Flavia Letícia De Mattos

# UM MÉTODO BASEADO EM APRENDIZAGEM PROFUNDA PARA CLASSIFICAÇÕES DE EMOÇÕES EM ÁUDIOS

Curitiba - PR, Brasil

2023

Flavia Letícia De Mattos

# UM MÉTODO BASEADO EM APRENDIZAGEM PROFUNDA PARA CLASSIFICAÇÕES DE EMOÇÕES EM ÁUDIOS

Dissertação apresentada ao Programa de Pós-Graduação em Informática da Pontifícia Universidade Católica do Paraná como requisito para obtenção do título de mestre em Informática.

Pontifícia Universidade Católica do Paraná - PUCPR

Programa de Pós-Graduação em Informática - PPGIa

Orientador: Marcelo Eduardo Pellenz

Coorientador: Alceu de Souza Britto Jr.

Curitiba - PR, Brasil

2023

Dados da Catalogação na Publicação
Pontifícia Universidade Católica do Paraná
Sistema Integrado de Bibliotecas – SIBI/PUCPR
Biblioteca Central
Luci Eduarda Wielganczuk – CRB 9/1118

M444m 2023

Mattos, Flávia Letícia De
Um método baseado em aprendizagem profunda para classificações de emoções em áudios / Flávia Letícia De Mattos ; orientador: Marcelo Eduardo Pellenz ; coorientador: Alceu de Souza Britto Jr. – 2023.
68 f. : il. ; 30 cm

Dissertação (mestrado) – Pontifícia Universidade Católica do Paraná, Curitiba, 2023
Bibliografia: f. 66-68

1. Informática. 2. Processamento de sinais – Técnicas digitais. 3. Inteligência computacional. 4. Aprendizado profundo (Aprendizado de máquina). 5. Redes neurais (Computação). I. Pellenz, Marcelo Eduardo. II. Britto Júnior, Alceu de Souza. III. Pontifícia Universidade Católica do Paraná. Programa de Pós-Graduação em Informática. IV. Título.

CDD. 20. ed. – 004

Pontifícia Universidade Católica do Paraná
Escola Politécnica
Programa de Pós-Graduação em Informática

Curitiba, 28 de agosto de 2023.

71-2023

# DECLARAÇÃO

Declaro para os devidos fins, que **FLAVIA LETICIA DE MATTOS** defendeu a dissertação intitulada "**UM MÉTODO BASEADO EM APRENDIZAGEM PROFUNDA PARA CLASSIFICAÇÕES DE EMOÇÕES EM ÁUDIOS**", na área de concentração Ciência da Computação no dia 30 de maio de 2023, a qual foi aprovada.

Declaro ainda, que foram feitas todas as alterações solicitadas pela Banca Examinadora, cumprindo todas as normas de formatação definidas pelo Programa.

Por ser verdade firmo a presente declaração.

gov.br Documento assinado digitalmente
EMERSON CABRERA PARAISO
Data: 28/08/2023 08:50:02-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Emerson Cabrera Paraiso
Coordenador do Programa de Pós-Graduação em Informática

# AGRADECIMENTOS

Começo agradecendo a minha família: meus pais, Sirlei e Flavio, ao meu irmão, João, e, ao meu namorado, Lucas, pelo apoio, parceria, carinho e incentivo, que, combinado com a minha educação base auxiliou no processo de desenvolvimento deste trabalho.

Continuo com um agradecimento especial aos meus orientadores, Marcelo Pellenz e Alceu Britto, pelos ensinamentos, esclarecimentos e atenção. Com certeza toda a trajetória deste trabalho e a oportunidade de aprender mais com vocês ficará marcada em mim.

Agradeço à PUCPR, ao Programa de Pós-Graduação em Informática (PPGIa) e a Fundação CAPES pela oportunidade e suporte para realização do meu mestrado. Este estudo foi financiado em parte pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código Financeiro 001.

E finalizo os agradecimentos a todos que contribuíram para a execução deste trabalho de maneira direta ou indireta.

*Dedico este trabalho aos meus pais, ao meu irmão e ao meu namorado, que durante o desenvolvimento deste trabalho sempre estiveram ao meu lado, me dando apoio, sempre me incentivando e me dando carinho em todos os momentos.*

# RESUMO

Nos últimos anos, as técnicas de reconhecimento de emoções de fala ganharam importância, principalmente em estudos e aplicações de interação humano-computador. Esta área de pesquisa tem diferentes desafios, incluindo o desenvolvimento de métodos de detecção novos e eficientes, extração eficiente de dados de áudio e estratégias de pré-processamento na análise temporal. Este artigo propõe um novo método para detectar a emoção da fala em dados de áudio brutos. A estratégia proposta usa dados de mel-espectrograma otimizados a partir de arquivos de áudio e combina algoritmos de aprendizado profundo para melhorar o desempenho da detecção. Essa combinação contou com os seguintes algoritmos: CNN (Convolutional Neural Network), VGG (Visual Geometry Group), ResNet (Residual neural Network) e LSTM (Long Short-Term Memory). O papel do algoritmo CNN é extrair as características presentes nas imagens dos mel-espectrogramas aplicados como entrada ao método. Essas características são combinadas com as redes VGG e ResNet, que são algoritmos pré-treinados. Por fim, o algoritmo LSTM recebe todas essas informações combinadas para identificar as emoções predefinidas. O método proposto foi desenvolvido utilizando o banco de dados RAVDESS e considerando oito emoções. Os resultados mostram uma melhoria na métrica da acurácia de 9% em comparação com estratégias na literatura que usam processamento de dados brutos.

**Palavras-chave**: Emoção, Detecção, Aprendizagem Profunda, Áudio, Processamento Digital de Sinais, Rede Neural, Características, Classificação.

# ABSTRACT

In recent years, speech emotion recognition techniques have gained importance, mainly in human-computer interaction studies and applications. This research area has different challenges, including the development of new and efficient detection methods, efficient extraction of audio data and pre-processing strategies in temporal analysis. This article proposes a new method to detect speech emotion in raw audio data. The proposed strategy uses optimized honey-spectrogram data from audio files and combines deep learning algorithms to improve detection performance. This combination relied on the following algorithms: CNN (Convolutional Neural Network), VGG (Visual Geometry Group), ResNet (Residual Neural Network) and LSTM (Long Short-Term Memory). The role of the CNN algorithm is to extract the features present in the honey-spectrogram images applied as input to the method. These features are combined with the VGG and ResNet networks, which are pre-trained algorithms. Finally, the LSTM algorithm receives all this information combined to identify the predefined emotions. The proposed method was developed using the RAVDESS database and considering eight emotions. The results show an improvement in the accuracy metric of 9% compared to strategies in the literature that use raw data processing.

**Keywords**: Emotion, Detection, Deep Learning, Audio, Digital Signal Processing, Neural Network, Characteristics, Classification.

# LISTA DE ILUSTRAÇÕES

# LISTA DE TABELAS

# LISTA DE ABREVIATURAS E SIGLAS

IHC – Interação Humano Computador

PLN – Processamento de Linguagem Natural

IA – Inteligência Artificial

CNN – *Convolutional Neural Network* | Rede Neural Convolucional

LSTM – *Long short-term memory* | Memória de Curto Prazo

VGG – *Visual Geometry Group*

RESNET – *Residual Neural Network*

RNN – *Recurrent Neural Network* | Redes Neurais Recorrentes

BPTT – *BackPropagation Through Time* | Retro-propagação ao Longo do Tempo

SOM – *Self-Organized Maps* | Mapa Auto-Organizado

BMU – *Best Match Unit* | Melhor Unidade Correspondente

RAVDESS – *Ryerson Audio-Visual Database of Emotional Speech and Song*

SAVEE – *Surrey Audio-Visual Expressed Emotion*

EMO-DB – *Berlin Database of Emotional Speech*

IEMOCAP – *Interactive Emotional Dyadic Motion Capture*

RML – *Ryerson Multimedia Research Laboratory*

TESS – *Toronto Emotional Speech Set*

MFCC – *Mel Frequency Cepstrum Coefficients* | Coeficientes Cepstrais de Frequência Mel

RMS – *Root Mean Square* | Raiz Quadrada Média

ZCR – *Zero Crossing Rate* | Taxa de Cruzamento Zero

DTFT – *Discrete Time Fourier Transform* | Transformada de Fourier de Tempo Discreto

IDTFT – *Inverse Discrete Time Fourier Transform* | Transformada Inversa de Fourier de Tempo Discreto

DFT – *Discrete Fourier Transform* | Transformada discreta de Fourier

STFT – *Short-Time Fourier Transform* | Transformada de Fourier de Curto Termo

FFT – *Fast Fourier Transform* | Transformada Rápida de Fourier

VP – Verdadeiro Positivo

VN – Verdadeiro Negativo

FP – Falso Positivo

FN – Falso Negativo

VN – Verdadeiro Negativo

Hz – Hertz

SR – *Sample Rate* | Taxa de Amostragem

# SUMÁRIO

# 1 INTRODUÇÃO

Atualmente, diferentes tecnologias que utilizamos com frequência no nosso cotidiano como, por exemplo, as assistentes virtuais de bancos, entre outros aplicativos para smartphones, utilizam a aprendizagem profunda para receber, interpretar e efetuar comandos a partir da nossa fala ou a partir de uma mensagem escrita. Porém, nesse processo também pode haver uma interpretação da emoção, e com isso a mesma pode ser avaliada para uma tomada de decisão ou uma resposta. Outro tipo de aplicação que utiliza a análise e interpretação das emoções é a avaliação da percepção de um determinado assunto através das redes sociais. Algumas empresas, por exemplo, utilizam essas aplicações para analisar a percepção dos consumidores sobre um determinado produto (RANJAN; SOOD; VERMA, 2018). Essas ações vem se tornando cada vez mais populares em diferentes tecnologias pois agregam ainda mais conhecimento as máquinas, acrescentam ainda mais facilidade ao nosso cotidiano e aumenta a gama de sistemas baseados na Interação Humano-Commputador (IHC) e no Processamento de Linguagem Natural (PLN).

O processamento de linguagem natural é uma subárea da Inteligência Artificial (IA) que tem como vertente estudar a capacidade e as limitações de uma máquina para entender a linguagem dos seres humanos. O objetivo do PLN é fornecer aos computadores a capacidade de entender um texto, sendo que, o termo entender significa a capacidade de reconhecer o contexto, fazer análise sintática, semântica, extrair informação, analisar sentimentos e até aprender conceitos após o processamento. Mas, com todo esse avanço da tecnologia, as análises e interpretações de linguagem enfrentam diversos desafios. Um destes desafios é a capacidade de criação de ferramentas de software capazes de realizar essa interpretação de maneira acessível, performática e garantir a qualidade ao final do processo.

O conceito de emoções tem uma longa história na ciência e na filosofia e vem se tornado um grande tema de estudo explorado por diferentes grupos de pesquisa, não somente da computação, mas, como por exemplo, da história, psicologia e da teologia. Conceitualmente, as emoções são divididas em dois grupos: básicas e complexas. As emoções básicas, ou emoções primárias, são um conjunto limitado de emoções que são tipicamente reconhecidas por acontecerem de maneira automática pelo ser humano e são associadas a expressões faciais. O primeiro a sugerir que tanto as emoções quanto suas expressões faciais universais são biológicas e adaptativas foi Charles Darwin (DARWIN, 1872). O psicólogo Paul Ekman possui um estudo do ano de 1971 (EKMAN, 1971) onde o mesmo apresenta um conjunto de 6 emoções básicas após observações e estudos em uma tribo indígena, sendo que as emoções são: felicidade, tristeza, medo, raiva, nojo e

surpresa. Existe um modelo chamado O Modelo Circumplex de Emoções proposto pelo psicólogo James Russell em 1989 (RUSSELL; LEWICKA; NIIT, 1989). O modelo circular bidimensional de Russell é dividido em quatro quadrantes com dois eixos cruzados e as emoções são colocadas neste plano bidimensional atráves de uma variável para representar cada eixo. As duas variáveis são a valência e a excitação. O eixo x (horizontal) representa os níveis de valência, ou a escala agradável a desagradável da emoção. O eixo y (vertical) representa os níveis de excitação ou a intensidade da emoção. Com esses estudos e modelos várias bases de dados de imagens, áudios ou músicas foram criadas seguindo as suas definições de emoções.

A proposta deste trabalho é desenvolver um método baseado em algoritmos de aprendizagem profunda que vão analisar, detectar e classificar um conjunto de emoções em arquivos de áudio. Para o desenvolvimento desse método foram definidos diferentes algoritmos, como: CNN (*Convolutional Neural Network*), LSTM (*Long Short-Term Memory*), VGG (*Visual Geometry Group*) e ResNet (*Residual Network*). Posteriormente serão explicados com mais detalhes no decorrer deste trabalho, mas, um dos objetivos de abranger diferentes algoritmos foi a necessidade de explorar os seus comportamentos e resultados no âmbito dos áudios, que é uma área bastante abrangente e nos últimos anos tem sido explorada por diferentes áreas acadêmicas. O conjunto de emoções e seus respectivos arquivos de áudio definidos para o método foram baseadas na base de dados chamada RAVDESS (*Ryerson Audio-Visual Database of Emotional Speech and Song*) (LIVINGSTONE; RUSSO, 2018). Essa base possui um total de 1440 de arquivos na sua base de fala, pois a base possui uma base de fala e outra de música, e possui 8 emoções pré classificadas, e, essas emoções são baseadas no estudo de Paul Ekman (EKMAN, 1971). Porém, o estudo de Paul Ekamn define 6 emoções, e com isso, a base adicionou as emoções de calmo e neutro como condições básicas do ser humano. Demais detalhes da base de dados serão explicados nos próximos capítulos do trabalho.

## 1.1 Motivação

Como mencionado anteriormente, o desenvolvimento de aplicações que analisam, detectam e interpretam emoções vem crescendo no decorrer dos últimos anos, em diferentes áreas. Porém, para o desenvolvimento dessas aplicações é necessário aprofundar os estudos para o desenvolvimento de métodos e aplicações mais eficientes e resultando em uma experiência satisfatória para os usuários finais destes sistemas. Algumas dificuldades inerentes ao uso de aprendizagem profunda nesta área de pesquisa incluem a necessidade de criação de bases de dados consistentes e que auxiliem no desenvolvimento de métodos eficientes. Também é necessário estudar e entender os diferentes algoritmos, como as suas vantagens e desvantagens e como a partir da base de dados realizar a interpretação das emoções desejadas. Sendo assim, a principal motivação deste trabalho é explorar os estudos

sobre as emoções e as técnicas existentes relacionadas a área de aprendizagem profunda, para aprimorar os métodos existentes para detecção de emoções em áudios.

Discorrendo melhor a partir das dificuldades citadas, outra motivação para o desenvolvimento do método foi abranger características não funcionais, ou seja, além da aplicação dos algoritmos e a base de dados, o autor teve como visão manter o método eficaz, pois, como uma das características do desenvolvimento do método foi a aplicação de diferentes algoritmos, se faz necessário ter uma otimização no método. Uma outra visão foi aplicar melhorias na extração das características dos áudios que podem trazer uma complexidade computacional. Outro ponto que se faz necessário relacionar com a motivação é que o desenvolvimento de um método eficaz relacionando o mesmo a ser simples, atingir os objetivos com qualidade e ser reprodutível para futuras pesquisas.

## 1.2 Objetivos

O objetivo principal deste trabalho é desenvolver um método eficiente baseado em um modelo preditivo, que, com auxílio das técnicas de aprendizagem profunda possa classificar emoções a partir de áudios, sendo esses áudios falas de atores gravadas em estúdio. Para isso, inicialmente foi desenhado a representação do problema explorando as principais características dos áudios e os métodos possíveis para extração e visualização das mesmas, e, em seguida, montar o modelo baseado em aprendizagem profunda.

### 1.2.1 Objetivos Específicos

A partir do objetivo principal, definimos três objetivos específicos para este trabalho:

- O primeiro objetivo é abranger as técnicas de processamento dos áudios, onde, se inicia com a leitura do áudio sem nenhum processamento, ou seja, manter a integridade original das informações. A partir disso, foi aplicado a extração das características, que, podem ter visualizações diferentes a partir da variação dos parâmetros, e por fim, aplicar a característica extraída no método.
- O segundo objetivo específico é explorar as diferentes técnicas de aprendizagem profunda combinado com algoritmos de rede neural e classificação, onde, cada um possui uma área de aplicação, vantagens e desvantagens, que, quando combinadas podem trazer ou não resultados relevantes ao desenvolvimento do método proposto.
- O terceiro e último objetivo específico é alcançar de resultado final em relação a métrica de acurácia valores superiores a 70%. Esse valor foi definido a partir das análises dos trabalhos relacionados.

## 1.3 Questões de Pesquisa

Para o desenvolvimento deste trabalho, além da sua motivação e objetivos, também existem algumas questões a serem respondidas com base no desenvolvimento do método proposto e os resultados obtidos. Sendo assim, as questões a serem respondidas estão listadas a seguir:

1. O pré-processamento dos sinais de áudio, como por exemplo a geração das características espectrais com diferentes parâmetros, influenciam significativamente nos resultados do método ?
2. A utilização de características espectrais em formato de imagem apresenta resultados relevantes, quando são entradas para o método ?
3. Os algoritmos de aprendizagem profunda combinados com algoritmos de rede neural seria a que retorna o melhor resultado para o método ?

## 1.4 Estrutura do Documento

Este documento está estruturado da seguinte forma:

- Capítulo 2 – Neste capítulo é apresentado todos os conceitos teóricos utilizados para o desenvolvimento deste trabalho, sendo eles: processamento dos áudios, características espectrais e temporais, apresentação da comparação de bases de áudios disponíveis para estudos e apresentação de conceitos sobre a aprendizagem profunda.
- Capítulo 3 – Neste capítulo o objetivo é explanar o estado da arte do presente trabalho, ou seja, apresentar os trabalhos relacionados estudados, explicações sobre os mesmos, comparações com o presente trabalho e as considerações finais.
- Capítulo 4 – Neste capítulo é descrito a metodologia para o desenvolvimento do método proposto por este trabalho, sendo, a aplicação do processamento dos áudios, normalização, truncamento, conceitos sobre o mel espectrograma, como foi desenvolvida a arquitetura do método e os algoritmos desenvolvidos.
- Capítulo 5 – Neste capítulo é apresentado os resultados obtidos pelo método desenvolvido, sendo, que os resultados são apresentados por diferentes visualizações: matriz de confusão, métricas da rede e comparação com o estado da arte.
- Capítulo 6 – Neste capítulo, o último deste documento tem como objetivo compilar todas as informações deste trabalho, e, concluir quais foram as dificuldades encontradas, responder as questões de pesquisa, compilar os resultados, apresentar os próximos passos, e com isso, fechar e concluir o documento e o trabalho.

# 2 FUNDAMENTAÇÃO TEÓRICA

Para o desenvolvimento deste trabalho, e, consequentemente do método proposto, foi necessário estudar diferentes conceitos relacionados aos tópicos base desta área de pesquisa como a visualização e o processamento de arquivos de áudio, processar a base de dados e entender sobre as diferentes técnicas de aprendizagem profunda. Sendo assim, o objetivo deste capítulo é apresentar todo o embasamento teórico do trabalho com a exploração dos tópicos mencionados.

