura 27 – Comparativos em WRA, dos métodos adotados no modelo FW2. *CL* vs. *BP*.



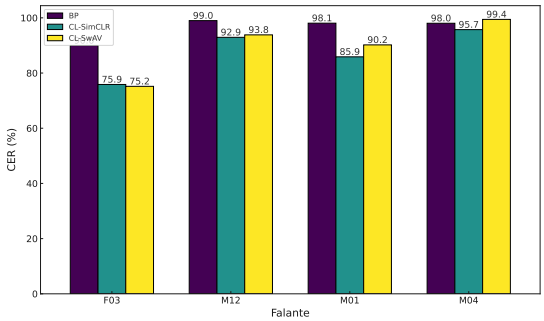
(a) - CER - Alta Intelig. - FW2



(b) - CER - Moderada Intelig. - FW2

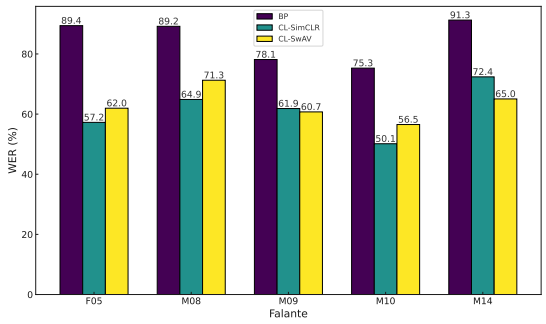


(c) - CER - Baixa Intelig. - FW2

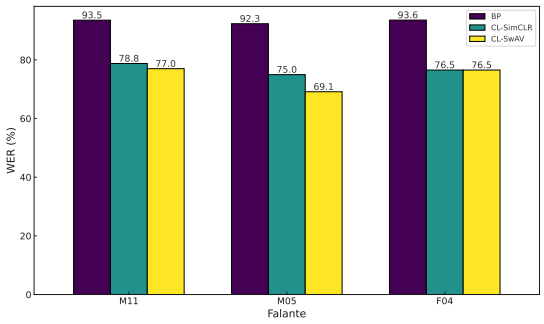


(d) - CER - Muito Baixa Intelig. - FW2

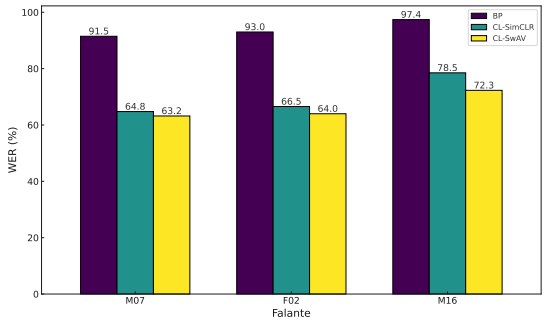
Figura 28 – Comparativos em CER, dos métodos adotados no modelo FW2. *CL* vs. *BP*.



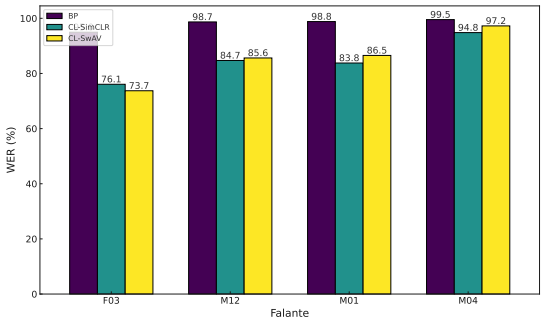
(a) - WER - Alta Intelig. - FW2



(b) - WER - Moderada Intelig. - FW2



(c) - WER - Baixa Intelig. - FW2



(d) - WER - Muito Baixa Intelig. - FW2

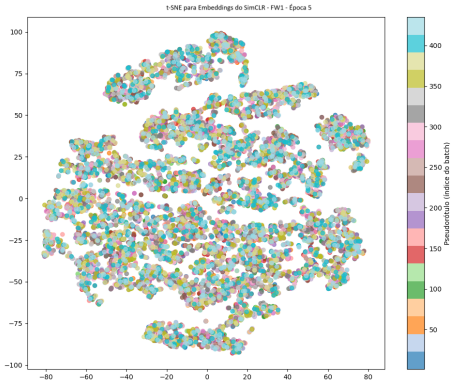
Figura 29 – Comparativos em WER, dos métodos adotados no modelo FW2. CL vs. BP.

dessa pesquisa.

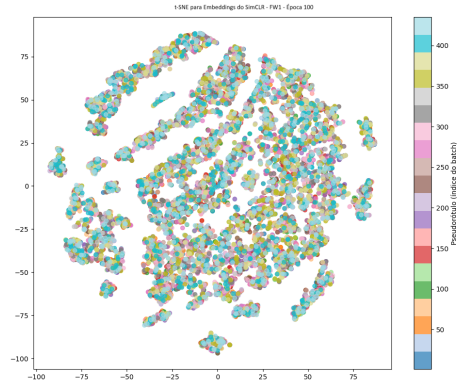
4.3.3 Análise dos Embeddings com t-SNE

Para investigar a evolução das representações aprendidas ao longo do treinamento, aplicamos a técnica de projeção, Embutimento Estocástico de Vizinhança Distribuída t (do inglês: *t-Distributed Stochastic Neighbor Embedding*) (t-SNE), aos *embeddings* extraídos pelos modelos contrastivos SimCLR e SwAV. As Figuras 30 e 31 apresentam visualizações bidimensionais dos *embeddings* nas épocas 5 e 100 para os modelos *FW1* e *FW2*. Observa-se que, ao longo do treinamento, os pontos tornam-se progressivamente mais organizados em regiões bem definidas, indicando uma melhor separação no espaço latente.

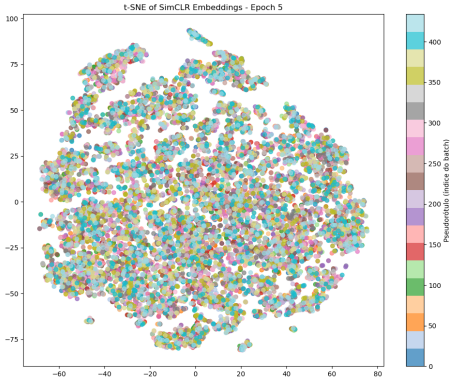
Nota-se que, mesmo nas primeiras épocas, os *embeddings* já apresentam certa organização, mas se tornam ainda mais estruturados à medida que o treinamento avança, com agrupamentos mais nítidos e coesos. Tal comportamento é consistente com o aprendizado de representações mais discriminativas. Em todas as projeções, as cores representam pseudorótulos (índices de *batch*), o que reforça a natureza não supervisionada dos métodos. A organização espacial crescente ao longo do tempo sugere que os modelos contrastivos foram bem-sucedidos em agrupar amostras semanticamente similares, o que é desejável para o uso posterior em tarefas *downstream* como o reconhecimento de fala.



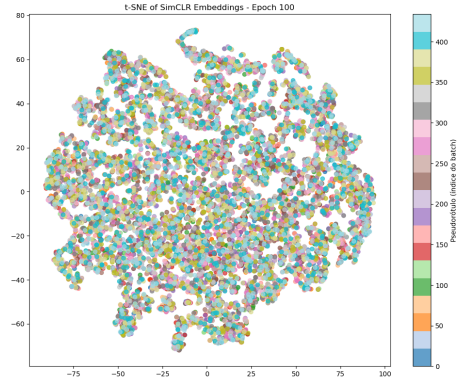
SimCLR - FW1 - Época 5 (a1)



SimCLR - FW1 - Época 100 (a2)



SwAV - FW1 - Época 5 (b1)



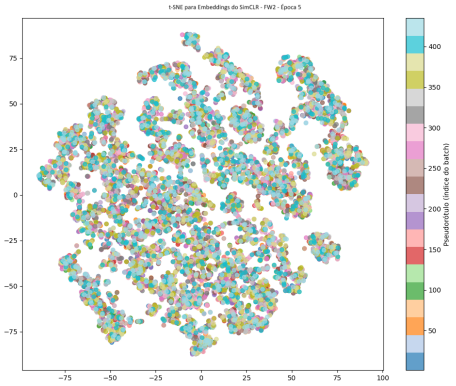
SwAV - FW1 - Época 100 (b2)

Figura 30 – Visualização t-SNE dos embeddings nas épocas 5 e 100 para FW1 - SimCLR e SwAV.

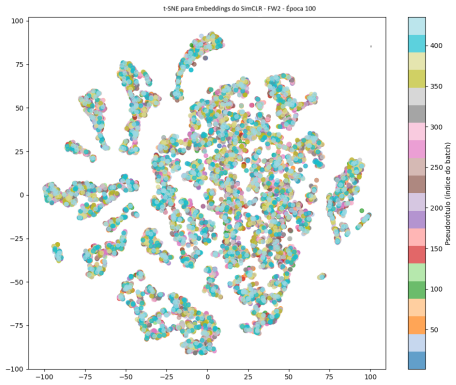
4.3.4 Comparação entre técnicas supervisionadas e contrastivas

As Figuras 32 e 33 sintetizam respectivamente, falante a falante, a métrica de inteligibilidade (WRA) e o erro de caracteres (CER) obtidos com o melhor aumento supervisionado — "AUG", "AP" ou "APP" — e o melhor pré-treinamento contrastivo — SimCLR ou SwAV — nas duas arquiteturas avaliadas. Como $WER = 100 - WRA$, apresentamos o WER apenas em forma gráfica, mais especificamente na Figura 34, por ser complementar e não acrescentar tendências que já não estejam refletidas no WRA.

