Reconhecimento de Emoções na Fala em Português usando Redes Neurais

Joelder Victor Antonino Aguiar

CENTRO DE INFORMÁTICA
UNIVERSIDADE FEDERAL DA PARAÍBA

João Pessoa, 2024

Joelder Victor Antonino Aguiar

Reconhecimento de Emoções na Fala em Português usando Redes Neurais

Monografia apresentada ao curso Engenharia de computação
do Centro de Informática, da Universidade Federal da Paraíba,
como requisito para a obtenção do grau de Bacharel em Engenharia de computação

Orientador: Alisson Vasconcelos de Brito

Outubro de 2024

Catalogação na publicação
Seção de Catalogação e Classificação

A282r Aguiar, Joelder Victor Antonino.
Reconhecimento de emoções na fala em português usando redes neurais / Joelder Victor Antonino Aguiar. - João Pessoa, 2024.
40 f. : il.

Orientação: Alisson Vasconcelos Brito.
TCC (Graduação) - UFPB/CI.

1. Reconhecimento de emoções. 2. Redes neurais convolucionais. 3. Redes de memória. 4. Computação afetiva. I. Brito, Alisson Vasconcelos. II. Título.

UFPB/CI CDU 004.852

Elaborado por Michelle de Kássia Fonseca Barbosa - CRB-738

AGRADECIMENTOS

Primeiramente, agradeço à minha família, que foi meu pilar e fonte constante de incentivo ao longo dessa jornada, com destaque especial para minha mãe e minha avó (in memoriam), por seu amor e apoio incondicional.

À minha namorada, Raquel, minha grande fonte de refúgio e motivação, especialmente nos momentos mais desafiadores dessa reta final.

Um agradecimento especial a Mateus Antônio, cuja simples postagem no *Instagram* foi o ponto de partida que me trouxe ao curso.

Aos meus amigos de turma, João Victor, Isaac Marinho, Egídio Neto, Miguel Elias e Yvson Nunes, cujas presenças e colaborações foram fundamentais para que eu pudesse chegar até aqui. A ajuda de vocês foi inestimável.

Expresso também minha gratidão a Laura Campos, Nelly Martins e Thaís Marques, pela amizade que transcendeu os muros da universidade.

Aos colegas de projetos, Samila Garrido, Luiz Fernando, Joan Vitor e Franklin Coelho, sou grato pela parceria e aprendizado.

Ao meu orientador, Alisson Brito, agradeço pela orientação, apoio e confiança, não apenas neste trabalho de conclusão, mas também em diversos outros projetos ao longo da graduação.

Por fim, estendo meus agradecimentos a todos os professores e professoras que tive o privilégio de conhecer, e que compartilharam seus valiosos ensinamentos. Agradeço especialmente aos professores Jorge Dias e Moisés Dantas, por sua orientação em projetos extracurriculares, que ampliaram meus horizontes acadêmicos e profissionais.

RESUMO

O reconhecimento de emoções na fala tem se tornado uma área de grande relevância dentro da computação afetiva, devido à sua aplicação em sistemas que buscam interações mais naturais entre humanos e máquinas. Esta monografia apresenta o desenvolvimento de um sistema para o reconhecimento de emoções em áudios em português, utilizando técnicas de aprendizado de máquina com redes neurais profundas, especificamente redes neurais convolucionais (*Convolutional Neural Networks - CNNs*, do inglês) e redes de memória de curto e longo prazo (*Long Short-Term Memory Networks* - LSTM, do inglês).

O principal desafio no reconhecimento automático de emoções reside na variabilidade das expressões emocionais entre indivíduos e culturas, além das dificuldades inerentes à extração e análise de características prosódicas e acústicas. A metodologia proposta busca superar essas limitações com a utilização da base de dados emoUERJ, que contém gravações em português com expressões de diferentes emoções. A partir da análise de espectrogramas e do uso de técnicas de processamento de sinais, o modelo desenvolvido foi testado em cenários com e sem ruído, atingindo resultados significativos.

Os experimentos realizados indicam que a combinação de CNNs e LSTMs oferece um desempenho robusto, permitindo a extração automática de características relevantes diretamente dos dados brutos, e demonstram a eficácia do modelo proposto na tarefa de reconhecimento de emoções na fala em português.

Palavras-chave: <Reconhecimento de emoções>, <Redes Neurais Convolucionais>, <Redes de Memória de Curto e Longo Prazo>, <Computação Afetiva>.

ABSTRACT

Speech emotion recognition has become an important area within affective computing due to its application in systems aiming for more natural human-machine interactions. This monograph presents the development of a system for recognizing emotions in Portuguese audio using deep neural networks, specifically Convolutional Neural Networks (CNN) and Long Short-Term Memory Networks (LSTM).

The main challenge in automatic emotion recognition lies in the variability of emotional expressions among individuals and cultures, in addition to the inherent difficulties in extracting and analyzing prosodic and acoustic features. The proposed methodology aims to overcome these limitations by utilizing the emoUERJ dataset, which contains Portuguese recordings expressing different emotions. Through the analysis of spectrograms and the use of signal processing techniques, the developed model was tested in both noise-free and noisy scenarios, achieving significant results.

The experiments indicate that the combination of CNNs and LSTMs offers robust performance, allowing for the automatic extraction of relevant features directly from raw data. The proposed model demonstrates effectiveness in the task of emotion recognition in Portuguese speech.

Keywords: <Emotion recognition>, < Convolutional Neural Networks>, <Long Short-Term Memory Networks>, <Affective Computing>.

LISTA DE FIGURAS

Figura 1 - Exemplo de forma de onda e mel-espectrograma para as emoções presentes no emoUERJ 23
Figura 2 - Exemplo do mel-espectrograma e log-mel espectrograma para a emoção "raiva" 24
Figura 3 - Distribuição da quantidade de *frames* por áudio 25
Figura 4 - Fluxo de camadas: Convolucional, Normalização em Lote, Ativação e Agrupamento 27
Figura 5 - Diagrama das camadas LFLB e seus filtros 31
Figura 6 - Gráfico obtido durante o treinamento do Experimento 1 32
Figura 7 - Gráfico obtido durante o treinamento do Experimento 2 33
Figura 8 - Matriz de confusão obtida no Experimento 1 35
Figura 9 - Matriz de confusão obtida no Experimento 2 37

LISTA DE TABELAS

Tabela 1 - Resumo da quantidade de áudios por emoção no conjunto de dados emoUERJ 22
Tabela 2 - Estatísticas de Duração e Taxa de Amostragem para os Conjuntos de Dados emoUERJ 22
Tabela 3 - Hiperparâmetros utilizados pela biblioteca *Hyperopt* 30
Tabela 4 - Número de Arquivos para Treinamento, Validação e Teste 30
Tabela 5 - Relatório de Classificação para o Experimento 1 34
Tabela 6 - Relatório de Classificação para o Experimento 2 36

LISTA DE ABREVIATURAS

ACC - *Accuracy*
BN - *Batch Normalization*
CNN - *Convolutional Neural Network*
DNN - *Deep Neural Network*
ELU - *Exponential Linear Unit*
emoDB - *Berlin Database of Emotional Speech*
FFT - *Fast Fourier Transform*
GPU - *Graphics Processing Unit*
IEMOCAP - *Interactive Emotional Dyadic Motion Capture*
LSTM - *Long Short-Term Memory Network*
MFCC - *Mel-Frequency Cepstral Coefficients*
RNN - *Recurrent Neural Network*
SNR - *Signal-to-Noise Ratio*
WA - *Weighted Accuracy*

Sumário

1 INTRODUÇÃO **12**

1.1 Definição do Problema 12

1.2 Objetivos 13

1.2.1 Objetivo Geral 13

1.2.2 Objetivos Específicos 13

1.3 Metodologia e Resultados 13

ANEXO A - ARTIGO **15**

REFERÊNCIAS **41**

1 INTRODUÇÃO

O reconhecimento de emoções por meio da fala tem se tornado uma área de interesse na computação afetiva, que visa desenvolver sistemas capazes de interpretar e responder às emoções humanas (PICARD, 1997). A habilidade de compreender estados emocionais é essencial para aprimorar a interação entre humanos e máquinas, tornando-a mais natural e eficiente (COWIE et al., 2001).

A fala é um dos meios mais ricos e naturais para a expressão de emoções, pois incorpora não apenas o conteúdo verbal, mas também características prosódicas e acústicas que refletem o estado emocional do locutor (SCHULLER; BATLINER, 2018). No entanto, esse reconhecimento automático dessas emoções apresenta desafios significativos devido à variabilidade inerente na expressão emocional entre diferentes indivíduos e culturas (AYADI; KAMEL; KARRAY, 2011).

Com o avanço do aprendizado de máquina (*machine learning*, do inglês), tornou-se possível abordar esses desafios por meio de modelos que aprendem a identificar padrões complexos nas características da fala (LEE et al., 2011). Técnicas tradicionais dependiam da extração manual de características e de classificadores estatísticos (VERVERIDIS; KOTROPOULOS, 2006), o que limitava a capacidade dos sistemas em capturar a riqueza dos sinais emocionais.

Com o aumento da capacidade de processamento dos computadores e o avanço das redes neurais profundas, revolucionou o campo ao permitir a extração automática de características relevantes diretamente dos dados brutos (HINTON; SALAKHUTDINOV, 2006). Arquiteturas como redes neurais convolucionais (CNNs) e redes neurais recorrentes (*Recurrent Neural Networks - RNNs*, do inglês), têm demonstrado desempenho superior em tarefas de reconhecimento de emoções na fala (TRIGEORGIS et al., 2018; SATT; ROZENBERG; HOORY, 2017). Essas redes são capazes de modelar relações não lineares complexas e capturar dependências temporais nos dados, essenciais para o entendimento das nuances emocionais (FAYEK; LECH; CAVEDON, 2017).