## 2.1 Características dos Sinais de Áudio

Os sinais de áudio são a representação do som que está na forma de sinais digitais e analógicos. Suas frequências variam entre 20Hz a 20kHz, e este é o limite inferior e superior do intervalo audível dos nossos ouvidos. Esses sinais podem ser armazenados em formato WAV que é um formato criado pela *Microsoft* e IBM (International Business Machines Corporation). Este é um formato digital de arquivo de áudio que possui a sequência de amostras quantizadas e codificadas do áudio analógico, não implementando nenhuma estratégia de compressão (KABAL, 2022).

Esses sinais possuem diferentes características que podem ser extraídas a partir do seu processamento digital. As características podem ser referentes ao domínio do tempo, da frequência ou mesmo uma composição de tempo/frequência. Essas caraterísticas podem ser representadas graficamente para diferentes aplicações, como por exemplo, aplicações em redes neurais. A Tabela 1 apresenta alguns exemplos das principais características que podem ser calculadas e analisadas a partir de um arquivo de áudio.

Tabela 1 – Exemplos de Características Extraídas dos Sinais de Áudio

| **Tipo de Característica** | **Frequência** | **Tempo** |
|---|---|---|
| Espectrograma | ✓ | ✓ |
| mel Espectrograma | ✓ | ✓ |
| MFCC (Mel Frequency Cepstrum Coefficients) | ✓ | ✓ |
| RMS (Root Mean Square) | | ✓ |
| ZCR (Zero Crossing Rate) | | ✓ |

O *espectrograma* é uma característica espectral que permite uma representação visual de frequências em função do tempo, indicando a intensidade de cada frequência que compõe o sinal de áudio. Para visualizar essas diferenças é utilizada uma representação usando várias cores que indicam a intensidade de cada componente espectral do sinal.

Matematicamente, os espectrogramas são criados usando-se a transformada de Fourier de curta duração (BADSHAH et al., 2017). O *espectrograma* e o *mel espectrograma* possuem características bastante semelhantes, no sentido de apresentarem a relação de tempo e frequência do sinal de áudio. Porém, o espectrograma trabalha com as frequências originais dos áudios e o mel espectrograma trabalha com a *escala mel* de frequências. Existem estudos que mostram que os humanos não percebem frequências em uma escala linear, ou seja, conseguimos detectar melhor as diferenças em frequências mais baixas do que em frequências mais altas. Com isso, a escala mel foi proposta em 1937 por Stevens, Volkmann e Newmann (STEVENS; NEWMAN, 1934). O principal objetivo da criação dessa escala era representar os componentes do som, por exemplo o tom que varia do grave ao agudo, dando importância aos componentes que são mais sensíveis ao ouvido humano, ou seja, que envolvem a faixa audível de 20Hz a 20kHz. Um objetivo subjacente da escala é reduzir a quantidade de informação que será representada, porém, sem perder as informações importantes. O cálculo para criação da escala envolve operações matemáticas, especificamente uma transformação logarítmica, nas frequências originais do som. Na seção de análise espectral de sinais os conceitos teóricos sobre as duas representações serão melhor detalhados.

O MFCC é uma característica que extrai coeficientes a partir de janelamentos pré-definidos. Essa representação tem como objetivo analisar a energia das componentes de frequência do áudio, em diferentes janelas de tempo. Na Figura 1 é apresentado um exemplo de representação do MFCC.

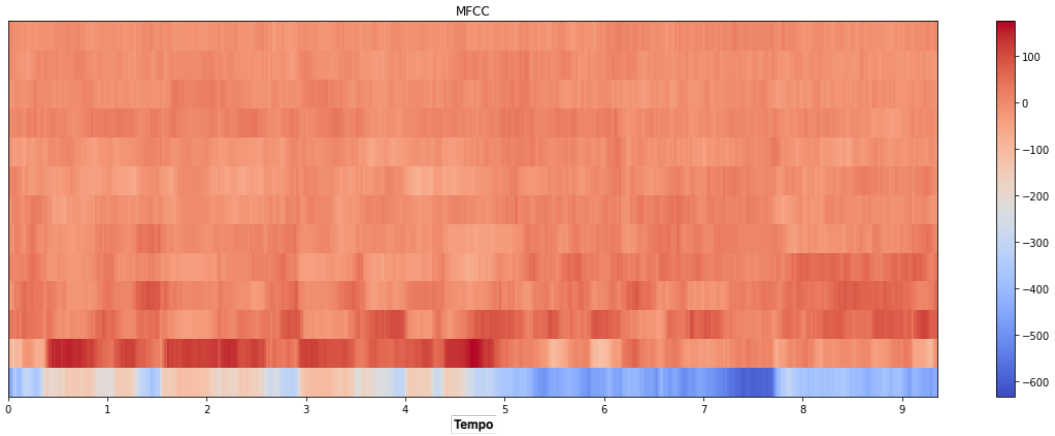


Figura 1 – Exemplo de MFCC (LIBROSA, 2022b)

O RMS é uma característica que representa a potência de cada segmento do áudio e tem como objetivo facilitar a visualização da variações que existem nos sinais de áudio. Essa representação não precisa de cálculo de transformada como o espectrograma, por exemplo. Na Figura 2 é apresentado um exemplo de representação do RMS.

A taxa de cruzamento por zero (ZCR) de um áudio é a taxa de mudanças de sinal

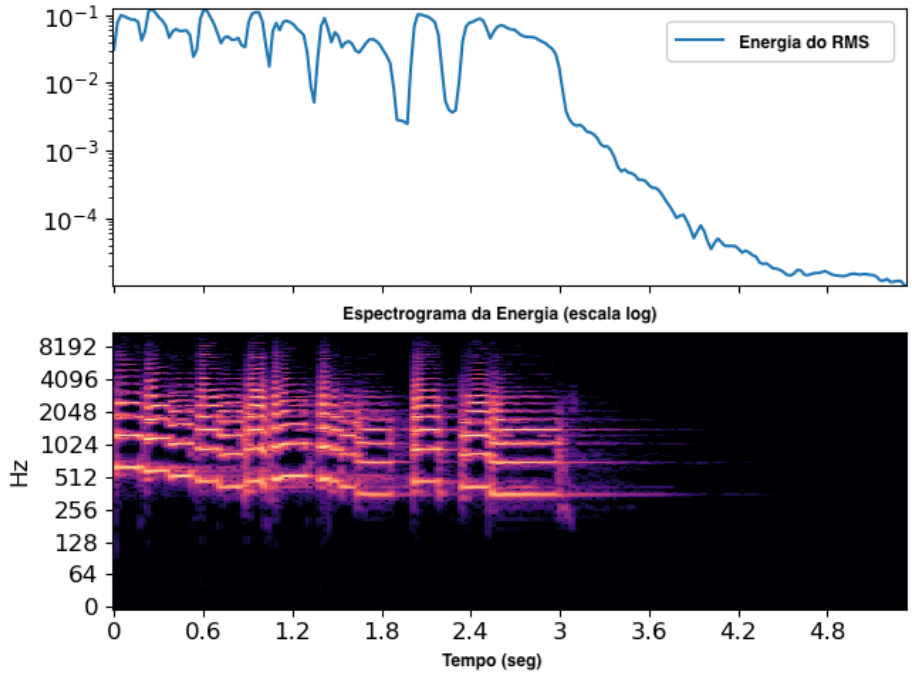


Figura 2 – Exemplo da Característica RMS (LIBROSA, 2022c)

durante um determinado tempo, ou seja, é o número de vezes que o sinal de áudio muda de valor, do positivo para negativo e vice-versa, dividido pelo comprimento do tempo. Essa característica pode ser armazenada e posteriormente utilizada para representar o áudio lido. Na Figura 3 é apresentado um exemplo de representação do ZCR.

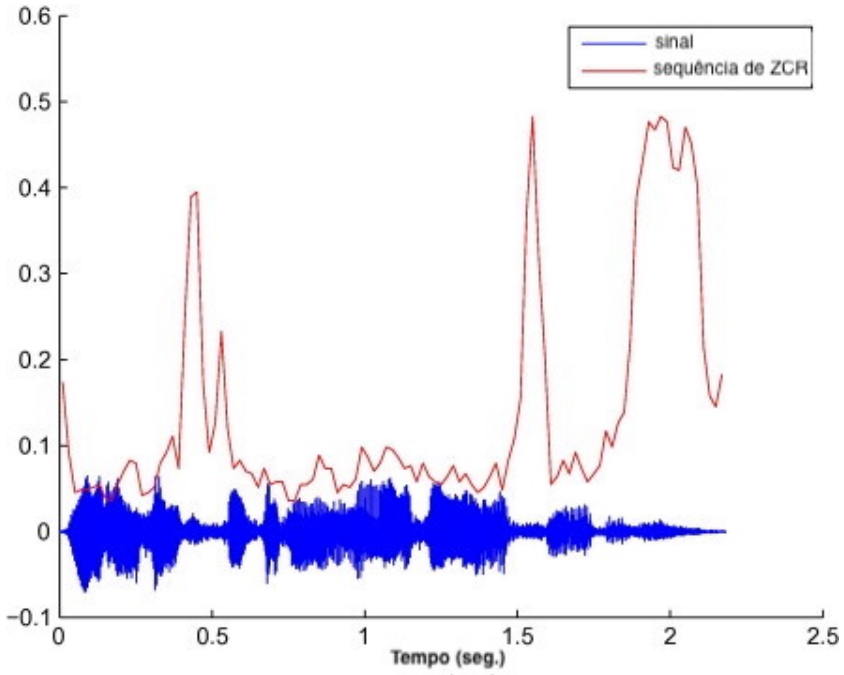


Figura 3 – Exemplo de ZCR (O AUTOR)

### 2.1.1 Análise Espectral de Sinais

Os sinais de áudio podem ser representados ou analisados no domínio da frequência a partir do gráfico de espectrograma. A análise através do espectrograma mostra a variação do conteúdo espectral do sinal em função do tempo. Basicamente, essa representação é possível a partir do cálculo da transformada discreta de Fourier, que é aplicado a diferentes segmentos temporais (janelas de amostras) do sinal de áudio original (EDU, 2021). Usualmente, os cálculos de espectro são realizados em segmentos de janela sobrepostos do sinal, conforme ilustrado na Figura 4. A seguir apresentamos as definições e formulações matemáticas para o cálculo de espectro de sinal em tempo discreto e do seu espectrograma.

Podemos representar matematicamente um sinal de áudio digitalizado como sendo uma sequência discreta, $x[n]$ de comprimento $N_x$, onde $n = 0, \ldots, N_x - 1$. O espectro de frequências de um sinal em tempo discreto, $x[n]$, é definido matematicamente a partir da Transformada de Fourier em Tempo Discreto (Discrete Time Fourier Transform - DTFT). As equações (2.1) e (2.2) definem matematicamente a DTFT e a sua transformada inversa, IDTFT.

$$X(e^{j\omega}) = \sum_{n=-\infty}^{+\infty} x[n] \cdot e^{-j\omega n} \qquad \text{DTFT} \tag{2.1}$$

$$x[n] = \frac{1}{2\pi} \int_{-\pi}^{+\pi} X(e^{j\omega}) \cdot e^{j\omega n}\, d\omega \qquad \text{IDTFT} \tag{2.2}$$

Computacionalmente realizamos o cálculo do espectro de apenas um conjunto de amostras do espectro completo da sequência discreta, que é definido pela DTFT. Neste caso definimos o conceito de Transformada Discreta de Fourier (Discrete Fourier Transform - DFT). As equações (2.3) e (2.4) definem matematicamente a DFT e a IDFT, respectivamente.

$$X[k] = X(e^{j\omega_k})|_{w_k = \frac{2\pi k}{N}} = \sum_{n=0}^{N-1} x[n] \cdot e^{-j\frac{2\pi k n}{N}} \qquad k = 0, 1, \ldots, N-1 \qquad \text{DFT} \tag{2.3}$$

$$x[n] = \frac{1}{N} \sum_{k=0}^{N-1} X[k] \cdot e^{j\frac{2\pi k n}{N}} \qquad n = 0, 1, \ldots, N-1 \qquad \text{IDFT} \tag{2.4}$$

A DFT representa um conjunto de $N$ amostras do espectro contínuo definido pela DTFT. As $N$ amostras do espectro são tomadas nas posições definidas por $w_k = \frac{2\pi k}{N}$, onde $w_k$ representa o valor da frequência normalizada onde a respectiva amostra do espectro foi calculada. A variável $k$ representa o índice da amostra do espectro. Na equação (2.3), a variável $X[k]$ representa os valores complexos (módulo e fase) das amostras do espectro. Na análise espectral de uma sequência discreta $x[n]$, com $N_x$ amostras, podemos calcular uma DFT de $N$ pontos, onde $N \geq N_x$. Quando desejamos aumentar a resolução em frequência da análise espectral da sequência discreta $x[n]$, devemos utilizar um valor $N > N_x$. Para

acelerar o cálculo da DFT se utiliza o algoritmo da FFT (Fast Fourier Transform) (DINIZ, 2014), que basicamente é uma implementação rápida da DFT. A FFT explora algumas propriedades de simetria no cálculo das amostras do espectro, reduzindo a complexidade de processamento.

A STFT (Short-Time Fourier Transform) considera apenas um segmento de curta duração da sequência completa do sinal original, para calcular a DFT. Usualmente isso é feito multiplicando-se a sequência $x[n]$ por um função de janela, $w[n]$ que é menor em duração ou comprimento. Tipicamente, duas janelas de duração finita são comumente utilizadas. A primeira é a janela *retangular*, que essencialmente extrai apenas a sequência desejada de menor comprimento, sem modificações adicionais no sinal. A segunda é a janela de *hamming*, que aplica uma atenuação nas extremidades do sinal janelado, com o objetivo de melhorar a representação espectral. Considerando a definição geral do espectro dado pela DTFT, a STFT é definida pela equação (2.5), onde $m$ representa o índice discreto do quadro (janela).

$$X(m, e^{j\omega}) = \sum_{n=-\infty}^{+\infty} w[m-n] \cdot x[n] \cdot e^{-j\omega n} \tag{2.5}$$

Na prática, a STFT é calculada em janelas finitas usando a DFT, conforme definido pela equação (2.6), onde $N$ denota o número de pontos do espectro (resolução em frequência) usado pela STFT e a variável $N_w$ representa o tamanho da janela de duração finita, $w[n]$.

$$X[m,k] = \sum_{n=m-(N_w-1)}^{m} w[m-n] \cdot x[n] \cdot e^{-j\omega_k n} = \sum_{n=m-(N_w-1)}^{m} w[m-n] \cdot x[n] \cdot e^{-j\frac{2\pi kn}{N}} \tag{2.6}$$

A variável $X[m,k]$ é função do tempo e da frequência, sendo agora o tempo e a frequência, ambas variáveis discretas. Podemos interpretar a STFT como representando a DFT da sequência discreta de duração finita $x[n] \cdot w[m-n]$, onde $m$ identifica a localização do segmento do sinal original extraído pelo janelamento $w[m-n]$.

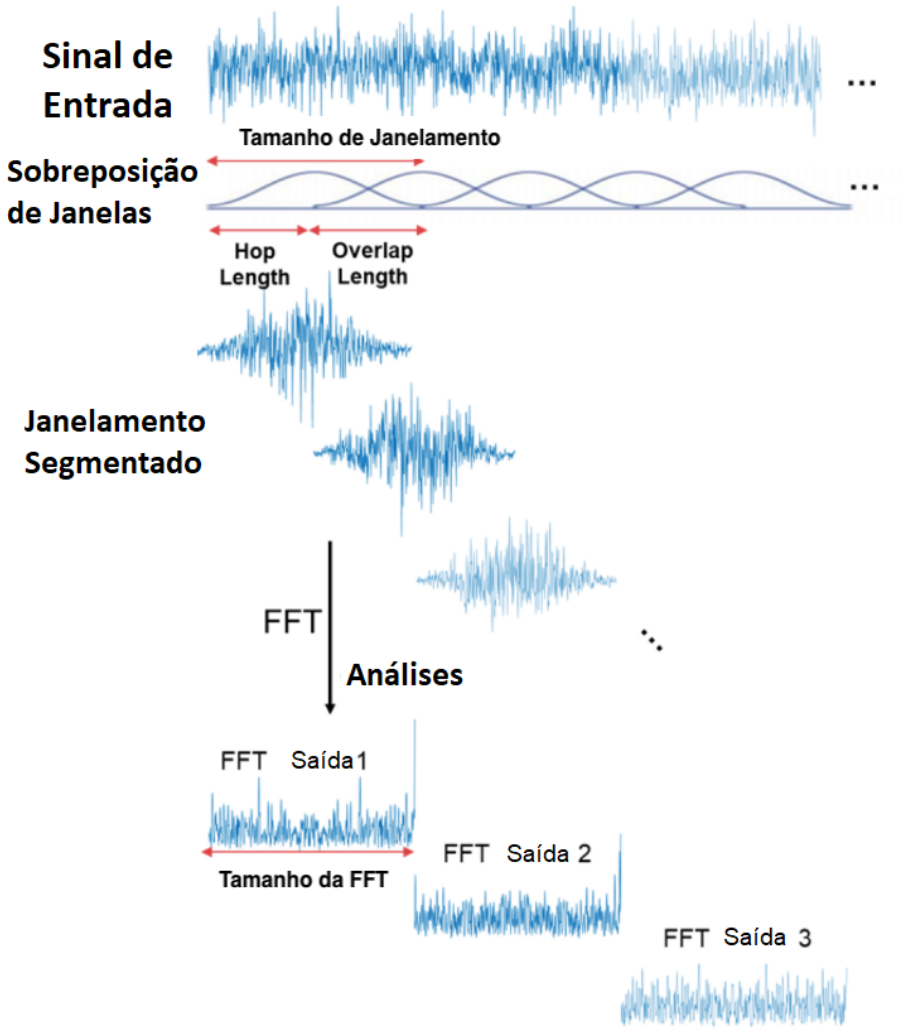


Figura 4 – Exemplo de Cálculo da STFT (EDU, 2021)

O espectrograma é uma maneira simples de se representar visualmente as intensidades das componentes espectrais (frequências) de um sinal ao longo do tempo. A Figura 5 apresenta um exemplo de sinal de áudio, amostrado com frequência de amostragem $F_s$=8kHz. A Figura 6 apresenta o gráfico do espectrograma para este sinal. O eixo $y$ representa a frequência (kHz) e o eixo $x$ representa o tempo. Podemos observar que isso nos permite identificar as variabilidades de frequência em função do tempo. A escala de cores representa a intensidade relativa de cada componente do espectro, conforme escala indicada no lado direito da figura. A frequência máxima do eixo $y$ é dada por $F_s/2$, conforme critério de Nyquist. Os parâmetros utilizados para o cálculo do espectrograma da Figura 6 foram 256 pontos para a FFT, tamanho da janela de 128 amostras (win length) e passo de deslocamento da janela (hop length) de 64 amostras. A escolha destes parâmetros afetam a resolução temporal e em frequência da STFT.