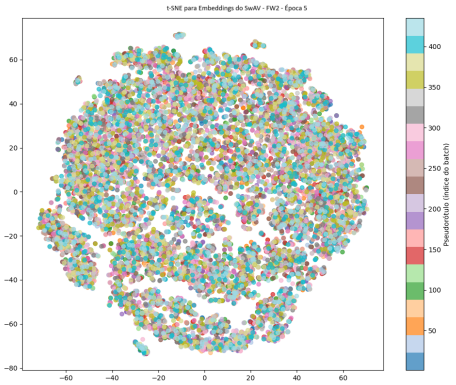
De modo geral, as técnicas de aumento supervisionado permanecem a referência de inteligibilidade: em média elevaram a WRA para 82 % no modelo raso (*FW1*) e para 50 % no modelo profundo (*FW2*), contra 26 % e 30 %, respectivamente, alcançados pelo melhor contraste. Mesmo assim o comportamento não é monolítico. No *FW1* o contraste superou a supervisão em oito das quinze vozes — casos como M16, F03, M12 e M07, todos com fala severamente comprometida, onde o SwAV ou o SimCLR acrescentaram em média 11 p.p. de WRA. Já no *FW2* a supervisão voltou a dominar em WRA para todos os falantes, mas a diferença reduziu-se a valores tão baixos quanto 3,8 p.p. (M16), sugerindo que arquiteturas mais profundas conseguem aproveitar melhor as representações contrastivas, ainda que não as tornem mais efetivas.



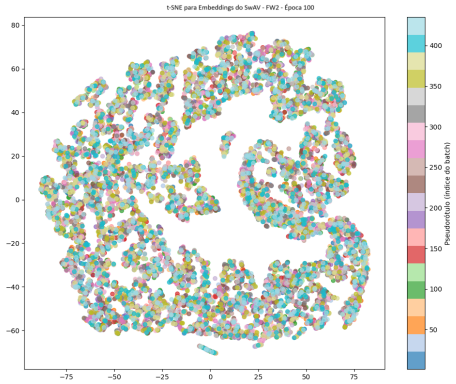
SimCLR - FW2 - Época 5 (a1)



SimCLR - FW2 - Época 100 (a2)



SwAV - FW2 - Época 5 (b1)

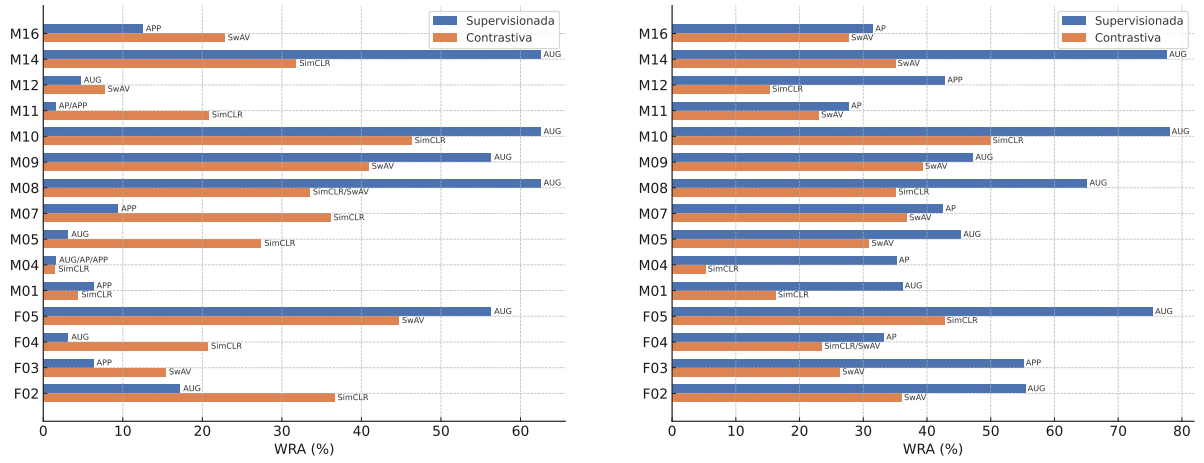


SwAV - FW2 - Época 100 (b2)

Figura 31 – Visualização t-SNE dos embeddings nas épocas 5 e 100 para FW2 - SimCLR e SwAV.

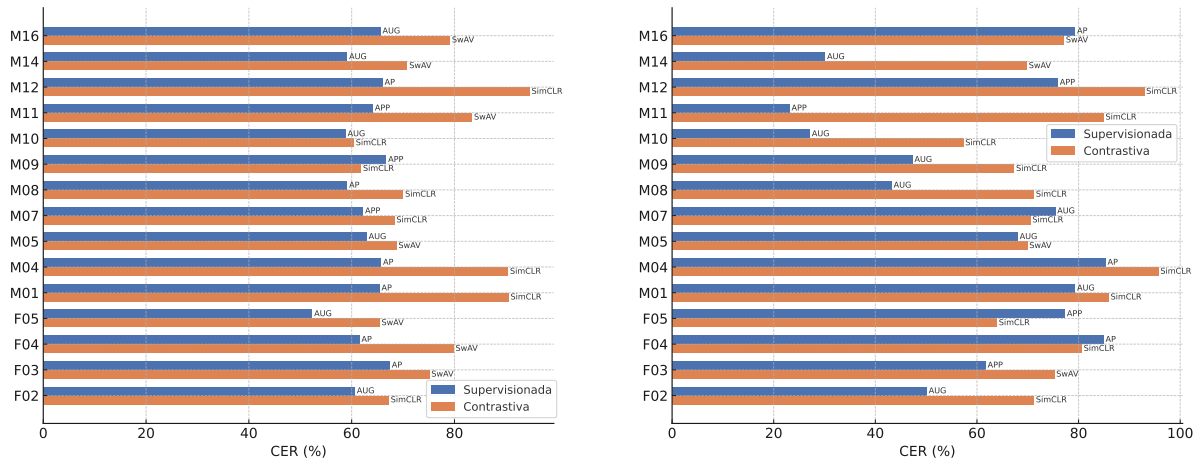
Quando se passa do nível lexical para o fonético, o panorama muda ligeiramente. No *FW1* o aumento supervisionado também possui o menor CER em quase toda a amostra, mas o SwAV obtém a melhor marca para o falante M09, reduzindo o erro em 4,9 p.p. No *FW2* o contraste vence em quatro grupos de amostras (F04, F05, M07, M16), com destaque para F05, em que o SwAV chega a diminuir o CER em 13,4 p.p. em relação ao "APP". Esses resultados indicam que, embora a supervisão continue imbatível em inteligibilidade global, o pré-treinamento contrastivo já produz representações fonéticas mais nítidas em parte dos casos — especialmente nas falas de inteligibilidade muito baixa.

Em síntese, para a configuração de treino utilizada (100 épocas, lote 32, GPU única) os aumentos supervisionados — sobretudo o "APP" seguido do "AUG" — entregam a melhor WRA/WER em praticamente todo o conjunto de falantes, enquanto o contraste (com ênfase no SwAV) se mostra promissor para reduzir o CER em vozes severas e, num número crescente de casos, aproximar-se da inteligibilidade obtida pela supervisão. Essas evidências sugerem que uma estratégia híbrida, que combine aumentos supervisionados com um pré-treinamento contrastivo mais longo ou executado em lotes maiores, pode ser o caminho para reduzir simultaneamente WER e CER em futuras versões do sistema.



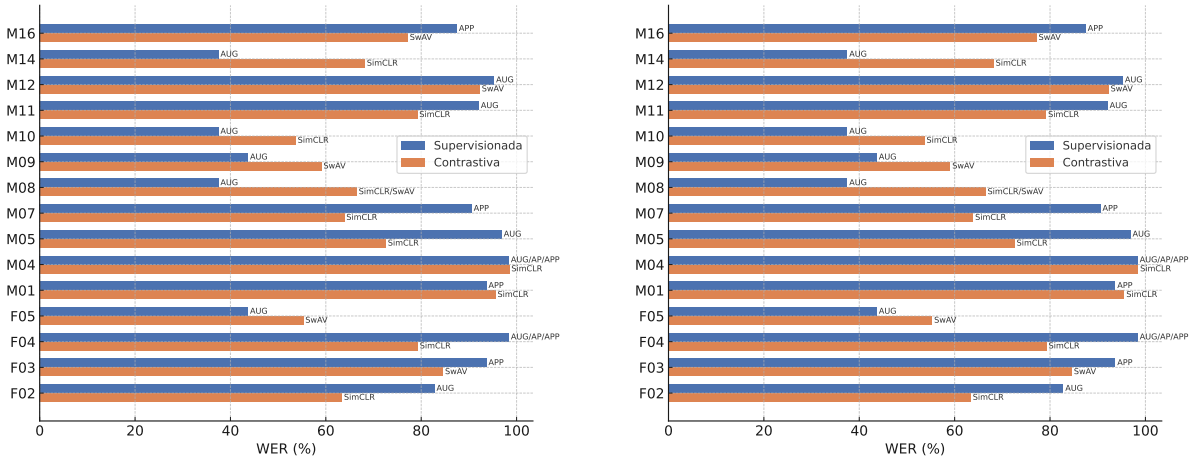
(a) FW1 (b) FW2

Figura 32 – Comparação em WRA, falante a falante, entre as melhores técnicas de aumento supervisionado e as melhores técnicas contrastivas para os modelos FW1 e FW2.



