UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Uso de Transformada de Pacotes Wavelet e Aprendizado Profundo no Reconhecimento de Emoções na Fala: Aplicações nas Bases CORAA e SofiaFala

Vinicius Rodrigues Costa

Monografia - MBA em Inteligência Artificial e Big Data

ICMC USP SÃO CARLOS

Vinicius Rodrigues Costa

Uso de Transformada de Pacotes Wavelet e Aprendizado Profundo no Reconhecimento de Emoções na Fala: Aplicações nas Bases CORAA e SofiaFala

Monografia apresentada ao Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, como parte dos requisitos para obtenção do título de Especialista em Inteligência Artificial e Big Data.

Área de concentração: Inteligência Artificial

Orientador: Profa. Dra. Alessandra Alaniz Macedo

Versão original

São Carlos

2024

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi, ICMC/USP, com os dados fornecidos pelo(a) autor(a)

Costa, Vinicius Rodrigues
S856m Uso de Transformada de Pacotes Wavelet e Aprendizado Profundo no Reconhecimento de Emoções na Fala: Aplicações nas Bases CORAA e SofiaFala / Vinicius Rodrigues Costa ; orientador Alessandra Alaniz Macedo. – São Carlos, 2024.
78 p. : il. (algumas color.) ; 30 cm.

Monografia (MBA em Inteligência Artificial e Big Data) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2024.

1. LaTeX. 2. abnTeX. 3. Classe USPSC. 4. Editoração de texto. 5. Normalização da documentação. 6. Tese. 7. Dissertação. 8. Documentos (elaboração). 9. Documentos eletrônicos. I. Macedo, Alessandra Alaniz, orient. II. Título.

Vinicius Rodrigues Costa

Use of Wavelet Packet Transform and Deep Learning in Speech Emotion Recognition: Applications on the CORAA and SofiaFala Datasets

Monograph presented to the Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, as part of the requirements for obtaining the title of Specialist in Artificial Intelligence and Big Data.

Concentration area: Artificial Intelligence

Advisor: Profa. Dra. Alessandra Alaniz Macedo

Original version

São Carlos

2024

Dedico este trabalho aos meus pais, Rogério e Luciana, que sempre priorizaram minha educação e me proporcionaram todo o apoio ao longo da minha vida, e à minha esposa, por todo o amor, carinho e imensurável suporte durante este período de dedicação e trabalho.

AGRADECIMENTOS

Expresso minha profunda gratidão à Profa. Dra. Alessandra Alaniz Macedo, por todo o apoio ao longo deste desafio e pela paciência demonstrada, mesmo nos momentos em que fui, por vezes, inconveniente.

Agradeço também aos meus mentores profissionais, Leonardo Faria e Rafael Ribeiro, por me incentivarem a cursar o MBA e por todo o auxílio oferecido na realização deste projeto.

Meus agradecimentos se estendem a todos os professores, técnicos e colegas do Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo, uma instituição da qual tenho muito orgulho de fazer parte.

"Prefiro ter perguntas que não podem ser respondidas a respostas que não podem ser questionadas."

Richard P. Feynman

RESUMO

Costa, V. R. **Uso de Transformada de Pacotes Wavelet e Aprendizado Profundo no Reconhecimento de Emoções na Fala: Aplicações nas Bases CORAA e SofiaFala**. 2024. 78p. Monografia (MBA em Inteligência Artificial e Big Data) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2024.

O reconhecimento de emoções na fala é fundamental para uma variedade de aplicações, desde interfaces de usuário mais empáticas até soluções assistivas para o suporte a tratamentos de fonoaudiologia e para o aprimoramento de ferramentas educacionais. A metodologia deste trabalho explora o uso da transformada de pacotes de Wavelet, aplicada para decompor o espectrograma Mel dos áudios em sub-bandas de frequência, combinada com redes neurais convolucionais para a classificação das emoções, visando o reconhecimento de emoções na fala, com foco na base de fala espontânea CORAA, composta por dados em português. O objetivo principal foi desenvolver um modelo capaz de lidar com as complexidades de dados de fala natural, com potencial aplicação em projetos como o SofiaFala, aplicativo assistivo projetado para apoiar o tratamento de pessoas com deficiências de fala. Os experimentos demonstraram que, embora a proposta tenha alcançado resultados comparáveis aos melhores obtidos com a base CORAA e próximos aos obtidos com a transformada discreta de Wavelet, o uso de redes neurais pré-treinadas ainda se mostrou superior. Adicionalmente, foram realizadas avaliações com outras bases de dados, como EMODB, SAVEE e RAVDESS, para verificar a generalização do modelo. Ao aplicar o modelo nos dados do SofiaFala, observou-se um possível viés de classificação em áudios de pessoas com deficiência de fala. O modelo atribuiu a mesma classificação a todos os áudios de um mesmo falante, independentemente do conteúdo emocional, tornando-o inadequado para esse tipo de aplicação. Este trabalho conclui que, apesar dos resultados promissores, há espaço para melhorias, incluindo o uso de modelos pré-treinados, técnicas adicionais de aumento de dados e ajustes finos nos parâmetros de extração de características. Além disso, seria interessante realizar uma marcação mais detalhada da base SofiaFala, o que poderia levar a uma melhoria nos resultados.

Palavras-chave: Reconhecimento de Emoções na Fala. Transformada de Pacotes Wavelet. Aprendizado Profundo. Rede Neural Convolucional. Fala Espontânea.

ABSTRACT

Costa, V. R. **Use of Wavelet Packet Transform and Deep Learning in Speech Emotion Recognition: Applications on the CORAA and SofiaFala Datasets**. 2024. 78p. Monograph (MBA in Artificial Intelligence and Big Data) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2024.

Speech emotion recognition is essential for a variety of applications, ranging from more empathetic user interfaces to assistive solutions supporting speech therapy treatments and the enhancement of educational tools. The methodology of this work explores the use of the Wavelet Packet Transform, applied to decompose the Mel spectrogram of audio signals into frequency sub-bands, combined with convolutional neural networks for emotion classification, aiming at speech emotion recognition, with a focus on the spontaneous speech corpus CORAA, which consists of data in Portuguese. The main goal was to develop a model capable of handling the complexities of natural speech data, with potential applications in projects like SofiaFala, an assistive app designed to support the treatment of people with speech disorders. The experiments demonstrated that, although the proposed model achieved results comparable to the best obtained with the CORAA corpus and close to those obtained with the Discrete Wavelet Transform, the use of pre-trained neural networks still proved to be superior. Additionally, evaluations were conducted with other datasets, such as EMODB, SAVEE, and RAVDESS, to verify the model's generalization. When applying the model to SofiaFala's data, a possible classification bias was observed in the audio of individuals with speech impairments. The model assigned the same classification to all audios from the same speaker, regardless of emotional content, making it unsuitable for this type of application. This study concludes that, despite promising results, there is room for improvement, including the use of pre-trained models, additional data augmentation techniques, and fine-tuning of feature extraction parameters. Furthermore, a more detailed annotation of the SofiaFala dataset could lead to better results.

Keywords: Speech Emotion Recognition. Wavelet Packet Transform. Deep Learning. Convolutional Neural Network. Spontaneous Speech.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AE	Autoencoder
CASIA	*Chinese Natural Emotional Database*
EESDB	*Elderly Emotional Speech Database*
EMODB	*Berlin Emotional Database*
IA	Inteligência Artificial
IEMOCAP	*Interactive Emotional Dyadic Motion Capture Database*
LSTM	*Long Short-Term Memory*
LSVM	SVM linear
MFCC	Coeficientes Cepstrais da Frequência Mel
PLN	Processamento de Linguagem Natural
RAVDESS	*Ryerson Audio-Visual Database of Emotional Speech and Song*
RCP	Rede de Crença Profunda
RNA	Rede Neural Artificial
RNC	Rede Neural Convolucional
RNR	Rede Neural Recorrente
RSVM	*SVM com função de base radial*
SAVEE	*Survey Audio-Visual Expressed Emotion*
SVM	*Support Vector Machine*
TCW	Transformada Contínua de Wavelet
TDW	Transformada Discreta de Wavelet
TF	Transformada de Fourier
TFCT	Transformada de Fourier de Curto Tempo
TPW	Transformada de Pacotes Wavelet
TW	Transformada de Wavelet
VAD	Detecção de Atividade de Voz

SUMÁRIO

1 INTRODUÇÃO

1.1 Contextualização e Motivação

Os estudos sobre linguagens são antigos, mas pertinentes até os dias de hoje. A própria origem da fala é um tema muito relevante e possui algumas hipóteses consideráveis, que podem descrevê-la de forma satisfatória. Da teoria de Chomsky da codificação da fala em nosso genoma até as teorias que defendem uma origem empírica das línguas, entender melhor o surgimento da nossa linguagem, assim como os assuntos relacionados, nos ajuda até mesmo a entender o que é ser humano (FONTANARI, 2009).

A linguagem é um instrumento projetado, de forma engenhosa, para descrever lugares, pessoas, objetos, eventos e até mesmo pensamentos e emoções. Utilizamos a linguagem para criar experiências compartilhadas com outras pessoas. Ao compartilhar nossas vivências e conhecimentos, podemos tornar a convivência, a aprendizagem, o ensino, enfim, a comunicação mais eficiente (CORBALLIS, 2002). A comunicação é o processo pelo qual as ideias da linguagem são transmitidas de uma pessoa a outra, seja de forma verbal, escrita, ou por meio de gestos. A relação entre linguagem e comunicação é interdependente: enquanto a linguagem fornece as ferramentas para organizar e expressar pensamentos complexos, a comunicação utiliza essas ferramentas para garantir que a mensagem seja transmitida com clareza, facilitando a compreensão e a interação entre as pessoas(FONTANARI, 2009).

A comunicação desempenha um papel crucial em todas as esferas da vida, sejam pessoal, profissional ou social. No ambiente de trabalho, a comunicação eficaz é fundamental para o sucesso profissional. Ela permite a colaboração, promove a inovação e ajuda a resolver conflitos. Assim como na vida pessoal, uma boa comunicação ajuda a manter relacionamentos saudáveis e duradouros.

Muitas pessoas enfrentam dificuldades de comunicação, especialmente em relação à fala, que transmite 38% das informações pelo tom de voz e 7% pelas palavras; enquanto os outros 55% são expressos por meio da linguagem corporal (The World of Work Project, 2019). Isso demonstra a essencialidade da fala, não apenas pelo seu conteúdo linguístico, mas também pelo impacto que exerce na transmissão de uma mensagem.

As dificuldades de fala podem resultar de diversas causas, como acidente vascular cerebral, lesão cerebral, perda auditiva, atrasos no desenvolvimento, fissura palatina, paralisia cerebral ou até questões emocionais ou do envelhecimento. Esses fatores estão associados a questões médicas, genéticas, neurodesenvolvimentais, miofuncionais e linguísticas. As dificuldades de fala podem ser categorizadas como funcionais (afetando aspectos motores ou linguísticos) ou orgânicas (relacionadas a problemas neurológicos, estruturais, sensoriais

ou perceptuais). Por exemplo, surdez e perda auditiva são impedimentos orgânicos na produção de sons da fala, causados por distúrbios sensoriais ou perceptuais (OLIVEIRA; GOULART; CHIARI, 2013). De modo geral, o termo transtorno de sons da fala refere-se a dificuldades na percepção, produção motora de sons da fala (articulação) ou representação fonológica dos sons e segmentos da fala.

Pessoas com histórico de transtornos de fala podem apresentar piores resultados em diversos domínios, tais como comunicação, realização educacional e *status* ocupacional, quando comparadas a seus pares sem deficiências (JOHNSON; BEITCHMAN; BROWNLIE, 2010). Problemas de fala afetam não só aqueles que os possuem, mas também podem afetar a vida de pessoas próximas, como chefes, colegas de trabalho, amigos de escola e os próprios pais de pessoas com distúrbios de fala (ARAS *et al.*, 2014). Até mesmo a autopercepção vocal do disfônico pode geral uma pior percepção sobre o impacto da disfonia em sua qualidade de vida (KASAMA; BRASOLOTTO, 2007). Assim, o tratamento respeitoso e personalizado de pessoas com transtornos de fala é crucial para promover a inclusão social, proporcionando-lhes meios adequados para expressar suas ideias e contribuir plenamente para a diversidade e riqueza da comunicação humana.

Procurar formas de auxiliar no tratamento de transtornos de fala é fundamental para auxiliar as pessoas com deficiência de fala, oferecendo-lhes oportunidades de melhorar sua comunicação, permitindo que expressem suas ideias de maneira mais inteligível. Os tratamentos destes transtornos devem ser feitos com fonoaudiólogos. Contudo, o treinamento com exercícios de fortalecimento de musculaturas da boca, da língua e de treinamento fonético devem ser feitos, além do consultório do terapeuta. Normalmente, esses exercícios são realizados em casa, no trabalho e em outros ambientes sem a presença do profissional e servem de reforço à terapia realizada em consultório. Dado que se tem parte do tratamento sendo feito longe da presença direta de um fonoaudiólogo, as tecnologias assistivas podem ser muito úteis para auxiliar no tratamento de transtornos da fala. Nesse contexto, o projeto o SofiaFala[1] propõe soluções para auxiliar os profissionais da saúde no acompanhamento de pessoas com deficiências de fala. O aplicativo SofiaFala foi utilizado em diversos pacientes, de forma a agregar acompanhamento e métricas aos trabalhos dos fonoaudiólogos, aumentando o potencial de análise em seus pacientes (RISSATO; MACEDO, 2021; MACEDO *et al.*, 2024).

Além dos exercícios tradicionais de fortalecimento e treinamento para fala, o uso de tecnologias assistivas pode desempenhar um papel fundamental no apoio à comunicação. Os sons capturados em áudio podem ser explorados de muitas formas, permitindo a extração de diversas características que, por sua vez, podem ser utilizadas para aplicações baseadas em fala humana. Essas ferramentas podem capturar a fala do usuário e permitir, por exemplo, o monitoramento da qualidade vocal. A capacidade de reconhecer emoções e

[1] http://dcm.ffclrp.usp.br/sofiafala/

suas flutuações em áudio podem auxiliar, por exemplo, na identificação de frustração, e assim, pode desempenhar um papel crucial no aprimoramento de produtos de *software* destinados a apoiar processos de aprendizagem e treinamento (KOŁAKOWSKA *et al.*, 2014). Assim, ferramentas educacionais podem ser aprimoradas, por meio da compreensão das emoções dos usuários, por exemplo. Um software poderia ser projetado para identificar sinais de emoções negativas durante uma atividade, adaptando dinamicamente o nível de dificuldade ou fornecendo sugestões personalizadas para ajudar o usuário. Isso poderia aumentar sua eficácia e também melhorar a experiência do usuário. Outras tarefas dessas aplicações incluem: reconhecimento de fala, aprimoramento de fala, reconhecimento de locutor (e/ou gênero), detecção de atividade vocal, análise de fala patológica e também reconhecimento de emoções (SHARMA; UMAPATHY; KRISHNAN, 2020).

Na literatura, existem diversos algoritmos de aprendizado de máquina que já foram utilizados para o reconhecimento de emoções na fala. Desde modelos mais tradicionais, como árvore de decisão, SVM (do inglês *Support Vector Machine*, ou Máquina de Vetores de Suporte) e modelo de mistura gaussiana, até modelos de aprendizado profundo, como redes neurais convolucionais, redes neurais recorrentes e redes neurais de *autoencoder*, entre outros (Shah Fahad *et al.*, 2021). A aplicação de redes neurais na identificação de sentimentos pode ser complementada por técnicas adicionais que auxiliam na extração de características dos áudios, tais como análise de frequência Mel, VAD (do inglês *Voice Activity Detection*, ou Detecção de Atividade de Voz), e transformadas de Fourier e Wavelet. Essas abordagens, assim como várias outras, podem aprimorar a qualidade do resultado na análise de sentimentos, gerando um potencial considerável para pesquisa nas áreas relacionadas ao reconhecimento de emoção na fala, com muitos trabalhos focados na utilização de estratégias diferentes, por meio das técnicas complementares citadas. Em uma pesquisa relacionada dentro do grupo no qual este trabalho está inserido, Vieira (2023) utiliza a transformada de Wavelet contínua para reconhecimento da emoção na fala. Em sua conclusão, Vieira (2023) conseguiu bons resultados (acurácia de 0,795 e *F1 score* de 0,566 para base CORAA) e sugere a exploração de outras estratégias de transformada de Wavelet.

1.2 Objetivos

O objetivo deste trabalho é desenvolver e avaliar um modelo de reconhecimento de emoções na fala, aplicando técnicas de aprendizado profundo e a transformada de pacotes Wavelet. Esse modelo foi avaliado, a partir do uso da base de dados CORAA (MARCACINI; JUNIOR; CASANOVA, 2022) e, posteriormente, será aplicado na base de dados de áudio do SofiaFala.

1.3 Organização do texto

O trabalho está organizado da seguinte forma: Capítulo 2 faz uma revisão de conceitos importantes relacionados ao tema do trabalho; Capítulo 3 faz um mapeamento sistemático de trabalhos da literatura relacionados; Capítulo 4 descreve a metodologia escolhida, as bases de dados utilizadas, pré-processamento e execução do modelo ; Capítulo 5 apresenta os testes realizados e seus resultados; Capítulo 6 encerra o trabalho destacando os resultados-chave e sugere direções para futuras pesquisas.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo introduz os conceitos teóricos que fundamentam este trabalho, apresentando-os da seguinte forma: Na seção 2.1, são abordados conceitos de aprendizado profundo; Na seção 2.2, são abordados conceitos de reconhecimento de emoção da fala e extração de características; por fim, na seção 2.3, a transformada de Wavelet é apresentada.

2.1 Aprendizado Profundo

Sistemas de Inteligência Artificial (IA) podem ser criados exclusivamente com conhecimento codificado e pré-programado, mas ao capacitar esses sistemas para adquirirem conhecimento próprio ao identificar padrões em dados brutos, estamos falando do conceito de aprendizado de máquina. A utilização dessa abordagem permite aos computadores lidar com problemas do mundo real e tomar decisões, que podem até mesmo parecer ou ser subjetivas (GOODFELLOW; BENGIO; COURVILLE, 2016).

O aprendizado de máquina é amplamente explorado na sociedade moderna em diversas tarefas, abrangendo desde buscas na web até a filtragem de conteúdo em redes sociais (LECUN; BENGIO; HINTON, 2015). Esses sistemas podem ser empregados para identificação de objetos em imagens, transcrições de fala para texto, gerando resultados relevantes e muitas outras aplicações (LECUN; BENGIO; HINTON, 2015). A resolução de tarefas de IA frequentemente envolve a criação de um conjunto apropriado de características a serem extraídas para a tarefa em questão, as quais são fornecidas a um algoritmo de aprendizado de máquina (GOODFELLOW; BENGIO; COURVILLE, 2016).