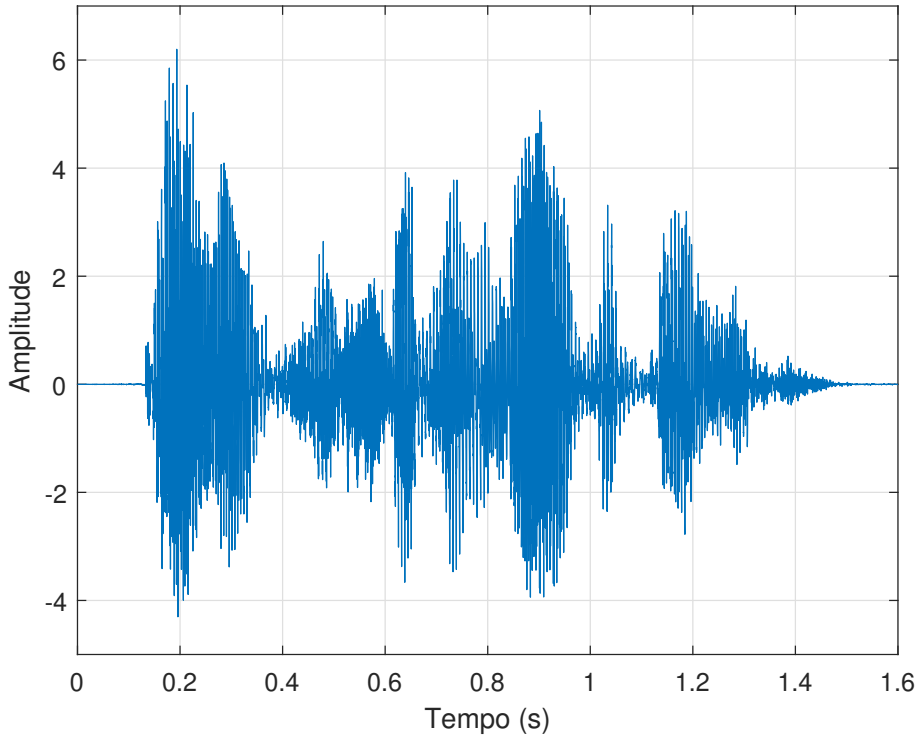


Figura 5 – Exemplo de sinal de áudio (O AUTOR)

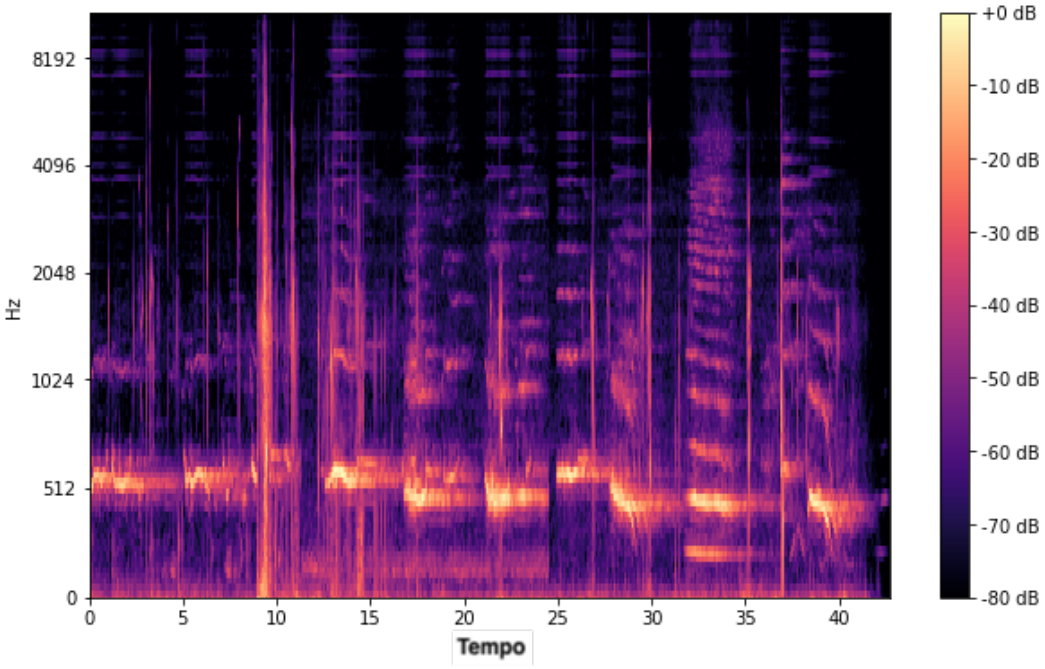


Figura 6 – Exemplo de espectrograma (O AUTOR)

Como mencionado anteriormente, o *espectrograma* e o *mel espectrograma* são exemplos de representação de um áudio. Estas representações são complementares, ou seja, a partir do espectrograma é possível entender o conceito da geração do mel espectrograma. O mel espectrograma é a representação de um áudio relacionando as suas variações

de frequência no decorrer do tempo, igual o espectrograma porém, a diferença é que a frequência do áudio é convertido para a escala mel. A escala mel é uma unidade de medida relacionada ao som (STEVENS; NEWMAN, 1934) e o principal objetivo da sua criação foi construir uma escala que refletisse como as pessoas ouvem os tons musicais. O nome desta escala, *mel*, faz referência à palavra *melodia*. Para a sua construção foram realizados experimentos com ouvintes e aplicado um método baseado em critérios perceptivos que é conhecido na psicofísica como diferença apenas perceptível (*Just-Noticeable Difference*-JND) ou limiar diferencial (*Differential Threshold*). Como mencionado anteriormente, a geração da escala mel envolve a aplicação de uma transformação logarítmica nas frequências originais do som. Esta transformação (O'SHAUGHNESSY, 1987) é definida pela equação (2.7) e ilustrada na Figura 7.

$$Mel = 2595 \cdot \log_{10}(1 + f/700) \tag{2.7}$$

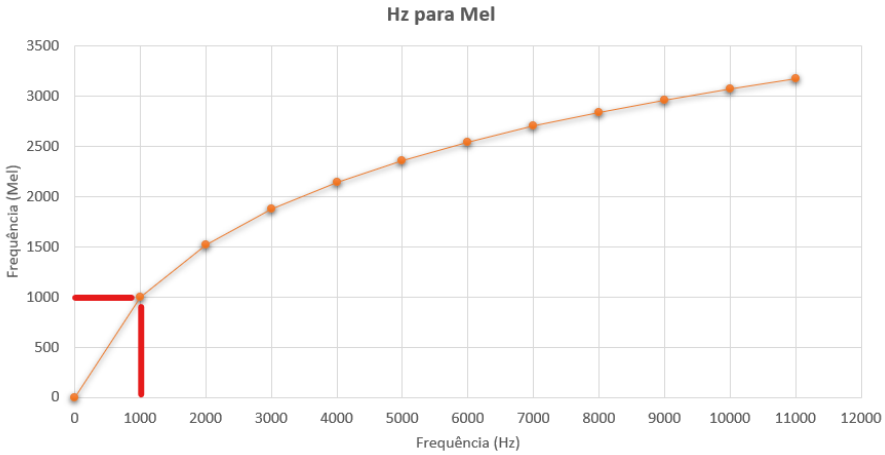


Figura 7 – Representação da Escala Mel (O AUTOR)

As Figuras 8 e 9 ilustram o processamento realizado para o cálculo do mel espectrograma, que como mencionado anteriormente, envolve cálculo de transformada de Fourier. De maneira geral, o mel espectrograma é obtido a partir do espectrograma usando-se um conjunto de filtros para separar as diferentes sub-bandas do espectro, representados na Figura 8. Os coeficientes do mel espectrograma representam a energia contida em cada sub-banda. A Figura 10 ilustra o gráfico de um mel espectrograma gerado.

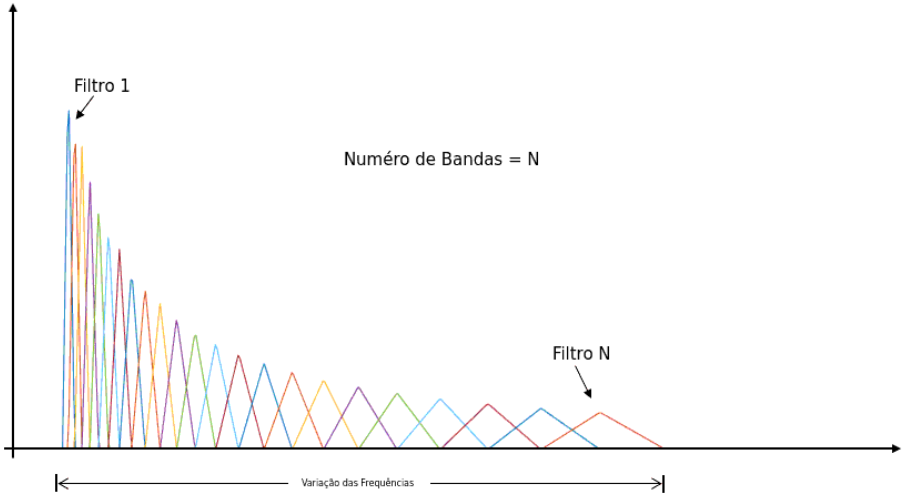


Figura 8 – Número de Filtros e Bandas da escala Mel (O AUTOR)

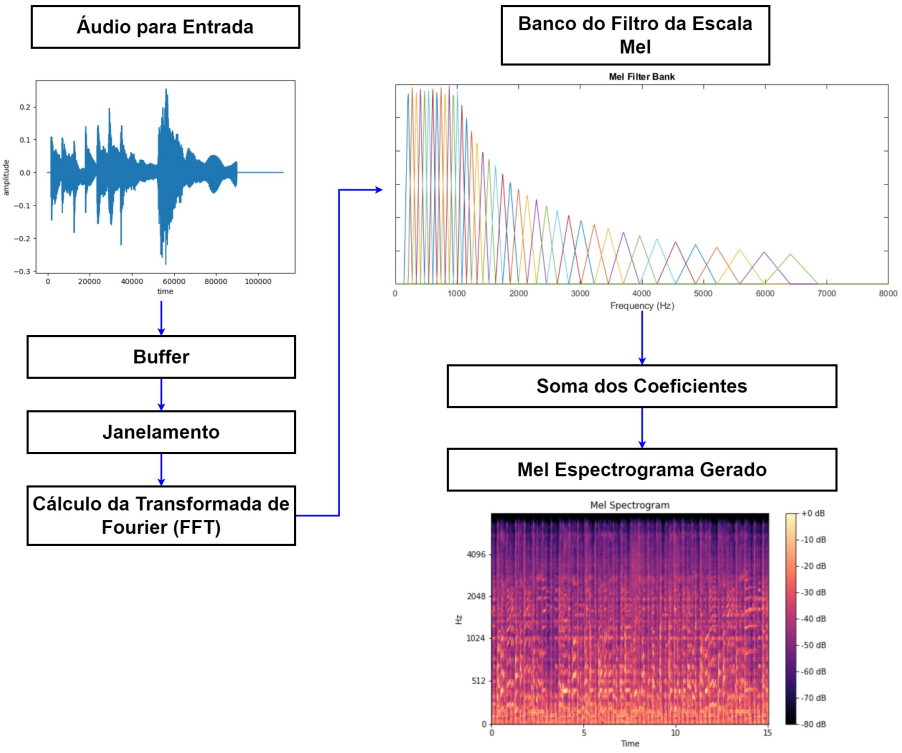


Figura 9 – Processamento para criação do mel espectrograma (O AUTOR)

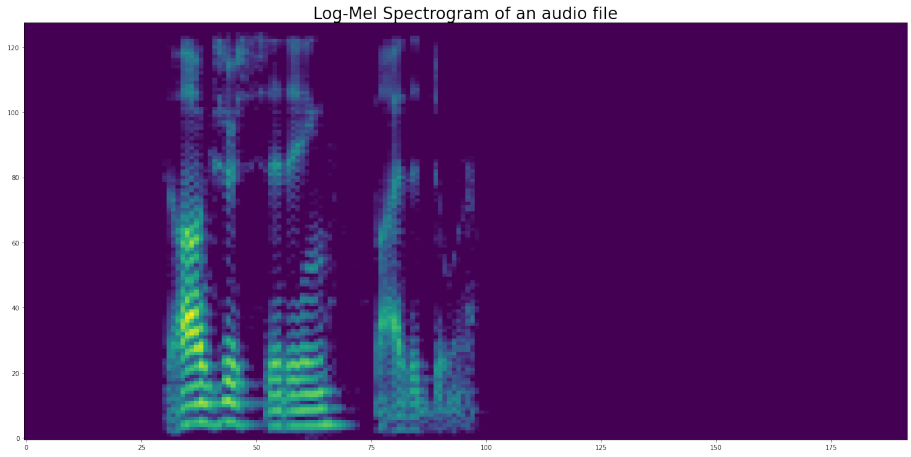


Figura 10 – Exemplo de mel espectrograma (O AUTOR)

## 2.2 Base de Dados

Para o desenvolvimento do método proposto neste trabalho foi necessário encontrar e definir uma base de dados de áudios disponíveis de forma digital, para a análise e detecção das emoções. Para isso, foi realizado um levantamento das bases disponíveis e mais utilizadas nos estudos acadêmicos, que, posteriormente, serão apresentados no estado da arte. A partir desses estudos foi levantado um total de 4 bases de dados disponíveis. Na Tabela 2 é apresentado as principais informações dessas bases pesquisadas, como, o idioma que as bases foram gravadas, quantas emoções, quantidade de atores ou atrizes, e entre outras informações. Analisando as principais informações, uma informação que possui uma diferença notável é na base IEMOCAP que possui mais emoções e arquivos ao ser comparada com as outras bases. Contudo, a base IEMOCAP não foi escolhida como base deste trabalho, pois a maioria dos trabalhos relacionados ao estado da arte utilizam as demais bases de dados. Outra comparação que foi realizada para definir qual seria a base de dados a ser aplicada no trabalho é a questão se a base utilizada sentenças faladas ou cantadas, pois, existem diferentes estudos que aplicam bases de músicas para classificar notas musicais ou genêros musicais. Sendo assim, como o objetivo principal deste trabalho é classificar emoções na fala, a aplicação da base SAVEE não teria nenhum fundamento. Com isso, dentro as opções de basesm a base RAVDESS foi definida como base de dados deste trabalho pois é uma base que possui várias aplicações na área acadêmica, inclusive nos trabalhos que serão abordados no estado da arte, ou, aplicam mais de uma base de dados incluindo a RAVDESS. Além disso, esta base possui uma estrutura simples de identificação das emoções, das sentenças que foram gravadas e como foram gravadas, o que facilita a manipulação da mesma.

Tabela 2 – Características das Principais Bases de Áudios

| Nome | Autor | Idioma da Base | Qtde. de Emoções | Qtde. de Sentenças | Qtde. de Locutores | Qtde. de Arquivos |
|---|---|---|---|---|---|---|
| RAVDESS | (LIVINGSTONE; RUSSO, 2018) | Inglês (Americano) | 8 | 2 | 24 | 1440 |
| IEMOCAP | (NG, 2020) | Inglês (Britânico) | 10 | 2 | 10 | 10040 |
| SAVEE | (JACKSON; HAQ, 2015) | Inglês (Americano) | 8 | 15 | 3 | 480 |
| EMO-DB | (AGNIHOTRI, 2020) | Alemão | 7 | 10 | 10 | 535 |

### 2.2.1 RAVDESS

A base RAVDESS (*Ryerson Audio-Visual Database of Emotional Speech and Song*) (LIVINGSTONE; RUSSO, 2018) é uma base que possui dois modelos de arquivos: fala e música. Para este trabalho foi utilizado somente a base de fala. Essa base de dados possui um total de 1440 arquivos de fala (arquivos de áudios). Esse total é a combinação de 60 gravações por ator de um total de 24 atores, sendo 12 atores e 12 atrizes. A base de fala possui 8 emoções distintas, porém, a quantidade de arquivos por emoção não é a mesma, ou seja, deixando a base desbalanceada. Na imagem é apresentada essa diferença de amostras. Os arquivos estão disponíveis na base no formato **WAV**. A identificação das informações nos arquivos de áudio, como por exemplo, se é um ator ou uma atriz, qual é o tipo de emoção, qual é a sentença que está sendo declarada, entre outras informações, estão identificadas no nome dos arquivos.

O nome dos arquivos possui **7** posições, e cada posição tem relação com uma informação, conforme estrutura apresentada na Tabela 3. Considere como exemplo o seguinte nome de um arquivo da base: **03-01-06-01-02-01-12.wav**. A partir da estrutura descrita da Tabela 3 podemos interpretar as informações do arquivo, conforme apresentado na Tabela 4. As informações apresentadas foram consultadas a partir do *website* (LIVINGSTONE; RUSSO, 2018) da base de dados onde é possível realizar o *download* da mesma.

Tabela 3 – Descrição do Nome dos Arquivos da Base RAVDESS

| Nome da Posição | Descrição da Posição |
|---|---|
| Modalidade | 01 = AV total<br>02 = somente vídeo<br>03 = somente áudio |
| Canal Vocal | 01 = fala<br>02 = música |
| Emoção | 01 = neutro, 02 = calmo<br>03 = feliz, 04 = triste<br>05 = zangado, 06 = medo<br>07 = nojo, 08 = surpreso |
| Intensidade | 01 = normal<br>02 = forte |
| Declaração | 01 = *"Kids are talking by the door"*<br>02 = *"Dogs are sitting by the door"* |
| Repetição | 01 = 1ª repetição<br>02 = 2ª repetição |
| Ator | Números de 01 ao 24<br>Ímpares são homens<br>Pares são mulheres |

Tabela 4 – Exemplo do Nome de um Arquivo da Base RAVDESS

| Nº. no Nome do Arquivo | Informação Relacionada ao Número | Descrição da Informação Relacionada |
|---|---|---|
| 03 | Modalidade | Somente Áudio |
| 01 | Canal Vocal | Fala |
| 06 | Emoção | Medo |
| 01 | Intensidade | Normal |
| 02 | Declaração | "Dogs are sitting by the door" |
| 01 | Repetição | 1ª Repetição |
| 12 | Ator | Mulher |

A Figura 11 exibe algumas fotos dos atores e atrizes quando gravaram os arquivos da base de maneira voluntária. As duas bases disponíveis, de fala e de música, trazem somente os arquivos de áudio no formato especificado e a imagem exibida é somente para ilustração. Sendo assim, as duas bases possuem diferenças linguísticas e de entonação.