(a) FW1 (b) FW2

Figura 33 – Comparação em CER, falante a falante, entre as melhores técnicas de aumento supervisionado e as melhores técnicas contrastivas para os modelos FW1 e FW2.



(a) FW1 (b) FW2

Figura 34 – Comparação em WER, falante a falante, entre as melhores técnicas de aumento supervisionado e as melhores técnicas contrastivas para os modelos FW1 e FW2.

4.4 Considerações finais

Os experimentos apresentados neste capítulo confirmaram, em parte, a eficácia do uso de aprendizado contrastivo como estratégia de pré-treinamento para o reconhecimento automático de fala disártrica. Foram testadas diferentes combinações de arquiteturas (*FW1* e *FW2*) e técnicas de aprendizado contrastivo (SimCLR e SwAV), utilizando métricas que abrangem desde a taxa de erro fonêmico (CER) e de palavras (WER) até a inteligibilidade efetiva das transcrições (WRA).

Esses resultados respondem diretamente à hipótese levantada na introdução deste trabalho. Em relação à eficácia do pré-treinamento contrastivo, os dados demonstram que todos os modelos aplicados à Fase CL superaram de forma consistente suas versões sem esse pré-treinamento (BP) em todas as métricas — confirmando a hipótese de que o uso de SimCLR e SwAV melhora significativamente o desempenho do sistema, principalmente se aplicados a modelos supervisionados treinados em banco de dados de vozes normotípicas não aumentadas.

No que diz respeito aos modelos mais simples, como o *FW1*, o uso do pré-treinamento contrastivo trouxe ganhos expressivos em WRA, com leve predominância do SimCLR, que obteve uma média global de WRA de 25,1%, frente a 24,3% com SwAV. As maiores elevações absolutas de WRA foram observadas em falantes severamente afetados, como M16, F03 e M12. Reduções médias de 23,9 pontos percentuais em WER e 15,9 em CER também foram observadas. Esses achados demonstram que modelos com menor capacidade paramétrica se beneficiam proporcionalmente mais do pré-treinamento CL, especialmente quando partem de desempenhos muito baixos, como no caso de falantes com fala severamente comprometida (ex: M07, F02, M16).

Comparando o desempenho entre os métodos SimCLR e SwAV, o modelo *FW2* com

SwAV obteve uma maior média global de WRA (27,95%), superando ligeiramente o SimCLR (27,6%), enquanto no *FW1* o SimCLR foi superior. No entanto, os resultados mostram que ambas as técnicas contrastivas foram eficazes de forma relativamente equilibrada no *FW2*, com o SwAV se destacando especialmente em falantes com inteligibilidade mais reduzida, e o SimCLR mantendo melhor desempenho médio em CER. Assim, observa-se que o SwAV proporciona parcialmente maior estabilidade semântica: a técnica oferece vantagens na inteligibilidade final, mas não necessariamente na redução de erros fonêmicos.

No que se refere ao impacto do grau de inteligibilidade dos falantes nos ganhos com a aplicação da Fase 2 (CL), observaram-se os seguintes resultados: falantes com inteligibilidade muito baixa — como M07, F02 e M01 — foram os que mais se beneficiaram do pré-treinamento, com ganhos percentuais superiores a 20 p.p. Em contraste, falantes com fala quase normotípica ou levemente alterada, como F05 e M14, apresentaram melhorias mais modestas, já que seus modelos base já alcançavam bons níveis de reconhecimento. Esses dados remetem à ideia de que o grau de severidade da disartria está diretamente relacionado à magnitude dos ganhos obtidos com técnicas de aprendizado contrastivo.

Além disso, os resultados reforçam que a arquitetura mais complexa (*FW2*) já parte de um desempenho mais elevado e, portanto, apresenta menores ganhos proporcionais — embora ainda obtenha ganhos absolutos expressivos. Isso mostra que arquiteturas simples e complexas podem ser igualmente beneficiadas por pre-treinamento contrastivo, mas de maneiras diferentes: a primeira por compensar limitações estruturais, a segunda por refinar representações já eficientes.

Ao cotejar os métodos contrastivos com os métodos de aumento de dados de forma supervisionada estudados no Capítulo 3, os métodos contrastivos apresentaram, em geral, um desempenho inferior, contudo, demonstraram uma tendência evolutiva se considerarmos melhor infraestrutura e parametrizações superiores.

Em síntese, este capítulo confirma a hipótese formulada para a segunda fase do trabalho e responde positivamente às questões que motivaram sua investigação. A combinação entre arquiteturas bem ajustadas e estratégias de pré-treinamento contrastivo mostra-se fundamental para o desenvolvimento de sistemas de ASR mais robustos, adaptáveis e inclusivos, capazes de lidar com a alta variabilidade observada na fala disártrica — especialmente em contextos clínicos de maior severidade articulatória.

Embora os resultados obtidos com o uso de aprendizado contrastivo nesta pesquisa não tenham superado, em termos absolutos, o desempenho do aprendizado supervisionado, os dados de evolução do treinamento sugerem fortemente que o modelo supervisionado já atinge um patamar de saturação com os dados disponíveis. Em outras palavras, as representações aprendidas via supervisão parecem esgotar sua capacidade de generalização frente ao conjunto limitado de amostras de fala disártrica, especialmente considerando a alta complexidade da arquitetura e a baixa variabilidade linguística do corpus.

Por outro lado, os modelos pré-treinados com técnicas de aprendizado contrastivo demonstraram um padrão contínuo de evolução, mesmo sob restrições significativas de recursos computacionais, número de épocas e tamanho de lote. Ao longo dos experimentos, observou-se que o aprendizado contrastivo ainda apresentava margem de melhoria, e possivelmente não atingiu sua capacidade plena devido às limitações impostas pelo ambiente experimental.

Essa tendência está em consonância com estudos fundamentais da área. No trabalho de Chen et al. (2020b), o SimCLR alcançou seus melhores desempenhos com tamanhos de lotes superiores a 4096 e mais de 800 épocas de treinamento, utilizando infraestrutura de larga escala com múltiplos TPUs. Da mesma forma, Caron et al. (2020), ao propor o SwAV, empregaram normalização de lote sincronizado, otimizadores sofisticados como LARS, e regimes de treino com centenas de épocas em bases massivas como ImageNet. Comparativamente, os experimentos realizados nesta dissertação foram conduzidos com tamanhos de lotes modestos (até 64), número reduzido de épocas (100) e infraestrutura limitada a GPUs unitárias de uso geral, o que impacta diretamente na eficiência de convergência dos métodos autossupervisionados.

Diante disso, os resultados aqui apresentados não devem ser interpretados como um limite superior do potencial do aprendizado contrastivo, mas sim como uma amostra inicial de sua aplicabilidade em contextos clínicos restritos, nos quais nem sempre é viável aplicar pipelines de pré-treinamento de larga escala. O fato de modelos como o *FW1*, mais simples e com menor capacidade representacional, terem apresentado ganhos proporcionais tão expressivos com o uso de contrastivo, reforça o valor dessas estratégias em cenários de baixa disponibilidade de dados rotulados e recursos limitados.

Assim, conclui-se que, embora o modelo supervisionado *FW2* ainda alcance o melhor desempenho final em determinadas métricas, o aprendizado contrastivo permanece como uma alternativa promissora e em evolução, especialmente para aplicações futuras que possam se beneficiar de pré-treinamentos mais longos, lotes maiores e melhorias na infraestrutura. A convergência entre estratégias supervisionadas e autossupervisionadas, em arquiteturas híbridas e ajustadas à realidade da fala disártrica, representa um caminho natural para futuras pesquisas na área.

Capítulo 5

Conclusão e Trabalhos Futuros

O objetivo deste estudo foi responder a duas perguntas complementares: primeiro, até que ponto aumentos de dados orientados pela patologia elevam o reconhecimento de fala disártrica e, segundo, em que medida o pré-treinamento contrastivo, alimentado por esses mesmos aumentos, gera representações capazes de superar os limites do treinamento puramente supervisionado.