1.1 Definição do Problema

Este trabalho apresenta no Anexo A o desenvolvimento de um modelo para reconhecimento de emoções em áudios em português, utilizando redes neurais convolucionais (CNN) e redes de memória de curto e longo prazo (LSTM). A base de dados utilizada será a *emoUERJ*, que contém gravações em português voltadas para o reconhecimento de emoções. Através da combinação dessas técnicas de aprendizado de máquina, espera-se alcançar resultados robustos, superando os desafios do reconhecimento de emoções em cenários com e sem ruído.

1.2 Objetivos

1.2.1 Objetivo Geral

Desenvolver um sistema de reconhecimento de emoções em áudios em português utilizando técnicas de redes neurais profundas, mais especificamente, redes neurais convolucionais (CNN) e redes de memória de curto e longo prazo (LSTM).

1.2.2 Objetivos Específicos

- Implementar um modelo de CNN e LSTM para reconhecimento de emoções em áudios em português;
- Realizar experimentos com áudios sem ruído e com adição de ruído, avaliando o desempenho do modelo em ambos os cenários;
- Utilizar a base de dados *emoUERJ* para treinar e testar o modelo, explorando as nuances de emoções presentes no idioma português;

1.3 Metodologia e Resultados

Neste trabalho, empregamos **redes neurais** (GOODFELLOW; BENGIO; COURVILLE, 2016) para a classificação de sinais de áudio, utilizando técnicas de processamento de sinais para preparar os dados de entrada. Redes neurais são modelos computacionais inspirados na estrutura do cérebro humano, capazes de aprender e reconhecer padrões complexos em grandes conjuntos de dados (LECUN; BENGIO; HINTON, 2015).

Para converter os sinais de áudio em um formato adequado para o treinamento das redes neurais, utilizamos **espectrogramas log-mel** (MCFEE et al., 2015). Um espectrograma log-mel é uma representação do espectro de frequências de um sinal de áudio, onde as frequências são mapeadas para a escala mel, que reflete a percepção humana de frequência. A aplicação do logaritmo na amplitude destaca componentes de baixa intensidade, tornando a representação mais robusta a variações dinâmicas (MCFEE et al., 2015).

Utilizamos dois tipos principais de arquiteturas de redes neurais em nossos experimentos: **Redes Neurais Convolucionais (CNN)** (KRIZHEVSKY; SUTSKEVER; HINTON, 2012) e **Redes Neurais de Memória de Longo Curto Prazo (LSTM)** (HOCHREITER; SCHMIDHUBER, 1997). As CNNs são eficazes na extração de características espaciais dos espectrogramas, enquanto as LSTMs são adequadas para capturar dependências temporais nos dados sequenciais.

Este trabalho envolveu a realização de dois experimentos, seguindo as etapas descritas abaixo:

1. **Experimento 1**: Conversão do áudio bruto em log-mel espectrogramas, seguida do treinamento do modelo.
2. **Experimento 2**: Adição de ruído ao áudio bruto, transformação para log-mel espectrogramas e, por fim, treinamento do modelo.

Os resultados obtidos foram os seguintes: no **Experimento 1**, o modelo alcançou uma acurácia global de 94,74% e uma acurácia ponderada de 95,24%. No **Experimento 2**, mesmo com a adição de ruído, foi registrada uma acurácia global de 89,47% e uma acurácia ponderada de 87,43%.

ANEXO A – ARTIGO

Reconhecimento de Emoções na Fala em Português Usando Redes Neurais

Artigo de acordo com as normas da Sociedade Brasileira de Computação - SBC

Reconhecimento de Emoções na Fala em Português usando Redes Neurais

Joelder Victor Antonino Aguiar[1], **Alisson Vasconcelos de Brito**[1]

[1]Centro de Informática – Universidade Federal da Paraíba (UFPB)
João Pessoa – PB – Brasil

joelder.antonino@gmail.com, alisson@ci.ufpb.br

Abstract. *Speech emotion recognition has emerged as a crucial area in applications such as virtual assistants and customer service systems. This paper investigates the application of Convolutional Neural Networks (CNN) and Long Short-Term Memory Networks (LSTM) for emotion recognition in Portuguese audio, using the emoUERJ dataset. The methodology involves the automatic extraction of features through Mel-spectrograms, conversion to the logarithmic decibel scale, and input resizing for standardization. Two experiments were conducted: the first with noise-free audio and the second with the addition of Gaussian noise to the test audio. The model achieved an accuracy of 94.74% in the noise-free scenario and 89.47% with noise, demonstrating robustness in the task of emotion classification. The difficulties observed in the "Anger"class under noisy conditions highlight the need for advanced data augmentation techniques and alternative architectures.*

Resumo. *O reconhecimento de emoções na fala tem se destacado como uma área importante em aplicações como assistentes virtuais e sistemas de atendimento ao cliente. Este trabalho investiga a aplicação de redes neurais convolucionais (Convolutional Neural Networks - CNNs, do inglês) e redes de memória de curto e longo prazo (Long Short-Term Memory Networks - LSTM, do inglês) para o reconhecimento de emoções em áudios em português, utilizando a base de dados emoUERJ. A metodologia envolve a extração automática de características por meio de Mel espectrogramas, a conversão para a escala logarítmica em decibéis e o redimensionamento das entradas para padronização. Dois experimentos foram realizados: o primeiro com áudios sem ruído e o segundo com a adição de ruído gaussiano. O modelo alcançou uma acurácia de 94,74% no cenário sem ruído e 89,47% com ruído, demonstrando robustez na tarefa de classificação emocional. As dificuldades observadas na classe "Raiva"em condições ruidosas indicam a necessidade de técnicas avançadas de aumento de dados e arquiteturas alternativas.*

1. Introdução

O reconhecimento de emoções por meio da fala tem se tornado uma área de grande interesse devido à sua relevância em aplicações como assistentes virtuais, sistemas de atendimento ao cliente e interfaces homem-máquina mais intuitivas [Schuller e Batliner 2018, Akçay e Oğuz 2020]. A capacidade de uma máquina interpretar corretamente as emoções humanas pode melhorar significativamente a qualidade da interação entre humanos e computadores, tornando-a mais natural e eficiente [Ayadi et al. 2011].

Tradicionalmente, o reconhecimento de emoções na fala envolvia a extração manual de características prosódicas e espectrais, seguida da classificação por modelos de aprendizado de máquina convencionais [Ververidis e Kotropoulos 2006]. No entanto, esse processo pode ser limitado pela dependência de seleção manual de características e pela dificuldade em capturar nuances emocionais complexas presentes no sinal de áudio [Schuller et al. 2011].

Com o avanço das redes neurais profundas, especialmente das redes neurais convolucionais (*Convolutional Neural Networks - CNNs*, do inglês) e redes neurais recorrentes (*Recurrent Neural Networks - RNNs*, do inglês), houve um progresso significativo na área [Trigeorgis et al. 2016]. Essas técnicas permitem a extração automática de características relevantes diretamente dos dados e podem ser potencializadas com o uso dos espectrogramas, que conseguem realçar as frequências ao longo do tempo. Além disso, as redes neurais são capazes de modelar relações não lineares complexas nos dados, o que é determinante para captar as sutilezas das emoções humanas [Han et al. 2014].

O uso de redes neurais no reconhecimento de emoções tem demonstrado melhorias notáveis no desempenho dos modelos, tornando-os mais robustos e precisos [Zhang et al. 2018]. Estudos recentes têm explorado arquiteturas híbridas que combinam CNNs e RNNs para capturar tanto características locais quanto dependências temporais no sinal de áudio [Satt et al. 2017]. Essas abordagens têm alcançado resultados promissores em diversos conjuntos de dados e idiomas, evidenciando a eficácia das redes neurais profundas nessa área [Latif et al. 2020].

Diante disso, este trabalho busca aplicar estudos já desenvolvidos para conjuntos de dados em inglês e aplica-las ao idioma português, utilizando rede neurais convolucionais e redes de memória de curto e longo prazo.

2. Trabalhos relacionados

Nesta pesquisa, foram analisados quatro artigos que abordam o reconhecimento de emoções por meio da fala. O objetivo foi realizar uma análise concisa do processo de desenvolvimento adotado em cada estudo, incluindo os algoritmos empregados e os resultados alcançados pelos autores.

O trabalho de [Wang et al. 2022] teve o propósito de desenvolver um modelo de reconhecimento de emoções em fala multilíngue utilizando diferentes bases de dados em três idiomas: IEMOCAP (inglês), EmoDB (alemão) e Cafe (francês). O experimento consistiu em extrair cinco características: Mel-Frequency Cepstrum MFCC, Allosaurus [Li et al. 2020], Wav2Vec [Schneider et al. 2019], GE2E [Wan et al. 2018], e BYOL [Grill et al. 2020]. Essas características foram passadas para um classificador que possuía torres separadas para cada idioma, cujos hiperparâmetros foram ajustados de acordo com o idioma.

As camadas iniciais do modelo, que incluem CNN, LSTM e *Self-Attention*, foram compartilhadas entre todos os idiomas. No experimento, os autores alcançaram uma acurácia média de 74,34% para o inglês, 90,70% para o francês e 95,56% para o alemão, demonstrando a eficácia do modelo no reconhecimento de emoções em múltiplos idiomas.

Já [Zhao et al. 2019] em seu trabalho desenvolveu e avaliou duas redes neurais convolucionais e de memória de longo prazo (CNN e LSTM) para o reconhecimento de emoções na fala, empregando tanto sinais de fala quanto espectrogramas para capturar características emocionais locais e globais.

Foram utilizadas duas bases de dados: EmoDB, que contém gravações em alemão, e IEMOCAP, que inclui interações emocionais em inglês. Os modelos foram testados tanto em áudios brutos quanto em espectrogramas , demonstrando que a inclusão de espectrogramas melhora significativamente a acurácia.

Os algoritmos empregados combinaram redes convolucionais e LSTM, com uma arquitetura composta por quatro blocos de aprendizado de características locais (*Local Features Learning Blocks* - LFLBs, do inglês) e uma camada LSTM. Os LFLBs são responsáveis por aprender correlações locais e hierárquicas, enquanto a camada LSTM captura dependências de longo prazo.