Dado que as técnicas tradicionais de aprendizado de máquina enfrentam limitações em lidar com dados de linguagem natural de forma bruta, as Redes Neurais Artificiais (RNA) surgiram como uma solução. Essas redes são compostas por neurônios artificiais organizados em camadas, capazes de aprender representações de dados complexos por meio de conexões otimizadas. Por muito tempo, uma das poucas formas de trabalhar com dados brutos foi através da construção de sistemas como as RNA, capazes de identificar padrões nesses dados e extrair suas características, criando uma representação mais adequada das informações (LECUN; BENGIO; HINTON, 2015). Uma solução para esse problema é usar o aprendizado de máquina para descobrir não apenas a saída, mas também a própria representação dos dados (GOODFELLOW; BENGIO; COURVILLE, 2016). As representações aprendidas frequentemente resultam em desempenho muito melhor do que pode ser obtido com representações projetadas manualmente, segundo (GOODFELLOW; BENGIO; COURVILLE, 2016). Assim, o aprendizado de representação é um conjunto de métodos que permite que uma máquina receba dados brutos e descubra automaticamente

as representações necessárias para tarefas como detecção e classificação (LECUN; BENGIO; HINTON, 2015).

Uma dificuldade em muitas aplicações de IA é que muitos fatores influenciam cada pedaço de dados observados. Ao analisar gravações de fala, pode-se observar, por exemplo, o sexo, a idade e até mesmo o sotaque do falante. Contudo, extrair estas características de alto nível de dados brutos não é uma tarefa trivial (GOODFELLOW; BENGIO; COURVILLE, 2016, pp.2-26).

O aprendizado profundo aborda o desafio de representar múltiplos fatores de variação, criando e permitindo que a máquina aprenda representações mais complexas a partir de mais simples (GOODFELLOW; BENGIO; COURVILLE, 2016, pp.2-26). O método pode ser descrito como um conjunto de etapas de aprendizado de representação com vários níveis de representação (LECUN; BENGIO; HINTON, 2015). Assim, o método começa com módulos não lineares simples, que transformam gradualmente a representação inicial dos dados brutos em níveis mais altos e abstratos, capazes de capturar características mais complexas e refinadas dos dados (LECUN; BENGIO; HINTON, 2015). Dessa forma, o aprendizado profundo é um subconjunto do aprendizado de máquina, que utiliza uma cascata de múltiplas camadas de transformações para aprender funções complexas (SHINDE; SHAH, 2018).

O aprendizado profundo é capaz de descobrir estruturas complexas em grandes conjuntos de dados, utilizando um algoritmo de retropropagação. Este algoritmo vai indicar como o computador deve alterar seus parâmetros internos, de modo a calcular a representação em cada camada com base na representação da camada anterior (LECUN; BENGIO; HINTON, 2015). Este tipo de IA ganhou destaque nos últimos tempos, não apenas por sua capacidade de trabalhar com conjuntos de dados complexos, mas também pelo aumento significativo de capacidade de processamento de microchips e redução considerável no custo de *hardware* de computação (SHINDE; SHAH, 2018). Esses avanços permitiram a criação e utilização de diversas técnicas diferentes de aprendizado profundo, como Autoencoder (AE), Rede de Crença Profunda (RCP), Rede Neural Convolucional (RNC), Rede Neural Recorrente (RNR) e muitas outras (SHINDE; SHAH, 2018). Existem também diversas estruturas que podem ser utilizadas para implementação de técnicas de Aprendizado Profundo como, por exemplo, TensorFlow (da empresa Google) e PyTorch (da empresa Facebook). As estruturas assim como as técnicas são selecionadas, de acordo com o tipo de problema a ser resolvido e a plataforma usadas para desenvolver as soluções (SHINDE; SHAH, 2018).

As aplicações convencionais de Aprendizado Profundo abrangem diferentes áreas da Ciência da Computação como visão computacional, predição, análise semântica e processamento de linguagem natural. Dentro da visão computacional, existem sub áreas como detecção de objetos, reconhecimento de imagens e também reconhecimento e pro-

cessamento de áudios (SHINDE; SHAH, 2018). A utilização de aprendizado profundo melhorou, de forma significativa, o estado da arte no reconhecimento de voz, uma vez que a utilização de RNC trouxe grande avanço para o processamento de voz e imagem (LECUN; BENGIO; HINTON, 2015).

2.1.1 Rede Neural Convolucional

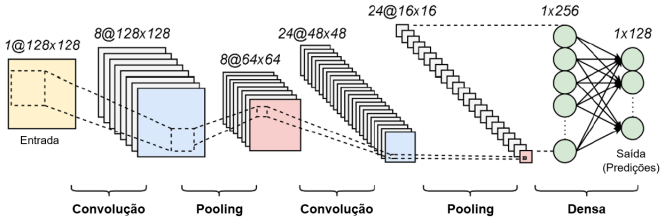
Uma RNC é um tipo de RNA bem conhecido, composto por neurônios que se otimizam automaticamente durante o processo de aprendizado, assim como outras RNAs.(O'SHEA; NASH, 2015). Da mesma forma que a RNA tradicional, os neurônios vão receber uma entrada e realizar uma operação, mas a RNC utiliza uma operação matemática linear entre as matrizes chamada de convolução (GOODFELLOW; BENGIO; COURVILLE, 2016). A RNC recebe esse nome por utilizar a convolução (em vez da multiplicação geral de matrizes) em pelo menos uma de suas camadas (ALBAWI; MOHAMMED; AL-ZAWI, 2017). De forma resumida, a convolução é uma operação matemática no qual um filtro (ou *kernel*) desliza sobre a entrada, multiplicando e somando os valores da entrada pelo filtro em cada posição em que ele foi aplicado (GOODFELLOW; BENGIO; COURVILLE, 2016).

Uma RNC é um tipo de rede neural especializada para processar dados com topologia semelhante a uma grade, como dados de séries temporais de uma dimensão ou mesmo imagens cujos os *pixels* formam uma grade bidimensional (GOODFELLOW; BENGIO; COURVILLE, 2016). Uma das grandes vantagens da utilização da RNC em relação às redes neurais tradicionais é a redução do número de parâmetros necessários, possibilitando abordar modelos maiores para resolver tarefas mais complexas (GOODFELLOW; BENGIO; COURVILLE, 2016).

Uma RNC é composta de três tipos de camadas: Convolucional, *Pooling* e Densa (totalmente conectada); ao empilhar estas camadas, uma arquitetura de RNC é criada (O'SHEA; NASH, 2015). A primeira camada da RNC, assim como outras RNAs, é composta apenas pelos dados de entrada. Os demais tipos de camadas são definidos, de acordo com a arquitetura desejada. Um exemplo de arquitetura de RNC pode ser visto na Figura 1, onde à esquerda está representada a entrada de dados correspondente a uma imagem de tamanho 128x128 com 1 canal. Na sequência, uma série de camadas convolucionais é aplicada para extrair as características mais relevantes da imagem, seguidas por camadas de pooling, responsáveis por reduzir a dimensionalidade dos dados. Na parte final da arquitetura, uma camada totalmente conectada (densa) é utilizada para realizar as predições finais.

A camada convolucional é fundamental para criação de uma RNC, seus parâmetros são definidos nos *kernels*, que funcionam como filtros de pequena dimensão que se estendem por toda profundidade de entrada (O'SHEA; NASH, 2015). Os parâmetros destes filtros são aprendidos, durante o treinamento, e, quando os dados alcançam essa camada, ela

Figura 1 – Exemplo de arquitetura de RNC

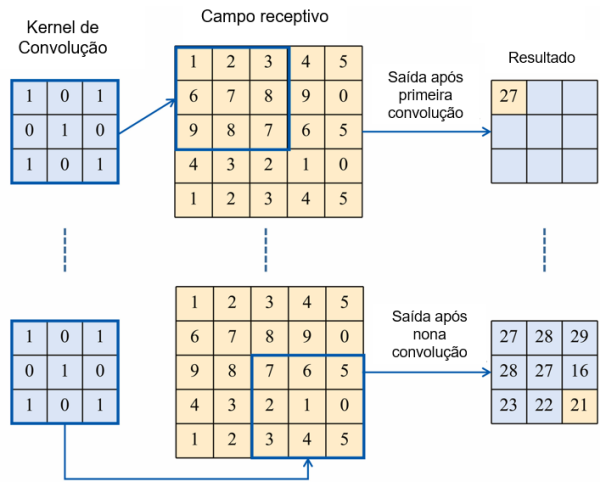


Fonte: Adaptado de Montalbo and Alon (2021)

aplica a convolução a cada filtro resultando na geração de um mapa de ativação (O'SHEA; NASH, 2015). Esse processo permite que cada neurônio em uma camada convolucional esteja conectado apenas a uma pequena região de entrada (campo receptivo), permitindo que a rede aprenda características específicas nessas posições (O'SHEA; NASH, 2015). Um exemplo do *kernel* de convolução em ação pode ser observado na Figura 2. À esquerda, tem-se um *kernel* de convolução de tamanho 3x3 que é aplicado sobre uma matriz de entrada representada no campo receptivo. Na primeira convolução, o *kernel* percorre a matriz de entrada e gera uma saída, obtida a partir da soma ponderada dos valores do *kernel* e da região correspondente da matriz de entrada. O processo continua até que o *kernel* tenha percorrido todo o campo receptivo, gerando uma matriz de resultados após a nona convolução.

Outra característica importante dessas camadas DA RNC é a redução significativa da complexidade do modelo, por meio dos seguintes três hiperparâmetros: profundidade, passo (*stride*) e *zero-padding*.

Figura 2 – Exemplo de *kernel* de convolucao em ação

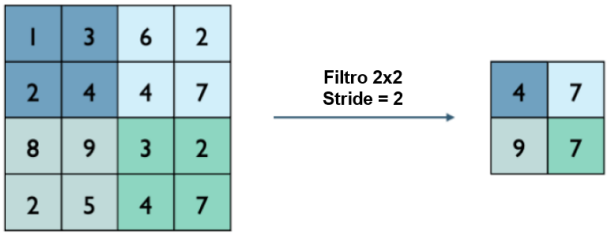


Fonte: Adaptado de Gu *et al.* (2021)

As camadas de *pooling* são utilizadas para reduzir os tamanhos das saídas (*down-sampling*) das camadas de convolução, otimizando a carga computacional (ZHANG; LEE; LIU, 2024). Visto de uma ótica de processamento de imagem, o uso da camada de *pooling*

realiza um efeito similar a reduzir a resolução de uma imagem (ALBAWI; MOHAMMED; AL-ZAWI, 2017). Um dos métodos mais comuns de *pooling*, o *max-pooling*, pode ser visto na Figura 3. Divide-se a imagem em sub-regiões e retorna o valor máximo encontrado em cada sub-região.

Figura 3 – Exemplo de *max-pooling* de filtro 2x2 e *stride* = 2



Fonte: Adaptado de Zhang, Lee and Liu (2024)

A última camada de *pooling* é responsável por transformar a matriz multidimensional em um vetor unidimensional. Esse vetor unidimensional segue para a camada densa (totalmente conectada), responsável por gerar um vetor que representa a probabilidade de uma característica pertencer a uma classe determinada (ZHANG; LEE; LIU, 2024). Essa camada recebe esse nome, pois todos os neurônios na camada densa são conectados aos neurônios da camada anterior. Essa camada funciona de forma análoga a redes neurais tradicionais, e também utiliza uma função de ativação para introduzir a não linearidade e aprender padrões mais complexos (O'SHEA; NASH, 2015). Um exemplo da camada densa pode ser observado no final da Figura 1, após a última camada de *pooling*. No exemplo citado, após a saída da última camada de *pooling*, os mapas de características são achatados em um vetor unidimensional. Esse vetor é então conectado diretamente aos neurônios da camada densa, formando uma rede neural tradicional responsável por gerar as predições.

2.2 Reconhecimento de Emoção da Fala

As emoções podem ser descritas como um estado psicológico da combinação entre a experiência subjetiva, a resposta física e comportamental (LALITHA *et al.*, 2015). Elas desempenham um papel muito importante nas tomadas de decisões diárias e são uma das formas de entender o estado psicológico de uma pessoa. As emoções podem ser expressadas, por meio de expressões faciais, linguagem corporal e também pela fala. Analogamente, a identificação de emoções pode ser abordada de diversas formas, como análise de expressões faciais, análise de fala ou mesmo uma abordagem em conjunto das duas (LALITHA *et al.*, 2015). Deste modo, o reconhecimento de emoções na própria fala pode ser feito extraindo determinadas características de um sinal de áudio ou de um sinal visual.

Os tipos de emoções reconhecidas podem ser definidos, por meio de distintos modelos emocionais. Um dos modelos mais influentes no campo das emoções é o de Ekman

(1992), que, com base em estudos de expressões faciais em diferentes culturas e países, propôs a caracterização de seis emoções universais: tristeza, raiva, alegria, medo, surpresa e nojo. No entanto, mesmo com uma definição de emoções universais, nem todos os modelos seguem fielmente essa classificação. A escolha de um conjunto específico de emoções é uma etapa essencial para qualquer análise emocional, permitindo que as emoções sejam organizadas em escores, classificações ou dimensões, conforme a necessidade do modelo (WANI *et al.*, 2021). Essas emoções podem ser agrupadas de acordo com a intensidade e o tipo, ou classificadas de forma mais detalhada, dependendo do objetivo da análise.

Outro aspecto a ser considerado, em relação à decisão sobre os tipos de emoções analisadas, é a dependência do corpus linguístico e de suas marcações. As bases de dados utilizadas para marcação de emoções podem sofrer variação por diversos motivos, como cultura, língua, gênero e situação. A emoção é um conjunto complexo de variáveis e sua análise depende diretamente da qualidade do conjunto de dados, e caso suas marcações estiverem comprometidas as conclusões tiradas dela podem ser incorretas (WANI *et al.*, 2021). Os bancos de dados criados para marcação de emoção na fala variam de acordo com seus objetivos, mas podem ser organizados em três tipos de fala: espontânea, atuada e provocada (WANI *et al.*, 2021). A fala espontânea, como o próprio nome já diz, é caracterizada pela espontaneidade do indivíduo que apresentou a emoção, em que muitas vezes o áudio é analisado sem que a pessoa que gravou o perceba. Já a atuada é um tipo de fala que atores e artistas profissionais simulam a emoção. Por fim, a fala provocada consiste em provocar ou induzir a emoção, por meio de alguma situação, e assim registrar a emoção.

2.2.1 Pré-processamento e Extração de Características de Emoção na Fala

O Processamento de Linguagem Natural (PLN) é uma área da Ciência da Computação que utiliza técnicas computacionais para interpretar e manipular o conteúdo da linguagem humana (HIRSCHBERG; MANNING, 2015). Entre os focos de estudo do PLN tem-se o desenvolvimento de tecnologias como tradução automática, a síntese e o reconhecimento de fala e de sentimentos (HIRSCHBERG; MANNING, 2015). Em alguns casos, a utilização dessas tecnologias pode ajudar a deixar a interação entre humanos e máquinas mais natural, uma vez que o computador pode reconhecer um estado emocional da mesma forma que um humano (KHAN, 2016).

A identificação do sentimento e emoções de uma pessoa em relação a determinados produtos ou serviços também pode ser muito importante para a melhoria dos mesmos. Além disso, essa abordagem pode ser utilizada para melhorar a experiência e o tratamento de pessoas que enfrentam dificuldades na comunicação.

Uma forma de abordar o reconhecimento de emoções em fala é por meio de técnicas de PLN, que permitem tratar e manipular dados de forma a identificar padrões emocionais.

Um sistema de reconhecimento de fala pode ser separado em três etapas: pré-processamento, extração de características e classificação da emoção (MADANIAN *et al.*, 2023). Após obter os dados, o pré-processamento é a primeira etapa do processo de reconhecimento de emoção da fala. Nessa etapa, os dados podem passar por algumas fases como: ajuste de frequências, normalização do volume, remoção de ruídos, aumento sintético de dados e outros processos responsáveis por normalizar as características (WANI *et al.*, 2021). Deste modo, as variações nas gravações de áudio não afetam o processo de reconhecimento da emoção.

Os ajustes de frequência são realizados, pois componentes de alta frequência são geralmente considerados redundantes, podendo ser utilizados filtros para remover frequências desnecessárias (Rí; CIARDI; CONCI, 2023). Já a normalização do volume é uma metodologia para ajustar o áudio de forma a padronizá-lo em todas as amostras (WANI *et al.*, 2021). Essa normalização também pode incluir ajustes de *trimming* ou *padding* para que todos os áudios fiquem do mesmo tamanho. *Trimming* refere-se ao corte do áudio para ajustá-lo a um comprimento específico, enquanto *padding* envolve adicionar silêncio (ou zeros) ao áudio para alcançar o comprimento desejado (Rí; CIARDI; CONCI, 2023).

Muitas vezes, o ambiente de geração dos áudios pode ter ruídos e outras interferências indesejadas. Consequentemente, a remoção de ruido é um processo muito comum na análise de sinais de fala e pode afetar de forma crítica a acurácia dos modelos (WANI *et al.*, 2021). Uma das formas de remover ruído, e melhorar a interpretação de emoções nos sinais de áudio, é por meio da divisão dos sinais em quadros fixos juntamente com uma função de janela. Estes processos são conhecidos como *framing* e *windowing* e ajudam a reduzir vazamentos espectrais e eliminar descontinuidades (MADANIAN *et al.*, 2023).

Em uma linha muito próxima da redução de ruído, a detecção de ativação por voz (*Voice Activity Detection*) também é muito utilizada para detectar silêncios, fala com voz e fala sem voz (partes da fala sem a utilização de cordas vocais) (WANI *et al.*, 2021). Como muitas vezes os dados são limitados, outro processo muito comum no pré-processamento é o aumento sintético de dados (*Data Augmentation*), nos sinais do conjunto de dados de treino (Rí; CIARDI; CONCI, 2023).

Uma das fases cruciais no reconhecimento de emoções é a seleção de características e seu dimensionamento (WANI *et al.*, 2021). A seleção de técnicas é fundamental para evitar problemas de alta dimensionalidade, aumento no tempo de treinamento e sobre-ajuste (*overfitting*) dos classificadores (WANI *et al.*, 2021). Assim, a eficiência e as taxas de predição do sistema vão depender do classificador utilizado, e também das características extraídas (KHAN, 2016).

Uma das técnicas clássicas mais utilizadas para extração de características de fala é o Coeficiente Cepstral de Frequência Mel (*MFCC*, na sigla em inglês). A técnica extrai

características espectrais de um sinal de áudio representativo do trato vocal, por meio da transformada de Fourier (WANI *et al.*, 2021). A utilização da escala Mel é importante, pois ela se aproxima melhor da percepção humana da voz em comparação a escalas lineares (VENKATARAMANAN; RAJAMOHAN, 2019). O mel-espectrograma é uma representação do sinal de áudio em uma escala Mel que corresponde à representação de tempo versus frequência log-mel, que foi obtida durante o cálculo dos *MFCC's* (VENKATARAMANAN; RAJAMOHAN, 2019). Uma vez que o espectrograma em escala Mel (assim como *MFCC's*) pode ser representado como imagem, essas imagens podem ser utilizadas como entrada de dados para redes de aprendizado profundo (VENKATARAMANAN; RAJAMOHAN, 2019).

A extração de características pode ser feita de diversas formas, podendo ser com métodos mais clássicos, obtendo características acústicas (frequência fundamental, intensidade, taxa de variação espectral e etc.) juntamente com características prosódicas (ritmo, pausas na fala e duração de fonemas) (MADANIAN *et al.*, 2023). A extração de características pode ser feita também com técnicas mais recentes, que trabalham com transformada de Fourier ou transformada de Wavelet para uma análise de tempo e frequência juntamente com análise de espectrogramas (MADANIAN *et al.*, 2023).