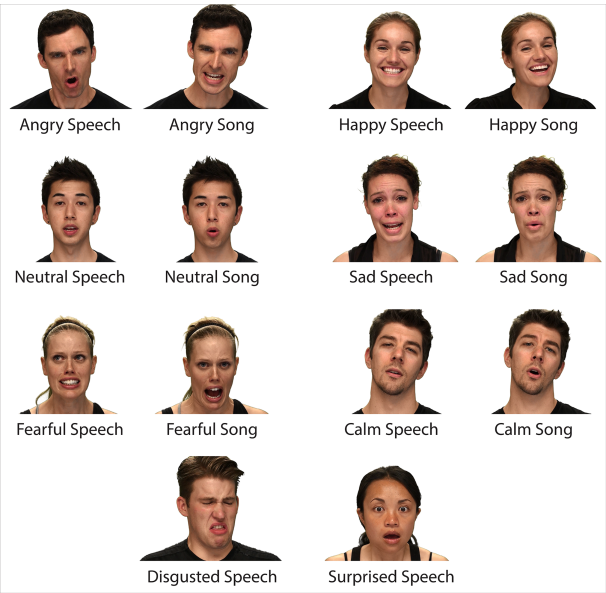


Figura 11 – Exemplo das expressões faciais da base RAVDESS (LIVINGSTONE; RUSSO, 2018)

## 2.3 Aprendizagem Profunda

A aprendizagem profunda (do inglês: *Deep Learning*) é um subconjunto da área da aprendizagem de máquina que, em simples termos, essa aprendizagem aplica diferentes algoritmos de redes neurais para permitir que sistemas aprendam e tomem decisões com base em dados não rotulados. Para o desenvolvimento desse tipo de aprendizagem a maioria dos métodos utilizam arquiteturas de rede neural, e, por isso, são conhecidos também como redes neurais profundas. De maneira geral o aprendizado profundo utiliza diferentes cascatas em várias camadas de processamento não linear para extração de dados e posteriormente os dividi entre as camadas para ao final do processo reunir todos os dados processados para tomar uma decisão, ou, demonstrar uma classificação. As camadas inferiores próximas à entrada de dados aprendem sobre os recursos e as camadas superiores aprendem sobre os recursos mais complexos, sendo que, esses recursos são derivados da camada inferior. A arquitetura de uma forma geral tem uma representação hierárquica, ou seja, significa que o aprendizado profundo é aplicado em analises que se faz necessário extrair conhecimento de grandes quantidades de dados e esses dados foram coletados de diferentes fontes. (ZHANG; LIU, 2017)

Esse tipo de aprendizagem possui duas sub áreas: aprendizagem profunda *supervisionada* e *não supervisionada*. A aprendizagem supervisionada é aplicável em problemas onde a sua solução, ou, o dado alvo é previamente rotulado junto com os demais dados que serão aplicados no treinamento da rede. Existem diferentes algoritmos que são baseadas na aprendizagem supervisionada, como redes neurais convolucionais e redes neurais recorrentes, esses dois modelos e suas variações serão abordados na sequência.

Uma CNN (do inglês: *Convolutional Neural Network*) é uma rede neural multicamadas, e, é extremamente útil em aplicações que utilizam processamento de imagens. A sua arquitetura envolve várias camadas profundas, e, as camadas iniciais são responsáveis por reconhecer os dados (arestas) e as camadas seguintes recombinam esses dados em formato de atributos do nível superior da entrada para realizar o seu processamento. Uma variação da CNN é a arquitetura LeNet CNN. Essa arquitetura é composta por várias camadas que realizam a extração de atributos e, em seguida, realiza a classificação (MADHAVAN; JONES, 2021). Na imagem da Figura 12 é apresentada uma representação da arquitetura LeNet e suas camadas.

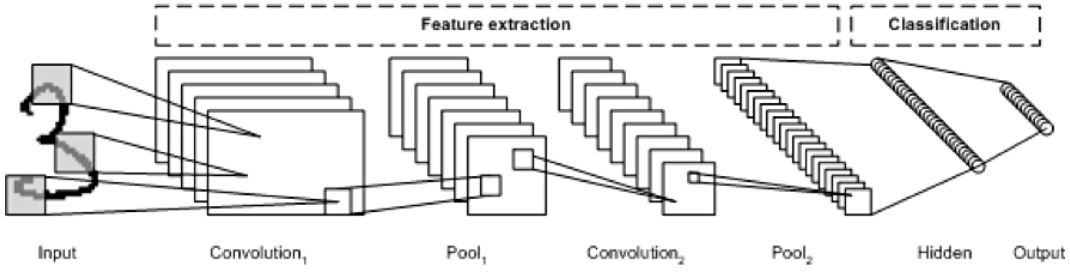


Figura 12 – Arquitetura básica LetNet CNN (MADHAVAN; JONES, 2021)

A imagem de entrada é dividida em campos que representam uma camada convolucional e após isso é realizada a extração das suas características. A etapa seguinte é o agrupamento que tem como objetivo reduzir a dimensionalidade das características extraídas, e ao mesmo tempo, retém as informações mais importantes por meio do agrupamento máximo. A próxima etapa de convolução e *pooling* são executadas, alimentando um *perceptron* de multicamadas (classificador linear) totalmente conectado. Por fim, a saída desta rede é um conjunto de nós que identificam as características da imagem. Caso seja necessário, um novo treinamento da rede pode ser aplicado a partir de uma retro propagação. Essa arquitetura que utiliza camadas profundas de processamento, convoluções, agrupamento e uma camada de classificação totalmente conectada serviu para o surgimento de várias novas aplicações das redes neurais de aprendizagem profunda, como por exemplo, a CNN pode ser aplicada no reconhecimento de vídeo e outras tarefas que envolvem no processamento de linguagem natural (MADHAVAN; JONES, 2021).

A RNN (do inglês: *Recurrent Neural Network* ou Redes Neurais Recorrentes) é uma das arquiteturas de rede fundamentais na área de aprendizagem profunda, e, a sua principal diferença com uma rede multicamada é que uma rede recorrente não possui conexões completamente interligadas e pode ter conexões que se retroalimentam nas camadas anteriores ou na mesma camada. Esse *feedback* permite que rede recorrente mantenha a memória das entradas anteriores e com isso modele outros problemas no tempo. As RNNs consistem em um rico conjunto de arquiteturas e de suas topologias mais populares é chamada de LSTM, que será apresentada na sequência. As RNNs podem ser

manipuladas em relação ao tempo e podem ser treinadas com retropropagação padrão ou usando uma variante de retropropagação chamada retropropagação no tempo (BPTT, do inglês: *BackPropagation Through Time*) (MADHAVAN; JONES, 2021). Na Figura 13 é apresentada uma arquitetura típica de uma RNN.

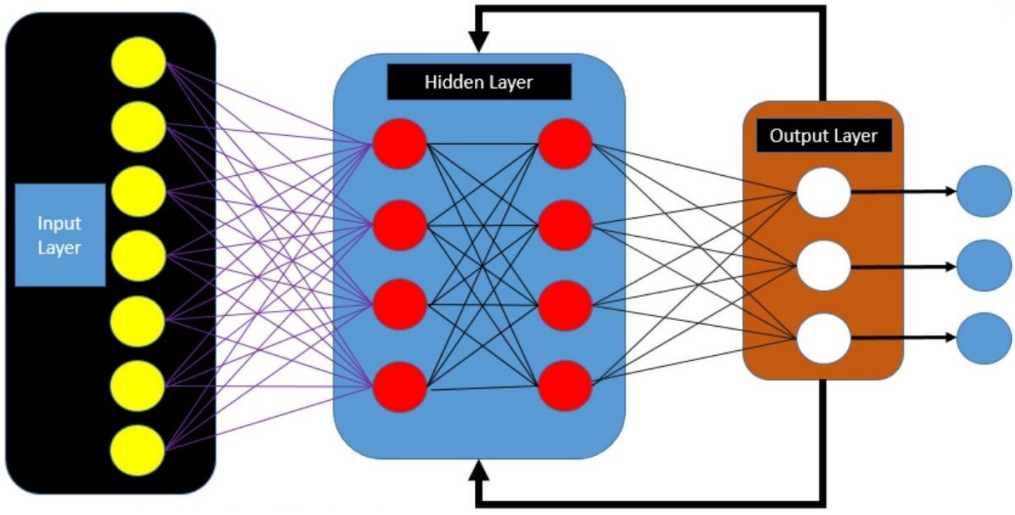


Figura 13 – Arquitetura RNN (ELECTRICAL; ELECTRONICS, 2022)

A LSTM (do inglês: *Long Short Term Memory* ou Memória de Curto Prazo) surgiu a partir da ideia de aprimorar as arquiteturas típicas de redes neurais baseadas em neurônios introduzindo o conceito de uma célula de memória. A célula de memória pode armazenar o seu valor por um curto ou longo período de tempo a partir das suas entradas, ou seja, a célula pode se lembrar do que é importante para a rede e não apenas o último valor. A célula de memória da LSTM contém três portões que controlam como as informações vão fluir para dentro ou para fora da mesma. O portão de entrada controla quando novas informações podem fluir para a memória, já o portão de esquecimento controla quando uma informação existente é esquecida, permitindo que a célula obtenha espaço para armazenar novos dados, e por fim, o portão de saída que tem como objetivo controlar quando as informações contidas na célula são usadas na saída da célula. Para todas essas trocas de informação a célula contém pesos que controlam cada portão mencionado. O algoritmo de treinamento, geralmente BPTT, mencionado anteriormente, otimiza esses pesos com base no erro de saída de rede (MADHAVAN; JONES, 2021). Na Figura 14 é possível visualizar a arquitetura sobre os portões da célula de memória que envolve o funcionamento do LSTM.

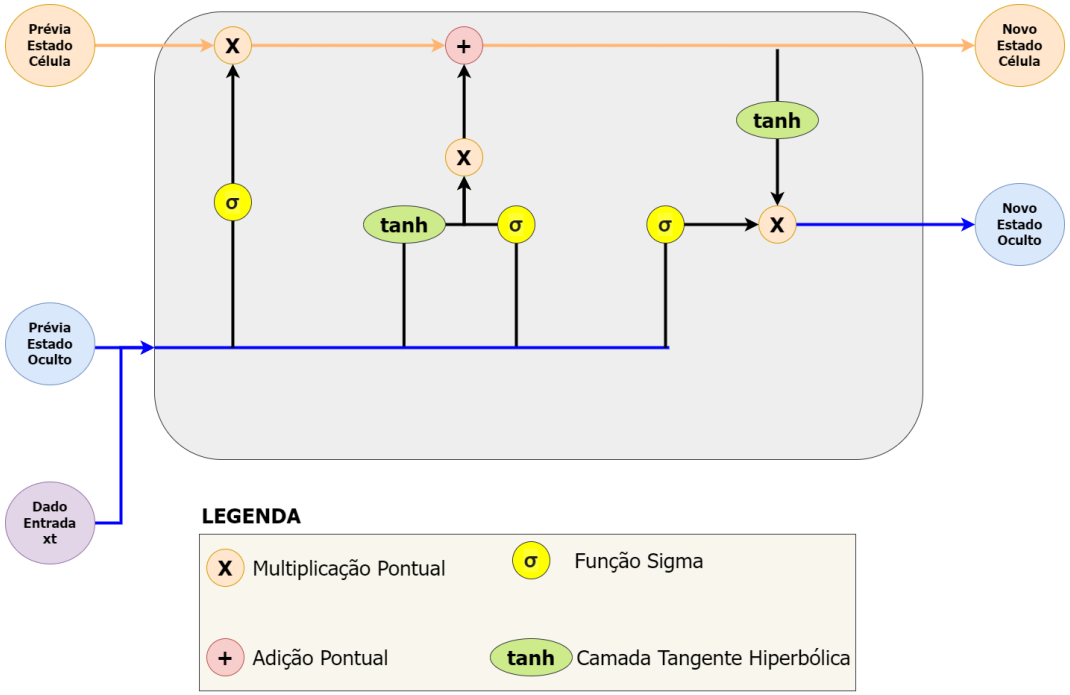


Figura 14 – Arquitetura da célula de memória do LSTM, O AUTOR

A aprendizagem não supervisionada refere-se a problemas em que não há rótulo nos dados usados para treinamento. Para esse tipo de aprendizagem existem diferentes algoritmos como mapas auto-organizados e máquinas boltzmann restritas. O SOM (do inglês: *Self-Organized Maps* ou mapa auto-organizado ) é uma rede neural não supervisionada que cria *clusters*, ou agrupamento, do conjunto de dados de entrada reduzindo a dimensionalidade da entrada. Esse modelo de rede neural possui diferenças quando comparado aos modelos tradicionais, como por exemplo, os pesos servem como uma característica do nó. As entradas são normalizadas e após esse processo uma das entradas é escolhida de maneira aleatória e os pesos inicializados e relacionados as entradas quando são próximos de zero e com isso os pesos representam o nó de entrada. Quando são aplicadas diferentes combinações desses pesos aleatórios os mesmos representam variações do nó de entrada. Um cálculo que é realizado após a definição dos nós é a distância euclidiana, que, entre cada um dos nós de saída com o nó de entrada é calculada e o nó com a menor distância é declarado como a representação mais precisa da entrada e é marcado como a melhor unidade correspondente ou BMU. Essas BMUs são os pontos centrais e os raios dos pontos ao redor dos pesos da BMU são atualizados conforme a proximidade e o objetivo é reduzir o raio. Por fim, ao final da rede, como nenhuma função de ativação é aplicada e não há rótulos de destino para comparação, não há conceito de cálculo de erro e retro propagação.

## 2.4 Avaliação de Desempenho da Aprendizagem de Máquina

Quando desenvolvemos um sistema com aplicação de algoritmos de aprendizagem profunda, é necessário realizar a avaliação de desempenho. Para ser possível avaliar o desempenho conforme o seu objetivo definido previamente, por exemplo, classificação de imagens de cachorros e gatos, é necessário realizar a extração de métricas. As métricas de uma rede neural são a base para diferentes avaliações como desempenho dos acertos da classificação, taxa de erro, tempo de processamento, entre outros. Como o exemplo citado anteriormente, para um problema de classificação é necessário definir quantas e quais classes você deseja que a rede interprete e apresente os resultados, e, a partir desse exemplo é definida a predição do modelo. Predição é a relação entre as classes, ou seja, quantas classes foram classificadas corretamente ou quantas classes foram classificadas erroneamente, e, para entender essa relação existem os seguintes conceitos (FERRARI, 2017):

- Verdadeiro Positivo (VP): quando a rede diz que a classe é positiva e a resposta é positiva;
- Verdadeiro Negativo (VN): quando a rede diz que a classe é negativa e a resposta é negativa;
- Falso Positivo (FP): quando a rede diz que a classe é positiva porém a resposta era negativa;
- Falso Negativo (FN): quando a rede diz que a classe é negativa porém a resposta era positiva;

A partir destes conceitos é possível definir métricas de avaliação para os algoritmos de aprendizagem profunda. A matriz de confusão é um modelo de resultado que possui como princípio exibir a distribuição dos dados em relação as suas classes atuais e das classes previstas pela rede. Isso visa indicar a qualidade da classificação do modelo desenvolvido (TRENTIN, 2015). Na Figura 15 é apresentado um exemplo da matriz de confusão.

| | | **Valor Predito** | |
|---|---|---|---|
| | | **Sim** | **Não** |
| **Real** | **Sim** | Verdadeiro Positivo (VP) | Falso Negativo (FN) |
| | **Não** | Falso Positivo (FP) | Verdadeiro Negativo (VN) |

Figura 15 – Exemplo de Matriz de Confusão (NOGARE, 2020)

A métrica de acurácia, $A$, é uma métrica que avalia o percentual de acertos do modelo, ou seja, ela é obtida pela razão entre a quantidade de acertos pelo total de entradas. Outro método para se obter essa métrica é a partir da matriz de confusão apresentada anteriormente, sendo definida como:

$$A = \frac{VP + VN}{VP + FN + VN + FP} \tag{2.8}$$

A métrica de *recall* ou sensibilidade é uma métrica que avalia o quanto a rede conseguiu obter sucesso com resultados classificados como positivos, sendo definida como:

$$Recall = \frac{VP}{VP + FN} \tag{2.9}$$

A métrica *precision* é uma métrica que avalia a quantidade de verdadeiros positivos sobre a soma de todos os valores positivos que a rede classificou. Essa métrica é definida como:

$$Precision = \frac{VP}{VP + FP} \tag{2.10}$$

*F-score* ou F1-score é uma métrica que envolve o cálculo da média harmônica da *precision* e do *recall*. Ela pode ser obtida com base na equação 2.11:

$$F-\text{score} = 2 \cdot \left( \frac{Precision \cdot Recall}{Precision + Recall} \right) \tag{2.11}$$

A métrica *F-score* é aplicada quando existe um desbalanceamento da base de dados, e, como mencionado na seção de base de dados, a base definida para esse trabalho possui essa característica e com isso, sendo necessário avaliar essa métrica pois se for avaliar pela acurácia não é correto. A matriz de confusão e as métricas apresentadas são utilizadas no Capítulo 5 para apresentação dos resultados do método proposto por este trabalho e também para realizar comparações com os trabalhos relacionados que estão apresentados no Capítulo 3.

# 3 ESTADO DA ARTE

O estado da arte tem como objetivo apresentar as pesquisas científicas relacionadas com todas as características e objetivos relacionados com os presentes trabalho, como, classificar diferentes emoções, aplicação dessa classificação a partir do processamento de áudios, incluir a base de dados RAVDESS, e entre outras informações presentes nas combinações de palavras chaves que serão apresentadas a seguir.

A relação das pesquisas com este trabalho pode envolver a metodologia utilizada para processamento dos arquivos, algoritmos de aprendizagem profunda, base de dados, entre outros pontos. Para encontrar essas pesquisas foi necessário realizar uma busca em diferentes bibliotecas digitais que armazenam esses trabalhos científicos. Essa pesquisa envolve combinar diferentes termos de busca relacionados ao tema central deste trabalho. A partir das pesquisas científicas encontradas, podemos extrair e expandir o conhecimento técnico utilizado no desenvolvimento desta pesquisa. Com isso é possível avaliar a necessidade de alterações ou melhorias no trabalho, e também auxiliar na avaliação dos resultados obtidos em termos da relevância para a área de pesquisa.

Para encontrar essas pesquisas foram consultadas três bibliotecas digitais principais: IEEE, ACM e MDPI. Como mencionado, para encontrar as pesquisas nas bases é necessário criar uma combinação de palavras chaves, sendo assim, o conjunto de palavras para a busca na IEEE foi: *("All Metadata":emotion) AND ("All Metadata":recognition) AND ("All Metadata":cnn) AND ("All Metadata":ravdess) AND ("All Metadata":audio)* com o intervalo de data entre 2012 e 2022. Essa pesquisa retornou um total de 20 trabalhos relacionados. A pesquisa na base ACM teve como conjunto de palavras a combinação: *[All: emotion] AND [All: audio] AND [All: recognition] AND [All: cnn] AND [All: ravdess] AND [Publication Date: (01/01/2012 TO 12/31/2022)]*. Essa pesquisa retornou um total de 18 trabalhos relacionados. Por fim, na base MDPI a combinação das palavras foi: *emotion/cnn/audio/recognition/ravdess*, e, essa pesquisa retornou um total de 3 trabalhos relacionados. A pré-seleção dos artigos para cada base mencionada se iniciou com a leitura dos resumos, pois, é possível levantar algumas informações relevantes dos trabalhos e se possuem relação com o presente trabalho. Após o término dessa leitura foi realizada a leitura completa dos trabalhos, onde, é possível avaliar a aplicação da base de dados, processamento dos áudios, algoritmos, métricas, entre outras informações relevantes. Por fim, com a leitura e o levantamento de informações finalizados, ao final do processo, ocorreu a seleção de 5 pesquisas relacionadas com o presente trabalho.

## 3.1 Trabalhos Relacionados

Nesta seção apresentamos uma síntese das propostas encontradas nos trabalhos relacionados e suas principais características. Estas características incluem, por exemplo, qual foi a base de dados utilizada, quais algoritmos foram definidos na arquitetura e qual foi a característica definida para extração após o processamento dos áudios. Na Tabela 5 são apresentadas essas características gerais dos trabalhos relacionados.