Para a base de dados EmoDB, a rede 2D CNN LSTM obteve acurácias de 95,89% em experimentos dependentes do falante e 95,33% em experimentos independentes do falante. Na base IEMOCAP, a rede alcançou 85,58% e 79,72% de acurácia.

No trabalho de [Demircan e Örnek 2020], foram desenvolvidas duas abordagens distintas para o reconhecimento de emoções na fala. A primeira abordagem utilizou espectrogramas como entrada para uma arquitetura de rede neural profunda baseada na AlexNet (2012), enquanto a segunda fez uso de coeficientes (*Mel-Frequency Cepstral Coefficients* - MFCC, do inglês) aplicados a uma rede neural profunda (*Deep Neural Network* - DNN, do inglês) mais simples.

Para os experimentos, foi utilizado o banco de dados EmoDB. A classificação baseada em espectrogramas, utilizando a arquitetura AlexNet, demonstrou melhor desempenho em comparação à abordagem com MFCC. A AlexNet alcançou uma acurácia ponderada de 88,46% ao classificar emoções como Raiva, Felicidade e Medo. Já para as

emoções Tédio, Neutro e Tristeza, a acurácia ponderada foi de 84,09%. Esses resultados indicam que o uso de imagens de espectrograma em conjunto com uma rede neural mais complexa pode oferecer ganhos significativos no reconhecimento emocional na fala.

Enquanto que no trabalho [Peixoto e Linhares 2023] empregou uma CNN para o reconhecimento de emoções em áudios em português, levando em consideração variações linguísticas das diferentes regiões do Brasil. A pesquisa foi estruturada em duas fases: treinamento e testes. Durante o treinamento, foi utilizada a base de dados EmoUERJ, enquanto os testes foram realizados com um conjunto de áudios extraídos de vídeos do YouTube. No conjunto de treinamento, o modelo atingiu acurácias de 91,51% para os dados sem distinção de gênero e 94,70% para os dados com distinção de gênero. No entanto, nos testes com dados externos, os resultados da rede não foram satisfatórios.

Em comparação com os trabalhos analisados, este estudo se diferencia ao investigar o reconhecimento de emoções na fala em português com o uso de uma combinação de CNN e LSTM. Embora [Peixoto e Linhares 2023] também tenha explorado o reconhecimento de emoções em português, eles se concentraram em variações linguísticas regionais e obtiveram resultados menos satisfatórios em testes externos. Nosso estudo, por outro lado, foca na robustez do modelo frente a ruído gaussiano, apresentando uma abordagem mais abrangente para lidar com condições adversas. Diferentemente de [Wang et al. 2022] e [Zhao et al. 2019], que abordaram reconhecimento multilíngue e arquiteturas híbridas com foco em outros idiomas, nosso trabalho busca adaptar e otimizar modelos especificamente para o contexto do português.

Quadro 1: Resumo sobre os trabalhos relacionados analisados.

Autor	Objetivo	Algoritmo	Dataset	Métrica
[Wang et al. 2022]	Desenvolver um modelo de reconhecimento de emoções em fala multilíngue	CNN, LSTM, Self-Attention com torres separadas para cada idioma; características: MFCC, Allosaurus, Wav2Vec, GE2E, BYOL	IEMOCAP (inglês), EmoDB (alemão), Cafe (francês)	Acurácia média: 74,34% (inglês), 90,70% (francês), 95,56% (alemão)
[Zhao et al. 2019]	Desenvolver e avaliar redes CNN LSTM para reconhecimento de emoções usando sinais de fala e espectrogramas	Combinação de CNN e LSTM com quatro blocos de aprendizado de características locais (LFLBs) e uma camada LSTM	EmoDB (alemão), IEMOCAP (inglês)	EmoDB: 95,89% (dependente do falante), 95,33% (independente do falante); IEMOCAP: 85,58% e 79,72%
[Demircan e Örnek 2020]	Desenvolver duas abordagens para reconhecimento de emoções na fala	1) Espectrogramas com AlexNet; 2) MFCC com DNN simples	EmoDB	AlexNet: Acurácia ponderada de 88,46% (Raiva, Felicidade, Medo); 84,09% (Tédio, Neutro, Tristeza)

Continua na próxima página

Quadro 1 – Continuação da página anterior

Autor	Objetivo	Algoritmo	Dataset	Métrica
[Peixoto e Linhares 2023]	Empregar uma CNN para reconhecimento de emoções em português do Brasil considerando variações regionais	CNN	Treinamento: EmoUERJ; Teste: áudios do YouTube	Treinamento: 91,51% (sem distinção de gênero), 94,70% (com distinção de gênero); Testes externos: resultados insatisfatórios
Este trabalho	Reconhecimento de emoções em português com foco na robustez frente a ruído	CNN e LSTM	emoUERJ	Acurácia 94,74% (sem ruído), 89,47% (com ruído)

3. Metodologia

Nesta seção, é descrita a metodologia utilizada ao longo deste trabalho, o que inclui a descrição da base de dados, as técnicas de pré-processamento realizadas, o ambiente de desenvolvimento utilizado, as métricas de avaliação e os modelos utilizados.

3.1. Base de dados

Os áudios utilizados neste trabalho foram extraídos do conjunto de dados: *emoUERJ* [Bastos Germano et al. 2021].

A base de dados *emoUERJ* foi desenvolvida na Universidade do Estado do Rio de Janeiro com o objetivo de criar modelos específicos de reconhecimento de emoções na fala em português, uma vez que há poucas bases de dados disponíveis nesse idioma. Para a construção do banco, foram gravadas dez frases por oito atores, divididos igualmente entre os gêneros, que escolheram livremente as frases para expressar quatro emoções: felicidade, raiva, tristeza e neutra. No total, foram gerados 377 áudios, sendo 91 áudios de felicidade, 94 áudios de raiva, 100 áudios de tristeza e 92 áudios neutros.

Resumo dos áudios presentes na base dados por emoção:

Tabela 1. Resumo da quantidade de áudios por emoção no conjunto de dados emoUERJ

Emoção	emoUERJ
Felicidade	91
Raiva	94
Tristeza	100
Neutro	92
Total	377

Tabela 2. Estatísticas de Duração e Taxa de Amostragem para os Conjuntos de Dados EmoUERJ

Conjunto de Dados	Duração Média (s)	Duração Mínima (s)	Duração Máxima (s)	Taxa de Amostragem Média (Hz)
EmoUERJ	3.01	1.19	7.09	44100.00

3.1.1. Pré-processamento

3.1.1.1. Mel espectrograma

Espectrogramas consiste na representação visual do espectro de frequências de um sinal variando no tempo[Lacerda et al. 2023]. Já o Mel Espectrograma é a conversão dos espectrogramas para a escala mel que busca enfatizar os sons produzidos pelos humanos e por isso seu amplo uso no reconhecimento de emoções por meio da fala.

Os arquivos de áudio brutos foram processados para a extração de características de frequência, com foco na geração do espectrograma na escala Mel. Utilizou-se 2048

pontos na Transformada Rápida de Fourier (*Fast Fourier Transform* - FFT, do inglês), um parâmetro também adotado por [Zhao et al. 2019] em seu estudo. Esse valor, combinado com a taxa de amostragem média dos áudios, resultou em janelas de análise de aproximadamente 46ms, como mostra a Equação (1). O deslocamento entre essas janelas conhecido como (*hop length*, do inglês) foi definido em 512 amostras, mantendo a consistência com a configuração de [Zhao et al. 2019]. A Figura 1 mostra as formas de ondas para cada emoção e o seu respectivo mel espectrograma.

$$\text{Janelas de Análise} = \frac{\text{Pontos da FFT}}{\text{Taxa de Amostragem}} \tag{1}$$