Os experimentos do Capítulo 3 mostraram que o ruído aditivo somado à dilatação temporal, de modo geral, já promoviam melhorias na transcrição nos cenários de inteligibilidade moderada a baixa. Quando se acrescentou a Oclusão Espectral ("AP") esses ganhos se estenderam, mas de forma pontual: a "AP" foi a técnica isoladamente vencedora em quatro dos quinze falantes em *FW2* — M11, M07, M16 e M04 — e surgiu como a técnica mais eficaz em três falantes em *FW1* — F04, M11 e M04. Já a combinação de todas as perturbações ("APP") mostrou abrangência similar: foi a melhor escolha para seis falantes em FW1 — M11, M07, M16, F03, M01 e M04 — e para 3 falantes em FW2 — F03, F04 e M12. Em vozes quase normais, porém, os métodos estudados evidenciaram sinais de "sobreaquecimento" espectral, sinalizando que a suas aplicações devem ser dosadas. Importa notar que, embora "APP" tenha se destacado em alguns falantes severos, seu impacto sobre o reconhecimento fonêmico — especialmente medido por CER em *FW2* — foi negativo, sugerindo uma penalização fonética indevida por excesso de variabilidade, condicionada a uma profundidade maior da arquitetura.

No Capítulo 4, SimCLR e SwAV, em comparação com o método *baseline* - denominado BP - elevaram a WRA média global de 8,8 % para 27,8 % (-20 p.p. em WER). O modelo raso *FW1* saltou de 1,4 % para 25,2 % (+23,8 p.p.), validando que arquiteturas enxutas lucram proporcionalmente mais com o contraste. Em contrapartida, o modelo profundo *FW2* treinado *exclusivamente* de forma supervisionada permaneceu no topo absoluto —

36,1 % de WRA contra 27,9 % na versão contrastiva.

Esse ponto merece considerações relevantes. O pré-treinamento contrastivo foi executado em infraestrutura de GPU única, com lote limitado a 32 e apenas 100 épocas; estudos de referência indicam que SimCLR e SwAV ganham força em lotes maiores do que 4 k e centenas de épocas. Já a curva de aprendizado supervisionada do *FW2* se aproximou de saturação após 30 épocas adicionais: perdas estabilizadas e melhorias marginais inferiores a 0,2 p.p. a cada dez épocas. Logo, o contraste exibe espaço de crescimento inexplorado, enquanto a via puramente supervisionada indica esgotamento — situação que, em tese, se inverteria plausivelmente em regimes de lote e tempo mais generosos.

Consolida-se, portanto, um quadro em duas camadas: por um lado, os aumentos espectrais e temporais — inclusive a SO — revelam-se indispensáveis sempre que a coleta de dados patológicos é restrita; por outro, o contraste, embora ainda não tenha superado a melhor configuração supervisionada sob recursos limitados, apresenta trajetória ascendente e tende a apresentar melhorias adicionais à medida que os regimes de treinamento se aproximem das práticas de larga escala, como sugerido na literatura.

Trabalhos Futuros

O próximo passo natural do CADSR é torná-lo executável em dispositivos de borda — smartphones, próteses auditivas e microcontroladores com aceleradores DSP — sem sacrificar inteligibilidade. Pretende-se transferir o pré-treinamento contrastivo para variantes ultracompactas do HuBERT, como o Modelo Compactado de Representação de Fala Autossupervisionado (do inglês: *Distilled Hidden-Unit Bidirectional Encoder Representations from Transformers*) (DistilHuBERT) (CHANG et al., 2022) e o Aprendizado Leve e Configurável de Representações de Fala (do inglês: *Lightweight and Configurable Speech Representation Learning with Hidden-Unit BERT*) (LightHuBERT) (WANG et al., 2022), combinando distilação camada-a-camada, quantização a oito bits e poda estruturada dos blocos de atenção menos relevantes. O desempenho desses modelos será aferido em termos de latência e consumo energético em hardware real; a meta prática consiste em manter transcrições abaixo de 150 ms por frase — limite percebido como "instantâneo" segundo Nielsen (1993) e já atingido no Gboard offline ($\approx$ 80 ms) (CHEN et al., 2019) e nas *EdgeSpeechNets* em aparelhos de baixo custo (LIN et al., 2018) — com potência inferior a 1 W, patamar compatível com processadores de voz de mili-watt como o Knowles IA8201 (Knowles Corporation, 2021) ou o Syntiant NDP120 (Syntiant Corporation, 2021). Esses valores garantem uso contínuo em terapia domiciliar sem dependência de nuvem.

Embora os ganhos quantitativos sejam robustos, é essencial explicar por que o sistema falha ou acerta cada palavra. A incorporação de camadas de Inteligência Artificial Explicável Aplicada a Áudio (do inglês: *Audio Explainable Artificial Intelligence*) (Audio-XAI) — gradientes espectrais e Mapeamento de Ativação por Gradiente para Classes (do in-

glês: *Gradient-weighted Class Activation Mapping*) (Grad-CAM) adaptado, conforme a síntese de Akman et al. (2024), além do Explicações Interpretáveis e Auditivas para Modelos de Áudio (do inglês: *Audio-based Local Interpretable Model-agnostic Explanations*) (AudioLIME) de Haunschmid, Manilow e Widmer (2020) — deverá revelar em tempo real quais bandas de frequência e janelas temporais sustentam cada decisão. Mapear essas regiões possibilitará oferecer *feedback* visual e auditivo ao fonoaudiólogo, que poderá relacionar "pontos quentes" a fraquezas articulatórias específicas, e ainda replicar tais perturbações em fala normotípica abundante, criando gêmeos patológicos sintéticos que ampliem o treinamento sem exigir novas coletas de pacientes.

A escala do CADSR pode ser ampliada ao pré-treinar, em larga escala, com megacorpora de fala normotípica como o Mozilla Common Voice (ARDILA et al., 2020), o LibriLight (KAHN et al., 2020) e o VoxPopuli (WANG et al., 2021), aplicando neles as perturbações espectrais mais suscetíveis à disartria. Paralelamente, pretendemos incorporar séries longitudinais de pacientes acompanhados durante a reabilitação, de modo que o sistema não apenas reconheça, mas rastreie a evolução articulatória individual em múltiplos idiomas e dispositivos de borda.

Outro desdobramento relevante envolve a exploração de grandes Modelos de Linguagem em Grande Escala (do inglês: *Large Language Models*) (LLMs), com suporte a áudio, capazes de receber simultaneamente o sinal de fala disártrica e as transcrições parciais produzidas pelo pipeline de ASR. Esses modelos poderiam atuar como mecanismos de inferência contextual, reconstituindo a palavra-alvo mesmo quando a inteligibilidade do sinal é muito baixa. Tal estratégia incluiria tanto o re-ranqueamento de hipóteses em um vocabulário fechado (como o da UA-Speech) quanto a "reparação" do áudio por meio de síntese assistida, ampliando a robustez e a utilidade prática em situações de comunicação real.

Como desdobramento deste trabalho, propõe-se ainda a investigação de estratégias avançadas de aumento de dados no processo de *downstream*, com o objetivo de aprimorar a robustez e a generalização dos modelos em tarefas de reconhecimento de fala disártrica. Entre as técnicas a serem exploradas, destacam-se o *jitter* (variação temporal irregular entre ciclos sucessivos da fala), o *shimmer* (variação na amplitude do sinal de voz) e o *pitch shift* (alteração sistemática da frequência fundamental). Essas transformações são particularmente relevantes para simular padrões de variabilidade típicos da fala disártrica, favorecendo a adaptação dos modelos às diferentes manifestações da condição.

Além disso, serão incorporadas abordagens de aprendizado contrastivo em nível fonêmico, com o intuito de fortalecer a capacidade do modelo em distinguir padrões acústicos sutis entre fonemas semelhantes. A intenção é tornar o modelo mais atento às sutilezas dos sons da fala, como a diferença entre fonemas semelhantes (/p/ e /b/, por exemplo), que muitas vezes se confundem em falas atípicas. Ao trabalhar contrastes nesse nível mais fino, espera-se que o sistema produza transcrições mais fiéis mesmo em condições de

articulação severamente comprometida.

A combinação dessas técnicas poderá contribuir significativamente para a evolução dos modelos de reconhecimento de fala assistida, ampliando sua eficácia em contextos de alta variabilidade vocal e escassez de dados rotulados. Estudos futuros também poderão considerar a integração dessas abordagens com mecanismos de atenção e adaptação personalizada ao falante, fortalecendo ainda mais o potencial de aplicações clínicas e comunicativas.

Além dos avanços metodológicos apresentados, é importante refletir sobre aspectos práticos e éticos relacionados ao uso de sistemas de reconhecimento automático de fala disártrica. Embora os resultados obtidos com as arquiteturas propostas indiquem ganhos significativos em inteligibilidade, ainda persiste o desafio da variabilidade intrafalantes — especialmente em casos de doenças progressivas como a ELA, em que o padrão de fala pode se deteriorar com o tempo. Nesse contexto, futuros estudos poderão explorar estratégias de adaptação contínua, permitindo que o sistema acompanhe a evolução da fala do paciente.

Adicionalmente, a viabilidade de uso clínico em dispositivos embarcados — como próteses auditivas ou sistemas móveis — requer atenção especial. Os experimentos indicaram que arquiteturas mais compactas, como o modelo *FW1*, com pré-treinamento contrastivo, podem oferecer um bom balanceamento entre desempenho e custo computacional, viabilizando aplicações em contextos com restrições de recursos computacionais.