A transformada de Fourier é uma das técnica mais utilizadas no processamento de sinais para revelar a composição de frequência de uma série temporal transformando-a do domínio do tempo para o domínio da frequência (GAO; YAN, 2010). Apesar de ser muito utilizada na Ciência da Computação e na Engenharia, a transformada tem suas limitações. Para a análise de sinais no domínio tempo-frequência, a transformada de Fourier não oferece informações locais sobre o sinal, o que é uma limitação significativa (SIFUZZAMAN; ISLAM; ALI, 2009). Para solucionar este problema, Dennis Gabor introduziu a transformada de Fourier com janela, também conhecida como Transformada de Fourier de Curto Tempo (TFCT) (SIFUZZAMAN; ISLAM; ALI, 2009). Essa abordagem envolve a utilização de uma janela de análise de tamanho específico que se move através do sinal ao longo do eixo do tempo, permitindo a realização de uma transformada de Fourier localizada temporalmente (GAO; YAN, 2010).

Mesmo sendo capaz de representar temporalmente as frequências, a transformada de Fourier de curto tempo também é limitada pelo princípio da incerteza de Heisenberg, que implica um compromisso entre a resolução temporal e a frequência (GAO; YAN, 2010). Dessa forma, quanto maior a precisão no tempo, menor será a precisão na frequência e vice versa. A escolha do tamanho da janela influencia diretamente na resolução temporal e de frequência, e escolher o tamanho ideal não é uma tarefa trivial. Por esse motivo, os pesquisadores têm buscado técnicas mais adequadas para analisar sinais não estacionários (GAO; YAN, 2010). Dentro deste cenário, a transformada de Wavelet se mostra como uma das possíveis soluções para esse problema.

2.3 Transformada de Wavelet

A Transformada de Wavelet (TW) é uma técnica matemática que converte um sinal em uma forma diferente, assim como a TFCT. O termo TW é um termo genérico, uma vez que existem diferentes formas de fazer essa transformada. Entre as mais conhecidas, tem-se a Transformada de Wavelet Continua (TWC) e a Transformada Discreta de Wavelet (TDW) (GUIDO *et al.*, 2020). A transformada utiliza uma função para converter um sinal de entrada $s(t)$, onde t é o tempo para o domínio de tempo-frequência. A função utilizada para fazer essa operação é conhecida como Wavelet mãe $\psi(x)$ (GUIDO *et al.*, 2020). Segundo Guido *et al.* (2020), uma TWC pode ser descrita pela seguinte equação:

$$TCW(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} s(t)\psi\left(\frac{t-b}{a}\right) dt, \quad a > 0, b \in \mathbb{R} \tag{2.1}$$

Na Equação 2.1, a é um parâmetro de escala de contração (ou dilatação), enquanto b é é um parâmetro de deslocamento. O princípio de funcionamento da TW é baseado na correlação entre o sinal de entrada e o número infinito de possibilidades de dilatação e translação da função de onda mãe, à medida que a e b variam (GUIDO *et al.*, 2020). Isso captura o suporte temporal das frequências presentes no sinal, permitindo uma análise detalhada do comportamento do sinal em diferentes escalas de tempo e frequência (GUIDO *et al.*, 2020).

As operações TFCT e TWC compartilham propriedades matemáticas semelhantes. Ambas as funções de base são localizadas em frequência e a matriz de transformação inversa é a transposta da matriz original (GRAPS, 1995). Isso as torna transformações que podem ser vistas como rotações no espaço de funções para um domínio diferente (GRAPS, 1995). No caso da TFCT, esse novo domínio contém funções de base que são senos e cossenos, enquanto a TW contém uma função de base mais complexa (GRAPS, 1995).

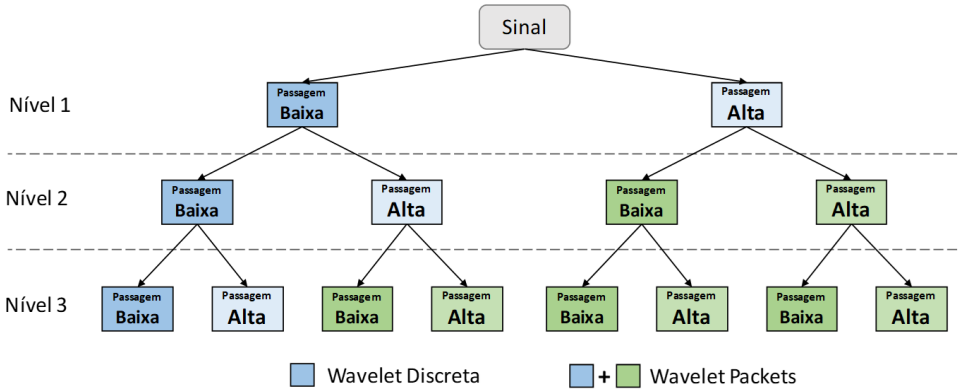
Outra informação importante da TDW é em relação a seus parâmetros de contração e deslocamento. Eles podem variar continuamente, gerando informações redundantes (GAO; YAN, 2010). Uma forma de reduzir a redundância nos coeficientes de wavelet, sem comprometer a informação do sinal original, é empregar uma abordagem que utilize valores discretos nos parâmetros de escala e translação (GAO; YAN, 2010). Essa transformação para valores discretos é responsável por gerar a TDW. Diferentemente da TWC, para a TDW, é necessário utilizar um par de filtros de alta e baixa passagem para separar o sinal contínuo (GUIDO *et al.*, 2020).

2.3.1 Transformada de Pacotes de Wavelet

Embora a TDW ofereça uma resolução flexível no domínio tempo-frequência, ela apresenta uma resolução relativamente baixa na região de alta frequência. Essa limitação

dificulta a diferenciação de componentes de alta frequência (GAO; YAN, 2010). Na transformada de pacotes de wavelet (TPW), o sinal é decomposto em um conjunto mais amplo de sub-bandas de frequência. Isso permite uma decomposição ainda maior na região de alta frequência, superando o problema de resolução da TDW (GAO; YAN, 2010). A diferença na decomposição do sinal pode ser observada na Figura 4, na qual as caixas de filtro azuis representam o máximo da decomposição da TDW, e as caixas de filtro verdes representam os ganhos de decomposição nas regiões de alta frequência, obtidos pela TPW.

Figura 4 – Exemplo de comparação entre TDW e TWP



Fonte: Autoria própria

A TPW proporciona uma análise em múltiplos níveis, decompondo o sinal tanto em componentes de baixa quanto de alta frequência, o que resulta em uma resolução mais precisa no domínio da frequência em comparação à TDW. Com a TPW, há uma flexibilidade adicional na escolha de funções wavelet em diferentes escalas e frequências, permitindo adaptar a base de decomposição ao conteúdo específico do sinal. Isso possibilita uma análise mais detalhada e eficiente para capturar padrões oscilatórios ou periódicos, particularmente úteis em áudios de fala, onde as variações frequenciais rápidas exigem um tratamento mais refinado (GOKHALE; KHANDUJA, 2010).

A TDW é capaz de destacar mudanças instantâneas na evolução espectral, sendo essa uma característica estendida pela TPW, tornando essas transformadas alternativas para decompor ou reconstruir sinais de áudio (Shah Fahad *et al.*, 2021). Essa técnica auxilia na análise mais adequada dos detalhes e mudanças rápidas em áudios de fala, o que pode ser difícil de ser realizado com a TFCT (Shah Fahad *et al.*, 2021). Ainda assim, a TW sofre devido à sua natureza linear e não adaptativa, possibilitando interpretações incorretas dos dados (Shah Fahad *et al.*, 2021). Dessa forma, características obtidas pela TW são úteis para classificação de emoções. No entanto, é necessário selecionar uma wavelet adequada para o reconhecimento de emoções, segundo (Shah Fahad *et al.*, 2021).

3 TRABALHOS RELACIONADOS

Na literatura, muitos estudos tratam do reconhecimento de emoções na fala, variando em suas abordagens, desde a escolha dos dados até o uso de modelos de inteligência artificial e atributos extraídos dos áudios para treinamento. O reconhecimento de emoção na fala pode ser abordado de diversas formas. A utilização da Transformada de Pacotes de Wavelet (TPW), dentro deste cenário relativamente recente, possui poucos resultados publicados na literatura. Para entender o uso de TPW em reconhecimento de fala na literatura, foi realizado um mapeamento sistemático.

O processo de mapeamento foi conduzido em várias etapas, visando uma compreensão abrangente do tema. Foi utilizada a ferramenta *on-line* Parsifal [1] para auxiliar na organização e desenvolvimento da pesquisa. Essa ferramenta foi utilizada na definição das perguntas de pesquisa, objetivando criar uma *string* de busca adequada e fazer a triagem dos trabalhos. Após essas etapas, os trabalhos mais relevantes foram analisados e os resultados apresentados nessa seção.

Para encontrar na literatura os trabalhos de reconhecimento de emoção na fala que utilizam a TPW juntamente com algoritmos de aprendizado profundo, foram criadas três perguntas principais:

- "Qual algoritmo de aprendizado de máquina foi utilizado?"
- "Qual família de wavelet mãe foi utilizada?"
- "Qual conjunto de dados foi utilizado para o desenvolvimento do modelo?"

Para encontrar trabalhos relacionados e relevantes ao tema, as bibliotecas digitais *ACM Digital Library*, *El Compendex (Engineering Village)*, *IEEE Digital Library*, *ISI Web of Science*, *PubMed*, *ScienceDirect* e *Scopus* foram selecionadas para as buscas. Os critérios para incluir ou excluir os trabalhos encontrados nas bibliotecas são mostrados na Tabela 1. Além desses critérios para validar os resultados encontrados pelas buscas, três artigos encontrados no mapeamento sistemático realizado por (VIEIRA, 2023) foram escolhidos para balizar o refinamento da *string* de busca. Os trabalhos escolhidos para balizamento foram (FENG; YANG, 2018), (MENG *et al.*, 2021), e (HUANG *et al.*, 2019). Eles se enquadram perfeitamente aos objetivos do mapeamento realizado, uma vez que ratam de reconhecimento de emoção da fala utilizando TPW em modelos de aprendizado profundo.

[1] https://parsif.al/

Tabela 1 – Critérios de inclusão e exclusão de trabalhos

Critério de Inclusão	**Critério de Exclusão**
Estudos que utilizam a TPW e aprendizado de máquina para reconhecimento de emoção na fala.	Capítulos de livro, pôsteres, páginas da web e slides.
	Estudos anteriores a 2014.
	Estudos duplicados.
	Estudos secundários.
	Fora do tema de pesquisa.
	Estudo que não utiliza áudio.
	Não trata de reconhecimento de emoção na fala.
	Não utiliza a TPW.
	Não utiliza aprendizado de máquina.
	Sem acesso.
	Teses, Monografias e Dissertações.

Fonte: Autoria própria

Para realizar as buscas, foi criada a seguinte *string* principal: *("speech emotion recognition" OR "ser" OR "speech emotion classification" OR "speech emotion detection") AND ("deep learning" OR "CNN" OR "DNN" OR "LSTM" OR "MLP" OR "deep network" OR "machine learning" OR "neural network") AND ("wavelet packet" OR "WPD" OR "wavelet packet analysis" OR "wavelet packet decomposition" OR "wavelet packet transform")*. Essa *string* foi refinada, levando em consideração os três artigos iniciais, e foi utilizada para a maioria das pesquisas, exceto para as bibliotecas *PubMed* e *ScienceDirect*. Essas bibliotecas possuem limitações quanto ao tamanho das *strings*, impedindo a utilização da *string* principal, e por esse motivo foram criadas duas novas *strings* para essas bibliotecas. Para *ScienceDirect* foi utilizada a *string ("speech emotion recognition" OR "ser" OR "speech emotion") AND ("CNN" OR "DNN" OR "LSTM" OR "MLP" OR "machine learning") AND ("wavelet packet" OR "WPD")*, e para *PubMed (speech emotion recognition OR ser OR emotion recognition) AND (wavelet packet OR WPD)*.

As *strings* utilizadas ajustam a abrangência da busca de acordo com a natureza de cada repositório. A *string* principal utilizada nas buscas é a mais abrangente e inclui termos relacionados ao reconhecimento de emoções na fala, a uma ampla gama de técnicas de aprendizado profundo e aprendizado de máquina, bem como à análise por TPW. Essa *string* é projetada para cobrir o maior espectro possível de pesquisas nas áreas de interesse. Dessa *string* principal derivam duas versões mais específicas. A *string* utilizada para *ScienceDirect* mantém o foco no reconhecimento de emoções na fala e aprendizado de máquina, além da análise de TPW. Por sua vez, a *string* para *PubMed* é ainda mais simplificada, focando

apenas em reconhecimento de emoções na fala e TPW, sem mencionar técnicas específicas de aprendizado de máquina. Além das *strings* outro critério de busca utilizado foi filtrar trabalhos publicados a partir de 2014. As buscas nas bibliotecas foram realizadas no dia 18/02/2024, e as quantidades de artigos recuperados por bibliotecas científicas digitais buscadas são apresentadas na Tabela 2. A diferença entre os números apresentados se justifica pela natureza de cada repositório e no caso de *PubMed* e *ScienceDirect*, pelas diferentes *strings* utilizadas. O maior número de estudos foi encontrado na *ScienceDirect*, base multidisciplinar com menos termos na *string* de busca.

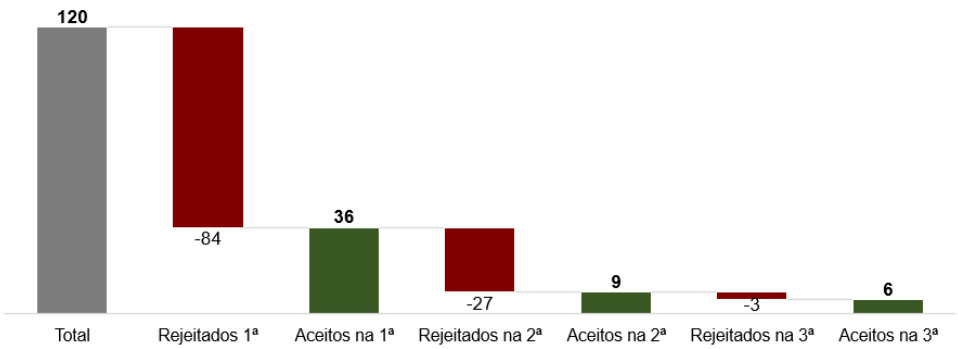
Tabela 2 – Número de trabalhos encontrados por biblioteca

Biblioteca	**N° de trabalhos encontrados**
ACM Digital Library	13
El Compendex	26
IEEE Digital Library	3
ISI Web of Science	7
PubMed	13
ScienceDirect	50
Scopus	8
Total	**120**

Fonte: Autoria própria

A primeira fase de triagem foi feita considerando critérios de exclusão que poderiam ser analisados sem necessidade de ler o resumo das publicações. Os critérios analisados para primeira etapa foram: estudos duplicados, estudos fora do assunto e estudos secundários. Na Figura 5, é possível observar que aproximadamente 30% dos trabalhos foram aceitos para a segunda etapa de triagem.

Figura 5 – Número de trabalhos aceitos e rejeitados por etapa de triagem

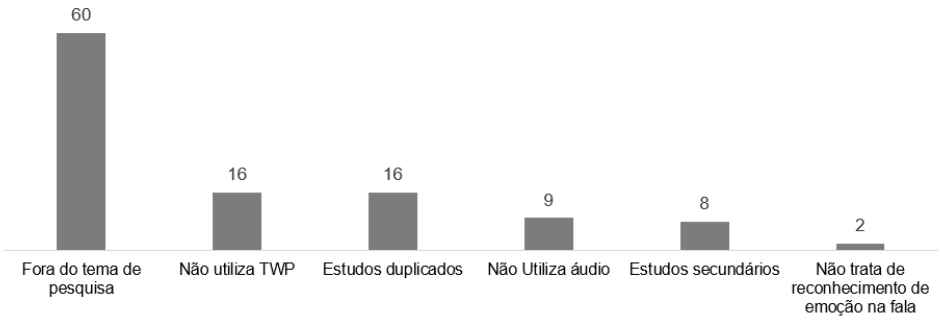


Fonte: Autoria própria

A segunda etapa de triagem foi realizada considerando critérios que necessitariam de mais informações sobre os trabalhos, portanto, foi necessária a leitura e a análise de

palavras-chave e o resumo de cada artigo. Os critérios analisados para segunda etapa foram: trabalhos que não utilizam TPW, trabalhos que não tratam de reconhecimento de emoção na fala, e trabalhos que não utilizam dados de áudio. Ao final da segunda triagem, também observada na Figura 5, somente 10 estudos (8% do total inicial) passaram pelos critérios. Na Figura 6, é mostrado o número de trabalhos rejeitados pelos seus respectivos critérios.

Figura 6 – Número de trabalhos rejeitados por critério de exclusão



Fonte: Autoria própria

No último processo de triagem, realizou-se mais um filtro de data, garantindo que os estudos encontrados não fossem precedentes aos três artigos utilizados para validação citados, reduzindo o período avaliado a trabalhos mais recentes. Isso implicou em um filtro de estudos não anteriores a 2018, resultando na remoção de três trabalhos. Uma observação importante sobre os três trabalhos iniciais foi a descoberta de que o trabalho de (HUANG *et al.*, 2019) havia sido publicado anteriormente em 2016 com pequenas mudanças no texto, e por este motivo esse trabalho acabou por não aparecer na última etapa de triagem. Somente dois trabalhos não identificaram a família de wavelet mãe utilizada. Por este motivo, somente os trabalhos anteriores ao novo filtro de data foram removidos, chegando a um total final de seis trabalhos (5% do total inicial). O resultado final do mapeamento sistemático é apresentado na Tabela 3 , mostrando os modelos utilizados, família de wavelet mãe e bancos de dados utilizados no treino.

Ficou evidente que existem poucos estudos relacionados ao tema de pesquisa deste trabalho. Dentre os trabalhos encontrados, a maioria utiliza a família de wavelet mãe Daubechies e diferentes tipos de modelos e bases de dados. Em relação ao aprendizado de máquina, são encontrados modelos de RCN, *Long Short-Term Memory* (LSTM), *Support Vector Machine* (SVM) e até mesmo árvore de decisão (junto com SVM). Entre as bases de dados de treino, a base *Berlin Emotional Database* (EMODB) (BURKHARDT *et al.*, 2005) ganhou destaque, pois foi utilizada em quatro dos seis estudos. Além da EMODB também são citadas e utilizadas as bases *Ryerson Audio-Visual Database of Emotional Speech and Song* (RAVDESS) (LIVINGSTONE; RUSSO, 2018), *Survey Audio-Visual Expressed Emotion* (SAVEE) (JACKSON; HAQ, 2011), *Interactive Emotional Dyadic*

Tabela 3 – Resultado final do mapeamento sistemático

Estudo	**Wavelet Mãe**	**Modelos utilizados**	**Banco de dados utilizado**
(BHANGALE; KOTHANDARAMAN, 2023)	Daubechies	RNC 1D	EMODB, RAVDESS
(FENG; YANG, 2018)	Daubechies	*Long Short-Term Memory* (LSTM)	CASIA
(WANG; HUO, 2019)	-	*Support Vector Machine* (SVM)	CASIA
(PALO; SUBUDHIRAY; DAS, 2023)	-	Árvore de Decisão, SVM	EMODB, SAVEE
(MENG *et al.*, 2021)	Daubechies	LSTM bidirecional com atenção	EMODB, IEMOCAP
(WANG *et al.*, 2020)	Daubechies	SVM linear (LSVM), SVM com função de base radial (RSVM)	EMODB, EESDB

Fonte: Autoria própria

Motion Capture Database (IEMOCAP) (BUSSO *et al.*, 2008), *Elderly Emotional Speech Database* (EESDB) (WANG, 2018) e *Chinese Natural Emotional Database* (CASIA) (BAO *et al.*, 2014).