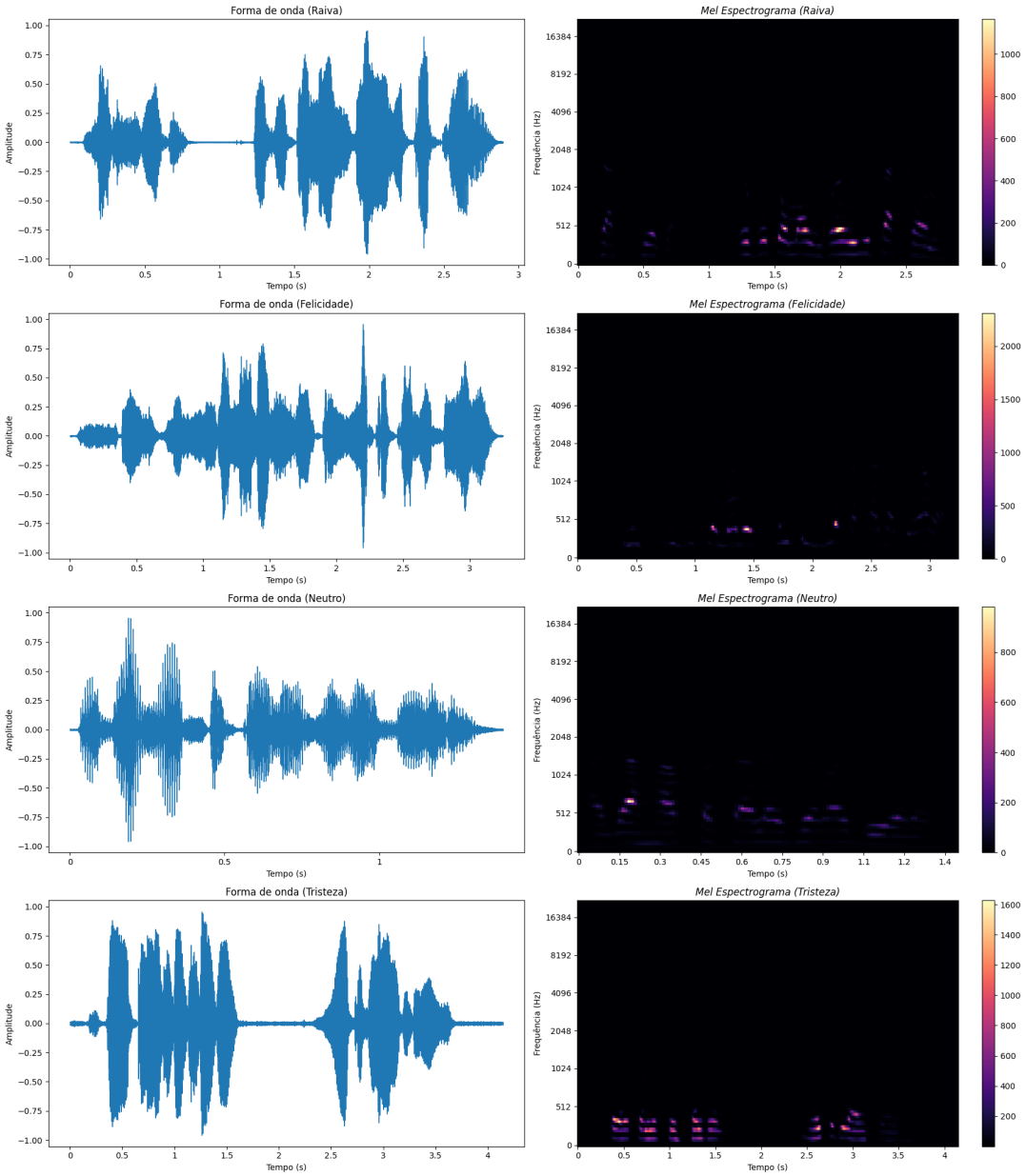

Figura 1. Exemplo de forma de onda e mel espectrograma para as emoções presentes no emoUERJ. Fonte: Autor.

3.1.1.2. Conversão para a Escala Logarítmica em Decibéis

Após a criação do mel espectrograma, ele é convertido para a escala de decibéis usando a Equação (2), e o resultado é apresentado na Figura 2. Essa transformação dos valores de potência para decibéis minimiza as variações extremas entre os valores, ajudando a atenuar ruídos e destacar nuances sutis no sinal de áudio. Esse processo é importante para o reconhecimento eficiente de emoções na fala, e por isso vem sendo amplamente utilizado como nos estudos de [Meng et al. 2019] e [Zhao et al. 2019].

$$\mathrm{dB} = 10 \log_{10} \left(\frac{S_{\mathrm{Mel}}}{\mathrm{ref}} \right) \tag{2}$$

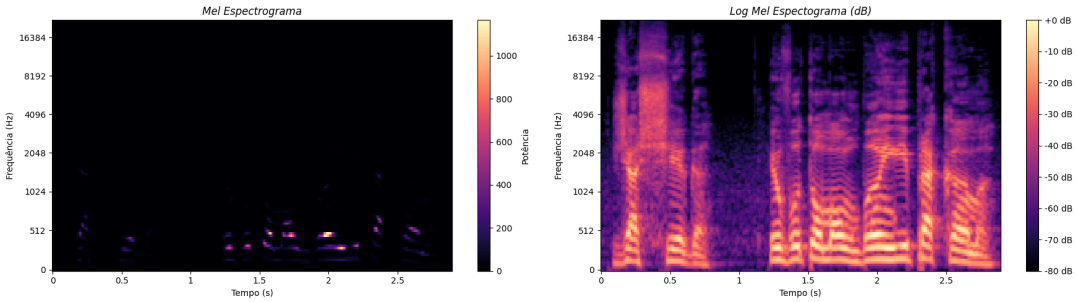

Figura 2. Exemplo do mel espectrograma e o log-mel espectrograma para a emoção raiva do emoUERJ. Fonte: Autor.

3.1.1.3. Redimensionamento

Dado que a base de dados utilizada contém arquivos de áudio com durações variadas, foi necessário padronizar as dimensões dos log-mel espectrogramas antes de alimentá-los na rede neural, que exige matrizes de tamanho constante. Para isso, todos os espectrogramas foram redimensionados para 256 x 140. Aqueles espectrogramas com menos de 256 *frames* foram complementados com zeros até alcançar a dimensão necessária. A escolha de 256 se deu ao observar a distribuição mostrada na Figura 3, em que a escolha do 256 abrange mais de 90% dos arquivos. Já o número do bandas de frequências foi definido empiricamente, sendo 140 a que alcançou melhor resultado nos testes.

3.2. Ambiente de desenvolvimento

Para o desenvolvimento desta pesquisa, foi utilizada a linguagem de programação Python em sua versão 3.10.12, acompanhada das seguintes bibliotecas:

- `audiomentations` - : Biblioteca *open-source* para *data augmentation* em sinais de áudio.
- `hyperopt` - versão 0.2.7: Biblioteca *open-source* para otimização de hiperparâmetros.
- `librosa` - versão 0.10.2: Biblioteca *open-source* para processamento de sinais de áudio;

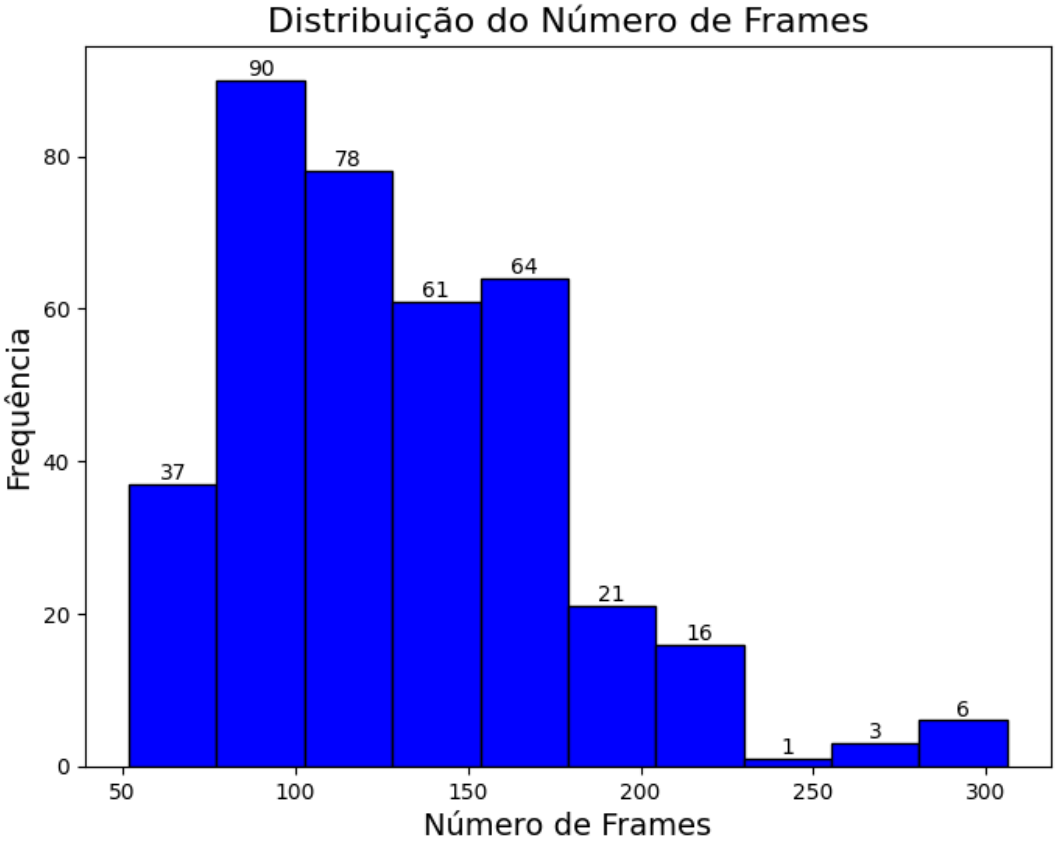

Figura 3. Distribuição da quantidade de *frames* por áudio. Fonte: Autor.

- `Matplotlib` - versão 3.7.1: Biblioteca *open-source* para plotagem de dados;
- `Numpy` - versão 1.26.4: Biblioteca *open-source* para processamento de dados multi-dimensionais e matrizes;
- `Scikit-learn` - versão 1.3.2: Biblioteca *open-source* para desenvolvimento de aplicações em *machine learning*;
- `Tensorflow` - versão 2.17.0: Biblioteca *open-source* para desenvolvimento de modelos e processamento via GPU.

O hardware utilizado foi o disponível no *Google Colaboratory*:

- Memória RAM: 12,7 GB
- Placa de vídeo: NVIDIA TESLA T4, 15 GB de VRAM
- Processador: Intel® Xeon® CPU @ 2.00 GHz

3.3. Métricas de avaliação

Neste estudo, a avaliação dos resultados é de suma importância para comprovar a capacidade de generalização do modelo. As métricas utilizadas durante a execução deste trabalho estão descritas a seguir.

3.3.1. Acurácia

Também conhecida como taxa global de sucesso do algoritmo (*Accuracy - ACC, do inglês*) é uma métrica bastante utilizada para problemas de classificação é o número

de classificações corretas dividido pelo número total de classificações, como mostra a Equação (3) [de Castro 2016].

$$\text{ACC} = \frac{\sum_{i=1}^{n} C_{ii}}{\sum_{i=1}^{n} \sum_{j=1}^{n} C_{ij}} \tag{3}$$

Em que C_{ii} representa o número de predições corretas para a classe i, correspondendo aos verdadeiros positivos na matriz de confusão de dimensão $n \times n$.