Por fim, o uso de ASR para pacientes com disartria envolve considerações éticas fundamentais, incluindo a proteção de dados sensíveis, a garantia de consentimento informado e a prevenção de usos indevidos em sistemas comerciais. Tais aspectos deverão ser cuidadosamente incorporados em futuros desdobramentos desta pesquisa, especialmente na transição para sistemas aplicados em ambientes clínicos ou domésticos.

Referências

ABAYOMI-ALLI, O. O. et al. Data augmentation and deep learning methods in sound classification: A systematic review. **Electronics**, v. 11, p. 3795, 2022.

AKMAN, B. et al. Audio explainable artificial intelligence: A review. **iScience**, v. 27, n. 2, p. 107148, 2024. Disponível em: <https://www.sciencedirect.com/science/article/pii/S2589004223019959>.

ALLEN, J. B.; RABINER, L. R. Short term spectral analysis, synthesis, and modification by discrete fourier transform. **IEEE Transactions on Acoustics, Speech, and Signal Processing**, IEEE, v. 25, n. 3, p. 235–238, 1977.

ALMADHOR, A. et al. E2e-dasr: End-to-end deep learning-based dysarthric automatic speech recognition. **Expert Systems with Applications**, Pergamon, v. 222, p. 119797, 7 2023. ISSN 0957-4174.

ARDILA, R. et al. Common voice: A massively-multilingual speech corpus. In: **Proceedings of LREC 2020**. [s.n.], 2020. p. 4218–4222. Disponível em: <https://commonvoice.mozilla.org/en/datasets>.

BAEVSKI, A. et al. wav2vec 2.0: A framework for self-supervised learning of speech representations. **Advances in Neural Information Processing Systems**, v. 33, p. 12449–12460, 2020.

BAHDANAU, D.; CHO, K.; BENGIO, Y. Neural machine translation by jointly learning to align and translate. In: **International Conference on Learning Representations (ICLR'15)**. [S.l.: s.n.], 2015.

BALL, M. et al. **Manual of Clinical Phonetics**. [S.l.]: Taylor & Francis, 2011.

CADET, X. F. et al. A study on the impact of self-supervised learning on automatic dysarthric speech assessment. **Journal Name**, Publisher, Volume, n. Number, p. Pages, 2024.

CARON, M. et al. Unsupervised learning of visual features by contrasting cluster assignments. In: **Advances in Neural Information Processing Systems**. [S.l.: s.n.], 2020. v. 33, p. 9912–9924.

CHAN, W. et al. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In: **Proc. ICASSP**. [s.n.], 2016. p. 4960–4964. Disponível em: <https://arxiv.org/abs/1508.01211>.

_____. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In: **ICASSP**. [S.l.: s.n.], 2016. p. 4960–4964.

CHANG, W.-C. et al. Distilhubert: Speech representation learning by layer-wise distillation of hidden-unit BERT. In: **Proceedings of Interspeech 2022**. [s.n.], 2022. p. 296–300. Disponível em: <https://www.isca-archive.org/interspeech_2022/chang22b_interspeech.html>.

CHEN, Q. et al. **An 80 MB On-device Speech Recognizer**. 2019. Google AI Blog, 12 mar. 2019. Disponível em: <https://ai.googleblog.com/2019/03/an-80mb-on-device-speech-recognizer.html>.

CHEN, T. et al. A simple framework for contrastive learning of visual representations. **International Conference on Machine Learning (ICML)**, 2020. Disponível em: <https://arxiv.org/abs/2002.05709>.

_____. A simple framework for contrastive learning of visual representations. In: PMLR. **International Conference on Machine Learning**. [S.l.], 2020. p. 1597–1607.

_____. A simple framework for contrastive learning of visual representations. In: **International Conference on Machine Learning**. [S.l.: s.n.], 2020.

CHOLLET, F. Xception: Deep learning with depthwise separable convolutions. In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. [S.l.: s.n.], 2017. p. 1251–1258.

CHOROWSKI, J. K. et al. Attention-based models for speech recognition. In: **Advances in Neural Information Processing Systems (NeurIPS'15)**. [S.l.: s.n.], 2015. p. 577–585.

COHEN, L. **Time–Frequency Analysis**. [S.l.]: Prentice Hall, 1995.

DARLEY, F. L.; ARONSON, A. E.; BROWN, J. R. Differential diagnostic patterns of dysarthria. **Journal of speech and hearing research**, American Speech-Language-Hearing Association, v. 12, p. 246–269, 1969. ISSN 00224685. Disponível em: <https://pubs.asha.org/doi/epdf/10.1044/jshr.1202.246>.

DAVIS, S. B.; MERMELSTEIN, P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. **IEEE Transactions on Acoustics, Speech, and Signal Processing**, IEEE, v. 28, n. 4, p. 357–366, 1980.

DEVLIN, J. et al. Bert: Pre-training of deep bidirectional transformers for language understanding. **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, Association for Computational Linguistics, p. 4171–4186, 2019.

_____. Bert: Pre-training of deep bidirectional transformers for language understanding. In: **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics**. [S.l.: s.n.], 2019.

DOSOVITSKIY, A. et al. An image is worth 16x16 words: Transformers for image recognition at scale. In: **International Conference on Learning Representations (ICLR'21)**. [S.l.: s.n.], 2021.

DUMOULIN, V.; VISIN, F. A guide to convolution arithmetic for deep learning. **arXiv preprint arXiv:1603.07285**, 2016.

FRACASSI, A. S. et al. Adjustment to the portuguese and application to patients with parkinson's disease of protocol within central origin dysarthrias' assessment. **Rev. CEFAC**, v. 13, n. 16, p. 1056–1065, 2011.

FU, J.; ZHENG, H.; MEI, T. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In: **IEEE Conference on Computer Vision and Pattern Recognition (CVPR'17)**. [S.l.: s.n.], 2017. p. 4438–4446.

GABOR, D. Theory of communication. **Journal of the IEE**, v. 93, n. 26, p. 429–457, 1946.

GOLDWATER, S.; JURAFSKY, D.; MANNING, C. D. Which words are hard to recognize? prosodic, lexical, and disfluency factors that increase speech recognition error rates. **Speech Communication**, v. 52, n. 3, p. 181–200, 2011.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep Learning**. MIT Press, 2016. Disponível em: <https://www.deeplearningbook.org/>.

_____. **Deep learning**. [S.l.]: MIT press, 2016.

GRACELLI, R.; ALMEIDA, J. Exploring alternative data augmentation methods in dysarthric automatic speech recognition. In: IEEE. **Proceedings of the IEEE 37th International Symposium on Computer-Based Medical Systems (CBMS)**. New York, NY, USA: IEEE, 2024. p. 1–6.

GRAVES, A. Sequence transduction with recurrent neural networks. **arXiv preprint arXiv:1211.3711**, 2012. Disponível em: <https://arxiv.org/abs/1211.3711>.

GRAVES, A. et al. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In: **Proceedings of the 23rd International Conference on Machine Learning (ICML)**. [S.l.]: ACM, 2006. p. 369–376.

_____. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. **Proceedings of the 23rd International Conference on Machine Learning (ICML)**, p. 369–376, 2006. Disponível em: <https://www.cs.toronto.edu/~graves/icml_2006.pdf>.

HANNUN, A. et al. Deep speech: Scaling up end-to-end speech recognition. **arXiv preprint arXiv:1412.5567**, 2014. Disponível em: <https://arxiv.org/abs/1412.5567>.

HAUNSCHMID, V.; MANILOW, E.; WIDMER, G. audioLIME: Listenable explanations using source separation. **arXiv preprint arXiv:2008.00582**, 2020. Disponível em: <https://arxiv.org/abs/2008.00582>.

HE, K. et al. Deep residual learning for image recognition. In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. [S.l.: s.n.], 2016. p. 770–778.

HSIEH, I.-T.; WU, C.-H. Dysarthric speech recognition using curriculum learning and articulatory feature embedding. In: ISCA. **Proceedings of Interspeech 2024**. 2024. Disponível em: <https://www.isca-archive.org/interspeech_2024/hsieh24_interspeech.pdf>.

HSU, W.-N. et al. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. In: **Proceedings of the 2021 IEEE/ACM International Conference on Acoustics, Speech and Signal Processing (ICASSP)**. [S.l.]: IEEE, 2021. p. 6413–6417.

_____. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. **IEEE/ACM Transactions on Audio, Speech, and Language Processing**, IEEE, v. 29, p. 3451–3460, 2021.

_____. Melhubert: Enhancing self-supervised speech representation learning with mel-spectrogram features. **arXiv preprint arXiv:2106.07447**, 2021. Disponível em: <https://arxiv.org/abs/2106.07447>.

HU, A.; PHADNIS, D.; SHAHAMIRI, S. R. Generating synthetic dysarthric speech to overcome dysarthria acoustic data scarcity. **Journal of Ambient Intelligence and Humanized Computing**, Springer, v. 14, p. 6751–6768, 2023. Disponível em: <https://doi.org/10.1007/s12652-021-03542-w>.