Tabela 5 – Características Gerais dos Trabalhos Relacionados

| Base | Referência | Base de Dados | Característica Extraída | Algoritmos |
|---|---|---|---|---|
| IEEE | (RAJAK; MALL, 2019) | RAVDESS | MFCC, Valência e Excitação | CNN 1D e CNN 3D |
| MDPI | (MUSTAQEEM; KWON, 2020) | IEMOCAP+RAVDESS | Espectrograma (128x128) | CNN 2D |
| ACM | (SLIMI et al., 2020) | RAVDESS+ TESS+RML+EMODB | Espectrograma (150x66) | *One Hidden-Layer Neural Network* |
| ACM | (GUPTA; CHANDRA, 2021) | RAVDESS+TESS | MFCC | *Wide Residual Network* |
| IEEE | (AYADI; LACHIRI, 2022a) | RAVDESS | MFCC | CNN 1D e LSTM |

1D=1 Dimensão - 2D=2 Dimensões - 3D=3 Dimensões

O conceito de *valence* (valência) e *arousal* (excitação) mencionados no trabalho (RAJAK; MALL, 2019) são uma técnica de pré-processamento dos áudios que envolve dividir as emoções em quadrantes a partir dessas duas características. A valência define se a emoção tem aspecto positivo ou negativo, e, a excitação define a força da emoção. Essas características e suas relações com os quadrantes são apresentadas na Figura 16.

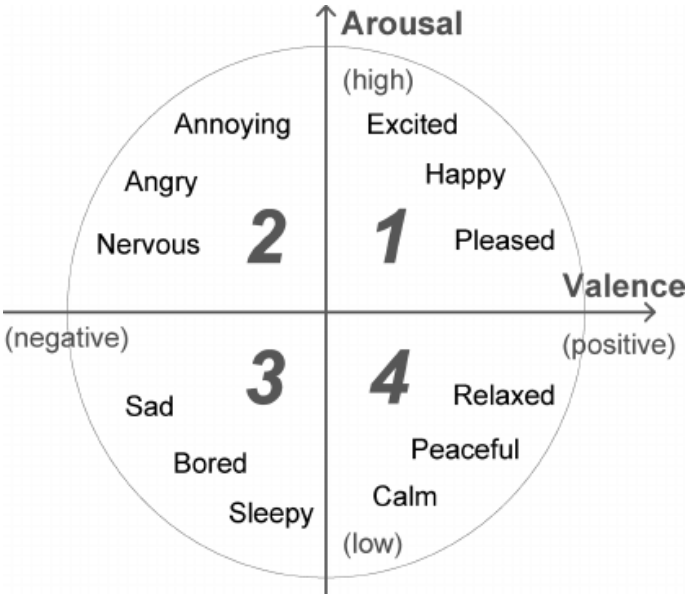


Figura 16 – Relação entre Valência e Excitação das Emoções (RUSSELL; WEISS; MENDELSOHN, 1989)

Para fins de compreensão dos diferenciais de cada trabalho relacionado existem características específicas de cada um, como por exemplo, qual foi a quantidade de emoções classificadas, qual foi o formato de entrada definida para a arquitetura do modelo, se existiu algum pré processamento dos áudios, entre outros pontos. Na Tabela 6 é apresentado as características específicas dos trabalhos relacionados.

Tabela 6 – Características Específicas dos Trabalhos Relacionados

| Referência do Trabalho | Quantidade de Emoções | Formato Entrada da Rede | Divisão Treinamento/Teste | Pré-Processamento |
|---|---|---|---|---|
| (RAJAK; MALL, 2019) | 4 | Vetor | 70% / 30% | Espectral |
| (MUSTAQEEM; KWON, 2020) | 8 | Imagem | 80% / 20% | Temporal+Espectral |
| (SLIMI et al., 2020) | 8 | Imagem | 80-85% / 20-15% | Espectral |
| (GUPTA; CHANDRA, 2021) | 8 | Imagem | 70% / 30% | Temporal+Espectral |
| (AYADI; LACHIRI, 2022a) | 6 | Imagem | 70% /30% | Espectral |

A partir das Tabelas 5 e 6 podemos observar que a maioria dos trabalhos utilizam a característica MFCC, relacionada com a escala mel, e utilizam o formato imagem como entrada da rede responsável por receber as informações e classificar as emoções. Com essas informações, a Tabela 7 tem como objetivo apresentar quais foram os parâmetros utilizados no pré-processamento dos trabalhos relacionados, sendo para MFCC ou para outra característica, com o objetivo de visualizar e compreender as diferenças entre os trabalhos e tirar possíveis conclusões.

Tabela 7 – Parâmetros de Pré-Processamento nos Trabalhos Relacionados

| Referência | $N_{fft}$ | Hop Lenght | $N_{MFCC}$ | Window Size | Overlap | $F_s$ |
|---|---|---|---|---|---|---|
| (RAJAK; MALL, 2019) | 2048 | 1024 | 12 | | | 44.1kHz |
| (MUSTAQEEM; KWON, 2020) | | | | | | 16kHz |
| (SLIMI et al., 2020) | | | | 2048 | 512 | 22.05kHz |
| (GUPTA; CHANDRA, 2021) | 2048 | 512 | 13 | | | 48kHz |
| (AYADI; LACHIRI, 2022a) | | | 12 | | | 48kHz |

Por fim, após a apresentação das principais características e etapas das propostas apresentadas nos trabalhos relacionados, a Tabela 8 apresenta a acurácia obtida pelos métodos propostos nestes trabalhos. É importante ressaltar que existem diferentes métricas possíveis que podem ser utilizadas para fins de comparação, porém, essas comparações serão realizadas com determinados trabalhos relacionados na seção de resultados.

Tabela 8 – Acurácia das Estratégias Propostas nos Trabalhos Relacionados

| Referência do Trabalho | Métricas dos Trabalhos |
|---|---|
| (RAJAK; MALL, 2019) | *F-score* = 50,00 % |
| (MUSTAQEEM; KWON, 2020) | Acurácia = 70,00 % |
| (SLIMI et al., 2020) | Acurácia = 77,50 % |
| (GUPTA; CHANDRA, 2021) | Acurácia = 90,00 % |
| (AYADI; LACHIRI, 2022a) | Acurácia = 53,32 % |

## 3.2 Considerações Finais

A representação das métricas atingidas por cada trabalho teve a sua visibilidade, pois, como abordado anteriormente, a base aplicada nesse trabalho não é balanceada, ou seja, a visibilidade da métrica de acurácia não possui fundamentação e sim pela métrica *F-score*. Sendo assim, as diferenças apresentadas pelos trabalhos visa salientar que os trabalhos que aplicaram somente a base RAVDESS deveriam apresentar o resultado a partir da *F-score* e não somente da acurácia. Os trabalhos que possuem mais base de dados possuem fundamentos para apresentar a acurácia pois algumas realizam combinações entre as bases aplicadas.

O trabalho (RAJAK; MALL, 2019) tem uma abordagem semelhante ao presente trabalho em termos de algoritmo e banco de dados, mas com outras características dos áudios e classificando menos emoções. O método em (MUSTAQEEM; KWON, 2020) foi a base para o presente trabalho, pois utiliza características espectrais, uma arquitetura CNN e oito emoções para classificação. Além disso, eles aplicam pré-processamento de áudio para melhorar o desempenho da CNN.

Em (SLIMI et al., 2020), uma arquitetura de algoritmo diferente foi apresentada, usando o espectrograma de áudio para extração de características. O trabalho apresenta três cenários de validação para classificação usando épocas de 2600, 2900 e 5000. Esse alto número de interações aumenta a complexidade computacional do método. Eles também segmentam o banco de dados de áudio com base na característica de intensidade de áudio (baixa ou alta) para a análise de desempenho. Avaliamos nossa estratégia proposta nos mesmos cenários para uma comparação justa. Os resultados serão apresentados e comparados na seção de resultados (Capítulo 5).

A abordagem em (GUPTA; CHANDRA, 2021) apresenta um algoritmo diferente e usa dois bancos de dados em seu método. No entanto, os autores não apresentam os resultados para as bases de dados separadas, apenas para a combinação. Os autores em (AYADI; LACHIRI, 2022a) apresentam uma arquitetura CNN com diferentes camadas e um LSTM. Eles usaram o espectrograma e suas análises empregam os dois formatos do banco de dados RAVDESS, os bancos de dados de música e fala. Portanto, não é possível comparar diretamente nossa estratégia com as apresentadas em (GUPTA; CHANDRA, 2021) e (AYADI; LACHIRI, 2022a). Considerando os protocolos experimentais aplicados em trabalhos relacionados, é importante ressaltar que o número de arquivos e emoções afetam diretamente o desempenho do método. Os resultados serão apresentados e comparados na seção de resultados (Capítulo 5).

# 4 METODOLOGIA

Este capítulo apresenta informações sobre a metodologia de desenvolvimento desta pesquisa. Isso inclui uma descrição dos principais estudos teóricos e suas aplicações, ferramentas e bibliotecas de simulação, além de outros pontos que foram necessários para o desenvolvimento da pesquisa. Estas etapas estão ilustradas na Figura 17.

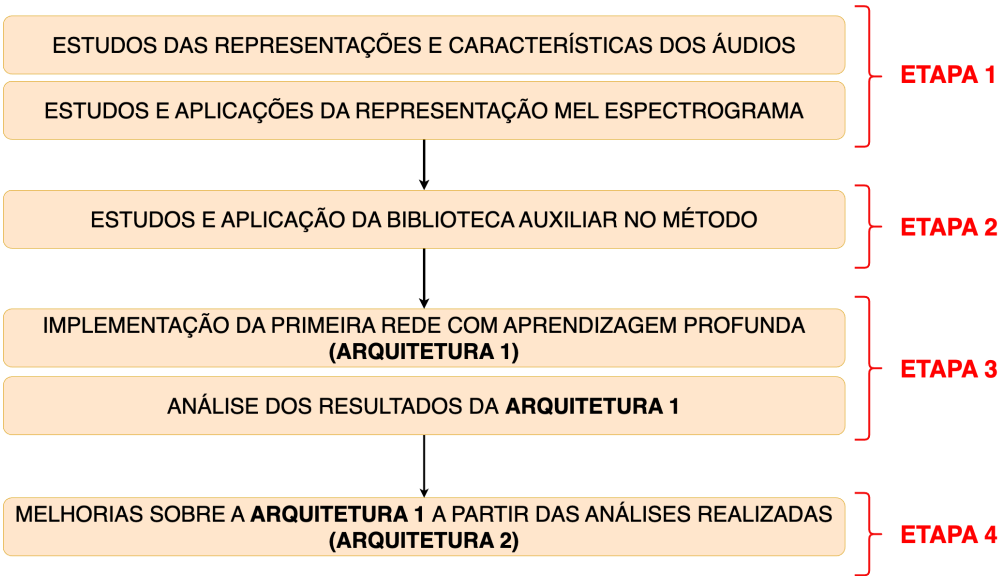


Figura 17 – Metodologia de desenvolvimento do trabalho.

A Etapa 1 teve como objetivo realizar estudos iniciais para entender mais sobre as principais características dos áudios, conforme apresentado na seção de fundamentação teórica. Após a finalização desses estudos e análises sobre as características, o *mel espectrograma* foi definido para ser a entrada de informações da arquitetura do trabalho, que, envolve ter uma representação gráfica do áudio.

A etapa 1 envolveu a realização de estudos inciais para entender mais sobre as principais características dos áudios, quias são as principais representações gráficas dos áudios, quais parâmetros são aplicados para obter essas representações, quais são as influências desses parâmetros nas representações gráficas, entre outros pontos. Esses estudos foram desenvolvidos de maneira prática, ou seja, a partir de criação de um *script* na linguagem *python* onde a entrada são os áudios da base de dados e no decorrer desse *script* os áudios são processados e com isso é apresentado as suas características brutas, ou seja, em formato numeral, e apresenta essas características de maneira gráfica, e também, as representações gráficas como o espectrograma e o mel espectrograma. Com esses estudos finalizados, a definição de aplicar o mel espectrograma como a entrada de informações da arquitetura do trabalho se decorreu pelo fato de ser uma representação

com bastante dado, ou seja, a imagem gerada a partir dessa representação gráfica possui bastante variação de cores e isso auxilia no processo de classificação das emoções, e, por ser uma representação aplicada em alguns trabalhos acadêmicos relacionados. Com a etapa 1 finalizada, a etapa 2 envolveu entender como os parâmetros que envolvem a representação definida são aplicados a partir da biblioteca auxiliar definida para o método, chamada *librosa* (LIBROSA, 2022a). No decorrer desse capítulo são apresentados mais detalhes sobre os parâmetros e a biblioteca auxiliar. A etapa 3 teve como objetivo reunir a parte prática do método, ou seja, desenvolver a rede com aplicação de aprendizagem profunda que recebe os *mel espectrogramas* gerados a partir dos áudios e posteriormente a rede apresenta as emoções classificadas a partir das informações. Essa primeira versão da rede do método proposto foi intitulada de *Arquitetura 1*, e, envolveu a especificação e implementação de uma arquitetura CNN, sem pré-treinamento. Após essa implementação foram analisados os resultados preliminares, com o objetivo de observar e propor possíveis melhorias do método. Algumas melhorias foram aplicadas na Arquitetura 1 previamente desenvolvida. A nova versão foi denominada de *Arquitetura 2*, e envolveu outros algoritmos além da CNN, outros modelos de entrada de informações, entre outras especificações que serão detalhadas no decorrer deste capítulo.

## 4.1 Etapa 1 - Análises de Características dos Áudios

Essa etapa envolveu o estudo dos áudios, com a definição de uma característica para aplicação no método e estudos sobre os efeitos dos parâmetros destas características. A primeira parte da etapa 1 foi um estudo geral a partir da fundamentação teórica, que incluiu a análise das características temporais e espectrais que podem ser analisadas nos áudios. Essas características tem diferentes funções, como representação da variação de frequências presentes no áudio, análise de diferentes tons, entre outros aspectos. Este estudo inicial auxiliou na definição da característica envolvida no método proposto, e também na avaliação dos trabalhos relacionados apresentados no capítulo do estado da arte. A maioria dos trabalhos utiliza a característica MFCC, essa característica possui relação com o mel espectrograma pois utilizam a mesma escala de representação, a escala mel. Contudo, a partir dos estudos e análises realizadas, a aplicação do mel espectrograma foi definida como entrada de informações para rede de classificação, que será apresentada na sequência.

Essa representação envolve alguns parâmetros específicos para sua geração e representação por meio de imagem. A imagem gerada a partir do mel espectrograma é aplicada na entrada da rede com aprendizagem profunda para a classificação de emoções, sendo, que para cada áudio da base de dados uma imagem é gerada e armazenada para ser aplicada na rede. No processo de geração dessas imagens foi necessário aplicar uma função para padronizar a duração dos mesmos e consequentemente ter o mesmo tamanho de imagem.

Esse processo garante que não ocorra perca de informações dos áudios pois processa os intervalos de silêncio que os áudios possuem originalmente. Inicialmente foi necessário entender sobre os parâmetros do mel espectrograma e como aplica-lós para a geração das imagens.

Os principais parâmetros que precisam ser especificados para o cálculo do mel espectrograma são a frequência ou taxa de amostragem ($F_s$), o número de pontos da FFT ($N$), o tamanho da janela ($N_w$), o número de amostras de deslocamento da janela ($H_w$) e a quantidade de bandas do filtro Mel que serão geradas ($N_{mel}$). Na função da biblioteca *librosa* estes parâmetros são denominados como *sample_rate*, *n_fft*, *win_length*, *hop_length* e *n_mel*, respectivamente. Na Tabela 9 são apresentados os conceitos relacionados aos parâmetros mencionados.

Antes de se iniciar as etapas do cálculo da transformada é necessário organizar o sinal em janelas. Essas janelas serão amostras da entrada de tamanho definido pelo parâmetro *win_length*, a partir do qual é calculada a FFT de *n_fft* pontos. A próxima janela é obtida a partir de um deslocamento definido pelo parâmetro *hop_length*, em amostras. Com o sinal organizado por janelas é possível aplicar a transformada de Fourier em cima das amostras. Após o cálculo da transformada é possível converter para a escala Mel todo o espectro de frequência em Hertz e separar em frequências uniformemente espaçadas definidas pelo *n_mel*. A expressão *uniformemente espaçados* não é em relação a distância na dimensão da frequência, mas sim em relação a distância como é ouvida pelo ouvido humano, definição da escala apresentada na fundamentação teórica. A partir da análise dos parâmetros, a próxima etapa envolveu os estudos da aplicação dos mesmos via biblioteca auxiliar, que, teve como um dos seus objetivos a geração das imagens do mel espectrograma.

Tabela 9 – Parâmetros do Mel Espectrograma (Biblioteca Librosa) (LIBROSA, 2022a)

| Parâmetro | Conceito | Exemplo de Valor |
|---|---|---|
| *sample_rate* | Taxa de Amostragem. Quantidade de amostras de um sinal analógico que será lida em uma determinada unidade de tempo e com isso é realizada a conversão em um sinal digital | 8kHz |
| *n_fft* | Especifica o número de pontos da FFT. Sendo normalmente utilizado em números de potência de 2 por questões de complexidade computacional. | 512 |
| *win_lenght* | Especifica o tamanho da janela que o áudio será lido e esse tamanho é definido na chamada da função e, caso necessário a janela será preenchida com zeros para corresponder a *n_fft* | 128 |
| *hop_lenght* | Especifica o número de amostras que serão lidas entre os quadros sucessivos da transformada de Fourier | 64 |
| *n_mel* | Especifica o número de bandas da escala Mel que serão geradas sobre o sinal de entrada | 128 |

## 4.2 Etapa 2 - Biblioteca Auxiliar

A etapa 2 envolveu entender sobre a aplicação da biblioteca auxiliar, chamada *librosa* (LIBROSA, 2022a), que, teve como objetivo auxiliar em diferentes partes do método proposto a partir de funções, como por exemplo, realizar a leitura dos áudios, realizar pré processamento dos mesmos e gerar as imagens do mel espectrograma. Essa biblioteca

possui diferentes funções que auxiliam no desenvolvimento de aplicações que realizam o processamento de áudio ou música, como por exemplo, a função que recebe o arquivo de áudio e os respectivos parâmetros para geração da imagem do mel espectrograma.

Com a função e os parâmetros definidos foi necessário entender como seria possível gerar a melhor imagem possível do mel espectrograma, e para isso acontecer, foi necessário entender a combinação desses parâmetros. Sendo assim, ocorreu um estudo prévio sobre as diferentes combinações possíveis dos parâmetros e seus efeitos. Inicialmente definimos faixas típicas de valores para cada parâmetro. A taxa de amostragem, *sample_rate*, foi definida em 8kHz e 16kHz. O parâmetro *n_fft* teve sua variação definida em função da taxa de amostragem. Para a taxa de amostragem de 8kHz os valores da *n_fft* foram definidos na faixa de 256 a 4096 pontos. Para a taxa de amostragem de 16kHz os valores da *n_fft* foram definidos na faixa de 512 a 8192 pontos. As variações dos valores da *n_fft* são especificados como potência de 2 por questões de complexidade computacional da FFT. O parâmetro *win_lenght* também teve sua variação definida em função da taxa de amostragem, seguindo a mesma variação mencionada do parâmetro *n_fft*. Para a taxa de amostragem de 8kHz, o parâmetro *hop_lenght* teve variação definida entre 64 até 4096, e para a taxa de amostragem de 16kHz iniciou em 128 e variou até 8192. Para o parâmetro *n_mel* definimos apenas um valor fixo de 128, que é um valor típico de análise usado em sinais de áudio. Estas combinações de parâmetros geram tamanhos de janela temporal de análise que variam de 16ms até 512ms.