3.3.2. Acurácia ponderada

Também conhecida como (*Weighted Accuracy - WA, do inglês*) é utilizada quando o conjunto de dados é balanceado e é dada pela Equação (4) [Demircan e Örnek 2020].

$$\text{WA} = \frac{1}{n} \sum_{i=1}^{n} \frac{CP_i}{T_i} \tag{4}$$

3.3.3. Precisão

Também conchecida como (*Precision -Pr*, do inglês) é quantidade de Verdadeiros Positivos - VP sobre a soma de Falsos Positivos - FP e Verdadeiros Positivos, demonstrada na Equação (5) [de Castro 2016].

$$\text{Pr} = \frac{VP}{FP + VP} \tag{5}$$

3.3.4. Revocação

A Revocação, também conhecida como (*Recall - Re, do inglês*) é quantidade de Verdadeiros Positivos sobre a soma de Falsos Negativos e Verdadeiros Positivos, como mostra a Equação (6) [de Castro 2016].

$$\text{Re} = \frac{VP}{FN + VP} \tag{6}$$

3.3.5. Medida - F

A Medida - F é outra métrica bastante utilizada para verificar a capacidade de generalização dos modelos de classificação, também conhecida como (*score-F, do inglês*) ou (*F1-score*, do inglês), estando contida no intervalo [0,1] é definida na Equação (7) [de Castro 2016]:

$$\text{F1-score} = 2 \times \frac{\text{Pr} \times \text{Re}}{\text{Pr} + \text{Re}} \tag{7}$$

3.3.6. Entropia Cruzada Categórica Esparsa

Também conhecida como (*Sparse Categorical Crossentropy*, do inglês) é usada com o intuito de calcular o quão bem um conjunto de probabilidades de classe estimadas corresponde às classes alvo e definida pela Equação (8). [Géron 2021]

$$L = -\sum_{c=1}^{C} y_c \log(\hat{y}_c) \tag{8}$$

3.4. Modelo

Para classificação das emoções através da fala, foi utilizado nesse estudo um modelo híbrido, baseada no estudo de [Zhao et al. 2019].

3.4.1. Bloco de Aprendizado Local

Foram utilizados 5 blocos convolucionais, que são responsáveis pela extração de características locais, (*Local Feature Learning Block - LFLB*, do inglês) compostos pelas seguintes camadas:

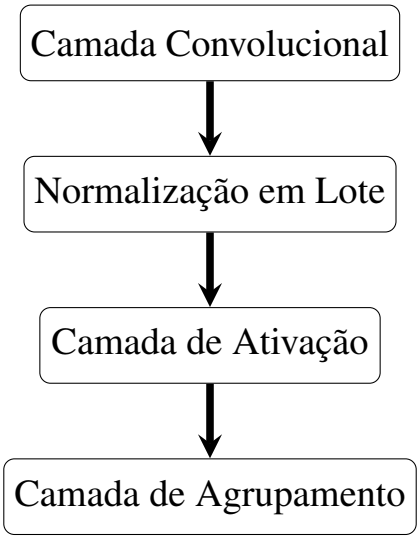

Figura 4. Fluxo de camadas: Camada Convolucional, Normalização em Lote, Camada de Ativação e Camada de Agrupamento

3.4.1.1. Camada Convolucional

Essa camada tem por objetivo extrair características locais dos dados de entrada. A camada aplica filtros que realizam operações de convolução sobre a entrada, permitindo que a rede identifique padrões, como bordas, texturas e formas. A saída é um conjunto de mapas de características que representam diferentes aspectos dos dados de entrada. Como resume a Equação (9) [Géron 2021] e [Zhao et al. 2019].

$$z(i,j) = x(i,j) * w(i,j) = \sum_{s=-a}^{a} \sum_{t=-b}^{b} x(s,t) \cdot w(i-s, j-t) \tag{9}$$

- Em que $z(i,j)$ é a saída do neurônio localizado na linha i e coluna j no mapa de características de uma camada convolucional.

3.4.1.2. Normalização em Lote

Conhecida como (*Batch Normalization - BN*, do inglês), tem por objetivo acelerar o treinamento e estabilizar a aprendizagem. A normalização em lote ajusta a saída da camada convolucional para que tenha uma média próxima de zero e uma variância próxima de um. Isso ajuda a mitigar o problema do desvanecimento do gradiente e permite que a rede aprenda de forma mais eficiente. Resumida na Equação (10) [Zhao et al. 2019].

$$z_i^l = \sigma \left(BN(b_i^l + \sum_j z_j^{l-1} * w_{ij}^l) \right) \tag{10}$$

- Em que z_i^l e z_j^{l-1} são, respectivamente, a i-ésima característica de saída na camada l e a j-ésima característica de entrada na camada $(l-1)$; w_{ij}^l denota o núcleo (*kernel*, do inglês) da convolução entre a i-ésima e a j-ésima característica.

3.4.1.3. Camada de Ativação

A função de ativação (*Exponential Linear Unit - ELU*, do inglês) é utilizada para permitir que a rede aprenda representações mais complexas. Ela ajuda a evitar problemas como a saturação que podem ocorrer com funções de ativação tradicionais, como a sigmoide ou a tangente hiperbólica, mostrada na Equação (11) [Zhao et al. 2019].

$$\sigma(x) = \begin{cases} x & \text{se } x \geq 0 \\ \alpha(e^x - 1) & \text{se } x < 0 \end{cases} \tag{11}$$

3.4.1.4. Camada de Agrupamento

Mais conhecida como (*max-pooling*, do inglês) opera selecionando o valor máximo em uma janela deslizante sobre os mapas de características, o que ajuda a preservar as informações mais relevantes enquanto reduz a quantidade de dados a serem processados nas camadas subsequentes. Isso também contribui para a invariância a pequenas translações nos dados de entrada. Representada na Equação (12) [Zhao et al. 2019].

$$z_k^l = \max_{\forall p \in \Omega_k} z_p^l \tag{12}$$

3.4.2. Memória longa e de curto prazo

As camadas dos blocos convolucionais, quando combinadas, conseguem extrair características locais significativas dos dados de entrada, que são então passadas para a camada LSTM.

Uma célula LSTM pode aprender a reconhecer uma entrada importante função da porta de entrada, conhecida como (*input gate*, do inglês), armazená-la no estado de longo prazo, preservá-la pelo tempo que for necessário esse é o papel da porta de esquecimento, mais conhecida (*forget gate*, do inglês) e extrai-la sempre que necessário. Isso explica por que essas células têm sido extremamente bem-sucedidas na identificação de padrões de longo prazo em séries temporais, textos longos, gravações de áudio e muito mais. Representadas nas Equações (13–17) [Géron 2021] e [Zhao et al. 2019].

$$f_t = \sigma_g(W_f z_{t-1}^l + U_f z_{t-1}^l + b_f) \tag{13}$$

$$i_t = \sigma_g(W_i z_{t-1}^l + U_i z_{t-1}^l + b_i) \tag{14}$$

$$o_t = \sigma_g(W_o z_{t-1}^l + U_o z_{t-1}^l + b_o) \tag{15}$$

$$c_t = f_t \circ c_{t-1} + i_t \circ \sigma_c(W_c z_{t-1}^l + U_c z_{t-1}^l + b_c) \tag{16}$$

$$z_t^l = o_t \circ \sigma_z c_t \tag{17}$$

- Onde z_{t-1}^l representa a entrada de uma unidade do LSTM e o z_t^l a saída.
- c_t representa uma unidade LSTM.
- W, U e b são matrizes e vetor de parâmetros, respectivamente.
- f_t, i_t e o_t os vetores portas;
- σ_g é uma função sigmoide, σ_c e σ_z são tangentes hiperbólicas;
- o operador $\circ$ representa o produto de Hadamard;
- o subscrito i, o, f, c representam a porta de entrada, porta de saída, *forget gate* e célula.
- O sobrescrito $l-1$ e l são os índices das características de entrada e saída;

3.5. Definição de Hiperparâmetros

Foram colocados como opções para escolha os seguintes hiperparâmetros, taxa de aprendizado (*learning rate*, do inglês), otimizador (*optimizer*, do inglês) e tamanho do lote (*batch size*, do inglês). Ao final a biblioteca `hyperopt` retornou os valores mostrados na Tabela 3.

3.6. Etapa de Treinamento

A seguir, será descrita a etapa de treinamento para os Experimentos 1 e 2.

Tabela 3. Hiperparametros utilizado a biblioteca `hyperopt`

Hiperparâmetro	Valor
Learning rate	0.02786
Batch size	32
optimizer	Adagrad

3.6.1. Experimento 1: Áudios sem Ruído

No Experimento 1, foi realizado o treino com a base de dados sem ruído. O conjunto foi dividido em treino, validação e teste, utilizando o método `train_test_split` da biblioteca `scikit-learn`, com 80% dos dados destinados ao treinamento e validação e 20% para testes.

As quantidades resultantes estão apresentadas na Tabela 4.

Tabela 4. Número de Arquivos para Treinamento, Validação e Teste

Tipo de Arquivo	Número de Arquivos
Treinamento	240
Validação	61
Teste	76

Sendo os dados de treinamento e validação passando pelo processo de (*data augmentation*, do inglês), de forma separada com objetivo de aumentar os dados artificialmente. Com funções para alteração de volume de forma aleatória e deslocamento no tempo.

A sequência de blocos convolucionais foi definida os seguintes parâmetros, exceto pelo número de filtros, que aumenta progressivamente a cada bloco, sendo reduzido na última camada, como mostra a Figura 5. Em todos os blocos, o *kernel* utilizado é de tamanho $(3, 3)$, com *strides* de $(1, 1)$, e as operações de *pooling* são realizadas com um tamanho de $(2, 2)$ e *strides* de $(2, 2)$.

Após passarem pelos blocos convolucionais, a imagem vai para uma camada LSTM com 256 unidades. Em seguida, a saída da LSTM entra em uma camada Densa com função de ativação *softmax*, gerando assim as probabilidades para cada *label*.

3.6.2. Experimento 2: Adicionando Ruído aos Áudios

O Experimento 2 seguiu os mesmos passos do Experimento 1, com a diferença de que agora foi adicionado ruído Gaussiano. O ruído gaussiano é caracterizado por sua distribuição normal, onde a maioria dos valores de ruído se concentra em torno da média, e a probabilidade de valores extremos diminui rapidamente. Isso pode dificultar a percepção da fala, pois o ruído pode mascarar os sinais de voz, tornando a tarefa de reconhecimento de fala mais desafiadora [Mitra et al. 2017].