HU, C. et al. Self-supervised speech representations for dysarthric speech recognition. **IEEE/ACM Transactions on Audio, Speech, and Language Processing**, 2024.

HU, I. et al. Gather-excite: Exploiting feature context in convolutional neural networks. In: **Advances in Neural Information Processing Systems (NeurIPS'18)**. [S.l.: s.n.], 2018. p. 9423–9433.

HUANG, G. et al. Densely connected convolutional networks. In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. [S.l.: s.n.], 2017. p. 4700–4708.

IOFFE, S.; SZEGEDY, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: PMLR. **International conference on machine learning**. [S.l.], 2015. p. 448–456.

IRSHAD, U. et al. Utran-dsr: A novel transformer-based model using feature enhancement for dysarthric speech recognition. **EURASIP Journal on Audio, Speech, and Music Processing**, Springer, v. 2024, n. 54, 2024.

ITO, K.; JOHNSON, L. **The LJ Speech Dataset**. 2017. <https://keithito.com/LJ-Speech-Dataset/>.

JIN, Z. et al. Personalized adversarial data augmentation for dysarthric and elderly speech recognition. **arXiv preprint arXiv:2205.06445**, 2022. Disponível em: <https://arxiv.org/abs/2205.06445>.

KAHN, J. et al. Libri-Light: A benchmark for ASR with almost no supervision. In: **Proceedings of ICASSP 2020**. [s.n.], 2020. p. 7669–7673. Disponível em: <https://arxiv.org/abs/2004.11572>.

KESHVARI, M. et al. Enhancement of dysarthric speech using contrastive speaker embeddings. In: **ICASSP**. [S.l.: s.n.], 2024.

KIM, H.-Y.; HASEGAWA-JOHNSON, M.; PERLMAN, E. Dysarthric speech database for universal access research. In: **INTERSPEECH**. [S.l.: s.n.], 2008. p. 1741–1744.

KIM, S.; HORI, T.; WATANABE, S. Joint ctc-attention based end-to-end speech recognition using multi-task learning. In: **Proceedings of ICASSP 2017**. [s.n.], 2017. p. 4835–4839. Disponível em: <https://arxiv.org/abs/1609.06773>.

KINGMA, D. P.; WELLING, M. Auto-encoding variational bayes. **arXiv preprint arXiv:1312.6114**, 2014. Disponível em: <https://arxiv.org/abs/1312.6114>.

Knowles Corporation. **IA8201 Product Brief**. 2021. Documento técnico. Disponível em: <https://www.knowles.com/docs/default-source/default-document-library/knowles-ia8201-product-brief.pdf>.

KO, T. et al. Audio augmentation for speech recognition. In: **Proceedings of Interspeech 2015**. [s.n.], 2015. p. 3586–3589. Disponível em: <https://www.isca-speech.org/archive/interspeech_2015/ko15_interspeech.html>.

KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. Imagenet classification with deep convolutional neural networks. In: **Advances in neural information processing systems**. [S.l.: s.n.], 2012. v. 25, p. 1097–1105.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. **Nature**, Nature Publishing Group, v. 521, n. 7553, p. 436–444, 2015.

LEE, J. et al. Dynamic phoneme-level contrastive learning for dysarthric asr. **IEEE Transactions on Neural Networks and Learning Systems**, 2025.

LEUNG, W.-Z.; LI, J.-H.; LI, M. Training data augmentation for dysarthric automatic speech recognition by text-to-dysarthric-speech synthesis. **arXiv preprint arXiv:2406.08568**, 2024. Disponível em: <https://arxiv.org/abs/2406.08568>.

LIN, Z. et al. EdgeSpeechNets: Highly efficient deep neural networks for speech recognition on the edge. **arXiv preprint arXiv:1810.08559**, 2018. Disponível em: <https://arxiv.org/abs/1810.08559>.

Ministério da Ciência, Tecnologia e Inovação. **Fator de emissão de CO na geração de energia elétrica no Brasil em 2023 é o menor em 12 anos**. 2024. Acesso em 15 maio 2025. Disponível em: https://www.gov.br/mcti/pt-br/acompanhe-o-mcti/noticias/2024/02/fator-de-emissao-de-co2-na-geracao-de-energia-eletrica-no-brasil-em-2023-e-o-menor-em-12-anos.

MNIH, V. et al. Recurrent models of visual attention. In: **Advances in Neural Information Processing Systems (NeurIPS'14)**. [S.l.: s.n.], 2014. p. 2204–2212.

NIELSEN, J. **Response Time Limits**. 1993. Acesso em 13 jun. 2025. Disponível em: <https://www.nngroup.com/articles/response-times-3-important-limits/>.

NVIDIA. **NVIDIA QuartzNet ASR Benchmark**. 2020. Acesso em 15 maio 2025. Disponível em: <https://developer.nvidia.com/blog/quartznet-a-fast-and-accurate-automatic-speech-recognition-model/>.

OORD, A. v. d.; LI, Y.; VINYALS, O. Representation learning with contrastive predictive coding. **arXiv preprint arXiv:1807.03748**, 2018.

ORTIZ, K. Z. **Distúrbios neurológicos adquiridos : fala e deglutição**. Manole, 2010. 510 p. ISBN 9788520428856. Disponível em: <https://ria.ufrn.br/jspui/handle/123456789/2783>.

PARK, D. S. et al. **SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition**. 2019. Disponível em: <https://arxiv.org/abs/1904.08779>.

PORTALETE, C. R. et al. Tratamento motor da fala na disartria flácida: um estudo de caso. **Audiology - Communication Research**, Academia Brasileira de Audiologia, v. 24, p. e2118, 9 2019. ISSN 2317-6431. Disponível em: <https://www.scielo.br/j/acr/a/wgNNz7ghTqwP5yJ7Q4WdNVw/>.

RAVANELLI, M. et al. Multi-task self-supervised learning for robust speech recognition. **arXiv preprint arXiv:2001.09239**, 2020.

SAAD, A.; AHMED, J.; ELARABY, A. Classification of bird sound using high-and low-complexity convolutional neural networks. **Technical Journal of Sound Classification**, Technical College of Management, Kufa, Al Furat Alawsat Technical University, 2024.

SCHNEIDER, S. et al. Wav2vec: Unsupervised pre-training for speech recognition. In: **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)**. [S.l.: s.n.], 2019. p. 3846–3850.

SCHUSTER, M.; SCHULLER, B.; WENINGER, F. Automatic recognition of dysarthric speech: A review. **Speech Communication**, v. 110, p. 1–17, 2019.

SCHWARTZ, R. et al. Green ai. **Communications of the ACM**, v. 63, n. 12, p. 54–63, 2020.

SHAHAMIRI, S. R. Speech vision: An end-to-end deep learning-based dysarthric automatic speech recognition system. **IEEE Transactions on Neural Systems and Rehabilitation Engineering**, v. 29, p. 852–861, 2021.

SHAHAMIRI, S. R.; LAL, V.; SHAH, D. Dysarthric speech transformer: A sequence-to-sequence dysarthric speech recognition system. **IEEE Transactions on Neural Systems and Rehabilitation Engineering**, v. 31, p. 3407–3416, 2023.

SOLEYMANPOUR, S. et al. Accurate recognition of dysarthric speech via synthetic data augmentation. In: **Proc. Interspeech**. [S.l.: s.n.], 2023.

STRUBELL, E.; GANESH, A.; MCCALLUM, A. Energy and policy considerations for deep learning in nlp. In: **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**. [S.l.: s.n.], 2019. p. 3645–3650.

STUMPF, A. et al. Speaker-independent dysarthria severity classification with latent regularization. In: **Proc. Interspeech**. [S.l.: s.n.], 2024.

Syntiant Corporation. **NDP120 Neural Decision Processor Runs Multiple AI Models for Under 1 mW**. 2021. Disponível em: <https://syntiant.com/news/ndp120>.

VACHHANI, B.; BHAT, C.; KOPPARAPU, S. K. Data augmentation using healthy speech for dysarthric speech recognition. In: **Annual Conference of the International Speech Communication Association (Interspeech'18)**. [S.l.: s.n.], 2018. p. 471–475.

VINOTHA, R. et al. Enhancing dysarthric speech recognition through sepformer and hierarchical attention network models with multistage transfer learning. **Scientific Reports**, Nature Publishing Group, v. 14, n. 1, p. 29455, 2024.

WANG, C. et al. VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In: **Proceedings of ACL 2021**. [s.n.], 2021. p. 993–1003. Disponível em: <https://arxiv.org/abs/2101.00390>.

WANG, F. et al. Residual attention network for image classification. In: **IEEE Conference on Computer Vision and Pattern Recognition (CVPR'17)**. [S.l.: s.n.], 2017. p. 3156–3164.

WANG, P.-H. et al. Lighthubert: Lightweight and configurable speech representation learning with once-for-all hidden-unit BERT. In: **Proceedings of Interspeech 2022**. [s.n.], 2022. p. 3618–3622. Disponível em: <https://www.isca-archive.org/interspeech_2022/wang22i_interspeech.html>.