Com os valores dos parâmetros definidos, foi necessário avaliar as suas combinações e qual seria a melhor para geração das imagens do mel espectrograma, e, para isso, foi desenvolvido a primeira rede do método proposto, intitulada Arquitetura 1.

## 4.3 Etapa 3 - Desenvolvimento da Arquitetura 1

A etapa 3 teve como objetivo implementar uma primeira versão de rede com aplicação de aprendizagem profunda, intitulada Arquitetura 1, e a partir dessa rede implementada realizar análises preliminares de desempenho em termos da acurácia das combinações dos parâmetros de geração do mel espectrograma. As combinações de parâmetros tem um grande impacto no processo de geração das imagens do mel espectrograma, pois, os mesmos são utilizados como entrada juntamente com o áudio lido da base de dados, processo mencionado anteriormente. Sendo assim, os mesmos possuem efeito no desempenho da rede, analisado a partir do valor da acurácia obtida na classificação das emoções, e com isso, é possível identificar qual seria a melhor combinação de parâmetros, podendo ser uma ou mais combinações. A partir dessa primeira rede e seus respectivos resultados é possível focar em melhorar a extração das características e suas respectivas imagens, sendo parte das melhorias a serem aplicadas para a segunda versão da arquitetura, denominada

Arquitetura 2.

A Tabela 10 apresenta todas as combinações dos parâmetros do mel espectrograma juntamente com seus respectivos resultados de acurácia obtidos na Arquitetura 1. As combinações serão apresentadas em duas partes, primeira parte mantendo a taxa de amostragem em 8kHz e variando os parâmetros em ordem crescente, e a segunda parte mantendo a taxa de amostragem em 16kHz e variando os parâmetros em ordem crescente.

Tabela 10 – Combinação de Parâmetros do Mel Espectrograma para Arquitetura 1

| $F_s$ | n_fft | win_length | hop_length | Janela Temporal (STFT) | Acurácia |
|---|---|---|---|---|---|
| 8kHz | 256 | 128 | 64 | 16ms | 61% |
| 8kHz | 256 | 128 | 128 | 16ms | 61% |
| 8kHz | 256 | 256 | 256 | 32ms | 63% |
| 8kHz | 512 | 256 | 128 | 32ms | 69% |
| 8kHz | 512 | 256 | 256 | 32ms | 61% |
| 8kHz | 512 | 512 | 512 | 64ms | 67% |
| 8kHz | 1024 | 512 | 256 | 64ms | 70% |
| 8kHz | 1024 | 512 | 512 | 64ms | 59% |
| 8kHz | 1024 | 1024 | 1024 | 128ms | 60% |
| 8kHz | 2048 | 1024 | 512 | 128ms | 63% |
| 8kHz | 2048 | 1024 | 1024 | 128ms | 55% |
| 8kHz | 2048 | 2048 | 2048 | 256ms | 61% |
| 8kHz | 4096 | 2048 | 1024 | 256ms | 64% |
| 8kHz | 4096 | 2048 | 2048 | 256ms | 63% |
| 8kHz | 4096 | 4096 | 4096 | 512ms | 58% |
| 16kHz | 512 | 256 | 128 | 16ms | 62% |
| 16kHz | 512 | 256 | 256 | 16ms | 63% |
| 16kHz | 512 | 512 | 512 | 32ms | 63% |
| 16kHz | 1024 | 512 | 256 | 32ms | 63% |
| 16kHz | 1024 | 512 | 512 | 32ms | 63% |
| 16kHz | 1024 | 1024 | 1024 | 64ms | 57% |
| 16kHz | 2048 | 1024 | 512 | 64ms | 68% |
| 16kHz | 2048 | 1024 | 1024 | 64ms | 63% |
| 16kHz | 2048 | 2048 | 2048 | 128ms | 59% |
| 16kHz | 4096 | 2048 | 1024 | 128ms | 64% |
| 16kHz | 4096 | 2048 | 2048 | 128ms | 59% |
| 16kHz | 4096 | 4096 | 4096 | 256ms | 58% |
| 16kHz | 8192 | 4096 | 2048 | 256ms | 48% |
| 16kHz | 8192 | 4096 | 4096 | 256ms | 61% |
| 16kHz | 8192 | 8192 | 8192 | 512ms | 60% |

A partir da tabela completa das combinações aplicadas na Arquitetura1, a próxima etapa foi analisar os resultados de acurácia, considerando os valores mais altos para definir melhorias no método proposto. A ideia inicial foi definir 3 conjuntos de combinações para aplicação na Arquitetura 2. Uma análise que vale ressaltar sobre essas melhorias é que a

Arquitetura 1 não foi considerada como versão final do método proposto pois a mesma envolve somente um algoritmo de aprendizagem profunda. Trabalhos relacionados utilizam em média 2 a 3 algoritmos em seus métodos e estratégias. Portanto, ainda é possível melhorar e acrescentar valor ao método proposto. As 3 combinações que obtiveram os melhores resultados na Arquitetura 1 são apresentadas na Tabela 11.

Tabela 11 – Parâmetros Selecionados

| $F_s$ | n_fft | win_length | hop_length | Janela Temporal (STFT) | Acurácia |
|---|---|---|---|---|---|
| 16kHz | 2048 | 1024 | 512 | 64ms | 68% |
| 8kHz | 1024 | 512 | 256 | 64ms | 70% |
| 8kHz | 512 | 256 | 128 | 32ms | 69% |

Ao analisar os resultados de acurácia da Tabela 11, as combinações que possuem a taxa de amostragem em 8kHz possuem o valor de acurácia muito próximos. Outra análise envolveu a aplicação de diferentes taxas de amostragem no método. A variação da taxa de amostragem em 8kHz ou 16kHz influencia na quantidade de dados da arquitetura, ou seja, a imagem possui diferenças significativas quando é aplicada na entrada da rede com aprendizagem profunda. Sendo assim, analisando esses dois pontos, e, avaliando a necessidade de agregar valor e complexidade ao método para os seus resultados serem relevantes para a área de estudo em questão, foi definido que a Arquitetura 2 seria avaliada somente com as duas primeiras combinações descritas na tabela.

Como mencionado, em paralelo ao processo de geração das imagens do mel espectrograma a primeira versão da rede que seria responsável por receber as imagens e classificar as emoções foi desenvolvida com o objetivo de avaliar se o método proposto seria efetivo, e, essa rede foi denominada de Arquitetura 1. Na Figura 18 é apresentada a Arquitetura 1 com todas as suas etapas para ao final ter a classificação das emoções. Na sequência cada etapa será melhor detalhada.

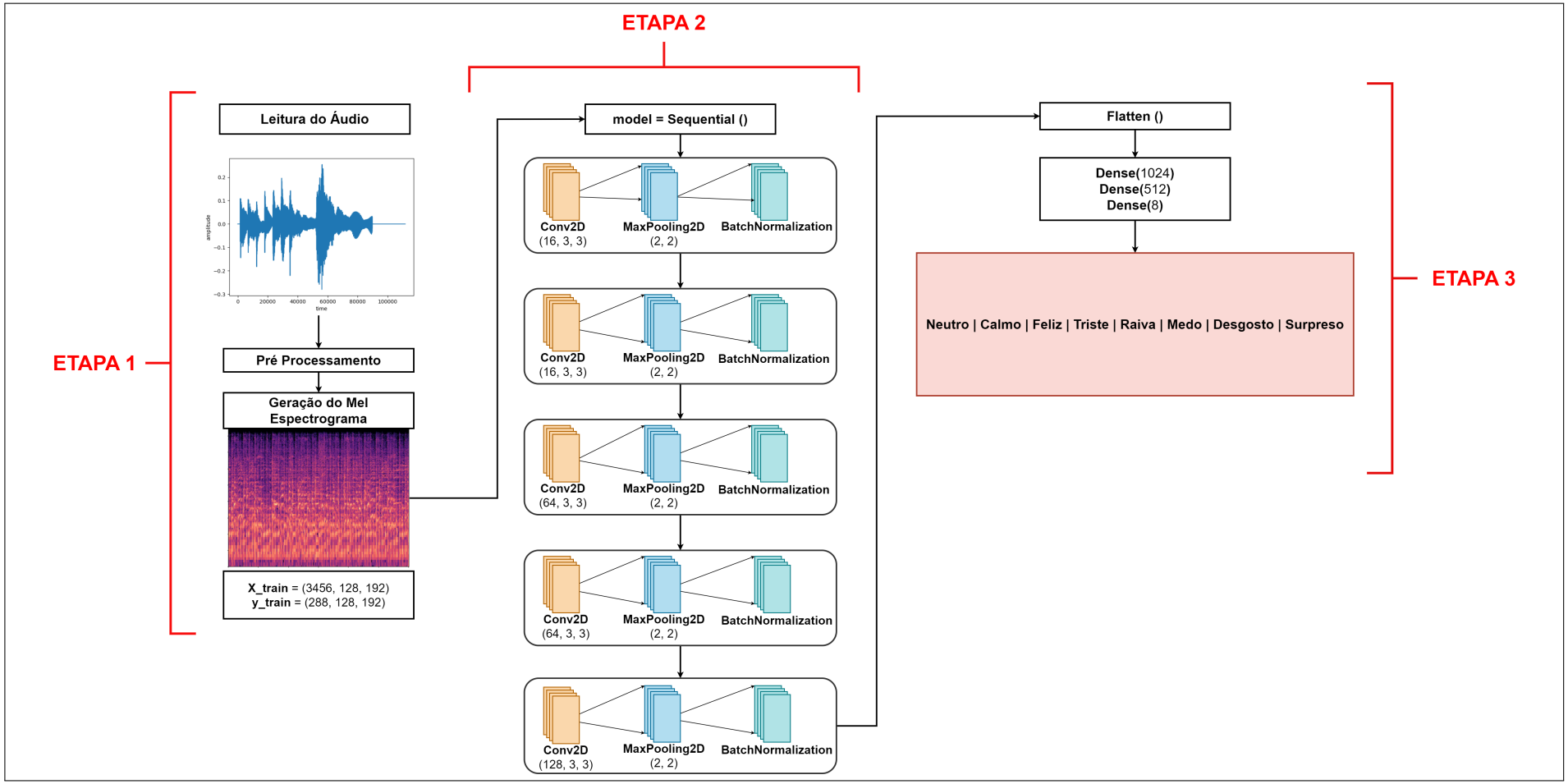


Figura 18 – Modelo da Rede - Arquitetura 1

Como apresentado na Etapa 2 desse capítulo, ocorreu a implementação com auxílio de uma biblioteca que teve várias aplicações durante o processo de desenvolvimento. Na Etapa 1 apresentada na Figura 18 envolveu a aplicação da biblioteca *librosa* para leitura do áudio, sendo esse áudio fornecido pela base de dados. Após essa leitura as informações desse áudio passam por uma etapa de pré-processamento que envolveu separar e armazenar os áudios por emoções, sendo que essa separação é baseada nas informações que estão presentes no nome dos arquivos da base de dados, conforme discutido anteriormente. Outro pré-processamento que é realizado nos áudios é a adição de ruído, que será detalhada na sequência.

A adição de ruído nos áudios lidos da base de dados teve como motivação o fato de os mesmos são gravados em estúdio, ou seja, não possuem nenhuma interferência de ruído que possa ampliar a complexidade para a rede classificar as emoções. Sendo assim, com essa ideia foi desenvolvido uma função para essa etapa que acrescenta ruído branco nos áudios no áudio original. Dessa forma a quantidade de informações da base é ampliada, agregando mais complexidade para a rede ao receber as imagens como entrada.

Após o pré-processamento dos áudios, entramos na etapa de geração das imagens do mel espectrograma. Essa geração é realizada utilizando uma função da biblioteca *librosa* e as mesmas são divididas em dois grandes vetores: treinamento e testes. Na figura 18 são representados os vetores de treinamento, pois, são os aplicados como entrada no algoritmo de aprendizagem para iniciar o processamento das informações. A relação entre os eixos $X$ e $y$ envolvem pré catalogar as informações dos áudios e das emoções, e, durante o algoritmo essas informações são combinadas e processadas conforme as camadas do algoritmo e ao final apresentar o resultado das classificações. O vetor de teste é aplicado ao final para validar e visualizar o resultado da rede após a aprendizagem.

A Etapa 2 da Figura 18 apresenta a rede criada baseada no algoritmo CNN, sendo o modelo *Sequential* ou Sequencial com 2 dimensões. Um modelo sequencial é utilizado no formato de pilha de camadas onde cada camada tem exatamente um *tensor* de entrada e um *tensor* de saída (KERAS, 2022e), ou seja, formato de entrada e saída. Na estrutura desse modelo as camadas são: $Conv2D$, $MaxPooling2D$ e $BatchNormalization$. A camada $Conv2D$ é baseada na operação matemática de convolução, e, seu objetivo e criar um conjunto de filtros em formato de matriz podendo ser de duas dimensões (2D) ou três dimensões (3D), por exemplo. Com isso, na Figura 18 o valor descrito abaixo dessas camadas $Conv2D$ é o filtro que será aplicado nas imagens, por exemplo (16,3,3), e será o novo formato das mesmas. Após a aplicação desses filtros da camada de convolução os *pixels* vizinhos nas imagens tendem a ter valores semelhantes, ou seja, muitas das informações contidas na saída de uma camada convolução são redundantes. Para resolver esse problema a camada $MaxPooling2D$ é aplicada na sequência. Essa camada irá executar o *pooling máximo*. Essa execução envolve percorrer a imagem de entrada em blocos 2x2.

Na Figura 18 é o valor logo abaixo da camada, e esse valor máximo divide a largura e a altura da imagem de entrada. Na sequência a camada $BatchNormalization$ recebe a saída da camada $MaxPooling2D$ e essa camada é uma técnica de normalização feita entre as camadas da rede em vez de processar diretamente os dados brutos. Isso é feito em lotes em vez do conjunto de dados completo, ou seja, melhorando a performance de processamento dentro da rede e agiliza o treinamento pois as taxas de aprendizado são altas.

A última etapa (Etapa 3) tem como objetivo ser o fechamento da classificação das imagens, ou seja, cada camada tem seu objetivo para isso. A camada *Flatten* (KERAS, 2022c) ou achatamento é aplicada na saída do algoritmo CNN e a mesma converte todas as matrizes bidimensionais resultantes em um único vetor linear contínuo e esse novo recurso é a entrada para a camada totalmente conectada para classificar a imagem, na próxima etapa. Após essa função as camadas densas ou *Dense* (KERAS, 2022a) são aplicadas. Essas camadas estão profundamente conectadas à camada anterior, ou seja, os neurônios da camada atual estão conectados aos neurônios da camada anterior. Com essa conexão entre neurônios é aplicado uma multiplicação entre esses neurônios e a sua saída será no formato do parâmetro aplicado na camada, ou seja, na Figura 18 a primeira camada densa terá o formato de saída (1, 1024), a segunda terá o formato (1, 512) e a última o formato em (1, 8). Essa última camada define a quantidade de emoções que deverão ser classificadas. Com isso, é exibido o resultado final do método de maneira geral e o resultado para cada emoção classificada.

Conforme apresentado na Tabela 11, os resultados desta rede ficaram próximos ou atingiram 70% de acurácia. Porém, como mencionado anteriormente, seria necessário avaliar melhorias na Arquitetura 1 com o objetivo melhorar seu desempenho. Portanto, a Etapa 4 da metodologia apresentada na Figura 17 foi desenvolvida.

## 4.4 Etapa 4 - Implementação da Arquitetura 2

Com as melhorias avaliadas do desenvolvimento da Arquitetura 1 ocorreu a implementação da Arquitetura 2, que é a versão final do método proposto. Vale ressaltar que o trabalho menciona somente duas arquiteturas diferentes com algoritmos de aprendizagem profunda desenvolvidos, porém, no decorrer do documento é possível visualizar outros meios de implementação foram necessários como os estudos preliminares das características temporais e espectrais dos áudios, visualização das imagens geradas do mel espectrograma, entre outros pontos da fundamentação teórica deste trabalho. Com isso, na Figura 19 é apresentada a Arquitetura 2 com todas as suas etapas, para ao final ter a classificação das emoções. Na sequência cada etapa será melhor detalhada.

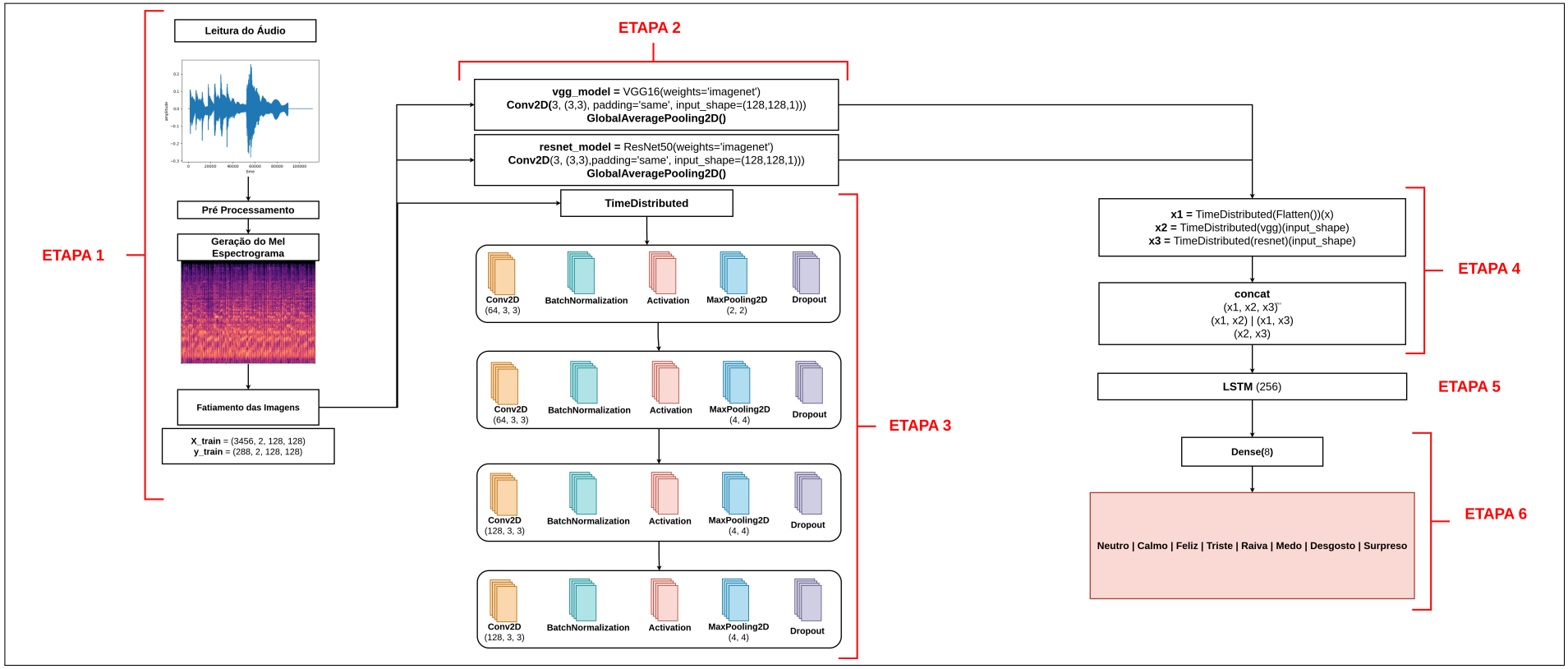


Figura 19 – Modelo da Rede - Arquitetura 2

A Etapa 1 da Arquitetura 2 se manteve muito parecida com a mesma etapa na Arquitetura 1, porém, houve uma adição de mm novo parâmetro na leitura do áudio, e, esse parâmetro foi definido com auxílio da biblioteca *librosa* que permite definir um *offset*, ou seja, o tempo em que será iniciado efetivamente a leitura do áudio. Isso permite reduzir os períodos de silêncio ou com informações não relevantes presentes no início das gravações da base de dados. Sendo assim, além das combinações de parâmetros selecionadas anteriormente, foram definidos três valores de *offset*: 0, 0,2 e 0,3 segundos. Sendo 0s indicando a leitura do áudio original. Esse parâmetro teve como objetivo acrescentar valor na rede implementada na Arquitetura 1, pois, como mencionado anteriormente na Etapa 3, a Arquitetura 1 foi desenvolvida com o objetivo de validar o método proposto porém era uma rede simples e que não agregaria valor para a área de aprendizagem profunda, que busca sempre ter novos métodos de processamento. A influência desse parâmetro será detalhada nos resultados do Capítulo 5.