Os resultados apresentados na Tabela 4 mostram as melhores acurácias encontradas em diferentes estudos e bases de dados. Destacam-se as altas acurácias alcançadas em alguns estudos, como 98,18% no SAVEE e 97,95% no EMODB, conforme (PALO; SUBUDHIRAY; DAS, 2023). Esses resultados indicam a eficácia das técnicas aplicadas em determinadas bases de dados. No entanto, é importante ressaltar que a comparação direta dos resultados é dificultada pela diversidade de técnicas empregadas em cada estudo. Cada abordagem pode utilizar diferentes algoritmos de aprendizado profundo, métodos de pré-processamento e técnicas de extração de características. Essa variabilidade dificulta determinar qual técnica ou configuração é superior de forma geral. Apesar dos resultados terem boa acurácia, a quantidade de trabalhos encontrados pode sugerir um desafio na aplicação da técnica TPW para reconhecimento de emoções na fala.

Tabela 4 – Melhores resultados de acurácia encontrados nos estudos

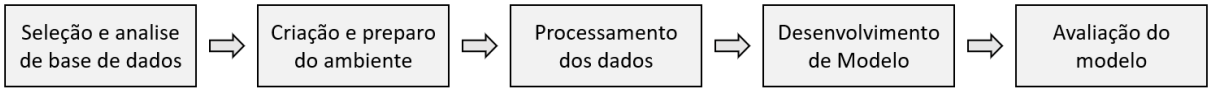
Estudo	**Banco utilizado**	**Melhor acurácia**
(BHANGALE; KOTHANDARAMAN, 2023)	RAVDESS	94,18%
	EMODB	93,31%
(FENG; YANG, 2018)	CASIA	86%
(WANG; HUO, 2019)	CASIA	95%
(PALO; SUBUDHIRAY; DAS, 2023)	SAVEE	98,18%
	EMODB	97,95%
(MENG *et al.*, 2021)	EMODB	82,26%
	IEMOCAP	66,9%
(WANG *et al.*, 2020)	EMODB	79,2%
	EESDB	71,3%

Fonte: Autoria própria

4 MATERIAIS E MÉTODOS

Esse capítulo apresenta, em mais detalhes, os conjuntos de dados utilizados, assim como descreve as etapas do desenvolvimento do trabalho. O fluxograma das etapas de desenvolvimento pode ser visualizado na Figura 7. O conteúdo do capítulo é apresentado da seguinte forma: na seção 4.1 são apresentadas as bases de dados utilizadas no desenvolvimento do trabalho; na seção 4.2, o ambiente de execução e ferramentas utilizadas em sua preparação são descritos; na seção 4.3, são mostradas as etapas de pré-processamento dos dados; na seção 4.4, os modelos de classificação utilizados discutidos; e, por fim, na Seção 4.4.1, são apresentadas as métricas utilizadas na avaliação dos modelos.

Figura 7 – Fluxograma de desenvolvimento do trabalho



Fonte: Autoria Própria

4.1 Conjuntos de Dados

Como visto na Seção 2.2 da fundamentação teórica, existem diferentes tipos de conjuntos de dados que podem ser utilizados para fazer o reconhecimento da emoção na fala. Existem também diferentes conjuntos de dados que podem ser utilizados para realizar o treino de modelos de reconhecimento de emoções. Observando os trabalhos relacionados, alguns conjuntos de dados se destacam pela sua ampla utilização. Assim, este trabalho utiliza as bases de dados EMODB (BURKHARDT *et al.*, 2005), RAVDESS (LIVINGSTONE; RUSSO, 2018), SAVEE (JACKSON; HAQ, 2011). Outra base muito importante utilizada para treinar os modelos neste trabalho é a base CORAA (MARCACINI; JUNIOR; CASANOVA, 2022), sendo esta uma base de dados na língua portuguesa do Brasil. Além dessas bases, o conjunto de dados do SofiaFala (RISSATO; MACEDO, 2021) é utilizado como teste de marcação do modelo desenvolvido.

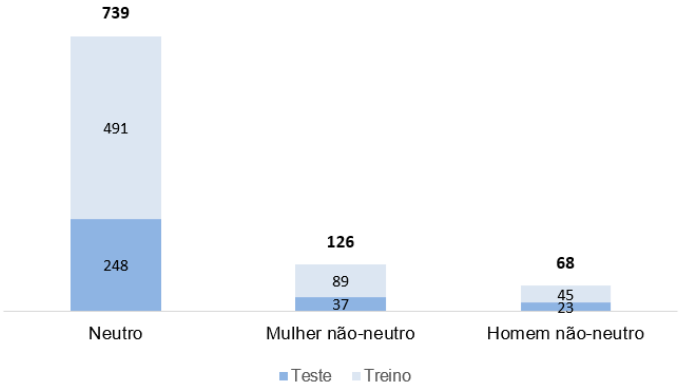
4.1.1 CORAA

Os dados deste conjunto foram apresentados no evento da PROPOR (*International Conference on Computational Processing of Portuguese Language*) de 2022, e podem ser acessados no *GitHub*[1]. O conjunto é constituído por 933 arquivos de áudios de fala espontânea no idioma português do Brasil. Os áudios são falas de homens e de mulheres sem a marcação de idade. As emoções são rotuladas nas seguintes três classes: homem não-neutro, mulher não neutro e neutro. As distribuições das emoções podem ser observadas

[1] https://github.com/rmarcacini/ser-coraa-pt-br/

na Figura 8, onde é possível observar que os dados não são balanceados e a categoria neutra possui um volume muito maior de ocorrências.

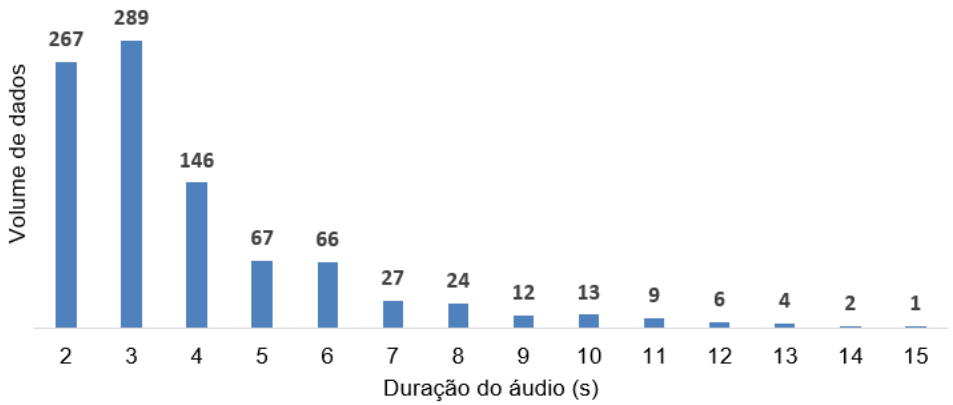
Figura 8 – Distribuição de classes de emoções na base CORAA



Fonte: Autoria Própria

A taxa de aquisição dos áudios é de 16kHz em formato mono e tem duração total de 1 hora e 11 segundos (21 minutos e 49 segundos de teste e 39 minutos e 22 segundos de treino). Como observado na Figura 9, a duração média dos áudios é de 3 segundos, podendo oscilar entre 2 e 14 segundos por arquivo. O conjunto de dados é dividido em dois, com 67% dos arquivos em treino e 33% em teste (625 arquivos de treino e 308 de teste). Outra observação importante sobre este conjunto de dados é a presença de ruídos e outros fatores, como a ocorrência de mais de um falante e a mistura de vozes masculinas e femininas em um mesmo áudio. Esses ruídos podem prejudicar a qualidade da classificação. Ao analisar algumas amostras de áudio, é possível observar a presença de múltiplas vozes em alguns arquivos.

Figura 9 – Distribuição de tempo de aúdio na base CORAA

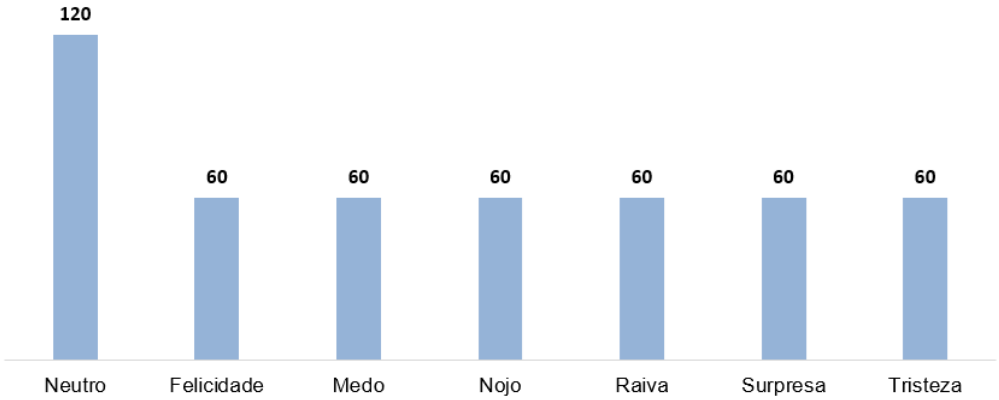


Fonte: Autoria Própria

4.1.2 SAVEE

Os dados SAVEE (*Surrey Audio-Visual Expressed Emotion*) foram desenvolvidos em 2007 pela Universidade Surrey e podem ser acessados em seu site[2]. O conjunto é constituído por 480 arquivos de áudios de fala atuada no idioma inglês da Inglaterra. Os áudios são de quatro atores homens, com idades entre 27 e 31 anos, provenientes de diferentes regiões do Reino Unido (um do País de Gales, dois do sul da Inglaterra e um da Escócia). As emoções são rotuladas nas seguintes sete classes: raiva, nojo, medo, felicidade, tristeza, surpresa e neutro. As distribuições das emoções podem ser observadas na Figura 10, onde os dados apresentam um certo equilíbrio, uma vez que as emoções estão distribuídas de forma uniforme. No entanto, nota-se uma predominância de marcações "neutras", que são duas vezes mais frequentes do que as outras emoções. A distribuição de falas entre os atores é balanceada, com 120 arquivos de áudio por ator. Outra informação importante é o fato de as marcações serem feitas, por meio da análise de expressões faciais, mas apenas os arquivos de áudio são disponibilizados no conjunto de dados.

Figura 10 – Distribuição de classes de emoções na base SAVEE

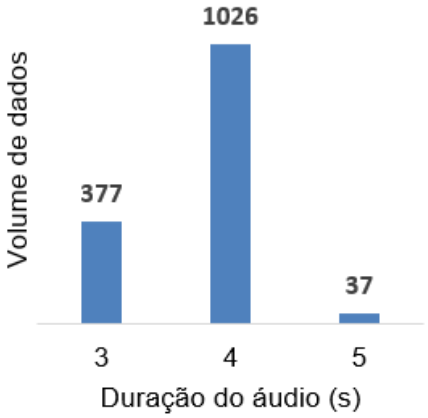


Fonte: Autoria Própria

A taxa de aquisição dos áudios é de 44.1kHz em formato mono e tem duração total 30 minutos e 42 segundos, com média de 3 segundos por arquivo, podendo oscilar entre 3 e 5 segundos por arquivo. Como observado na Figura 11, a duração média dos áudios é de 3 segundos. O conjunto de dados é composto por fala atuada, o que pode influenciar a naturalidade das expressões emocionais capturadas. Outra observação importante deste conjunto de dados é a diversidade regional dos falantes, que pode introduzir variações sutis nas características fonéticas dos áudios.

[2] http://kahlan.eps.surrey.ac.uk/savee/

Figura 11 – Distribuição de tempo de aúdio na base SAVEE



Fonte: Autoria Própria

4.1.3 RAVDESS

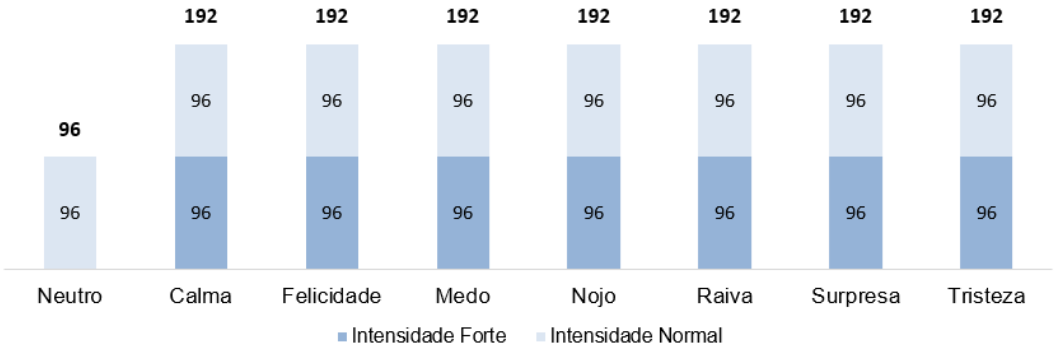
Os dados RAVDESS (*Ryerson Audio-Visual Database of Emotional Speech and Song*) estão disponíveis e publicados no site Zenodo[3]. O conjunto original é constituído por 7356 arquivos de dados, incluindo faces e voz, faces e apenas voz. A coleção somente de áudios é formada por 2452 arquivos, sendo segmentada em músicas (1012 arquivos) e fala (1440 arquivos). Este trabalho tem como foco a utilização apenas dos áudios de fala.

Os áudios são de fala atuada e foram criados por 24 atores profissionais, 12 homens e 12 mulheres, na língua inglesa norte-americana. Os critérios de avaliação para os áudios foram validação emocional, intensidade e genuinidade. As emoções são rotuladas nas seguintes oito classes: calma, alegria, tristeza, raiva, medo, surpresa, nojo e neutro; com intensidades normal ou forte. Como observado na Figura 12, as emoções são distribuídas de forma balanceada, apesar de haver um número menor de registros neutros em comparação com as outras emoções (metade da quantidade das demais emoções). O menor número de neutros ocorre devido à ausência de variação de intensidade forte no neutro.

Os arquivos de áudio têm uma taxa de aquisição de 48kHz e são disponibilizados em formato mono e estéreo. Cada ator realizou 60 falas, totalizando 1440 arquivos de áudio apenas de fala. A duração total dos áudios é de 88 minutos e 48 segundos. Como obsevado na Figura 13, os arquivos tem duração média de 3 segundos variando entre 2 e 7 segundos por arquivo.

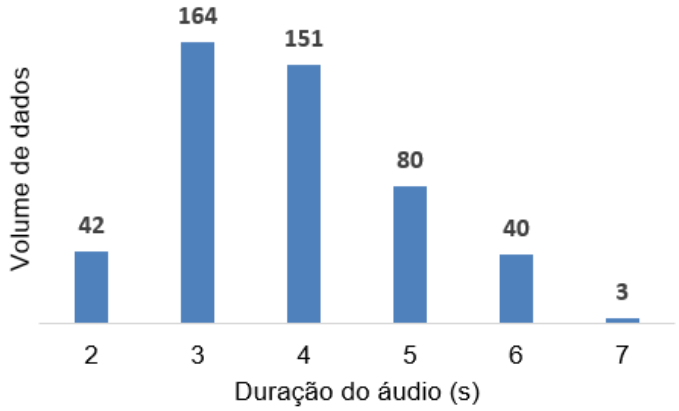
[3] https://zenodo.org/records/1188976

Figura 12 – Distribuição de classes de emoções na base RAVDESS



Fonte: Autoria Própria

Figura 13 – Distribuição de tempo de aúdio na base RAVDESS



Fonte: Autoria Própria

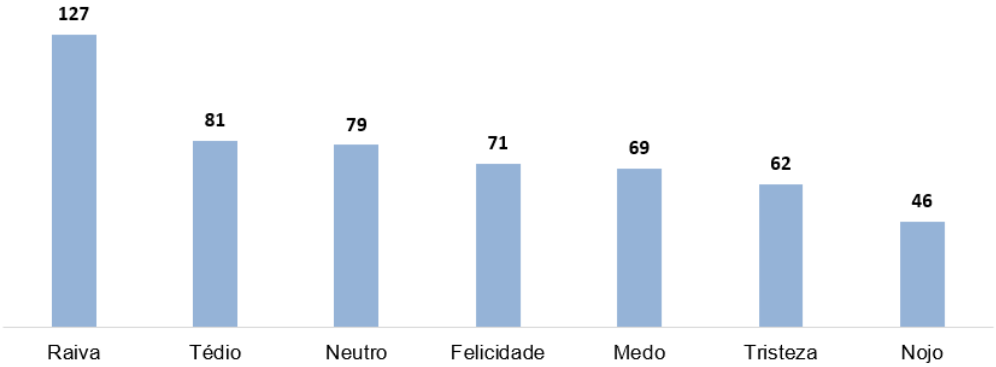
4.1.4 EMODB

Os dados EMODB (*Berlin Database of Emotional Speech*) foram desenvolvidos pelo Instituto de Fala e Comunicação da Universidade Técnica de Berlim e podem ser acessados em seu site[4]. O conjunto é constituído por 535 arquivos de áudios de fala atuada no idioma alemão. Os áudios são de dez atores, cinco homens e cinco mulheres, com idades entre 21 e 35 anos. As emoções são rotuladas nas seguintes sete classes: raiva, tédio, nojo, medo/ansiedade, felicidade, tristeza e neutro. As distribuições das emoções podem ser observadas na Figura 14, onde é possível notar que os dados não são balanceados, com algumas emoções representadas mais frequentemente do que outras. A emoção com maior representação é a raiva, que constitui 24% do conjunto de dados, enquanto o nojo (a menor classe) compõe apenas 9%.