As Equações (18–22) descrevem resumidamente o ruído Gaussiano.

$$y[n] = x[n] + w[n] \tag{18}$$

Onde:

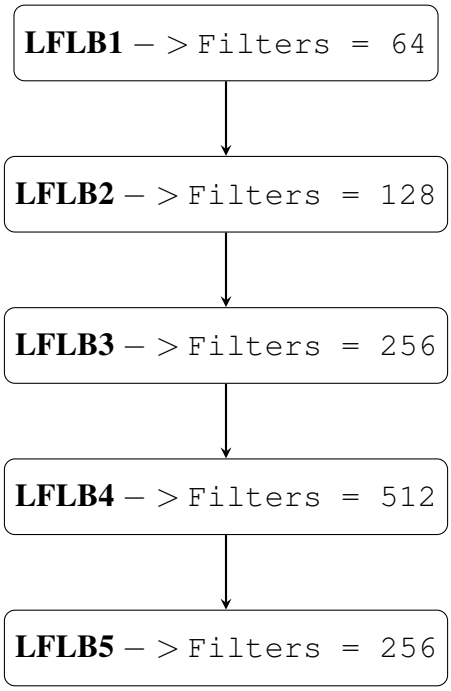

Figura 5. Diagrama das camadas LFLB e seus filtros

- $y[n]$ é o sinal com ruído no instante n,
- $x[n]$ é o sinal original no instante n,
- $w[n]$ é o ruído gaussiano adicionado no instante n.

O ruído gaussiano é extraído de uma distribuição normal com média zero e variância σ_w^2:

$$w[n] \sim \mathcal{N}(0, \sigma_w^2) \tag{19}$$

A potência do sinal $x[n]$ é dada por:

$$P_x = \frac{1}{N} \sum_{n=1}^{N} x[n]^2 \tag{20}$$

A potência do ruído P_w é ajustada de acordo com a Relação Sinal-Ruído (*Signal-to-Noise Ratio - SNR*, do inglês) desejada:

$$P_w = \frac{P_x}{10^{\frac{\text{SNR}}{10}}} \tag{21}$$

A variância do ruído σ_w^2 é dada por:

$$\sigma_w^2 = P_w \tag{22}$$

Para aplicação do ruído foi utilizado o método `AddGaussianSNR` da biblioteca `audiomentations` com os seguintes parâmetros:

- `min_snr_db=5.0`: Define o valor mínimo da SNR como 5 dB.
- `max_snr_db=40.0`: Define o valor máximo da SNR como 40 dB.
- `p=1.0`: Define a probabilidade de 100% (ou seja, essa transformação será aplicada sempre que o áudio for processado)

3.7. Etapa de Avaliação

A seguir, será descrita a etapa de avaliação para os Experimentos 1 e 2.

3.7.1. Experimento 1: Áudios sem Ruído

A avaliação ocorreu em paralelo com o treinamento, sendo a avaliada a acurácia a cada época e a entropia cruzada categórica esparsa, definida como nossa função de perda. Como está resumido na Figura 6. A partir da avaliação foi escolhido o modelo a ser salvo, garantindo que pudéssemos utilizar o modelo com a menor perda para os testes.

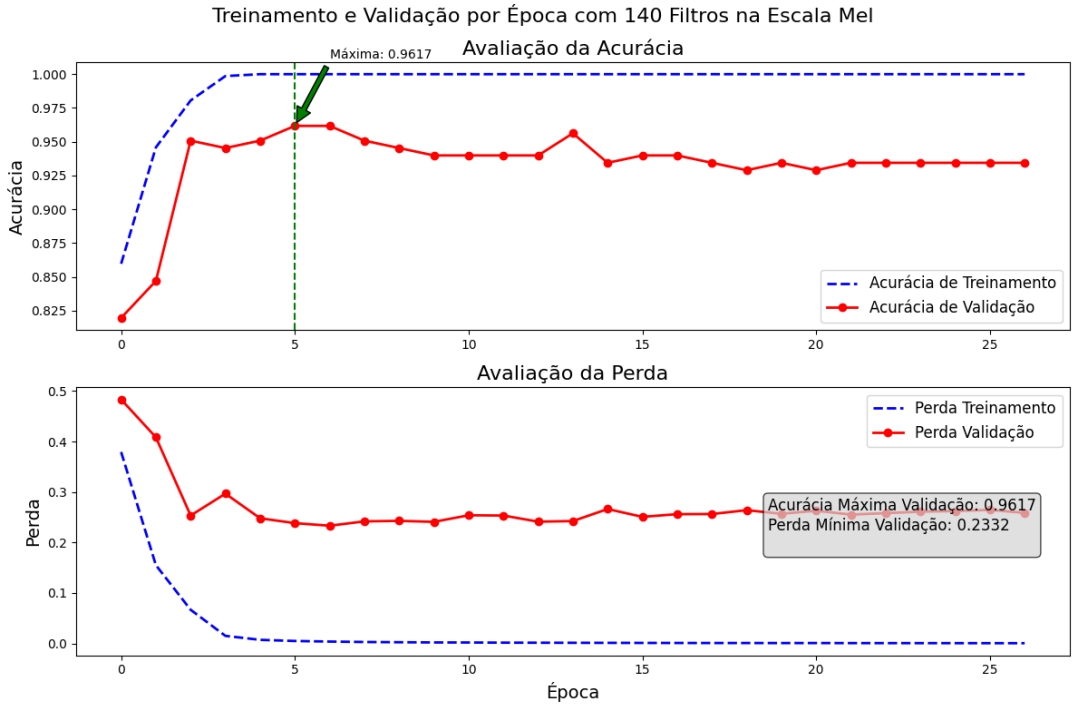

Figura 6. Gráfico obtido durante treinamento do Experimento 1. Fonte: Autor.

3.7.2. Experimento 2: Áudios com Ruído

A avaliação no experimento 2 ocorreu da mesma forma do experimento 1, sendo o seu resumo mostrado na Figura 7.

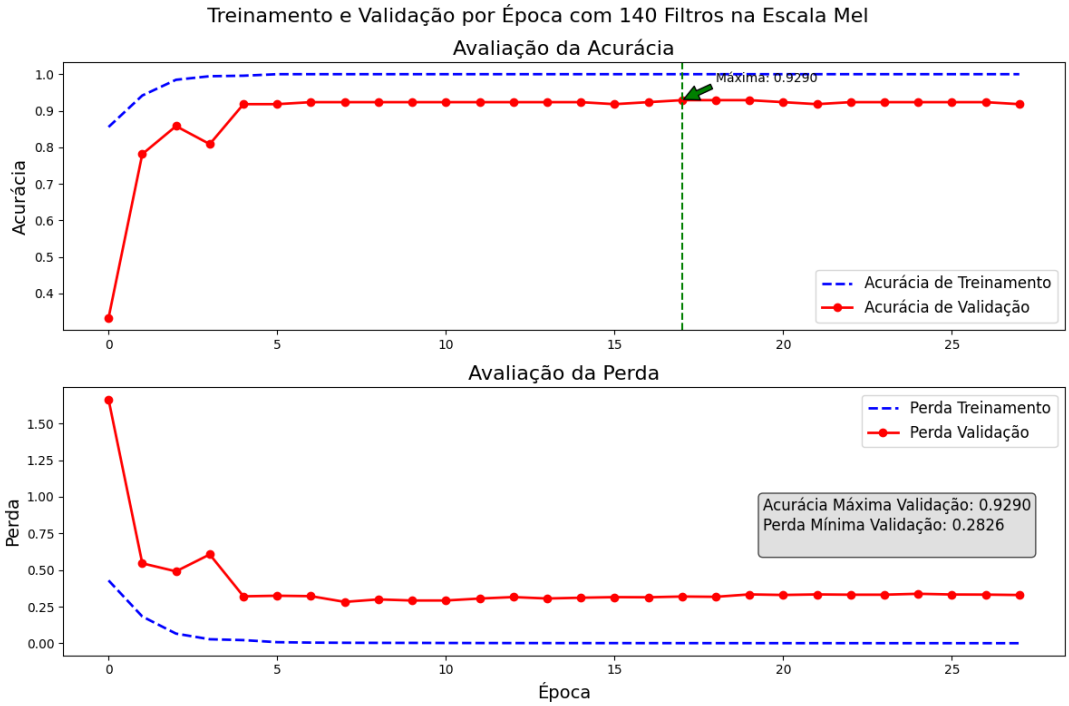

Figura 7. Gráfico obtido durante o treinamento do Experimento 2. Fonte: Autor.

4. Resultados e discussões

Nesta seção, são apresentados os resultados e as discussões dos Experimentos 1 e 2.

4.1. Experimento 1: Classificação com Áudios sem Ruído

Para avaliação dos resultados do modelo foram geradas as seguintes métricas.

Os resultados obtidos demonstram um desempenho robusto do modelo de classificação proposto na tarefa de reconhecimento de emoções. A acurácia geral alcançada foi de 94,74%, já a acurácia ponderada atingiu 95,24%, mostrando que o modelo mantém um desempenho consistente mesmo quando considera a distribuição das classes.

A precisão global do modelo foi de 95,57%, e a revocação global foi de 94,74%, resultando em um F1-Score global de 94,69%. Esses valores sugerem que o modelo é eficaz tanto em identificar corretamente as emoções presentes (revocação) quanto em evitar falsos positivos (precisão).

Analisando as métricas por classe:

- **Felicidade**: Apresentou uma precisão perfeita de 100%, porém com uma revocação de 80,95%, demonstrando que, embora todas as previsões de felicidade estejam corretas, nem todos os casos reais de felicidade foram identificados. O F1-Score de 89,47% reflete esse equilíbrio.
- **Raiva**: Obteve uma precisão de 80% e uma revocação de 100%, sugerindo que todos os casos de raiva foram corretamente identificados, mas houve algumas previsões de raiva que não eram verdadeiras. O F1-Score de 88,89% indica um bom desempenho geral nesta classe.
- **Tristeza**: Apresentou altos índices tanto de precisão (96,43%) quanto de revocação (100%), resultando em um F1-Score de 98,18%. Isso demonstra que o modelo é altamente eficaz em identificar e classificar corretamente a tristeza.
- **Neutro**: Alcançou precisão e revocação perfeitas de 100%, com um F1-Score de 100%, indicando desempenho ideal na classificação desta classe.