WANG, S. et al. Enhancing dysarthric speech recognition for unseen speakers via prototype-based adaptation. **arXiv preprint arXiv:2407.18461v1**, 2024. Disponível em: <https://arxiv.org/abs/2407.18461v1>.

WATANABE, S. et al. Hybrid ctc/attention architecture for end-to-end speech recognition. **IEEE Journal of Selected Topics in Signal Processing**, v. 11, n. 8, p. 1240–1253, 2017. Disponível em: <https://arxiv.org/abs/1706.02737>.

WEI, S. et al. A comparison on data augmentation methods based on deep learning for audio classification. **Journal of Physics: Conference Series**, v. 1453, p. 012085, 2020.

WOO, S. et al. Cbam: Convolutional block attention module. In: **European Conference on Computer Vision (ECCV'18)**. [S.l.: s.n.], 2018. p. 3–19.

WU, J.; LI, X.; ZHOU, Y. An infrared image detection of power equipment based on super-resolution reconstruction and YOLOv4. **The Journal of Engineering**, Wiley Online Library, v. 2022, n. 10, p. 1006–1016, 2022. Disponível em: <https://arxiv.org/abs/1603.07285>.

XU, J. et al. Show, attend and tell: Neural image caption generation with visual attention. In: **International Conference on Machine Learning (ICML'15)**. [S.l.: s.n.], 2015. p. 2048–2057.

YORKSTON, K. et al. **Management of Speech and Swallowing in Degenerative Diseases**. [S.l.]: Pro-Ed, 2010.

Apêndices

APÊNDICE A

Custo Computacional e Consumo Energético

Além das métricas tradicionais de desempenho, como WER e CER, também avaliamos o custo computacional e o impacto ambiental associado ao treinamento completo dos modelos propostos (*upstream* + *downstream*). Esta análise torna-se cada vez mais relevante diante das crescentes preocupações com a sustentabilidade computacional no campo do aprendizado profundo.

O modelo *FW1* apresentou um total de 4.534,4 MFLOPs (2.267,2 × 2) e uma potência média da GPU de aproximadamente 77,52 Watts (38,76 × 2). Por outro lado, o modelo *FW2*, embora tenha exigido menos operações (4.294,38 MFLOPs), apresentou um consumo energético superior: 85,94 Watts (42,97 × 2).

Considerando o tempo total de treinamento — 100 épocas de *upstream* com duração de 25 minutos cada, e 100 épocas de *downstream* com duração de 30 segundos — estimamos aproximadamente 42,5 horas de uso computacional por modelo. Com base no fator médio nacional de emissão de CO_2 de 0,0385 kg/kWh, conforme publicado pelo MCTI em 2024 (Ministério da Ciência, Tecnologia e Inovação, 2024), as estimativas de consumo energético e emissão de carbono são apresentadas na Tabela 8.

Tabela 8 – Resumo do custo computacional, energético e ambiental dos modelos treinados (fator de emissão brasileiro de 2024)

Modelo	**FLOPs Totais (MFLOPs)**	**Potência GPU (W)**	**Energia (kWh)**	**CO_2 Est. (kg)**
FW1	4.534,40	77,52	3,296	0,127
FW2	4.294,38	85,94	3,659	0,141

Esses resultados revelam que a eficiência energética nem sempre está diretamente cor-

relacionada à quantidade de operações executadas, sendo também influenciada por fatores como a estrutura do modelo, o grau de utilização da GPU e as estratégias de treinamento. Em particular, o modelo *FW1* demonstrou melhor equilíbrio entre desempenho e consumo energético.

Mesmo considerando o treinamento completo, as emissões estimadas de CO_2 ficaram abaixo de 150 gramas por modelo, demonstrando alta eficiência ambiental em comparação com grandes modelos de referência da literatura.

Para contextualizar, a Tabela 9 apresenta dados de consumo energético estimado para modelos amplamente utilizados na literatura, com base em fontes primárias e cálculos derivados.

Tabela 9 – Comparativo com modelos de referência da literatura

Modelo	**FLOPs (estimado)**	**Energia (kWh)**	**Potência GPU típica**
BERT-Large (Strubell et al.)	$> 10^{20}$	∼1500	8×300W
SimCLR ResNet50 (Chen et al.)	$> 10^{20}$	∼400	8×250W
DeepSpeech (Mozilla/NVIDIA)	10^{16}–10^{18}	∼50–150	1×200W
FW1 (nosso)	$4{,}5 \times 10^9$	3,296	1×40W
FW2 (nosso)	$4{,}3 \times 10^9$	3,659	1×43W

Para reforçar a confiabilidade dos dados, a Tabela 10 resume as fontes utilizadas, os métodos de estimativa adotados e o fator de emissão original aplicado para os modelos de terceiros. Apenas os modelos *FW1* e *FW2* foram convertidos diretamente com base no fator nacional de 2024.

Tabela 10 – Fontes, métodos de estimativa e padronização de emissão de CO_2 (valores originais)

Modelo	**Fonte principal**	**Tipo de dado**	**Observações sobre conversão e confiabilidade**
BERT-Large	Strubell et al. (STRUBELL; GANESH; MCCALLUM, 2019)	Medição empírica (1500 kWh)	Fator de emissão dos EUA (0,432 kg/kWh); sem conversão nesta tabela.
SimCLR (ResNet50)	Chen et al. (CHEN et al., 2020b), Schwartz et al. (SCHWARTZ et al., 2020)	Estimativa indireta (400 kWh)	Baseado em uso de TPUv3-32; valor mantido como original.
DeepSpeech (QuartzNet)	Hannun et al. (HANNUN et al., 2014), NVIDIA (NVIDIA, 2020)	Medições práticas (50–150 kWh)	Estimativas consistentes com uso de GPU 200W por 2–3 dias; sem conversão nesta tabela.
FW1 / FW2 (nosso)	Experimento atual	Medido via tempo × potência	Consumo convertido com fator nacional de emissão (0,0385 kg/kWh) (Ministério da Ciência, Tecnologia e Inovação, 2024).

Para permitir uma comparação justa com os modelos nacionais, a Tabela 11 apresenta as emissões convertidas para o fator de emissão médio brasileiro (0,0385 kg CO_2/kWh), evidenciando que mesmo os modelos internacionais mais eficientes (como o DeepSpeech) apresentam emissões entre 15 e 45 vezes maiores do que os modelos *FW1* e *FW2*.

Tabela 11 – Comparativo de consumo energético e emissão de CO_2 com fator brasileiro de 0,0385 kg/kWh

Modelo	**Energia (kWh)**	**CO_2 EUA (kg)**	**CO_2 Brasil (kg)**
BERT-Large (Strubell et al.)	1500	648,00	57,75
SimCLR ResNet50 (estimado)	400	172,80	15,40
DeepSpeech (benchmark)	50–150	21,60–64,80	1,93–5,78
FW1 (nosso)	3,296	*1,42*	0,127
FW2 (nosso)	3,659	*1,58*	0,141

Como se observa, os modelos utilizados neste trabalho estão diversas ordens de magnitude abaixo das arquiteturas de grande porte em termos de demanda computacional e impacto ambiental. Isso demonstra que, mesmo considerando o treinamento completo (pré-treinamento e ajuste fino), os modelos *FW1* e *FW2* mantêm excelente eficiência energética, sendo particularmente adequados para contextos de pesquisa com recursos computacionais limitados.

APÊNDICE B

Tabelas

Tabela 12 – Comparação dos métodos CL e BP nos modelos FW1 e FW2

Modelo	**Falante**	WRA_{CL}	WRA_{BP}	**Δ WRA (p.p.)**
FW1-SimCLR	M04	0,015	0,003	1,2
FW1-SimCLR	F03	0,123	0,008	11,5
FW1-SimCLR	M12	0,076	0,005	7,1
FW1-SimCLR	M01	0,044	0,001	4,3
FW1-SimCLR	M07	0,361	0,016	34,5
FW1-SimCLR	F02	0,366	0,015	35,1
FW1-SimCLR	M16	0,202	0,006	19,6
FW1-SimCLR	M05	0,274	0,015	25,9
FW1-SimCLR	M11	0,208	0,010	19,8
FW1-SimCLR	F04	0,207	0,004	20,3
FW1-SimCLR	M09	0,406	0,031	37,5
FW1-SimCLR	M14	0,318	0,015	30,3
FW1-SimCLR	M10	0,463	0,038	42,5
FW1-SimCLR	M08	0,335	0,013	32,2
FW1-SimCLR	F05	0,374	0,024	35,0
FW1-SwAV	M04	0,0102	0,003	0,7
FW1-SwAV	F03	0,1541	0,008	14,6
FW1-SwAV	M12	0,0771	0,005	7,2