A taxa de aquisição dos áudios é de 48kHz, posteriormente convertida para 16kHz,

[4] http://www.emodb.bilderbar.info

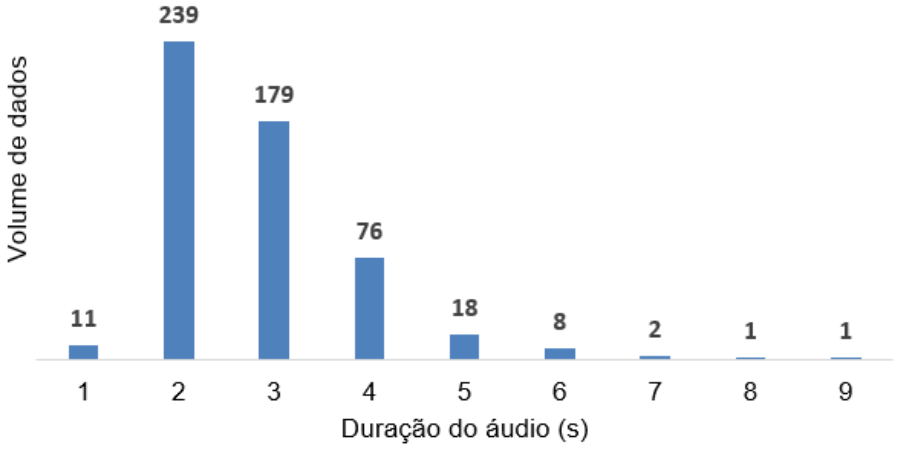
Figura 14 – Distribuição de classes de emoções na base EMODB



Fonte: Autoria Própria

em formato mono, com duração total de 24 minutos e 47 segundos. Como observado na Figura 15, a duração média dos áudios é de 2 segundos por arquivo, podendo oscilar entre 1 e 9 segundos. O conjunto de dados é composto por fala atuada, com dez frases repetidas de diversas formas, o que pode influenciar a naturalidade das expressões emocionais capturadas.

Figura 15 – Distribuição de tempo de aúdio na base EMODB



Fonte: Autoria Própria

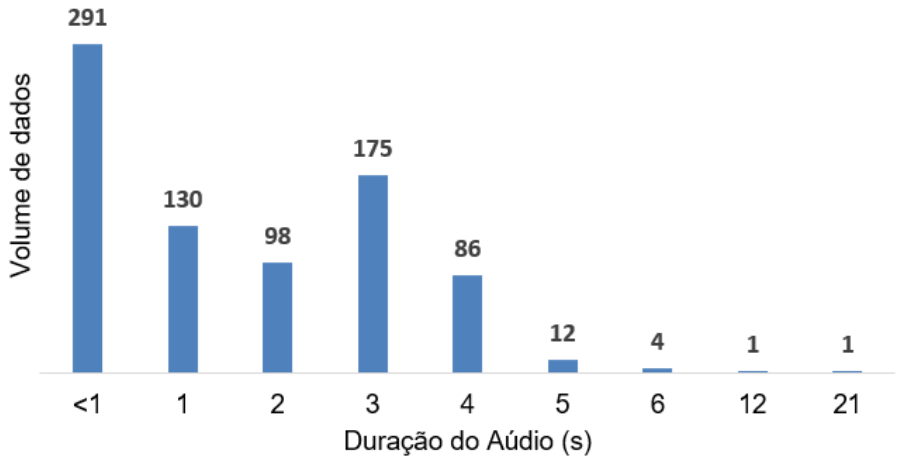
4.1.5 Base de Áudios do SofiaFala

A base de dados de fala do SofiaFala foi criada com o objetivo de apoiar o desenvolvimento de soluções assistivas para terapias de fonoaudiologia. Ela é composta por um total de 1.387 arquivos de áudio coletados ao longo de vários períodos de tempo, sendo que a última coleta, realizada entre fevereiro e junho de 2024. A base de dados tem 808 arquivos de áudio, os quais foram utilizados neste trabalho. Esses áudios estão no

formato .wav e foram gravados em língua portuguesa por homens, mulheres e crianças com deficiência de fala.

Os arquivos de áudio possuem uma taxa de aquisição de 48KHz, e a base apresenta uma duração total de 23 minutos e 3 segundos. A duração dos áudios varia entre 0,1 segundos e 21,3 segundos, com uma média de 1,7 segundos. O tempo de áudio está distribuído de maneira desigual, como pode ser observado na Figura 16. A maioria dos arquivos de áudio (89%) é mono, com apenas 11% dos arquivos sendo estéreo. Um aspecto importante desta base é a ausência de rótulos ou informações explícitas sobre as emoções presentes nos áudios, uma vez que esse conjunto de dados não é formado por fala espontânea, mas por repetição de frases indicadas em texto e em áudio.

Figura 16 – Distribuição de tempo de aúdio na base SofiaFala



Fonte: Autoria Própria

4.2 Ambiente de Execução e Ferramentas Utilizadas

O trabalho foi realizado utilizando o *Google Colab Pro*[5], um ambiente de desenvolvimento baseado em nuvem. A configuração utilizada incluía uma GPU NVIDIA T4 com 15 GB de memória dedicada e 12,7 GB de RAM do sistema, permitindo o processamento eficiente de operações de aprendizado profundo e cálculos intensivos. Os arquivos do projeto foram armazenados no *Google Drive*. O ambiente de desenvolvimento foi construído sobre o *Python* 3.8.19.

Com suporte da GPU, o sistema utilizou o *CUDA* 11.7. O desenvolvimento dos modelos de aprendizado profundo (RNC) foi realizado com o *framework Torch* 1.13.1, juntamente com a versão compatível do *cuDNN*, otimizando o desempenho das operações de aprendizado profundo. A biblioteca *torchaudio* 0.13.1+cu117 foi utilizada para manipulação

[5] https://colab.research.google.com/

dos áudios, enquanto *torch-audiomentations* 0.11.0 e *audiomentations* 0.30.0 foram usadas para técnicas de *data augmentation*.

Além disso, as bibliotecas *numpy* 1.22.3 e *pandas* 1.5.2 foram empregadas para manipulação de dados, enquanto *PyWavelets* 1.4.1 foi usada para implementar a TPW no processamento de sinais. Para visualização de dados, foram utilizadas as bibliotecas *matplotlib* 3.6.3 e *seaborn* 0.12.2. O desenvolvimento foi conduzido em *Jupyter Notebook* 1.0.0 e *notebook* 6.5.2.

Para avaliar o desempenho dos modelos desenvolvidos, utilizou-se a biblioteca *Scikit-learn* 1.5.0, que oferece diversas métricas de avaliação, como precisão (*precision*), revocação (*recall*) e medida F (*F1 score*), para análise e comparação de diferentes abordagens de aprendizado de máquina.

4.3 Pré-processamento de dados

Na primeira etapa do processo, os áudios foram pré-processados para otimizar o uso de processamento na GPU. Todos os áudios foram reamostrados para uma taxa de 8kHz e, quando necessário, transformados para o formato mono, garantindo consistência na entrada dos dados. Em seguida, foi realizada a normalização dos áudios, ajustando-os para terem o mesmo comprimento, preenchendo com zeros aqueles que eram mais curtos, para garantir uniformidade no tamanho das entradas durante o treinamento dos modelos.

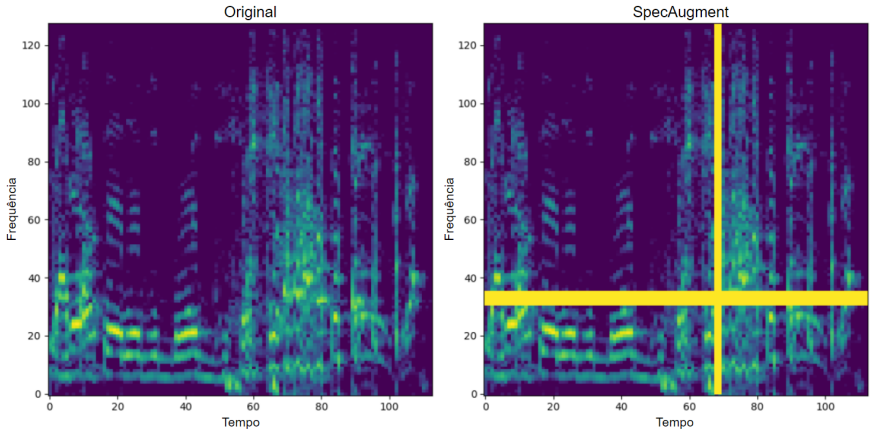
Durante a etapa de extração de características, os áudios passaram por um processo de transformação em espectrograma na escala mel, que é uma representação visual das frequências do áudio ao longo do tempo, cujas altas frequências foram representadas com mais detalhes do que as baixas. Em seguida, os espectrogramas foram submetidos a um processo de *data augmentation*, utilizando a técnica de SpecAugment, que consiste em aplicar perturbações aleatórias nos espectrogramas, como mascaramento de frequências e durações (como observado na Figura 17), a fim de aumentar a robustez e a variabilidade dos dados durante o treinamento dos modelos.

Para extrair características adicionais, utilizou-se o pacote PyWavelets, com a wavelet mãe de Daubechies diretamente sobre o espectrograma na escala mel. Isso permitiu decompor o espectrograma em componentes de frequência e escala diferentes, fornecendo uma representação mais abrangente e detalhada dos dados de áudio.

4.4 Modelo de Classificação

Para este trabalho foi utilizada uma arquitetura de RNC (RNC-10) desenvolvida por Kong *et al.* (2020) e empregada por Gauy and Finger (2022) e por Vieira (2023) em seus estudos de classificação da base CORAA. A aplicação da RNC-10 nos trabalhos citados

Figura 17 – Exemplo de aplicação de *SpecAugment*

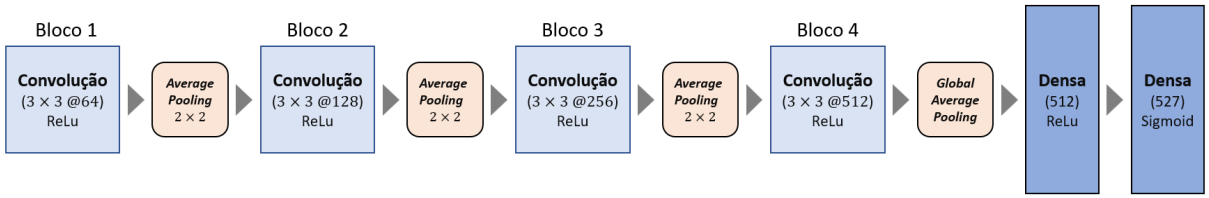


Fonte: Autoria Própria

apresentou desempenho superior em comparação com outras redes neurais, destacando-se como uma arquitetura mais eficiente para essa tarefa.

A arquitetura RNC-10 é composta por vários blocos convolucionais seguidos por camadas densas. Como observado na Figura 18, a rede é composta por 4 blocos, E cada bloco contém duas camadas convolucionais de *kernel* 3x3 seguidas por normalização de *batch*, uma função de ativação ReLU e uma camada de *pooling* médio de *kernel* 2x2. O número de canais de saída das camadas convolucionais aumenta progressivamente em cada bloco: o primeiro bloco tem 64 canais, o segundo 128, o terceiro 256 e o quarto 512. Após os blocos convolucionais, a rede inclui uma camada de *global pooling*, que permite agregar as características espaciais de forma global antes de passar para a camada densa. Em seguida, a rede possui uma camada densa com 512 unidades e função de ativação ReLU. A camada final é uma camada densa com um número de unidades igual ao número de classes da tarefa de classificação, utilizando a função de ativação sigmoide para produzir as probabilidades de cada classe.

Figura 18 – Arquitetura do modelo RNC-10



Fonte: Autoria Própria

4.4.1 Métricas de Avaliação de Modelos

As métricas de avaliação de acurácia do F1-*Score* são adotadas neste estudo e são fundamentais para a análise e validação dos resultados obtidos. A acurácia, apesar

de transmitir uma falsa impressão de desempenho para dados desbalanceados, também é considerada como uma métrica de avaliação, pois é a mais amplamente utilizada nos trabalhos relacionados encontrados durante o mapeamento sistemático. A acurácia fornece uma visão geral do desempenho do modelo, representando a proporção de previsões corretas em relação ao total de previsões realizadas. Sua fórmula pode ser observada na Equação 4.1 como uma relação de falsos positivos, falsos negativos, verdadeiros positivos e verdadeiros negativos.

$$Acurácia = \frac{\text{Verdadeiro Positivo} + \text{Verdadeiro Negativo}}{\text{Total de previsões}} \tag{4.1}$$

O F1-*Score* é uma das principais métricas empregadas, especialmente devido à sua eficácia em conjuntos de dados desbalanceados. Essa métrica também foi utilizada no evento da PROPOR para avaliação da base CORAA. Sua fórmula pode ser vista na Equação 4.4 como uma relação de Precisão e Revocação, que por sua vez também utilizam os mesmos argumentos usados para o cálculo da acurácia.

$$Precisão = \frac{\text{Verdadeiro Positivo}}{\text{Verdadeiro Positivo} + \text{Falso Positivo}} \tag{4.2}$$

$$Revogação = \frac{\text{Verdadeiro Positivo}}{\text{Verdadeiro Positivo} + \text{Falso Negativo}} \tag{4.3}$$

$$F1\ score = \frac{2 * (Precisão * Revogação)}{Precisão + Revogação} \tag{4.4}$$

O estudo utilizou o *F1 score* e a acurácia como métricas de avaliação em todos os conjuntos de dados, garantindo uma análise mais abrangente e validação dos resultados. A acurácia oferece uma visão geral do desempenho do modelo, embora tenha limitações com dados desbalanceados, uma vez que ela mede a proporção de previsões corretas, mas em um conjunto desbalanceado, um modelo pode obter uma alta acurácia simplesmente prevendo a classe majoritária na maioria dos casos. Por exemplo, se 90% dos dados pertencem a uma classe, um modelo que sempre prevê essa classe terá 90% de acurácia, mas não terá aprendido a identificar a classe minoritária. O *F1 score* complementa essa análise, proporcionando uma avaliação mais precisa nesses cenários. *F1 score* é útil quando há desequilíbrio entre as classes. O *F1 score* dá uma visão melhor do desempenho do modelo nesses casos, já que não favorece a classe majoritária e dá mais ênfase ao equilíbrio entre precisão e revocação.

No entanto, o *F1 score* não é sempre a melhor escolha em todos os contextos. Em algumas situações, outras métricas como a AUC-ROC podem ser mais adequadas, especialmente se o objetivo for avaliar a capacidade do modelo de distinguir entre classes

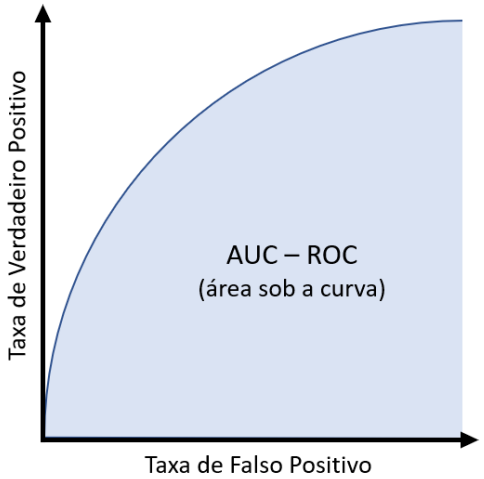
positivas e negativas. A curva ROC faz a relação entre a taxa de verdadeiros positivos (Equação 4.5) e a taxa de falsos positivos (Equação 4.6).

$$\text{Taxa de Verdadeiro Positivo} = \frac{\text{Verdadeiro Positivo}}{\text{Verdadeiro Positivo} + \text{Falso Negativo}} \tag{4.5}$$

$$\text{Taxa de Falso Positivo} = \frac{\text{Falso Positivo}}{\text{Verdadeiro Negativo} + \text{Falso Positivo}} \tag{4.6}$$

Conforme ilustrado na Figura 19, quanto mais a curva se aproxima do canto superior esquerdo do gráfico, maior é a capacidade discriminativa do modelo. O valor da AUC, que varia de 0 a 1, representa a área sob a curva ROC. Um valor de 1 indica uma separação perfeita entre as classes, enquanto um valor de 0,5 sugere que o modelo está classificando de forma aleatória.

Figura 19 – Exemplo de curva ROC



Fonte: Autoria Própria

5 EXPERIMENTOS E RESULTADOS

Este capítulo apresenta os experimentos realizados e os resultados obtidos no contexto desta pesquisa. O foco dos experimentos foi a análise da base de dados CORAA, seguindo uma adaptação da metodologia proposta por Vieira (2023). A partir dos experimentos conduzidos, o melhor resultado obtido foi utilizado como referência de parâmetros para a análise da performance do modelo nas demais bases de dados.

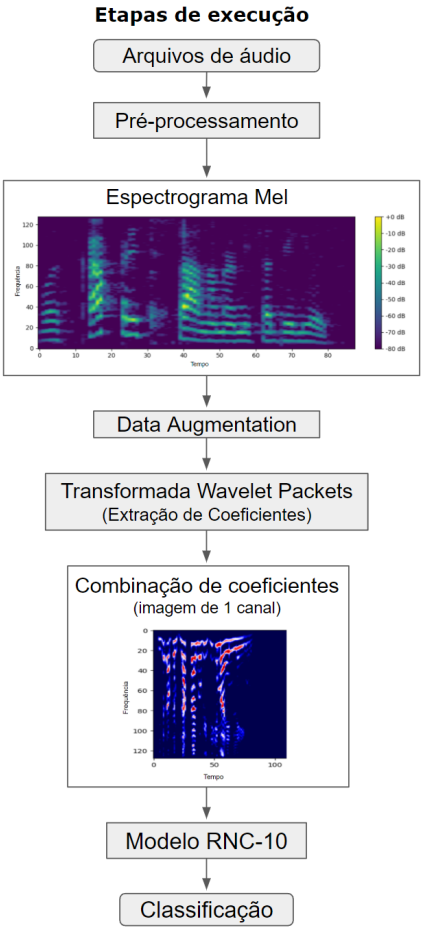
Os dados de treinamento foram divididos em cinco partes utilizando a técnica de validação cruzada estratificada (*k-fold*), de forma a manter a mesma proporção de classes do conjunto original em cada divisão. No caso da base CORAA, a separação dos conjuntos de treino e validação foi realizada utilizando um subconjunto previamente definido dentro da própria base, como observado anteriormente na Figura 8. Para os demais conjuntos de dados, foi adotada uma divisão de 64% para treino, 16% para validação e 20% teste. Em todas as bases, os arquivos foram divididos aleatoriamente entre os diferentes falantes utilizando uma mesma semente para a divisão dos dados. Conforme detalhado anteriormente, os dados da base CORAA consistem em 1.352 arquivos, contendo gravações de fala espontânea em português. As classes esperadas de saída estão divididas em três categorias: “homem com emoção”, “mulher com emoção” e “neutro”.

Para a realização dos testes, utilizou-se uma adaptação da metodologia proposta por Vieira (2023), aplicando inicialmente os mesmos parâmetros de teste que apresentaram os melhores resultados em sua proposta, uma vez que esta também utiliza a transformada Wavelet. A adaptação consistiu na implementação da transformada de pacotes Wavelet (TPW) como etapa de extração de características, enquanto as demais etapas de pré-processamento e o modelo RNC-10 foram mantidos sem alterações. As etapas de execução do código, conforme ilustradas na Figura 20, seguiram a seguinte ordem: inicialmente, realizou-se a extração do espectrograma Mel dos áudios, seguida pela etapa de aumento de dados (*data augmentation*). Posteriormente, aplicou-se a transformada Wavelet, e os coeficientes resultantes dessa transformada foram combinados para formar uma imagem de um canal. Em seguida, esse resultado foi alimentado no modelo RNC-10, que realizou a classificação final.