A etapa de pré-processamento desenvolvida anteriormente que envolve adição de ruído se manteve igual a Arquitetura 1. Outra parte chamada *Fatiamento das Imagens* foi adicionada na Etapa 1 da Arquitetura 2. Essa parte tem como objetivo fatiar as imagens geradas do mel espectrograma pois o algoritmo LSTM foi adicionado, e, esse algoritmo funciona com o armazenamento de informações em tempo curtos, ou seja, o seu desempenho e memória são influenciados por esses pontos, sendo assim, o objetivo de fatiar as imagens geradas a partir do mel espectrograma auxiliam em utilizar pouca memória do algoritmo, ter um bom desempenho e sem perder a qualidade do processamento. As leituras das informações é realizada feita em ordem temporal, ou seja, as fatias são lidas conforme a ordem original dos áudios. Essas imagens são fatiadas em uma função desenvolvida após a geração das imagens e esse fatiamento é baseado nas seguintes equações:

$$hop\_time = hop\_length/2 \tag{4.1}$$

$$window\_time = hop\_time/2 \tag{4.2}$$

$$n\_frames = 1 + ((formato\_imagem - hop\_time)/window\_time) \tag{4.3}$$

O parâmetro *hop_length* é o parâmetro apresentado na Etapa 1 e seus valores descritos na Etapa 3, ou seja, para cada variação do seu valor ocorreu uma variação no fatiamento das imagens, e, consequentemente todas as novas imagens geradas foram armazenadas e posteriormente serviram de entrada de informação para a nova rede desenvolvida. O parâmetro *hop_time* que está associado ao parâmetro *hop_lenght* recebe o valor do mesmo dividido por 2, ou seja, conforme a variação apresentada na Tabela 10

ocorreu a variação desse parâmetro. O parâmetro *window_time* é associado ao parâmetro *hop_time* que possui a mesma estratégia de divisão sofrendo alterações conforme as suas variações, mencionado anteriormente. Com esses parâmetros definidos ocorre o cálculo para definir qual seria o tamanho das fatias que as imagens serão reprocessadas e posteriormente as mesmas são armazenadas nesse novo formato.

Após essa nova geração de imagens as mesmas são divididas em treinamento e teste, mesma estrutura da Arquitetura 1, porém, o formato do $X\_train$ e do $y\_train$ é em 4 dimensões e não em 3 como anteriormente. Essa diferença foi necessária pelo novo modelo de CNN que foi implementada na Arquitetura 2. Anteriormente o seu modelo era *Sequential* e agora o modelo é *TimeDistributed*. Como mencionado, o fatiamento das imagens ocorreu principalmente pelo fato da adição do algoritmo LSTM ao método. Mas, quando queremos trabalhar com informações que possuem uma ordem cronológica, ou seja, pedaços de vídeos, pedaços de imagem, o modelo *TimeDistributed* foi disponibilizado pela biblioteca KERAS (KERAS, 2022f) para auxiliar nesse trabalho. Conectando o formato do vetor $X\_train$ ao modelo *TimeDistributed*, o segundo parâmetro do vetor é utilizado pelo algoritmo para internamente processar a quantidade de imagens definidas, ou seja, no nosso caso a cada camada definida do algoritmo o mesmo era processar 2 imagens. O terceiro e o quarto parâmetro do vetor $X\_train$ define qual será o tamanho dessas imagens que serão processadas. Vale ressaltar que o tamanho definido nos algoritmos VGG16 e ResNet50 seguem o padrão da literatura e com isso temos um padrão de tamanho das imagens, o que não acontecia na Arquitetura 1.

Com as informações organizadas na Etapa 1, a Etapa 2 da Figura 19 existem duas definições: do algoritmo VGG16 e a ResNet50. Esses dois algoritmos foram adicionados no método pois são algoritmos pré treinados e usualmente utilizados em aplicações que utilizam processamento de imagens para classificação de dados, e, essa informação está conectada com a declaração dos dois algoritmos com a expressão: $weights =' imagenet'$, que, expressa que os pesos, ou seja, a classificação do algoritmo deve levar em consideração que o formato da entrada será imagem. Os dois utilizam uma camada de convolução pois como explicado anteriormente essa camada recebe informações e seus formatos para continuar para a classificação. Nos dois casos essa terá o tamanho 3, filtro (3x3), ou seja, quantos pedaços será processado, a expressão $padding =' same'$ significa que o algoritmo irá devolver uma saída mantendo o formato da entrada, caso seja necessário, será preenchido com zeros a esquerda ou a direita. Ao final das duas declarações a expressão $GlobalAveragePooling2D$ possui o mesmo objetivo da camada *pooling* definida na arquitetura 1, que é reduzir as informações a partir do cálculo da média entre eles e enviar essas novas informações para a próxima etapa do algoritmo.

Após essas definições da Etapa 2, a Etapa 3 se inicia com a definição do modelo *TimeDistributed* da CNN, explicado anteriormente, e, na sequência a rede inicia seu

processamento em 4 etapas com diferentes camadas como: *Conv, Batch Normalization, Activation, MaxPooling* e *Dropout*. Na figura 19 essas duas definições não possuem entrada, pois como mencionado, essa etapa é para a definição dos algoritmos, e, os mesmos possuem informações internas que não precisam ser relacionadas nesse momento da arquitetura mas serão relacionadas quando os mesmos forem aplicados na arquitetura na etapa 4. As camadas convolução, normalização e *pooling* (*Conv, Batch Normalization e MaxPooling*) possuem o mesmo objetivo da aplicação na Arquitetura 1. A camada *Activation* (KERAS, 2022d) ou ativação não é necessariamente uma camada que precisa ser explicita na arquitetura do algoritmo pois ela deve ocorrer internamente, porém, essa etapa decide se um neurônio deve ser ativado ou não a partir do cálculo da soma ponderada e adicionando viés a ela. O seu principal objetivo é introduzir não linearidade na saída de um neurônio, ou seja, as informações que percorrerem uma rede com aprendizagem profunda, ou rede neural, vão atualizando os pesos dos seus neurônios com base no erro da saída da camada anterior, chama-se isso de retro-propagação, e, essa parte só é possível de acontecer a partir da ativação. A camada *dropout* (KERAS, 2022b) evita o *overfitting* que é um comportamento indesejável da rede que ocorre quando o modelo de aprendizagem fornece a previsão de classificação de maneira precisa para dados de treinamento, ou seja de entrada, mas não para novos dados, por exemplo, os dados de teste.

Com as camadas definidas das Etapas 2 e 3, a Etapa 4 tem como objetivo reunir as informações de saída de todos os algoritmos. A primeira definição da etapa 4 que é $x1 = TimeDistributed(Flatten)$ é a aplicação de uma camada *Flatten* ou achatamento na saída do algoritmo CNN, variável $x$. A segunda definição $x2 = TimeDistributed(vgg)$ é chamar o algoritmo VGG16 para processar as imagens seguindo o formato de entrada definido na Etapa 1, e, a terceira e última definição $x3 = TimeDistributed(resnet)$ tem o mesmo objetivo mas para o algoritmo ResNet50. Com as informações organizadas em $x1, x2$ e $x3$ agora ocorre a concatenação, ou seja, combinar par a par ou os três algoritmos: $x = (x1, x2) \mid (x1, x3) \mid (x2, x3) \mid (x1, x2, x3)$, e, essa concatenação leva em consideração os pesos de cada classificador e transforma em um vetor único com as informações para prosseguir para a Etapa 5.

A Etapa 5, como apresentado anteriormente, recebe as informações organizadas da Etapa 4, porém, quem recebe essas informações é o algoritmo LSTM mencionado anteriormente nesse capítulo. Nesta etapa ele tem como objetivo receber as informações concatenadas na variável $x$ e ele recebe essas informações levando em consideração 256 parâmetros de entrada, sendo essa definição por meio da literatura do algoritmo, e a camada final vai possui a mesma largura de parâmetros.

A última etapa, Etapa 6, possui a mesma estratégia da Arquitetura 1, que é a aplicação de camada densa ou *Dense* que tem como objetivo apresentar para a rede que ela vai precisar identificar 8 classes. Nessa arquitetura as camadas densas foram reduzidas

pois como as informações foram processadas por diferentes algoritmos as mesmas já estão em um formato de conexão entre neurônios que somente uma camada densa conseguiria processar. Por fim, após todo esse processo é obtido o resultado da classificação de maneira geral e para cada emoção.

## 4.5 Considerações Finais

A implementação de duas arquiteturas trouxeram ao método proposto deste trabalho uma grande visibilidade dos objetivos e a motivação que foram definidos. A partir da versão definida como arquitetura 2 foi possível avaliar diferentes benefícios como o aumento das camadas de processamento com a aplicação dos algoritmos LSTM, VGG16 e ResNet50, que, tiveram como principal objetivo agregar ao método no âmbito computacional e também trazer o método a aplicação real do conceito de aprendizagem profunda.

Ao observar a Arquitetura 1 a mesma não envolveu complexidade no sentido de diferentes algoritmos e por fim combinar os mesmos, como na arquitetura 2. Discutindo tecnicamente, a arquitetura 2 trouxe ao autor a necessidade de entender diferentes algoritmos e suas aplicações, e também, entender como aplicar corretamente as informações de entrada dos algoritmos. Essas análises técnicas garantem que ao final do processo seja possível concluir que a combinação dos algoritmos influenciou positivamente no método proposto trazendo um resultado que agregue na área de estudos e apresentando uma abordagem diferenciada.

# 5 RESULTADOS

Este capítulo tem como objetivo apresentar os resultados obtidos pelo método proposto. Alguns resultados preliminares já foram apresentados no Capítulo 4, pois os mesmos fizeram parte dos estudos iniciais e melhorias deste trabalho, como a acurácia do primeiro modelo da rede, estudos dos parâmetros relacionados ao mel espectrograma, entre outros pontos. Conforme discutido no Capítulo 3, os resultados obtidos pelo método serão comparados com os trabalhos relacionados. A razão é que existem diferenças de metodologia, base de dados, tratamento dos dados, entre outras características, que influenciam na comparação. Portanto não seria totalmente justa uma comparação direta dos resultados. A apresentação dos resultados será feita utilizando-se a *matriz de confusão* e a *acurácia* conforme mencionado no Capítulo 2. A matriz de confusão é um modelo que tem como objetivo apresentar a qualidade da classificação da rede relacionando as classes utilizadas no processo e a acurácia apresenta isso em números conforme a equação (2.8).

## 5.1 Desempenho da Arquitetura 2

Conforme detalhado anteriormente, a execução do método proposto, denominado Arquitetura 2 no Capítulo 4, envolveu várias técnicas e ferramentas, parâmetros que são necessários para criação dos dados de entrada e que influenciam no processamento dos mesmos, e na classificação das emoções. A etapa final, que envolve a apresentação dos melhores resultados obtidos pelo método desenvolvido, está atrelada aos parâmetros de processamento do áudio, como a taxa de amostragem (*sample rate*) e ao *offset*. O parâmetro *offset* possui variações entre as tabelas de resultados, pois, é um parâmetro aplicado pela biblioteca auxiliar para iniciar a leitura do áudio, ou seja, ele garante que o áudio será lido a partir de 0,2s ou 0,3s por exemplo. Esse parâmetro foi apresentado pois influenciou nos resultados e o mesmo garante que períodos silenciosos não iriam influenciar no método. As tabelas que são apresentadas na sequência com os melhores resultados obtidos pelo método desenvolvido incluem as métricas *F1-Score*, *Precision*, *Recall* e acurácia.

Os resultados que serão apresentados nessa etapa não possuem divisão pré estabelecida, como por exemplo, dividir por gênero ou intensidade do áudio. Esses resultados serão apresentados na sequência. Os resultados são apresentados em ordem decrescente com relação a métrica de acurácia obtida para a combinação dos parâmetros, pois, essa métrica será utilizada para comparação com os trabalhos relacionados e somente o maior valor será apresentado. Na Tabela 12 é apresentado o primeiro resultado.

Visualizando os resultados da tabela 12 podemos avaliar como o modelo se comportou em relação a cada emoção classificada. Observando a métrica de *F1-score*, que é a

Tabela 12 – Resultados para $SR$ = 8kHz + *Offset* = 0,3

| **Emoção** | **Precision** | **Recall** | **F1-Score** |
|---|---|---|---|
| Raiva | 0.67 | 0.69 | 0.69 |
| Feliz | 0.77 | 0.83 | 0.79 |
| Neutro | 0.65 | 0.68 | 0.67 |
| Triste | 0.70 | 0.59 | 0.68 |
| Medo | 0.87 | 0.75 | 0.83 |
| Surpreso | 0.90 | 0.85 | 0.89 |
| Desgosto | 0.88 | 0.84 | 0.86 |
| Calmo | 0.85 | 0.98 | 0.91 |
| **Acurácia** | | | **0.79** |

SR = *Sample Rate*

média entre as duas métricas, as melhores emoções classificadas em ordem decrescente são: **Calmo**, **Surpreso**, **Desgosto**, **Medo**, **Feliz**, **Raiva**, **Triste** e **Neutro**. É importante avaliar cada emoção pois o modelo foi proposto para avaliar as 8 emoções que compõem a base RAVDESS. A observação pela métrica *F1-score* se deve ao fato da base não ser balanceada. Observadno a ordem das classificação a emoção neutro teve a menor métrica entrea as emoções, mas, esse fato tem uma relação direta com a quantidade de amostras, pois, a emoção neutro não possui variação de intensidade, e com isso, as suas amostras em relação as demais emoções são menores.

Na Tabela 13 é apresentado o terceiro resultado obtido pelo método.

Tabela 13 – Resultados para $SR$ = 8kHz + *Offset* = 0,2

| **Emoção** | **Precision** | **Recall** | **F1-Score** |
|---|---|---|---|
| Raiva | 0.65 | 0.65 | 0.65 |
| Feliz | 0.77 | 0.81 | 0.79 |
| Neutro | 0.61 | 0.68 | 0.67 |
| Triste | 0.70 | 0.55 | 0.62 |
| Medo | 0.87 | 0.75 | 0.81 |
| Surpreso | 0.90 | 0.85 | 0.87 |
| Desgosto | 0.88 | 0.84 | 0.86 |
| Calmo | 0.85 | 0.98 | 0.91 |
| **Acurácia** | | | **0.77** |

SR = *Sample Rate*

Realizando a mesma análise pela métrica *F1-score*, as emoções classificadas em ordem decrescente são: **Calmo**, **Surpreso**, **Desgosto**, **Medo**, **Feliz**, **Neutro**, **Raiva** e **Triste**. Comparando a ordem das emoções entre o primeiro e o segundo resultado é notável que o método obteve o mesmo comportamento para as 5 primeiras emoções com valores

altos da métrica, porém, as 3 últimas emoções invertidas, e com isso, a argumentação sobre a quantidade de amostras não se manteve válida aqui para a emoção neutro. Uma influência notável é a diferença dos valores de *offset*, pois, a taxa de amostragem (*sample rate*) se manteve igual entre resultados porém o primeiro resultado aplicou 0,3s nos áudios e o segundo aplicou 0,2s. Essa diferença influencia no tempo de leitura do áudio, conforme mencionado anteriormente, e com isso, é possível afirmar que essa diferença pode ter agregado na compilação de informações, ou até mesmo, ter retirando mais períodos de silêncio para os áudios da emoção de neutro e auxilinado na classificação da emoção.

Na Tabela 14 é apresentado o terceiro resultado.

Tabela 14 – Resultados para $SR$ = 16kHz + $Offset$ = 0

| **Emoção** | **Precision** | **Recall** | **F1-Score** |
|---|---|---|---|
| Raiva | 0.80 | 0.55 | 0.65 |
| Feliz | 0.91 | 0.71 | 0.80 |
| Neutro | 0.94 | 0.92 | 0.93 |
| Triste | 0.62 | 0.59 | 0.60 |
| Medo | 0.82 | 0.58 | 0.68 |
| Surpreso | 0.68 | 0.94 | 0.79 |
| Desgosto | 0.61 | 0.88 | 0.72 |
| Calmo | 0.91 | 0.90 | 0.91 |
| **Acurácia** | | | **0.76** |

SR = *Sample Rate*

Mantendo a estratégia de análise da métrica de *F1-score*, para o terceiro resultado temos a seguinte ordem de maneira decrescente das emoções são: **Neutro**, **Calmo**, **Feliz**, **Surpreso**, **Desgosto**, **Medo**, **Raiva** e **Triste**. Comparando a ordem das emoções do terceiro resultado em comparação ao primeiro resultado, nenhuma emoção se manteve na ordem, ou seja, a influência do *sample rate* maior e o *offset* em 0s, que é a leitura do áudio no seu inicio original, agregou e muito no resultado para o método ter um diferente comportamento entre as classificações. Já realizando as comparações entre o segundo e o terceiro resultado, a emoção **Triste** se manteve como a menor métrica, e com isso, trazendo um padrão ao comportamento do método que essa emoção não é muito bem classificada, trazendo futuras melhorias para o método.

Concluindo sobre os três resultados apresentados, em relação a acurácia o modelo não obteve grandes variações sobre os valores. Contudo, ampliando a análise para as métricas e para as emoções, é possível avaliar que para cada combinação de parâmetros ocorreu um comportamento diferente. Vale lembrar que além dos parâmetros de taxa de amostragem (*sample rate*) e o *offset*, também ocorreu a combinação dos parâmetros para a geração do mel espectrograma apresentados no Capítulo 4. De maneira geral o modelo atingiu o objetivo de classificar 8 emoções diferentes, com valores altos de acurácia. Para

uma análise mais detalhada do desempenho do método, na sequência é gerada a matriz de confusão para cada resultado.

### 5.1.1 Análise pelo Modelo de Matriz de Confusão

Na seção anterior apresentamos os resultados comparativos relacionados a acurácia do método. Nesta seção analisamos os resultados de forma mais detalhada, através da matriz de confusão. A matriz de confusão tem como objetivo apresentar a quantidade de informações que foram classificadas, corretamente ou não. Essa análise é realizada a partir da linha diagonal da matriz. Essa linha apresenta as combinações de dados classificados de maneira correta, ou seja, o par correto de classes na vertical e na horizontal. Os demais dados são considerados os falsos positivos, ou seja, a classe esperada são as listadas na vertical porém foram classificadas em outra classe apresentada na horizontal.