O resumo desses resultados podem ser observados na Tabela 5.

Tabela 5. Relatório de Classificação para Experimento 1

Classe	Precisão	Revocação	F1-Score	Suporte
Felicidade	1.00	0.81	0.89	21
Raiva	0.80	1.00	0.89	12
Tristeza	0.96	1.00	0.98	27
Neutro	1.00	1.00	1.00	16
Métricas Globais				
Acurácia			0.95	76
Média Macro	0.94	0.95	0.94	76
Média Ponderada	0.96	0.95	0.95	76

O relatório de classificação reforça esses achados, com a classe Neutro destacando-se pelo bom desempenho. A média macro das métricas mostra uma consistência no desempenho

entre as classes, com valores próximos a 94% para precisão, revocação e F1-Score. Já a média ponderada reflete a influência das classes com maior número de amostras, resultando em métricas ligeiramente superiores, próximas a 95%.

A Figura 8, mostra a quantidade de acertos e erros para cada classe em números absolutos.

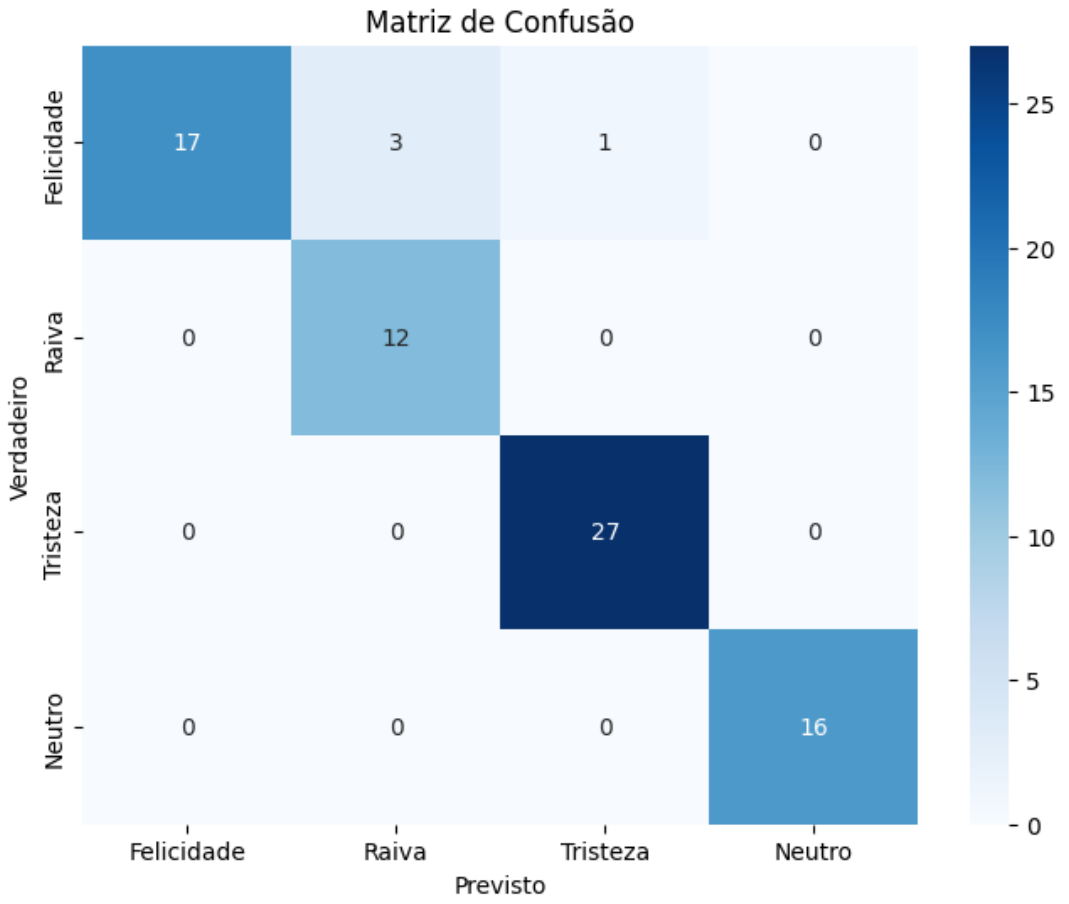

Figura 8. Matriz de Confusão obtida no experimento 1. Fonte: Autor.

4.2. Experimento 2: Classificação com Ruído

Os resultados do Experimento 2 indicam um desempenho sólido do modelo de classificação na tarefa de reconhecimento de emoções, mesmo após a adição de ruído gaussiano. A acurácia geral alcançada foi de 89,47%, enquanto a acurácia ponderada foi de 87,43%, sugerindo que o modelo mantém um desempenho consistente ao considerar a distribuição das classes.

A precisão global do modelo foi de 89,29%, e a revocação global foi de 89,47%, resultando em um F1-Score global de 89,24%. Esses valores indicam que o modelo é eficaz tanto em identificar corretamente as emoções presentes (revocação) quanto em evitar falsos positivos (precisão), mesmo em condições ruidosas.

Analisando as métricas por classe:

- **Felicidade**: Apresentou uma precisão de 90,00% e uma revocação de 85,71%, resultando em um F1-Score de 87,80%. Isso indica que o modelo identifica corretamente a maioria das instâncias de felicidade, embora ainda haja espaço para melhorar a identificação completa desta classe.

- **Raiva**: Obteve uma precisão de 100,00%, mas uma revocação de apenas 50,00%, levando a um F1-Score de 66,67%. Isso sugere que, embora todas as previsões de raiva estejam corretas, o modelo não está identificando todas as ocorrências reais dessa emoção, indicando a necessidade de melhorar a sensibilidade para esta classe.

- **Tristeza**: Apresentou uma precisão de 79,41% e uma revocação de 100,00%, resultando em um F1-Score de 88,52%. Isso demonstra que o modelo é eficaz em identificar todas as instâncias de tristeza, embora haja algumas previsões incorretas nesta classe.

- **Neutro**: Alcançou tanto precisão quanto revocação de 87,50%, com um F1-Score de 87,50%. O modelo mantém um desempenho consistente na classificação de emoções neutras, apesar da presença de ruído.

O resumo das métricas pode ser observado na Tabela 6.

Tabela 6. Relatório de Classificação Experimento 2

Classe	Precisão	Revocação	F1-Score	Suporte
Felicidade	0.90	0.86	0.88	21
Raiva	1.00	0.50	0.67	12
Tristeza	0.79	1.00	0.89	27
Neutro	0.88	0.88	0.88	16
Métricas Globais				
Acurácia			0.86	76
Média Macro	0.89	0.81	0.83	76
Média Ponderada	0.87	0.86	0.85	76

O relatório de classificação confirma essas observações, com a média macro das métricas mostrando valores de 89% para precisão, 81% para revocação e 83% para F1-Score. A média ponderada reflete a influência das classes com maior número de amostras, resultando nas métricas, 87% para precisão, 86% para revocação e 85% para F1-Score.

Portanto, mesmo com a adição de ruído gaussiano no Experimento 2, o modelo demonstra robustez e capacidade de generalização na tarefa de classificação emocional. Embora haja uma leve diminuição no desempenho em comparação com os resultados sem ruído, especialmente nas classes Felicidade e Raiva, o modelo continua eficaz.

Finalizado a Figura 8, mostra a quantidade de acertos e erros para cada classe em números absolutos.

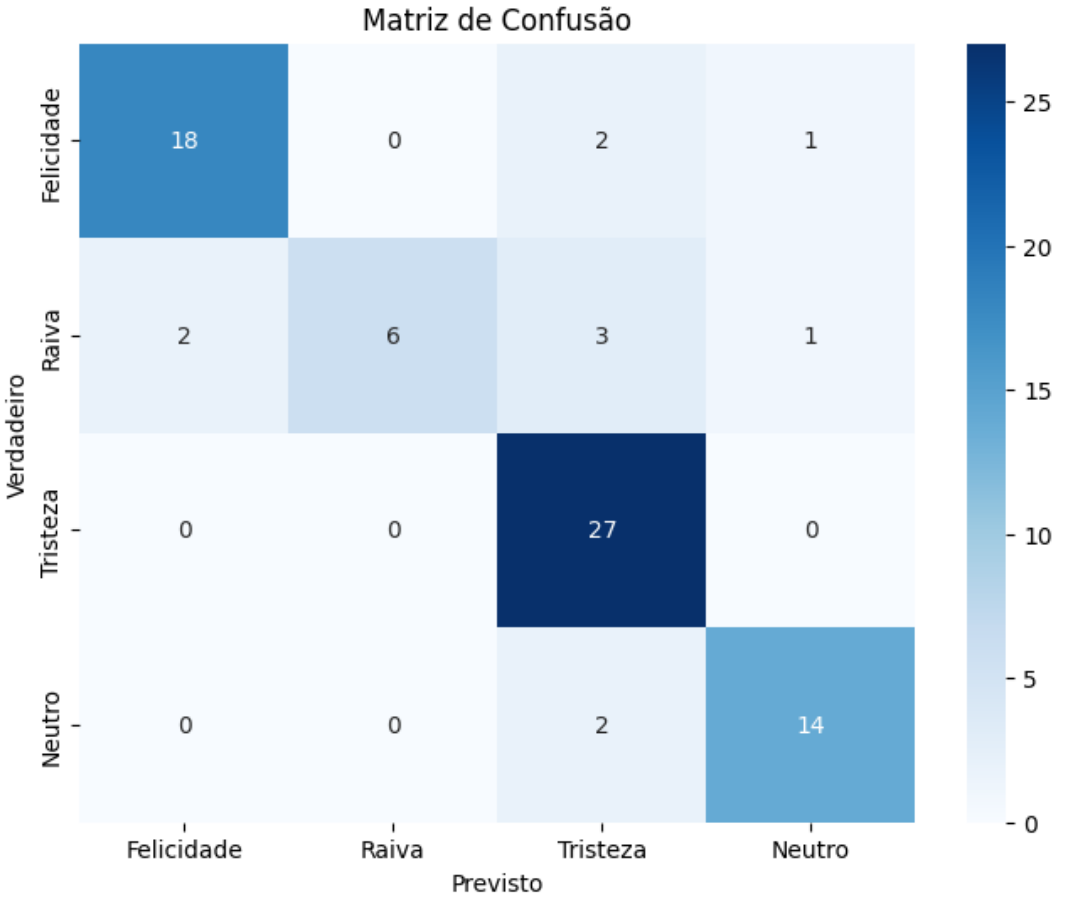

Figura 9. Matriz de Confusão obtida no experimento 2. Fonte: Autor.