5.1 Realizando os experimentos

O primeiro teste teve como objetivo replicar o melhor resultado do trabalho de referência, utilizando os mesmos parâmetros, mas aplicando a TPW em vez da transformada discreta de Wavelet (TDW). Para isso, foi utilizado o espectrograma Mel com 128 mels, janela de 400 pontos e deslocamento de 200. A Wavelet escolhida foi a “db4” com nível de decomposição 3. Os resultados de *F1 score* macro podem ser observados na Tabela 5, cuja

Figura 20 – Etapas de execução



Fonte: Autoria Própria

replicação utilizando a transformada discreta de Wavelet obteve um *F1 score* de 0,578 para validação e 0,527 para teste. Já o teste inicial com TPW apresentou um *F1 score* de 0,506 para validação e 0,439 para teste. O tempo médio de processamento para cada *fold* foi semelhante em ambos os casos, levando cerca de quatro horas por *fold*, totalizando aproximadamente 20 horas para a execução completa.

Tabela 5 – Resultados de *F1 score* macro obtidos nos testes iniciais com espectrograma Mel de 128 mels, janela de 400 pontos e deslocamento de 200

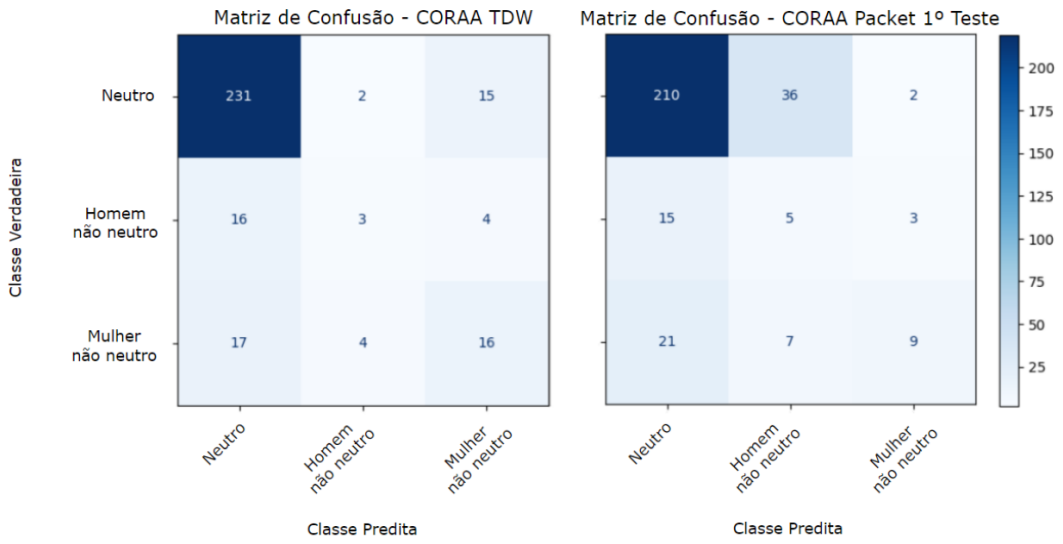
Método	***F1 score* Validação**	***F1 score* Teste**
Replica de referência com TDW	0,578	0,527
Teste inicial com TPW	0,506	0,439

Fonte: Autoria própria

Conforme mostrado na Figura 21, que apresenta as matrizes de confusão do primeiro teste com os dados CORAA, observa-se que a replicação com a TDW tem um índice de acerto semelhante ao da TPW para a classe “Homem não neutro”. Entretanto, a TPW

apresenta um desempenho inferior nas classes “Neutro” e “Mulher não neutro”, indicando uma menor precisão para essas categorias. Essas diferenças de desempenho indicam a necessidade de ajustes finos nos parâmetros para melhorar o desempenho do modelo em todas as classes.

Figura 21 – Matrizes de confusão comparando TDW e TWP



Fonte: Autoria Própria

Embora o valor exato do trabalho de referência não tenha sido alcançado, os resultados ficaram próximos, com uma diferença de 0,039 em relação ao *F1 score* de teste do trabalho de referência (0,566). No entanto, a análise das matrizes de confusão e dos resultados de *F1 score* indicou que a TDW apresentou melhor desempenho com os parâmetros testados. Por essa razão, decidiu-se explorar diferentes configurações, realizando variações na etapa de extração do espectrograma na escala Mel e nos parâmetros da extração da TWP para buscar melhores resultados.

Para os testes de variação de parâmetros, foram alterados os valores de janela (TF) e deslocamento para a extração do espectrograma Mel, enquanto o número de mels foi mantido em 128. Em relação a TPW, variou-se o tipo de Wavelet da família Daubechies (db), utilizando diferentes ordens, como “db4”, “db6” e “db8”, mantendo o nível de decomposição fixo em 3. As combinações de parâmetros testadas estão detalhadas na Tabela 6. Essas combinações foram baseadas na estratégia utilizada pelo trabalho de referência, que aplicou variações de parâmetros semelhantes para testar diferentes resoluções temporais e analisar o impacto das diferentes ordens na decomposição dos sinais. Esses ajustes visaram explorar configurações alternativas que pudessem melhorar o desempenho do modelo, uma vez que os resultados iniciais indicaram a necessidade de ajustes finos nos parâmetros.

A Tabela 7 apresenta os melhores resultados obtidos para cada uma das combinações

Tabela 6 – Combinações de parâmetros testadas para a extração do espectrograma Mel e TPW

Combinação	**Mels**	**TF**	**Deslocamento**	**Daubechies**	**Nível**
1	128	400	200	db4	3
2	128	400	200	db6	3
3	128	400	200	db8	3
4	128	1024	320	db4	3
5	128	1024	320	db6	3
6	128	1024	320	db8	3

Fonte: Autoria própria

de parâmetros testados. Analisando os resultados, a combinação 3 se destacou como a mais promissora, mostrando um equilíbrio entre o *F1 score* na validação (0,515) e no teste (0,494), com uma diferença mínima de 0,021 entre eles. Essa pequena variação sugere que o modelo alcançou um bom nível de generalização, sem sofrer de *overfitting*, o que é um indicador positivo para o desempenho em dados não vistos. Além disso, a combinação 3 apresentou um valor de acurácia de 0,740 e um ROC de 0,701, o que reforça sua robustez e equilíbrio em termos de sensibilidade e especificidade, quando comparado às demais combinações.

Tabela 7 – Melhores resultados obtidos para cada combinação de parâmetros testados

Combinação	***F1 score* Validação**	***F1 score* Teste**	**Acurácia**	**ROC**
1	0,510	0,448	0,727	0,655
2	0,554	0,436	0,672	0,677
3	0,515	0,494	0,740	0,701
4	0,546	0,479	0,818	0,666
5	0,493	0,470	0,766	0,633
6	0,484	0,524	0,672	0,677

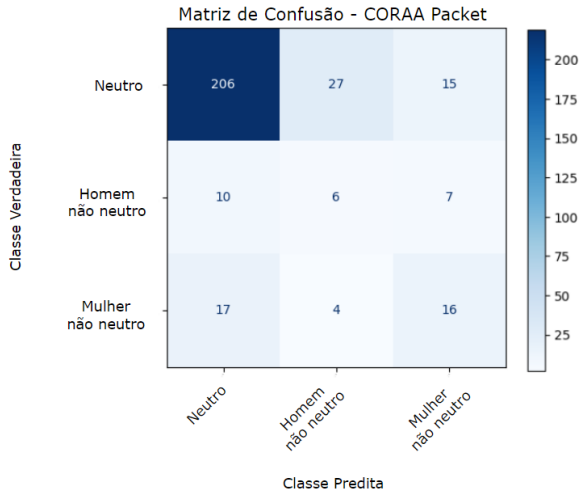
Fonte: Autoria própria

Apesar das variações realizadas nos parâmetros de extração de características, o tempo de processamento permaneceu relativamente constante entre todas as combinações. Essa estabilidade de processamento indica que as mudanças não impactaram significativamente o custo computacional, mantendo o tempo de *fold* de aproximadamente quatro horas.

Conforme mostrado na Figura 22, que apresenta a matriz de confusão da melhor combinação, observa-se uma melhoria na classificação da classe “Mulher não neutro” em

comparação ao primeiro teste, aproximando-se dos resultados obtidos com a TDW. No entanto, a matriz ainda indica um índice de erro mais elevado na classe “Neutro”.

Figura 22 – Matriz de confusão obtida com parametros da combinação 3



Fonte: Autoria Própria

Mesmo após os ajustes realizados, os resultados não superaram o desempenho alcançado com a TDW, que apresentou um *F1 score* superior tanto na validação quanto no teste. Diante disso, outras abordagens foram exploradas na tentativa de melhorar os resultados. Nesse sentido, aplicou-se técnicas adicionais como o *SpecAugment* e o VAD, no pré-processamento dos áudios.

Foi testada a técnica de aumento de dados (*SpecAugment*) para verificar seu impacto no desempenho do modelo, utilizando os mesmos parâmetros de extração de características da melhor combinação identificada anteriormente. No entanto, os resultados obtidos com o *data augmentation* foram inferiores aos alcançados sem o uso desta técnica. Como mostrado na Tabela 8, a combinação com *SpecAugment* resultou em um *F1 score* de 0,494 para validação e 0,414 para teste, com acurácia de 0,744 e ROC de 0,599, indicando uma redução no desempenho em comparação aos testes anteriores.

O VAD também foi testado utilizando os mesmos parâmetros de extração de características. A técnica foi aplicada aos áudios brutos, empregando o algoritmo do WebRTC. Durante os testes, foram apresentadas dificuldades em detectar áudio em alguns arquivos: em seis deles, não houve detecção de áudio, embora, após uma análise rápida, tenha sido constatada a presença de som. Mesmo após ajustes nos parâmetros do VAD, o problema persistiu, e o uso desta técnica resultou em um desempenho pior do que o esperado. Conforme apresentado na Tabela 8, a aplicação do VAD resultou em um *F1 score* de 0,457 para validação e 0,456 para teste, com acurácia de 0,779 e ROC de 0,619.

A Figura 23 representa as matrizes de confusão. É possível observar que tanto o *data augmentation* quanto o VAD resultaram em uma redução na capacidade do modelo

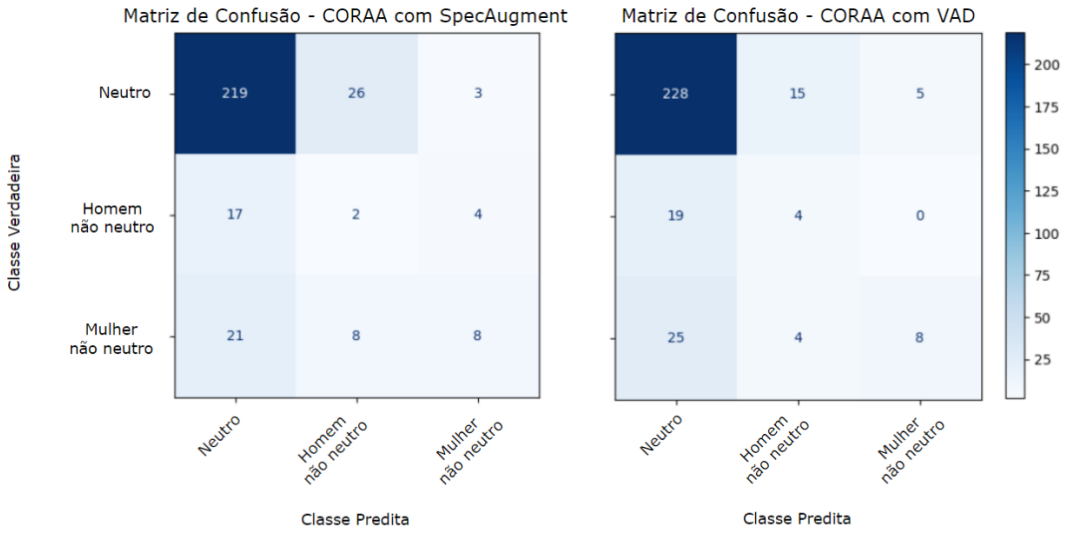
Tabela 8 – Resultados de desempenho do modelo com *SpecAugment* e VAD

Método	***F1 score* Validação**	***F1 score* Teste**	**Acurácia**	**ROC**
SpecAugment	0,494	0,414	0,744	0,599
VAD	0,457	0,456	0,779	0,619

Fonte: Autoria própria

de identificar as classes “Homem não neutro” e “Mulher não neutro”. Em contrapartida, houve um pequeno ganho na classificação da classe “Neutro”, especialmente com o uso do VAD. Esses resultados indicam que as técnicas testadas não proporcionaram a melhoria desejada no desempenho do modelo e, em alguns casos, até prejudicaram a capacidade de classificação de determinadas classes. Devido ao seu desempenho insatisfatório, essas técnicas não foram utilizadas na etapa subsequente de análise do modelo em outras bases de dados.

Figura 23 – Matrizes de confusão para o modelo com *SpecAugment* e VAD



Fonte: Autoria Própria

A melhor combinação de parâmetros identificada nos testes anteriores foi utilizada para avaliar o desempenho do modelo em outras bases de dados, seguindo a divisão discutida anteriormente. Os melhores resultados obtidos para cada base de dados são apresentados na Tabela 9.

Os resultados mostraram que a melhor performance foi alcançada com a base de dados EMODB, com um *F1 score* de 0,724 na validação e 0,669 no teste, bem como uma acurácia de 0,689 e um ROC de 0,897. Por outro lado, as bases de dados SAVEE e RAVDESS apresentaram uma performance inferior. A RAVDESS obteve o pior resultado, com um *F1 score* de 0,394 na validação, 0,332 no teste, acurácia de 0,398 e um ROC de 0,817. Esse resultado já era esperado, uma vez que a base EMODB é considerada a mais

Tabela 9 – Melhores resultados obtidos para as bases EMODB, SAVEE e RAVDESS utilizando a melhor combinação de parâmetros

Base de Dados	***F1 score* Validação**	***F1 score* Teste**	**Acurácia**	**ROC**
EMODB	0,724	0,669	0,689	0,897
SAVEE	0,663	0,533	0,579	0,851
RAVDESS	0,394	0,332	0,398	0,817

Fonte: Autoria própria

simples dentre as avaliadas, sendo composta por dados menos complexos do que as outras bases.

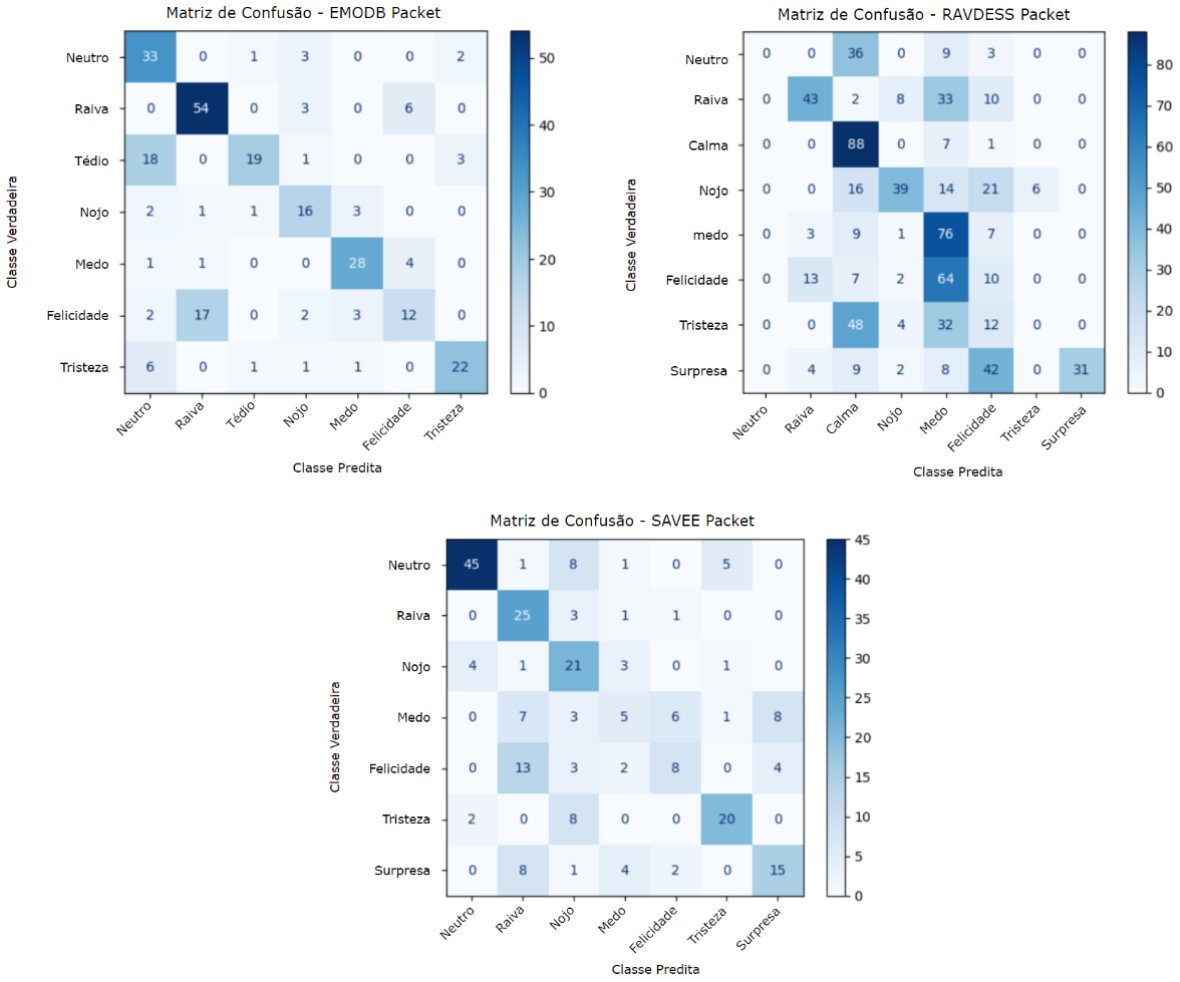
Esses resultados inferiores eram previsíveis, dado que a SAVEE e a RAVDESS possuem marcações feitas também, por meio da análise de expressões faciais (análise multimodal), além da análise do áudio. O modelo criado neste trabalho foca exclusivamente em características acústicas. A ausência da componente visual no processo de classificação pode ter contribuído para o desempenho inferior nessas bases.

Ao analisar a Figura 24, que apresenta as matrizes de confusão para as bases SAVEE, EMODB e RAVDESS, é possível identificar padrões específicos de desempenho do modelo para cada base de dados.

Os resultados de acurácia por emoção estão detalhados na Tabela 10. O desempenho do modelo pode variar significativamente conforme a base de dados e as características emocionais de cada conjunto, o que justifica as diferenças observadas nas classificações. Na base EMODB, os arquivos de áudio em alemão têm as classes de saída divididas em sete categorias: “Raiva”, “Medo”, “Tristeza”, “Felicidade”, “Nojo”, “Surpresa” e “Neutro”. A matriz de confusão e tabela de acurácias indicam que o modelo tem um bom desempenho na classificação de “Neutro”(85%), “Raiva”(86%) e “Medo”(82%), com um desempenho razoável também para “Nojo”(70%) e “Tristeza”(71%). No entanto, observa-se maior confusão em classes como “Tédio”(46%) e “Felicidade”(33%), cujas previsões se dispersam entre várias outras classes. Isso reflete que, embora a base EMODB seja mais simples, o modelo ainda enfrenta dificuldades em classificar algumas emoções.