Continua na próxima página

Modelo	**Falante**	WRA_{CL}	WRA_{BP}	**Δ WRA (p.p.)**
FW1-SwAV	M01	0,0314	0,001	3,0
FW1-SwAV	M07	0,3216	0,016	30,6
FW1-SwAV	F02	0,2908	0,014	27,7
FW1-SwAV	M16	0,2281	0,006	22,2
FW1-SwAV	M05	0,2639	0,015	24,9
FW1-SwAV	M11	0,1307	0,010	12,1
FW1-SwAV	F04	0,1496	0,004	14,6
FW1-SwAV	M09	0,4090	0,031	37,8
FW1-SwAV	M14	0,3064	0,015	29,1
FW1-SwAV	M10	0,3664	0,038	32,8
FW1-SwAV	M08	0,3350	0,013	32,2
FW1-SwAV	F05	0,4471	0,024	42,3
FW2-SimCLR	M04	0,050	0,011	3,9
FW2-SimCLR	F03	0,239	0,057	18,2
FW2-SimCLR	M12	0,153	0,022	13,1
FW2-SimCLR	M01	0,162	0,024	13,8
FW2-SimCLR	M07	0,352	0,090	26,2
FW2-SimCLR	F02	0,335	0,080	25,5
FW2-SimCLR	M16	0,215	0,039	17,6
FW2-SimCLR	M05	0,250	0,079	17,1
FW2-SimCLR	M11	0,212	0,067	14,5
FW2-SimCLR	F04	0,235	0,064	17,1
FW2-SimCLR	M09	0,382	0,224	15,8
FW2-SimCLR	M14	0,276	0,088	18,8
FW2-SimCLR	M10	0,499	0,250	24,9
FW2-SimCLR	M08	0,351	0,117	23,4
FW2-SimCLR	F05	0,428	0,108	32,0
FW2-SwAV	M04	0,027	0,011	1,6
FW2-SwAV	F03	0,262	0,057	20,5
FW2-SwAV	M12	0,144	0,022	12,2
FW2-SwAV	M01	0,135	0,024	11,1
FW2-SwAV	M07	0,368	0,090	27,8
FW2-SwAV	F02	0,360	0,080	28,0
FW2-SwAV	M16	0,277	0,039	23,8
FW2-SwAV	M05	0,309	0,079	23,0

Continua na próxima página

Modelo	**Falante**	WRA_{CL}	WRA_{BP}	**Δ WRA (p.p.)**
FW2-SwAV	M11	0,230	0,067	16,3
FW2-SwAV	F04	0,235	0,064	17,1
FW2-SwAV	M09	0,393	0,224	16,9
FW2-SwAV	M14	0,350	0,088	26,2
FW2-SwAV	M10	0,435	0,250	18,5
FW2-SwAV	M08	0,287	0,117	17,0
FW2-SwAV	F05	0,380	0,108	27,2

Tabela 13 – Comparação de CER e WER para os modelos FW1 e FW2 com e sem upstream, incluindo SwAV (ordem por inteligibilidade)

Modelo	**Falante**	**CER**$_{BP}$	**CER**$_{CL}$	**ΔCER**	**WER**$_{BP}$	**WER**$_{CL}$	**ΔWER**
FW1-SimCLR	M04	1,0354	0,9040	-12,69	0,9992	0,9851	-1,41
FW1-SwAV	M04	1,0354	0,9418	-9,36	0,9992	0,9898	-0,94
FW1-SimCLR	F03	0,8630	0,7793	-9,71	0,9916	0,8768	-11,59
FW1-SwAV	F03	0,8630	0,7506	-11,24	0,9916	0,8459	-14,57
FW1-SimCLR	M12	0,9279	0,9456	1,91	0,9954	0,9242	-7,15
FW1-SwAV	M12	0,9279	0,9574	2,95	0,9954	0,9229	-7,25
FW1-SimCLR	M01	0,9694	0,9054	-6,60	1,0065	0,9556	-5,06
FW1-SwAV	M01	0,9694	0,9478	-2,16	1,0065	0,9686	-3,79
FW1-SimCLR	M07	0,9183	0,6830	-25,61	0,9843	0,6387	-35,12
FW1-SwAV	M07	0,9183	0,7443	-17,40	0,9843	0,6784	-30,59
FW1-SimCLR	F02	0,8769	0,6716	-23,42	0,9849	0,6342	-35,63
FW1-SwAV	F02	0,8769	0,7759	-10,10	0,9849	0,7092	-27,57
FW1-SimCLR	M16	0,9267	0,8031	-13,34	0,9941	0,7980	-19,72
FW1-SwAV	M16	0,9267	0,7915	-13,52	0,9941	0,7719	-22,22
FW1-SimCLR	M05	0,8953	0,6997	-21,84	0,9910	0,7261	-26,72
FW1-SwAV	M05	0,8953	0,6865	-20,88	0,9910	0,7361	-25,49
FW1-SimCLR	M11	0,8871	0,8378	-5,56	0,9902	0,7922	-20,01
FW1-SwAV	M11	0,8871	0,8341	-5,30	0,9902	0,8693	-12,09
FW1-SimCLR	F04	0,9687	0,8146	-15,91	0,9961	0,7933	-20,35
FW1-SwAV	F04	0,9687	0,7985	-17,02	0,9961	0,8504	-14,57
FW1-SimCLR	M09	0,8827	0,6183	-29,93	0,9692	0,5944	-38,70
FW1-SwAV	M09	0,8827	0,6620	-22,07	0,9692	0,5910	-37,82
FW1-SimCLR	M14	0,8970	0,7156	-20,24	0,9849	0,6818	-30,77
FW1-SwAV	M14	0,8970	0,7079	-18,91	0,9849	0,6936	-29,13
FW1-SimCLR	M10	0,8976	0,6038	-32,76	0,9625	0,5367	-44,24

Continua na próxima página

Modelo	**Falante**	**CER$_{BP}$**	**CER$_{CL}$**	**ΔCER**	**WER$_{BP}$**	**WER$_{CL}$**	**ΔWER**
FW1-SwAV	M10	0,8976	0,7531	-14,45	0,9625	0,6336	-32,89
FW1-SimCLR	M08	0,9034	0,6997	-22,58	0,9866	0,6650	-32,60
FW1-SwAV	M08	0,9034	0,7198	-18,36	0,9866	0,6650	-32,16
FW1-SimCLR	F05	0,8874	0,6665	-24,88	0,9759	0,6263	-35,85
FW1-SwAV	F05	0,8874	0,6536	-23,38	0,9759	0,5529	-42,30
FW2-SimCLR	M04	0,9807	0,8587	-12,45	0,9882	0,8379	-15,20
FW2-SwAV	M04	0,9804	0,9944	1,43	0,9953	0,9725	-2,29
FW2-SimCLR	F03	0,8996	0,7586	-15,69	0,9496	0,7608	-19,87
FW2-SwAV	F03	0,8996	0,7518	-16,42	0,9496	0,7373	-22,37
FW2-SimCLR	M12	0,9899	0,9294	-6,11	0,9869	0,8471	-14,20
FW2-SwAV	M12	0,9899	0,9383	-5,21	0,9869	0,8562	-13,25
FW2-SimCLR	M01	0,9807	0,8587	-12,45	0,9882	0,8379	-15,20
FW2-SwAV	M01	0,9807	0,9020	-8,02	0,9882	0,8654	-12,42
FW2-SimCLR	M07	0,8714	0,7048	-19,12	0,9148	0,6476	-29,26
FW2-SwAV	M07	0,8714	0,7076	-18,81	0,9148	0,6319	-30,90
FW2-SimCLR	F02	0,8764	0,7120	-18,76	0,9300	0,6655	-28,47
FW2-SwAV	F02	0,8764	0,7285	-16,87	0,9300	0,6398	-31,24
FW2-SimCLR	M16	0,9142	0,7988	-12,61	0,9739	0,7850	-19,39
FW2-SwAV	M16	0,9142	0,7711	-15,66	0,9739	0,7229	-25,94
FW2-SimCLR	M05	0,8911	0,7542	-15,38	0,9232	0,7496	-18,80
FW2-SwAV	M05	0,8911	0,6991	-21,56	0,9232	0,6913	-25,13
FW2-SimCLR	M11	0,8918	0,8487	-4,84	0,9353	0,7876	-15,81
FW2-SwAV	M11	0,8918	0,9045	1,42	0,9353	0,7699	-17,69
FW2-SimCLR	F04	0,9760	0,8058	-17,45	0,9356	0,7653	-18,21
FW2-SwAV	F04	0,9760	0,8310	-14,83	0,9356	0,7653	-18,18
FW2-SimCLR	M09	0,7430	0,6735	-9,34	0,7815	0,6185	-20,85
FW2-SwAV	M09	0,7430	0,6928	-6,75	0,7815	0,6073	-22,29
FW2-SimCLR	M14	0,9354	0,7152	-23,53	0,9126	0,7238	-20,67
FW2-SwAV	M14	0,9354	0,6979	-25,39	0,9126	0,6504	-28,71
FW2-SimCLR	M10	0,7547	0,5727	-24,13	0,7529	0,5014	-33,43
FW2-SwAV	M10	0,7547	0,6614	-12,37	0,7529	0,5653	-24,92
FW2-SimCLR	M08	0,8834	0,7123	-19,35	0,8919	0,6487	-27,30
FW2-SwAV	M08	0,8834	0,7905	-10,51	0,8919	0,7126	-20,10
FW2-SimCLR	F05	0,8767	0,6395	-27,05	0,8941	0,5725	-35,97
FW2-SwAV	F05	0,8767	0,6734	-23,18	0,8941	0,6196	-30,70