5. Conclusão

Este estudo apresentou a aplicação de redes neurais convolucionais e de memória de curto e longo prazo para o reconhecimento de emoções na fala em português, utilizando a base de dados emoUERJ. Os resultados indicaram um desempenho robusto do modelo em cenários com e sem ruído, com métricas de acurácia, precisão e F1-Score alcançando valores satisfatórios em ambos os experimentos.

No primeiro experimento, em condições ideais sem ruído, o modelo obteve uma acurácia global de 94,74% e uma acurácia ponderada de 95,24%, destacando-se na classificação de emoções como ”Tristeza”e ”Neutro”. Já no segundo experimento, com a adição de ruído gaussiano, o desempenho do modelo manteve-se consistente, alcançando uma acurácia geral de 89,47%.

Apesar dos bons resultados, algumas limitações foram observadas, especialmente na classe ”Raiva”, que apresentou maior dificuldade em condições ruidosas. Dessa maneira, há espaço para melhorias, como a utilização outros conjuntos de dados em português e a exploração de arquiteturas alternativas para lidar melhor com ruído nos sinais de áudio.

5.1. Trabalhos Futuros

- Exploração de modelos baseados em *Transformers*, que têm demonstrado resultados promissores em tarefas de processamento de linguagem natural e podem melhorar o desempenho no reconhecimento de emoções em fala.
- Expansão do conjunto de dados em português, incluindo maior diversidade de falantes e variações regionais, visando aumentar a robustez e a capacidade de generalização dos modelos.
- Aplicação de modelos de *Large Language Models* (LLMs) para reconhecimento de emoções, aproveitando sua capacidade de lidar com grandes quantidades de dados e capturar nuances semânticas no processamento de áudio.

Referências

[Akçay e Oğuz 2020] Akçay, M. B. e Oğuz, K. (2020). Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Communication*, 116:56–76.

[Ayadi et al. 2011] Ayadi, M. E., Kamel, M. S., e Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3):572–587.

[Bastos Germano et al. 2021] Bastos Germano, R. G., Pompeu Tcheou, M., da Rocha Henriques, F., e Pinto Gomes Junior, S. (2021). emoUERJ: an emotional speech database in Portuguese.

[de Castro 2016] de Castro, D. G. F. L. N. (2016). *Introdução à Mineração de Dados: Conceitos Básicos, Algoritmos e Aplicações.* Grupo GEN, Rio de Janeiro.

[Demircan e Örnek 2020] Demircan, S. e Örnek, H. (2020). Comparison of the effects of mel coefficients and spectrogram images via deep learning in emotion classification. *Traitement du Signal*, 37(1):51–57.

[Grill et al. 2020] Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P. H., Buchatskaya, E., Doersch, C., Pires, B. A., Guo, Z. D., Azar, M. G., Piot, B., Kavukcuoglu, K., Munos, R., e Valko, M. (2020). Bootstrap your own latent: A new approach to self-supervised learning.

[Géron 2021] Géron, A. (2021). *Mãos à Obra: Aprendizado de Máquina com Scikit-Learn, Keras TensorFlow: Conceitos, Ferramentas e Técnicas para a Construção de Sistemas Inteligentes*. Editora Alta Books, Rio de Janeiro, e-book[a] edição.

[Han et al. 2014] Han, K., Yu, D., e Tashev, I. (2014). Speech emotion recognition using deep neural network and extreme learning machine. In *Proceedings of Interspeech*, pages 223–227.

[Lacerda et al. 2023] Lacerda, T. B., Miranda, P., Câmara, A., e Furtado, A. P. C. (2023). Deep learning and mel-spectrograms for physical violence detection in audio. In *Proceedings of the Conference on Advanced Studies and Systems*, Recife, PE, Brazil. CESAR.School, Recife, PE, Brazil and Universidade Federal Rural de Pernambuco, Recife, Pernambuco, Brazil. Email: `tbl@cesar.school`, {`pericles.miranda`, `andre.camara`, `anapaula.furtado`}`@ufrpe.br`.

[Latif et al. 2020] Latif, S., Qayyum, A., Usman, M., e Qadir, J. (2020). Speech technology for healthcare: Opportunities, challenges, and state of the art. *IEEE Reviews in Biomedical Engineering*, 14:342–356.

[Li et al. 2020] Li, X., Dalmia, S., Li, J., Lee, M., Littell, P., Yao, J., Anastasopoulos, A., Mortensen, D. R., Neubig, G., Black, A. W., e Metze, F. (2020). Universal phone recognition with a multilingual allophone system.

[Meng et al. 2019] Meng, H., Yan, T., Yuan, F., e Wei, H. (2019). Speech emotion recognition from 3d log-mel spectrograms with deep learning network. *IEEE Access*, 7:125868–125881.

[Mitra et al. 2017] Mitra, V., Franco, H., Stern, R., Hout, J., Ferrer, L., Graciarena, M., Wang, W., Vergyri, D., Alwan, A., e Hansen, J. (2017). *Robust Features in Deep-Learning-Based Speech Recognition*, pages 187–217.

[Peixoto e Linhares 2023] Peixoto, G. d. S. e Linhares, J. E. B. d. S. (2023). Reconhecimento de emocões através da fala utilizando rede neural convolucional. Technical report, Instituto Federal de Educação, Ciência e Tecnologia do Amazonas (IFAM) Campus Manaus Zona Leste, Manaus, AM, Brasil. Trabalho técnico.

[Satt et al. 2017] Satt, A., Rozenberg, S., e Hoory, R. (2017). Efficient emotion recognition from speech using deep learning on spectrograms. In *Proceedings of Interspeech*, pages 1089–1093.

[Schneider et al. 2019] Schneider, S., Baevski, A., Collobert, R., e Auli, M. (2019). wav2vec: Unsupervised pre-training for speech recognition.

[Schuller e Batliner 2018] Schuller, B. e Batliner, A. (2018). *Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing*. John Wiley & Sons.

[Schuller et al. 2011] Schuller, B., Steidl, S., Batliner, A., et al. (2011). The interspeech 2011 speaker state challenge. In *Proceedings of the Annual Conference of the International Speech Communication Association*, pages 3201–3204.

[Trigeorgis et al. 2016] Trigeorgis, G., Bousmalis, C., Zafeiriou, S., e Schuller, B. (2016). Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5200–5204.

[Ververidis e Kotropoulos 2006] Ververidis, D. e Kotropoulos, C. (2006). Emotional speech recognition: Resources, features, and methods. *Speech Communication*, 48(9):1162–1181.

[Wan et al. 2018] Wan, L., Wang, Q., Papir, A., e Moreno, I. L. (2018). Generalized end-to-end loss for speaker verification. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4879–4883.

[Wang et al. 2022] Wang, Z., Meng, Q., Lan, H., Zhang, X., Guo, K., e Gupta, A. (2022). Multilingual speech emotion recognition with multi-gating mechanism and neural architecture search.

[Zhang et al. 2018] Zhang, S., Zhang, Y., Tan, T., e Gao, W. (2018). Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching. *IEEE Transactions on Multimedia*, 20(6):1576–1590.

[Zhao et al. 2019] Zhao, J., Mao, X., e Chen, L. (2019). Speech emotion recognition using deep 1d 2d cnn lstm networks. *Biomedical Signal Processing and Control*, 47:312–323.

REFERÊNCIAS

AYADI, M. E.; KAMEL, M. S.; KARRAY, F. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, v. 44, n. 3, p. 572–587, March 2011.

COWIE, R. et al. Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine*, v. 18, n. 1, p. 32–80, January 2001.

FAYEK, H. M.; LECH, M.; CAVEDON, L. Evaluating deep learning architectures for speech emotion recognition. *Neural Networks*, v. 92, p. 60–68, October 2017.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. *Deep Learning*. [S.l.]: MIT Press, 2016.

HINTON, G. E.; SALAKHUTDINOV, R. R. Reducing the dimensionality of data with neural networks. *Science*, v. 313, n. 5786, p. 504–507, July 2006.

HOCHREITER, S.; SCHMIDHUBER, J. Long short-term memory. *Neural Computation*, MIT Press, v. 9, n. 8, p. 1735–1780, 1997.

KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. Imagenet classification with deep convolutional neural networks. In: *Advances in Neural Information Processing Systems*. [S.l.: s.n.], 2012. p. 1097–1105.

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

LEE, C.-C. et al. Emotion recognition using a hierarchical binary decision tree approach. *Speech Communication*, v. 53, n. 9–10, p. 1162–1171, November–December 2011.

MCFEE, B. et al. librosa: Audio and music signal analysis in python. In: *Proceedings of the 14th Python in Science Conference*. [S.l.: s.n.], 2015. p. 18–25.

PICARD, R. W. *Affective Computing*. [S.l.]: MIT Press, 1997.

SATT, A.; ROZENBERG, S.; HOORY, R. Efficient emotion recognition from speech using deep learning on spectrograms. In: *Proceedings of Interspeech*. [S.l.: s.n.], 2017. p. 1089–1093.

SCHULLER, B.; BATLINER, A. *Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing*. [S.l.]: John Wiley & Sons, 2018.

TRIGEORGIS, G. et al. Deep canonical time warping for simultaneous alignment and representation learning of sequences. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 40, n. 5, p. 1128–1138, May 2018.

VERVERIDIS, D.; KOTROPOULOS, C. Emotional speech recognition: Resources, features, and methods. *Speech Communication*, v. 48, n. 9, p. 1162–1181, September 2006.