Para a base RAVDESS, os arquivos de áudio em inglês têm como classes de saída oito categorias: “Calmo”, “Felicidade”, “Tristeza”, “Raiva”, “Medo”, “Surpresa”, “Nojo” e “Neutro”. O valor de ROC foi relativamente alto, mesmo com um *F1 score* mais baixo. Isso sugere que o modelo consegue discriminar razoavelmente bem entre algumas classes específicas. A análise da tabela de acurácias para a base RAVDESS revela um desempenho mais variado do modelo. As classes “Calmo”(92%) e “Medo”(79%) têm o melhor desempenho (ajudando a elevar o valor do ROC), enquanto há significativa confusão entre outras classes, como “Neutro”(2%), “Felicidade”(10%) e “Tristeza”(1%). A

Figura 24 – Matrizes de confusão para as bases EMODB, SAVEE e RAVDESS utilizando TPW



Fonte: Autoria Própria

distribuição homogênea de classes na RAVDESS, exceto para "Neutro," que possui menos exemplos, pode ter contribuído para o desempenho inferior nessa classe. Comparando com a base EMODB, o modelo apresentou um desempenho significativamente inferior para "Raiva" e "Nojo". Embora o modelo tenha mostrado um bom desempenho para a classe "Medo", semelhante ao observado na EMODB, seu desempenho foi significativamente pior nas outras classes.

A base SAVEE, composta por arquivos de áudio em inglês, apresentou um desempenho intermediário entre EMODB e RAVDESS. Para esta base, as classes de saída dividem-se em sete categorias: "Neutro", "Raiva", "Tristeza", "Medo", "Felicidade", "Surpresa" e "Nojo". O modelo demonstra uma boa capacidade de identificar a classe "Neutro"(75%), que é mais prevalente nesta base, o que pode ter facilitado o aprendizado para essa classe específica. No entanto, o desempenho para "Medo"(17%) foi consideravelmente inferior, especialmente quando comparado com as bases EMODB e RAVDESS. Além disso, o modelo também teve dificuldades em identificar "Felicidade" (27%), que apresentou resultados

Tabela 10 – Acurácia por emoção nas bases EMODB, SAVEE e RAVDESS, utilizando TPW

Emoção	**EMODB**	**SAVEE**	**RAVDESS**
Neutro	85%	75%	2%
Raiva	86%	83%	45%
Nojo	70%	70%	41%
Medo	82%	17%	79%
Felicidade	33%	27%	10%
Tristeza	71%	67%	1%
Surpresa	-	50%	32%
Calma	-	-	92%
Tédio	46%	-	-

Fonte: Autoria própria

inferiores ao da EMODB, mas ligeiramente superiores ao da RAVDESS. Embora a base SAVEE tenha oferecido um desempenho razoável em algumas classes, o modelo ainda enfrenta desafios consideráveis na diferenciação de certas emoções.

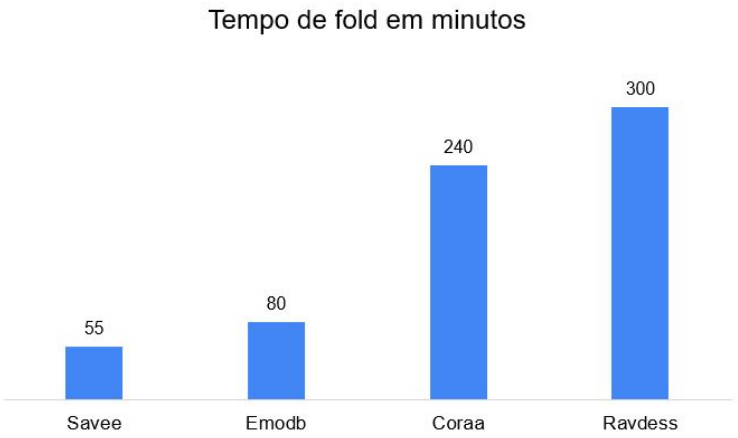
Também foram analisados os tempos de processamento de cada *fold* para as diferentes bases de dados, conforme ilustrado na Figura 25. Os tempos de processamento variaram significativamente entre as bases: para a base SAVEE (480 arquivos), o tempo médio por *fold* foi de 55 minutos; para EMODB (535 arquivos) , 80 minutos; para CORAA (933 arquivos), 240 minutos; e para RAVDESS (1440 arquivos), 300 minutos. Essas diferenças refletem a complexidade e o tamanho das bases de dados, com as bases maiores e mais complexas (RAVDESS e CORAA) demandando tempos de processamento consideravelmente mais longos.

Além dos tempos de processamento, também foi observado o consumo de recursos de *hardware* durante os experimentos. O maior uso de RAM do sistema registrado foi de 5,4 GB, dentro de um total disponível de 12,7 GB, enquanto o maior uso de RAM da GPU atingiu 5,8 GB, com um total de 15 GB disponíveis. Os testes foram realizados em uma GPU NVIDIA T4, fornecida pelo Google Colab, utilizando uma unidade de computação por hora de 1,76, um recurso computacional alocado dinamicamente para a execução dos notebooks.

5.2 Comparando a performance dos modelos

Finalmente, os resultados obtidos neste trabalho foram comparados com os valores reportados por Vieira (2023), que utilizou a TDW, conforme mostrado na Figura 26.

Figura 25 – Tempos de processamento médio por *fold* para as bases SAVEE, EMODB, CORAA e RAVDESS



Fonte: Autoria Própria

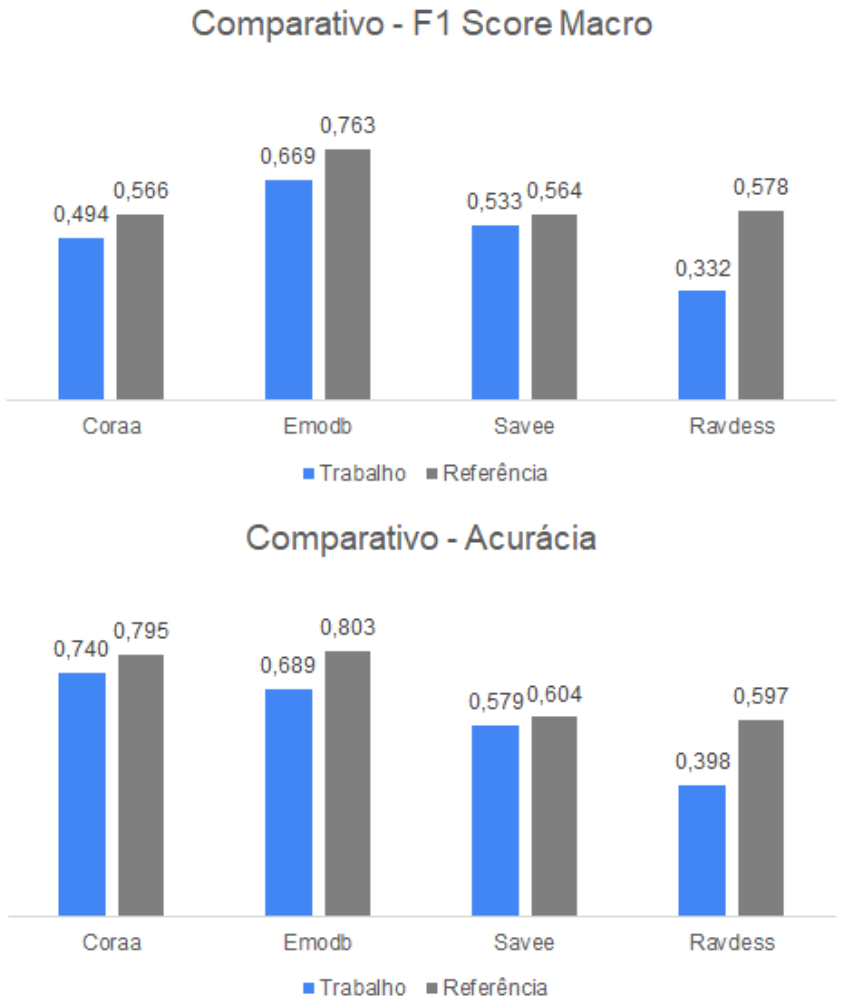
Foi observado em alguns casos que os resultados de *F1 score* e acurácia obtidos com a TPW ficaram próximos aos encontrados na referência. Para as bases de dados CORAA, EMODB e SAVEE, os valores de *F1 score* e acurácia apresentaram diferenças relativamente pequenas em relação aos resultados do trabalho de referência, sugerindo que nestas bases a estratégia de TPW teve um desempenho comparável ao da TDW. Mesmo com desempenhos semelhantes para determinadas emoções, é possível identificar tanto acertos em comum quanto diferenças significativas no desempenho do modelo nas bases de dados.

Entretanto, para a base de dados RAVDESS, os resultados obtidos com a TPW ficaram significativamente abaixo dos alcançados pela TDW, tanto em *F1 score* Macro (0,332 comparado a 0,578) quanto em Acurácia (0,398 comparado a 0,597). Esse resultado indica que a estratégia de TPW não conseguiu igualar ou se aproximar do desempenho alcançado pela TDW em uma base de dados de maior complexidade e com características multimodais. Apesar da TPW mostrar resultados próximos aos da TDW nas bases de dados CORAA, EMODB e SAVEE, a performance global da estratégia de TPW não superou os valores estabelecidos pelo trabalho de referência.

Os resultados de acurácia por emoção comparando o modelo baseado em TPW com o modelo de referência utilizando TDW estão ilustrados na Figura 27. Observa-se que, na base EMODB, o modelo com TPW apresentou um desempenho superior em “Medo”, “Nojo” e “Neutro”, superando o modelo com TDW nessas categorias. No entanto, o modelo com TDW teve um desempenho significativamente melhor em “Raiva” e “Tédio”. Além disso, a TDW também apresentou uma maior acurácia em “Felicidade” (64%), enquanto o TPW alcançou apenas 33% de acurácia para essa emoção.

Já na base de dados SAVEE, ambos os modelos, usando TPW e TDW, foram eficazes em identificar a classe “Neutro”, embora o modelo com TDW tenha sido mais

Figura 26 – Comparativo de desempenho entre TPW e TDW nas diferentes bases de dados



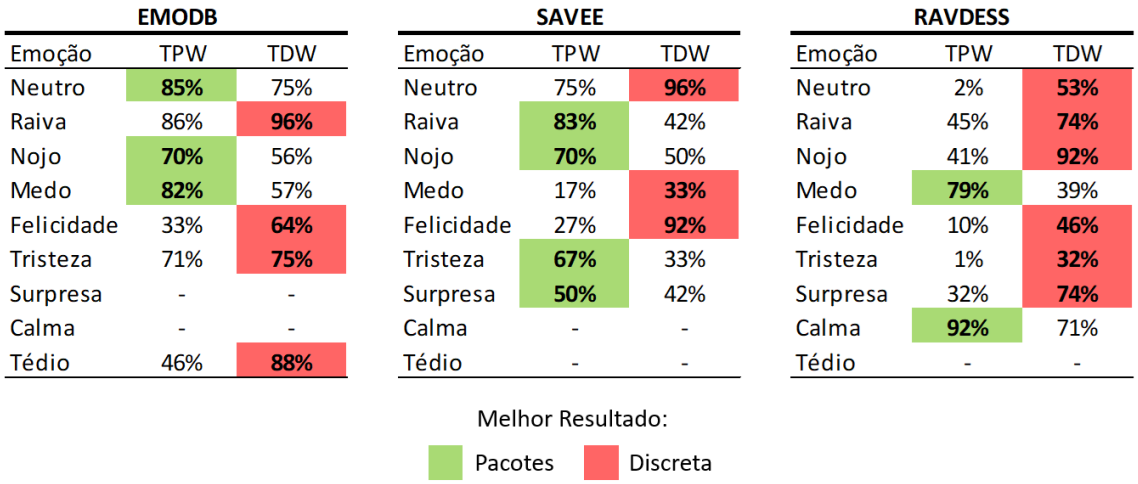
Fonte: Autoria Própria

eficiente, com uma acurácia de 96%. O modelo com TPW, por outro lado, apresentou um desempenho significativamente superior nas classes “Raiva” e “Nojo” em comparação a TDW. Ambos os modelos enfrentaram dificuldades na classificação de “Medo”, enquanto a TDW obteve um resultado melhor em “Felicidade”. Já para a classe “Tristeza”, a TPW teve um desempenho mais favorável, e para “Surpresa”, ambos os modelos alcançaram valores semelhantes.

Para a base RAVDESS, o modelo utilizando TPW apresentou um bom desempenho apenas nas classes “Calmo” e “Medo”, enquanto o trabalho de referência utilizando TDW obteve melhores resultados em outras emoções, como “Raiva”, “Nojo”, “Surpresa” e “Neutro”. Diferentemente deste trabalho, o modelo baseado em TDW teve um desempenho inferior na classe “Medo”, com acurácia de 39%.

Conforme mostrado na Tabela 11, os melhores resultados de *F1 score* obtidos na PROPOR 2022 variam entre 0,509 e 0,728. O melhor resultado alcançado neste trabalho,

Figura 27 – Comparativo de acurácia entre TPW e TDW nas bases EMODB, SAVEE e RAVDESS por emoção



Fonte: Autoria Própria

com um *F1 score* de 0,494, ficou abaixo dos resultados apresentados naquele evento, mas se aproximou dos valores obtidos pelos participantes da quarta posição, com uma diferença de 0,015. É importante destacar que o grande diferencial do trabalho que obteve o primeiro lugar (GAUY; FINGER, 2022), com um *F1 score* de 0,728, foi o uso de um modelo semelhante ao empregado neste estudo, mas com uma rede neural já pré-treinada em um conjunto de dados de áudio com mais de 5 mil horas. Esse pré-treinamento em um grande volume de dados contribuiu significativamente para o desempenho superior observado. Explorar modelos semelhantes pode ser uma direção para trabalhos futuros de pesquisa, visando melhorar os resultados alcançados.

Tabela 11 – Melhores resultados de *F1 score* obtidos na PROPOR 2022

Trabalho	**F1 score**
(GAUY; FINGER, 2022)	0,728
(ALVES *et al.*, 2022)	0,535
(PERIN; MATSUBARA, 2022)	0,525
(SCARANTI *et al.*, 2022)	0,509

Fonte: Autoria própria

5.3 Aplicando o modelo no SofiaFala

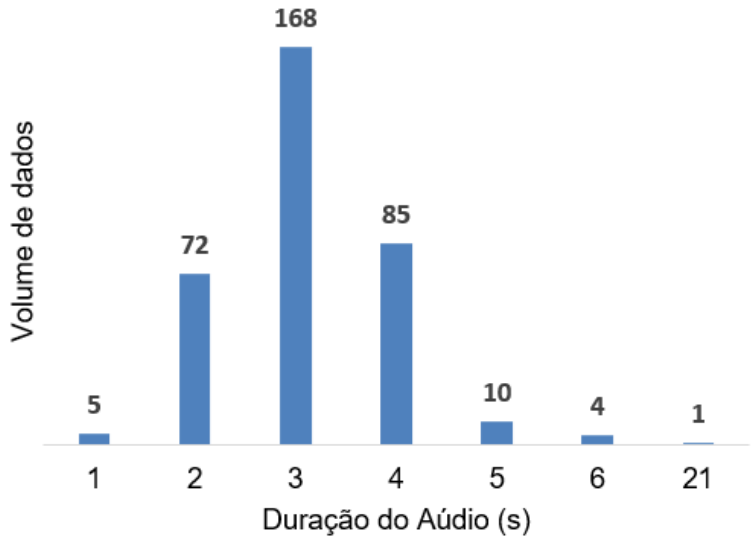
Por fim, na última etapa deste trabalho, foram realizados os testes do modelo desenvolvido utilizando os dados da base de áudios do SofiaFala. Durante essa fase, foi identificado um problema inicial relacionado ao formato dos arquivos de áudio. Embora os arquivos estivessem no formato .wav, não era possível abri-los corretamente, pois um erro

indicava incompatibilidade de formato. Após uma análise mais detalhada, descobriu-se que os arquivos estavam, na verdade, em um contêiner WebM, e não no formato .wav conforme esperado. Todos os arquivos foram, então, convertidos corretamente para o formato .wav. Durante o processo de conversão, foram identificados 10 arquivos vazios, que foram desconsiderados no restante do trabalho.

Feita a conversão, foi realizada uma análise detalhada da base de dados. Nessa análise, 465 arquivos (58% do total) foram considerados inadequados para os testes, devido a problemas como áudios sem som, ruídos excessivamente altos que impediam a identificação de qualquer trecho de voz, ou distorções causadas por interrupções involuntárias, tornando impossível a compreensão das frases. Apenas 343 áudios (42% do total) foram considerados adequados para serem utilizados nos testes.

Analisando os 343 áudios válidos, foi verificado que todos os arquivos possuíam apenas um canal de áudio (mono). A duração total desse subconjunto foi de 18 minutos e 16 segundos, com os áudios variando entre 1,0 e 21,3 segundos, e uma duração média de 3,2 segundos. A distribuição dos tempos dos áudios, como mostrado na Figura 28, agora apresenta uma distribuição mais próxima de uma normal, ao contrário da distribuição do conjunto original.

Figura 28 – Distribuição de tempo de aúdio no subconjunto de áudios validados do Sofia-Fala

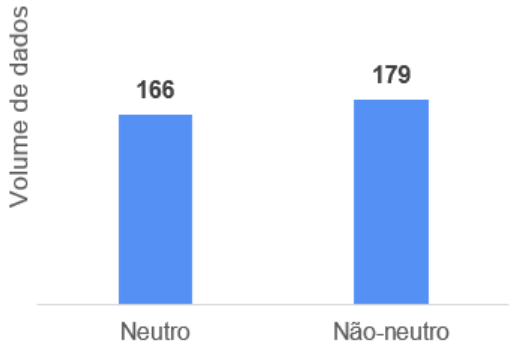


Fonte: Autoria Própria

Após a análise e seleção dos arquivos adequados, foi realizado o teste utilizando o melhor modelo desenvolvido para a base CORAA, aplicando-o ao novo subconjunto de dados do SofiaFala. Para simplificar a análise, as classes de saída do modelo foram agrupadas em apenas duas categorias: “neutro” e “não neutro”, com o objetivo de verificar a presença de emoções nos áudios, independentemente do tipo de emoção.

Os resultados das classificações podem ser observados na Figura 29, onde 179 áudios (52%) foram classificados como “não neutro” e 166 áudios (48%) como “neutro”. O número de áudios identificados como contendo emoção foi relativamente alto. No entanto, como a base SofiaFala não possui rótulos de emoção, os dados, após serem classificados pelo modelo, foram analisados de forma qualitativa.

Figura 29 – Distribuição das classificações de áudios em neutro e não neutro no conjunto de dados SofiaFala



Fonte: Autoria Própria

Ao realizar uma análise dos áudios classificados como “não neutro”, foi difícil identificar claramente emoções presentes nos áudios. A maioria dos áudios classificados como contendo emoção não apresenta características emocionais evidentes, podendo ser, na realidade, considerados neutros. Isso se deve ao fato de a base ter sido criada, a partir da repetição de frases ou áudios por pessoas com a fala comprometida, não podendo ser considerada uma base de fala espontânea, conforme descrito anteriormente.

Outro ponto observado é que os arquivos de áudio possuem identificadores no início do nome, o que permite a identificação de diferentes falantes. Ao analisar os resultados por identificador de falante, percebeu-se que a classificação de “neutro” ou “não neutro” se mantém constante para a maioria dos áudios de um mesmo falante. Em outras palavras, o modelo tende a classificar todos os áudios de uma mesma pessoa da mesma forma, o que sugere que o modelo está associando a classe mais à pessoa do que ao conteúdo emocional do áudio. Esse comportamento pode ser problemático para a utilização do modelo nesse conjunto de dados, uma vez que os falantes possuem deficiência de fala, e isso pode estar influenciando de forma inadequada as classificações do modelo. Esses testes foram realizados devido ao envolvimento do autor no grupo de desenvolvimento da tecnologia SofiaFala. A base de áudios foi originalmente criada para um propósito diferente, voltado ao acompanhamento de tratamentos de fala e não à identificação de emoções. No entanto, a base foi utilizada para testar a proposta do modelo com o intuito de explorar uma possível alternativa para analisar a neutralidade da fala coletada.

6 CONCLUSÕES

Muitas pessoas enfrentam dificuldades de comunicação pela fala, impactando sua interação social, qualidade de vida e sucesso profissional. Nesse contexto, o reconhecimento de emoções na fala é essencial para o desenvolvimento de tecnologias assistivas que ofereçam suporte a pessoas com deficiências de fala. O objetivo do trabalho foi avaliar a aplicação da TPW combinada com aprendizado profundo para o reconhecimento de emoções, focando na base de fala espontânea CORAA. A escolha dessa base se deve à sua relevância para o desenvolvimento de soluções que possam auxiliar projetos como o SofiaFala[1], que se beneficiariam de modelos capazes de lidar com as complexidades da fala espontânea.

Durante o desenvolvimento, diversos desafios forma enfrentado, começando pela dificuldade de replicar o modelo utilizado como referência de (VIEIRA, 2023). Esse problema reforça a importância de garantir todos os requisitos e configurações necessários para a correta execução de um modelo, já que pequenos detalhes podem impactar significativamente os resultados ou até mesmo impedir a execução do programa. Outra dificuldade encontrada foi a escassez de trabalhos na literatura que abordassem o reconhecimento de emoções utilizando a TPW ou outros tipos de Wavelets.

A classificação dos dados da base CORAA se mostrou desafiadora, não apenas pela tarefa de identificar emoções em fala espontânea, mas também pela necessidade de distinguir entre falantes masculinos e femininos. O conjunto de dados apresenta três classes: homem com emoção, mulher com emoção e neutro, exigindo uma sub-tarefa de identificação do sexo do falante, o que aumenta a complexidade do problema. Além disso, os ruídos e perturbações presentes nos áudios da base CORAA representaram outro desafio significativo, conforme constatado em uma análise qualitativa e corroborado por Alves *et al.* (2022), que destacou a presença de vozes sobrepostas, ruídos e variabilidade de gênero nas vozes.

Além da base CORAA, os resultados nas bases EMODB, SAVEE e RAVDESS variaram conforme a emoção. Na EMODB, o modelo com a TPW teve bom desempenho em “Neutro”, “Raiva” e “Medo”, superando a TDW em “Neutro” e “Medo”, mas a TDW foi melhor em “Raiva” e “Tédio”. Na SAVEE, a TPW se destacou em “Raiva” e “Nojo”, enquanto a TDW foi superior em “Felicidade”. Ambos os modelos tiveram dificuldades em “Medo”, com desempenho mais baixo na TPW. Na RAVDESS, a TPW teve bom

[1] O Projeto SofiaFala é uma tecnologia desenvolvida com o intuito de melhorar a qualidade de vida de pessoas com dificuldades de fala através da captação e análise de sons e imagens produzidos durante a execução do exercício fonoaudiológico para individualizar a intervenção terapêutica (RISSATO; MACEDO, 2021). Site oficial do projeto http://dcm.ffclrp.usp.br/sofiafala/

resultado em “Calmo” e “Medo”, mas a TDW foi melhor em “Raiva”, “Nojo”, “Surpresa” e “Neutro”. Esses resultados indicam que, embora o modelo com a TPW tenha mostrado potencial em algumas classes, ele enfrenta desafios em outras e não supera o desempenho do modelo com a TDW.

Quando comparados aos trabalhos discutidos no Capítulo 3, os resultados obtidos neste estudo são significativamente inferiores. No entanto, esses estudos utilizaram uma combinação diversificada de técnicas, como modelos pré-treinados em grandes conjuntos de dados, métodos de pré-processamento sofisticados e técnicas avançadas de extração de características. Essa variedade de abordagens não só dificulta a comparação direta entre os estudos, mas também a comparação com os resultados obtidos neste trabalho de conclusão de curso. Além disso, é importante destacar que, dos trabalhos encontrados na literatura, nenhum utilizou uma base de dados de fala espontânea, realizando o reconhecimento de emoções em condições que podem não ser iguais às da fala espontânea.

Embora os resultados deste trabalho sejam satisfatórios, com um *F1 score* de 0,494 e uma acurácia de 0,744 na base CORAA, há muitas formas que podem ser exploradas para melhorar o desempenho, conforme mencionado anteriormente. Técnicas adicionais de aumento de dados, o uso de algoritmos alternativos de VAD, e a experimentação com diferentes parâmetros para o espectrograma Mel e a transformada de Wavelet são alguns dos caminhos possíveis. No entanto, o maior destaque deve ser dado à possibilidade de utilizar um modelo pré-treinado, como evidenciado pelo melhor resultado alcançado para a base CORAA. Mesmo com todas as variações de parâmetros testadas neste estudo, ou comparando a TWP com a TDW, nenhum dos ajustes resultou em um ganho tão significativo quanto o trabalho que utilizou uma rede neural pré-treinada em um grande conjunto de dados.

Ao testar o modelo nos dados do SofiaFala, os resultados indicaram um desempenho insatisfatório, evidenciando um possível viés de classificação. O modelo foi treinado de forma dependente do falante, permitindo que o mesmo falante apareça tanto no conjunto de treino quanto no de teste, o que pode ter influenciado os resultados. Esse viés é preocupante, pois o modelo tende a atribuir uma mesma classe a todos os áudios de um mesmo falante, independentemente do conteúdo emocional, tornando-o inadequado para aplicação em pessoas com deficiência de fala.

Para evitar esses problemas e aprimorar o modelo em trabalhos futuros, seria interessante realizar uma marcação mais detalhada da base SofiaFala, seja por meio de especialistas humanos ou utilizando outros modelos que possam fornecer resultados mais robustos. Com essas marcações, seria possível conduzir estudos adicionais para identificar e entender o viés do modelo ao classificar áudios de pessoas com deficiência de fala. Além disso, uma marcação mais precisa permitiria o desenvolvimento de soluções independentes do falante, possibilitando o treinamento de modelos que generalizem melhor

para diferentes indivíduos, sem depender de dados específicos de cada falante. Essa abordagem abriria caminho para futuras pesquisas e melhorias no reconhecimento de emoções em fala espontânea, especialmente no contexto de pessoas com deficiência de fala. Futuros trabalhos também podem explorar o uso de modelos pré-treinados com potencial para alcançar melhorias substanciais no reconhecimento de emoções na fala. Embora seja possível investigar outros parâmetros de wavelet, técnicas de extração de espectrograma Mel e até testar novos algoritmos de detecção de atividade de voz, o uso de modelos pré-treinados em grandes volumes de dados foi o que apresentou os resultados mais promissores entre todas as soluções analisadas.

REFERÊNCIAS

ALBAWI, S.; MOHAMMED, T. A.; AL-ZAWI, S. Understanding of a convolutional neural network. *In*: **2017 International Conference on Engineering and Technology (ICET)**. [*S.l.: s.n.*], 2017. p. 1–6.

ALVES, C. *et al.* Transfer learning and data augmentation techniques applied to speech emotion recognition in se&r 2022. *In*: **SE&R@ PROPOR**. [*S.l.: s.n.*], 2022. p. 25–36.

ARAS, I. *et al.* Health related quality of life in parents of children with speech and hearing impairment. **International Journal of Pediatric Otorhinolaryngology**, v. 78, n. 2, p. 323–329, 2014. ISSN 0165-5876. Available at: <https://www.sciencedirect.com/science/article/pii/S0165587613006320>.

BAO, W. *et al.* Building a chinese natural emotional audio-visual database. *In*: **2014 12th International Conference on Signal Processing (ICSP)**. [*S.l.: s.n.*], 2014. p. 583–587.

BHANGALE, K.; KOTHANDARAMAN, M. Speech emotion recognition based on multiple acoustic features and deep convolutional neural network. **Electronics**, v. 12, p. 839, 02 2023.

BURKHARDT, F. *et al.* A database of german emotional speech. *In*: . [*S.l.: s.n.*], 2005. v. 5, p. 1517–1520.

BUSSO, C. *et al.* Iemocap: interactive emotional dyadic motion capture database. **Language Resources and Evaluation**, Springer, v. 42, n. 4, p. 335–359, 2008. ISSN 1574-0218.

CORBALLIS, M. **From Hand to Mouth: The Origins of Language**. Princeton University Press, 2002. 1–5 p. ISBN 9780691116730. Available at: <https://books.google.com.br/books?id=yEd_FchjDDMC>.

EKMAN, P. An argument for basic emotions. **Cognition and Emotion**, Routledge, v. 6, n. 3-4, p. 169–200, 1992. Available at: <https://doi.org/10.1080/02699939208411068>.

FENG, T.; YANG, S. Speech emotion recognition based on lstm and mel scale wavelet packet decomposition. *In*: **Proceedings of the 2018 International Conference on Algorithms, Computing and Artificial Intelligence**. New York, NY, USA: Association for Computing Machinery, 2018. (ACAI '18). ISBN 9781450366250. Available at: <https://doi.org/10.1145/3302425.3302444>.

FONTANARI, J. F. Reflexões sobre a origem e evolução da linguagem. **Ciências & Letras**, p. 247–258, 2009.

GAO, R.; YAN, R. **Wavelets: Theory and Applications for Manufacturing**. [*S.l.: s.n.*], 2010. 1-81 p. ISBN 978-1-4419-1544-3.

GAUY, M. M.; FINGER, M. Pretrained audio neural networks for speech emotion recognition in portuguese. **arXiv preprint arXiv:2210.14716**, 2022.

GOKHALE, M.; KHANDUJA, D. Time domain signal analysis using wavelet packet decomposition approach. **IJCNS**, v. 3, p. 321–329, 01 2010.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep Learning**. [*S.l.: s.n.*]: MIT Press, 2016. <http://www.deeplearningbook.org>.

GRAPS, A. "an introduction to wavelets". **IEEE Comp. Sci. Engi.**, v. 2, p. 50–61, 02 1995.

GU, Y. *et al.* A survey of computer-aided diagnosis of lung nodules from ct scans using deep learning. **Computers in Biology and Medicine**, v. 137, p. 104806, 2021. ISSN 0010-4825. Available at: <https://www.sciencedirect.com/science/article/pii/S0010482521006004>.

GUIDO, R. C. *et al.* Cwt x dwt x dtwt x sdtwt: clarifying terminologies and roles of different types of wavelet transforms. **International Journal of Wavelets, Multiresolution and Information Processing**, 2020.

HIRSCHBERG, J.; MANNING, C. D. Advances in natural language processing. **Science**, v. 349, n. 6245, p. 261–266, 2015. Available at: <https://www.science.org/doi/abs/10.1126/science.aaa8685>.

HUANG, Y. *et al.* Feature fusion methods research based on deep belief networks for speech emotion recognition under noise condition. **Journal of Ambient Intelligence and Humanized Computing**, v. 10, 05 2019.

JACKSON, P.; HAQ, S. ul. **Surrey Audio-Visual Expressed Emotion (SAVEE) database**. 2011.

JOHNSON, C. J.; BEITCHMAN, J. H.; BROWNLIE, E. Twenty-year follow-up of children with and without speech-language impairments: Family, educational, occupational, and quality of life outcomes. ASHA, 2010.

KASAMA, S. T.; BRASOLOTTO, A. G. Percepção vocal e qualidade de vida. **Pró-Fono Revista de Atualização Científica**, SciELO Brasil, v. 19, p. 19–28, 2007.

KHAN, N. A review on speech emotion recognition. **SMART MOVES JOURNAL IJOSTHE**, v. 3, n. 2, p. 6, Apr. 2016. Available at: <https://ijosthe.com/index.php/ojssports/article/view/84>.

KOŁAKOWSKA, A. *et al.* Emotion recognition and its applications. *In*: _____. **Human-Computer Systems Interaction: Backgrounds and Applications 3**. Cham: Springer International Publishing, 2014. p. 51–62. ISBN 978-3-319-08491-6. Available at: <https://doi.org/10.1007/978-3-319-08491-6_5>.

KONG, Q. *et al.* Panns: Large-scale pretrained audio neural networks for audio pattern recognition. **IEEE/ACM Transactions on Audio, Speech, and Language Processing**, IEEE, v. 28, p. 2880–2894, 2020.

LALITHA, S. *et al.* Speech emotion recognition using dwt. *In*: **2015 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)**. [*S.l.: s.n.*], 2015. p. 1–4.

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

LIVINGSTONE, S. R.; RUSSO, F. A. **The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)**. Zenodo, 2018. Available at: <https://doi.org/10.5281/zenodo.1188976>.

MACEDO, A. A. *et al.* A mobile application and system architecture for online speech training in portuguese: design, development, and evaluation of sofiafala. **Multimedia Tools and Applications**, 2024. ISSN 1573-7721. Available at: <https://doi.org/10.1007/s11042-024-19980-5>.

MADANIAN, S. *et al.* Speech emotion recognition using machine learning — a systematic review. **Intelligent Systems with Applications**, v. 20, p. 200266, 2023. ISSN 2667-3053. Available at: <https://www.sciencedirect.com/science/article/pii/S2667305323000911>.

MARCACINI, R. M.; JUNIOR, A. C.; CASANOVA, E. Overview of the automatic speech recognition for spontaneous and prepared speech speech emotion recognition in portuguese (ser) shared-tasks at propor 2022. *In*: **International Conference on Computational Processing of the Portuguese Language - PROPOR**. [*S.l.: s.n.*]: CEUR-WS, 2022.

MENG, H. *et al.* Speech emotion recognition using wavelet packet reconstruction with attention-based deep recurrent neutral networks. **Bulletin of the Polish Academy of Sciences Technical Sciences**, v. 69, n. No. 1, p. e136300, 2021. Available at: <http://journals.pan.pl/Content/119177/PDF/14_01872_Bpast.No.69(1)_24.02.21_K1_A_TeX.pdf>.

MONTALBO, F. J.; ALON, A. Empirical analysis of a fine-tuned deep convolutional model in classifying and detecting malaria parasites from blood smears. **KSII Transactions on Internet and Information Systems**, v. 15, p. 147–165, 01 2021.

OLIVEIRA, L. N. d.; GOULART, B. N. G. d.; CHIARI, B. M. Distúrbios de linguagem associados à surdez. **Rev. bras. crescimento desenvolv. hum**, p. 41–45, 2013.

O'SHEA, K.; NASH, R. An introduction to convolutional neural networks. **arXiv preprint arXiv:1511.08458**, 2015.

PALO, H. K.; SUBUDHIRAY, S.; DAS, N. The amalgamation of wavelet packet information gain entropy tuned source and system parameters for improved speech emotion recognition. **Speech Communication**, v. 149, p. 11–28, 2023. ISSN 0167-6393. Available at: <https://www.sciencedirect.com/science/article/pii/S0167639323000468>.

PERIN, E. S.; MATSUBARA, E. T. Transductive ensemble learning with graph neural network for speech emotion recognition. *In*: **SE&R@ PROPOR**. [*S.l.: s.n.*], 2022. p. 42–48.

RISSATO, P. H. D. G.; MACEDO, A. A. Sofiafala: Software inteligente de apoio à fala. *In*: **Anais Estendidos do XXVII Simpósio Brasileiro de Sistemas Multimídia e Web**. Porto Alegre, RS, Brasil: SBC, 2021. p. 91–94. ISSN 2596-1683. Available at: <https://sol.sbc.org.br/index.php/webmedia_estendido/article/view/17620>.

Rí, F. A. D.; CIARDI, F. C.; CONCI, N. Speech emotion recognition and deep learning: An extensive validation using convolutional neural networks. **IEEE Access**, v. 11, p. 116638–116649, 2023.

SCARANTI, A. *et al.* Speech emotion recognition in portuguese for sofiafala: Ser sofiafala. 2022.

Shah Fahad, M. *et al.* A survey of speech emotion recognition in natural environment. **Digital Signal Processing**, v. 110, p. 102951, 2021. ISSN 1051-2004. Available at: <https://www.sciencedirect.com/science/article/pii/S1051200420302967>.

SHARMA, G.; UMAPATHY, K.; KRISHNAN, S. Trends in audio signal feature extraction methods. **Applied Acoustics**, v. 158, p. 107020, 2020. ISSN 0003-682X. Available at: <https://www.sciencedirect.com/science/article/pii/S0003682X19308795>.

SHINDE, P. P.; SHAH, S. A review of machine learning and deep learning applications. *In*: IEEE. **2018 Fourth international conference on computing communication control and automation (ICCUBEA)**. [*S.l.: s.n.*], 2018. p. 1–6.

SIFUZZAMAN, M.; ISLAM, M. R.; ALI, M. Z. Application of wavelet transform and its advantages compared to fourier transform. Vidyasagar University, Midnapore, West-Bengal, India, 2009.

The World of Work Project. **Mehrabian's 7-38-55 Communication Model: It's More Than Words**. 2019. Acessado em Sep/5/2024. Available at: <https://worldofwork.io/2019/07/mehrabians-7-38-55-communication-model/>.

VENKATARAMANAN, K.; RAJAMOHAN, H. R. **Emotion Recognition from Speech**. 2019.

VIEIRA, R. G. **Reconhecimento de Emoção da Fala utilizando Aprendizado Profundo e Transformada Wavelet**. 9 2023 — Universidade de São Paulo - ICMC/USP, São Carlos, 9 2023. Tese de Especialização em Inteligência Artificial e Big Data.

WANG, K. A database of elderly emotional speech. *In*: . [*S.l.: s.n.*], 2018.

WANG, K. *et al.* Wavelet packet analysis for speaker-independent emotion recognition. **Neurocomputing**, v. 398, p. 257–264, 2020. ISSN 0925-2312. Available at: <https://www.sciencedirect.com/science/article/pii/S0925231220302812>.

WANG, Y.; HUO, H. Speech recognition based on genetic algorithm optimized support vector machine. *In*: **2019 6th International Conference on Systems and Informatics (ICSAI)**. [*S.l.: s.n.*], 2019. p. 439–444.

WANI, T. M. *et al.* A comprehensive review of speech emotion recognition systems. **IEEE Access**, v. 9, p. 47795–47814, 2021.

ZHANG, X.; LEE, V. C.; LIU, F. **From Data to Insights: A Comprehensive Survey on Advanced Applications in Thyroid Cancer Research**. 2024.