Na Tabela 15 é representada a matriz de confusão em relação ao primeiro resultado do método. A linha diagonal apresenta o resultado e desempenho do método desenvolvido, e, está em negrito para facilitar a visualização da mesma. Essa padrão será seguido em todas as tabelas.

Tabela 15 – 1º Resultado da Arquitetura 2

| | Raiva | Feliz | Neutro | Triste | Medo | Surpreso | Desgosto | Calmo |
|---|---|---|---|---|---|---|---|---|
| Raiva | **34** | 0 | 0 | 0 | 1 | 0 | 0 | 2 |
| Feliz | 1 | **25** | 2 | 6 | 0 | 3 | 4 | 5 |
| Neutro | 4 | 0 | **18** | 7 | 1 | 2 | 2 | 5 |
| Triste | 1 | 0 | 1 | **35** | 0 | 1 | 3 | 3 |
| Medo | 6 | 0 | 3 | 1 | **9** | 0 | 0 | 4 |
| Surpreso | 0 | 4 | 4 | 1 | 0 | **22** | 2 | 1 |
| Desgosto | 0 | 0 | 0 | 1 | 0 | 0 | **28** | 0 |
| Calmo | 2 | 0 | 1 | 6 | 2 | 1 | 1 | **23** |

Como explicado no Capítulo 2 a matriz de confusão relaciona vários conceitos conforme as classes no sentido de avaliar se as mesmas tiveram a sua relação entre classe real e a classe previstas. Analisando os valores da linha diagonal, é possível avaliar que a emoção triste foi a que teve mais dados relacionados corretamente, e, a emoção medo teve menos dados classificados corretamente. Relembrando que as métricas analisadas e apresentadas na Tabela 12 que são baseadas na matriz de confusão, porém, não devemos levar em consideração somente a linha diagonal pois as métricas também levam em consideração as amostras classificadas erroneamente, ou seja, o resto dos valores da linha horizontal, e, isso influencia nas equações das métricas. Na Tabela 16 é representada a matriz de confusão em relação ao segundo resultado do método.

Tabela 16 – 2º Resultado da Arquitetura 2

| | Raiva | Feliz | Neutro | Triste | Medo | Surpreso | Desgosto | Calmo |
|---|---|---|---|---|---|---|---|---|
| Raiva | **32** | 0 | 0 | 0 | 4 | 0 | 0 | 1 |
| Feliz | 2 | **31** | 0 | 2 | 1 | 4 | 3 | 3 |
| Neutro | 3 | 1 | **21** | 4 | 2 | 6 | 1 | 1 |
| Triste | 2 | 0 | 5 | **33** | 1 | 1 | 1 | 1 |
| Medo | 8 | 0 | 1 | 0 | **11** | 0 | 0 | 3 |
| Surpreso | 1 | 2 | 1 | 2 | 0 | **25** | 3 | 0 |
| Desgosto | 1 | 0 | 0 | 2 | 0 | 0 | **26** | 0 |
| Calmo | 4 | 4 | 1 | 3 | 5 | 2 | 0 | **17** |

Relembrando que as métricas analisadas e apresentadas na Tabela 13 nesse modelo a emoção triste também se manteve como a classificação mais alta e a emoção calmo que teve a classificação mais baixa. Porém, vale ressaltar que a emoções com classificação mais alta pela matriz de confusão possuem seu valor absoluto alto, ou seja, relação entre as classes, porém, em relação a taxa de acerto não é o melhor valor entre as emoções. Na Tabela 17 é representada a matriz de confusão em relação ao segundo resultado do método.

Tabela 17 – 3º Resultado da Arquitetura 2

| | Raiva | Feliz | Neutro | Triste | Medo | Surpreso | Desgosto | Calmo |
|---|---|---|---|---|---|---|---|---|
| Raiva | **34** | 0 | 0 | 0 | 1 | 0 | 0 | 2 |
| Feliz | 1 | **25** | 2 | 6 | 0 | 3 | 4 | 5 |
| Neutro | 4 | 0 | **18** | 7 | 1 | 2 | 2 | 5 |
| Triste | 1 | 0 | 1 | **35** | 0 | 1 | 3 | 3 |
| Medo | 6 | 0 | 3 | 1 | **9** | 0 | 0 | 4 |
| Surpreso | 0 | 4 | 4 | 1 | 0 | **22** | 2 | 1 |
| Desgosto | 0 | 0 | 0 | 1 | 0 | 0 | **28** | 0 |
| Calmo | 2 | 0 | 1 | 6 | 2 | 1 | 1 | **23** |

Mantendo a mesma estratégia de relembrar os resultados das métricas na Tabela 14 nesse modelo a emoção *triste* na classificação mais alta e a emoção *calmo* também se manteve na classificação mais baixa, levando em consideração o segundo resultado.

Algumas conclusões a partir da matriz de confusão são possíveis de se observar. Como a base é desbalanceada, não podemos levar em consideração os valores absolutos atingidos pelas mesmas quando apresentadas no formato de matriz de confusão, pois, essa visualização visa agregar ao método como foi seu comportamento para relacionar as classes reais com as classes previstas. Sendo assim, para uma base desbalanceada devemos focar nas análises pela métrica *F1-Score* apresentada anteriormente. Como mencionado, as métricas (*precision*) e *recall* utilizam os valores (VP, FN, FP e TN) nas suas equações (2.10) e (2.9). A métrica *F1-score*, definida na equação (2.11), é uma média das duas

métricas, ou seja, a relação matemática entre elas é direta.

## 5.2 Desempenho da Arquitetura 2 - Divisão por Intensidade

Uma abordagem apresentado por um dos trabalhos relacionados foi dividir a base de dados RAVDESS pela sua intensidade, baixa e alta, que é uma característica do áudio apresentado pela base que pode ser utilizada como cenário de teste. Sendo essa informação presente no nome do arquivo como citado no Capítulo 2. A comparação com o trabalho relacionado em questão será apresentado na seção a seguir. O resultado obtido pelo método proposto para intensidade alta é apresentado na Tabela 18. A acurácia com a divisão da intensidade alta comparado ao primeiro melhor resultado sem nenhuma divisão da base de dados, obteve um resultado bem mais elevado. O resultado obtido pelo método proposto para intensidade baixa é apresentada na Tabela 19.

Tabela 18 – Resultados para $SR$ = 8kHz + *Offset* = 0,2 + Intensidade = Alta

| **Emoção** | **Precision** | **Recall** | **F1-Score** |
|---|---|---|---|
| Raiva | 0.77 | 0.71 | 0.74 |
| Feliz | 0.79 | 0.81 | 0.79 |
| Neutro | 0.73 | 0.69 | 0.74 |
| Triste | 0.75 | 0.65 | 0.69 |
| Medo | 0.88 | 0.79 | 0.84 |
| Surpreso | 0.92 | 0.87 | 0.87 |
| Desgosto | 0.88 | 0.86 | 0.89 |
| Calmo | 0.89 | 0.98 | 0.94 |
| **Acurácia** | | | **0.83** |

SR = *Sample Rate*

Tabela 19 – Resultados para $SR$ = 8kHz + *Offset* = 0,2 + Intensidade = Baixa

| **Emoção** | **Precision** | **Recall** | **F1-Score** |
|---|---|---|---|
| Raiva | 0.77 | 0.56 | 0.65 |
| Feliz | 0.77 | 0.83 | 0.8 |
| Neutro | 0.70 | 0.80 | 0.75 |
| Triste | 0.66 | 0.73 | 0.69 |
| Medo | 0.78 | 0.73 | 0.75 |
| Surpreso | 0.82 | 0.86 | 0.84 |
| Desgosto | 0.82 | 0.71 | 0.76 |
| Calmo | 0.74 | 0.98 | 0.85 |
| **Acurácia** | | | **0.76** |

SR = *Sample Rate*

Realizando uma comparação entre os resultados das intensidades pela métrica de acurácia, relembrando que é uma comparação apenas quantitativa pelo fato da base ser desbalanceada, a intensidade alta fica com o resultado maior em relação a intensidade baixa. Comparando os resultados obtidos pela divisão de intensidade com o primeiro resultado obtido pelo método sem nenhuma divisão na base de dados, a intensidade alta possui uma acurácia maior que o primeiro resultado, e a intensidade baixa possui uma acurácia menor. Com essas diferenças entre os resultados, notamos que é relevante considerar esses resultados para o método desenvolvido.

Observando os resultados das intensidades no quesito emoção, podemos observar que as emoções *surpreso* e *calmo* se destacam nas métricas, mantendo um padrão com os resultados apresentados na base sem nenhuma divisão. Essa visão reforça a performance do método para essas emoções e com essa visibilidade é possível futuramente ocorrer melhorias para igualar ou outras emoções terem mais destaques nos resultados das métricas.

## 5.3 Comparação com Trabalhos Relacionados

Como detalhado no Capítulo 3, dos seis trabalhos que possuem relação com o tema desta pesquisa, três foram utilizados para fins de comparação com o objetivo de avaliar a efetividade do método desenvolvido, comparar se os resultados foram relevantes para a área de pesquisa, e com isso, identificar pontos de melhoria e sugestões para a evolução desta pesquisa. No final do Capítulo 3 foram feitas diversas considerações com o objetivo de avaliar todos os trabalhos relacionados e garantir que a comparação entre o presente trabalho e os trabalhos selecionados poderiam trazer melhorias e conclusões relevantes para o método desenvolvido. Reforçando as considerações do Capítulo 3, devido a divergência e falta de informação para comparação, as métricas dos trabalhos (GUPTA; CHANDRA, 2021) e (AYADI; LACHIRI, 2022a) não são apresentadas para fins de comparação. A primeira comparação foi em termos da acurácia dos trabalhos relacionados com o método proposto, conforme apresentado na Tabela 20.

Tabela 20 – Comparação da acurácia deste trabalho com os trabalhos relacionados

| Estratégia | Métrica |
|---|---|
| Método Proposto | Acurácia = 79.00 % |
| (RAJAK; MALL, 2019) | Acurácia = 76.20 % |
| (MUSTAQEEM; KWON, 2020) | Acurácia = 70.00 % |
| (SLIMI et al., 2020) | Acurácia = 77.50 % |

Ressaltando as acurácias dos trabalhos relacionados, o trabalho (RAJAK; MALL, 2019) trabalha com a emoção relacionando por quadrantes conforme o modelo de (RUSSELL; LEWICKA; NIIT, 1989), e com isso, a métrica de acurácia é na classificação dos quadrantes e não diretamente das emoções. Mas, para essa seção o resultado obtido com

as emoções da base de dados foi apresentado e será discutado na sequência. O trabalho (MUSTAQEEM; KWON, 2020) apresenta dois resultados de acurácia do seu método, *clean spectrogram* e *raw spectrogram*, e, na comparação foi apresentado o resultado do *raw spectrogram* pois é aplicado um processo semelhante ao método do presente trabalho, enquanto o *clean spectrogram* é aplicado um processo com diversas diferenças, e com isso, não seria uma comparação que possa agregar. O trabalho (SLIMI et al., 2020) possui diferentes valores de épocas para o seu algoritmo de aprendizagem, e, os valores são muito superiores ao aplicado pelo presente trabalho. Sendo assim, para a comparação conseguir agregar para o trabalho foi utilizado a métrica atingida com a menor quantidade de épocas que é 2600.

O trabalho (RAJAK; MALL, 2019) possui algumas diferenças em relação ao método proposto conforme mencionado anteriormente. Porém, o trabalho apresenta um resultado para base de dados completa, ou seja, sem divisão por gênero ou intensidade, por exemplo, utilizando algoritmos também aplicados no método desenvolvido pelo presente trabalho. Porém, o trabalho relacionado não aplicou as 8 emoções e sim 4 emoções (Felicidade, Raiva, Triste e Neutro). O método proposto atingiu 77% com quatro emoções comparados ao 76,20% do trabalho relacionado. Uma diferença pequena, porém, vale para acrescentar mais detalhes de comparação. Para essa comparação não foi aplicado o parâmetro *offset* apresentado anteriormente, e a *sample rate* se manteve a original do áudio para manter a comparação.

A comparação mais direta é realizada com as métricas do trabalho (MUSTAQEEM; KWON, 2020), que, como mencionado anteriormente, foi um trabalho com bastante relevância para a base do método desenvolvido. Na Tabela 21 é apresentado a comparação das métricas do método proposto com o trabalho relacionado. A avaliação entre os trabalhos será pela métrica *F1-score*, mesma métrica utilizada na primeira seção deste capítulo para avaliar os resultados do método proposto.

Os valores em negrito na tabela visam ressaltar os valores mais altos atingidos pelo método proposto e pelo trabalho relacionado, a fim de ressaltar quem atingiu a métrica mais alta para cada emoção classificada. Visualizando as diferenças, o método proposto é superior na maioria das emoções, excluindo a emoção triste, que, possui uma métrica bem alta no trabalho relacionado. A matriz de confusão do trabalho (AYADI; LACHIRI, 2022b) não é apresentado e o trabalho (AYADI; LACHIRI, 2022a) apresenta somente a matriz de confusão para 6 emoções, pois o trabalho utilizou um outro formato da base RAVDESS, conforme detalhado no Capítulo 3. Com isso, não é possível fazer uma comparação direta. Sendo assim, a matriz do método proposto será comparada com o trabalho (MUSTAQEEM; KWON, 2020). Na Tabela 22 é apresentada a comparação das matrizes de confusão.

Tabela 21 – Comparação das Métricas

(a) Método Proposto

| Emoção | precisão | recall | *f1-score* |
|---|---|---|---|
| Raiva | 0.65 | 0.65 | **0.65** |
| Feliz | 0.77 | 0.81 | **0.79** |
| Neutro | 0.61 | 0.68 | **0.67** |
| Triste | 0.70 | 0.55 | 0.62 |
| Medo | 0.87 | 0.75 | **0.81** |
| Surpreso | 0.90 | 0.85 | **0.87** |
| Desgosto | 0.88 | 0.84 | **0.86** |
| Calmo | 0.85 | 0.98 | **0.91** |

(b) (MUSTAQEEM; KWON, 2020)

| Emoção | precisão | recall | *f1-score* |
|---|---|---|---|
| Raiva | 0.40 | 1.00 | 0.57 |
| Feliz | 0.92 | 0.29 | 0.44 |
| Neutro | 0.91 | 0.42 | 0.57 |
| Triste | 0.98 | 0.98 | **0.98** |
| Medo | 0 | 0 | 0 |
| Surpreso | 0.90 | 0.46 | 0.61 |
| Desgosto | 0.92 | 0.86 | 0.89 |
| Calmo | 0.82 | 0.75 | 0.78 |

Um análise visível entre as duas matrizes de confusão é a diferença de valores na linha diagonal. Como mencionado anteriormente, a linha diagonal apresenta a relação positiva entre a classe real e a classe preditiva. O modelo do trabalho (MUSTAQEEM; KWON, 2020) não apresenta a quantidade de informações que foi dividida em treinamento e teste. Com isso, quando a matriz de confusão é analisada cada linha possui a sua quantidade de dados classificadas, e, essa informação de teste está relacionada com a matriz de confusão na coluna **Dados**. Ao somar a coluna de dados é possível alcançar o valor 288. Cada linha do trabalho (MUSTAQEEM; KWON, 2020) possui 100 dados e ao total eles utilizaram 800 dados de teste. Sem ter o conhecimento da quantidade de amostras aplicada ao total do método é difícil analisar se essa maior quantidade de amostras poderia influenciar positivamente na acurácia. Ressaltando que essa informação foi buscada por diferentes meios de comunicação, e não somente pela leitura do trabalho, porém, sem sucesso.

Em (SLIMI et al., 2020), o autor do trabalho apresenta a divisão do banco de dados de áudio com base na característica de intensidade de áudio (baixa ou alta) para a análise de desempenho. Também realizamos a comparação com este método e os resultados são apresentados na Tabela 23.

Tabela 22 – Comparação das Matrizes de Confusão

(a) Metodo Proposto

| | Raiva | Feliz | Neutro | Triste | Medo | Surpreso | Desgosto | Calmo | Dados |
|---|---|---|---|---|---|---|---|---|---|
| Raiva | **32** | 0 | 0 | 0 | 4 | 0 | 0 | 1 | 37 |
| Feliz | 2 | **31** | 0 | 2 | 1 | 4 | 3 | 3 | 46 |
| Neutro | 3 | 1 | **21** | 4 | 2 | 6 | 1 | 1 | 39 |
| Triste | 2 | 0 | 5 | **33** | 1 | 1 | 1 | 1 | 44 |
| Medo | 8 | 0 | 1 | 0 | **11** | 0 | 0 | 3 | 23 |
| Surpreso | 1 | 2 | 1 | 2 | 0 | **25** | 3 | 0 | 34 |
| Desgosto | 1 | 0 | 0 | 2 | 0 | 0 | **26** | 0 | 29 |
| Calmo | 4 | 4 | 1 | 3 | 5 | 2 | 0 | **17** | 36 |

(b) (MUSTAQEEM; KWON, 2020)

| | Raiva | Feliz | Neutro | Triste | Medo | Surpreso | Desgosto | Calmo | Dados |
|---|---|---|---|---|---|---|---|---|---|
| Raiva | **82** | 15 | 0 | 0 | 0 | 3 | 0 | 0 | 100 |
| Feliz | 4 | **87** | 0 | 0 | 2 | 4 | 0 | 2 | 100 |
| Neutro | 0 | 0 | **95** | 0 | 0 | 3 | 0 | 3 | 100 |
| Triste | 0 | 2 | 0 | **94** | 0 | 0 | 0 | 4 | 100 |
| Medo | 21 | 7 | 11 | 0 | **43** | 0 | 0 | 14 | 100 |
| Surpreso | 0 | 0 | 0 | 0 | 0 | **98** | 0 | 2 | 100 |
| Desgosto | 18 | 0 | 5 | 0 | 0 | 9 | **52** | 16 | 100 |
| Calmo | 0 | 0 | 15 | 0 | 0 | 0 | 0 | **85** | 100 |

Tabela 23 – Comparação com o Trabalho Relacionado (SLIMI et al., 2020)

| Referência | Intensidade do Áudio | Épocas | Acurácia |
|---|---|---|---|
| Método Proposto | Baixa | 150 | 76.00% |
| (SLIMI et al., 2020) | Baixa | 2900 | 68.23% |
| Método Proposto | Alta | 150 | 83.00% |
| (SLIMI et al., 2020) | Alta | 5000 | 86.31% |

Com a Tabela 23 é possível observar que o método proposto alcançou uma melhor acurácia com muito menos épocas usando as amostras de áudio de baixa intensidade. Vale ressaltar que são arquiteturas diferentes. No entanto, o número de épocas é um ponto importante para questões computacionais. Para amostras de áudio de alta intensidade, nossa estratégia proposta alcançou um valor de acurácia interessante, porém, não superior ao do trabalho relacionado, mas, vale ressaltar que também seguir a abordagem de utilizar menos iterações. O valor das épocas foram variados para fins de comparação, porém, esse foi o melhor resultado alcançado pelo método proposto.

## 5.4 Considerações Finais

Em vista dos resultados apresentados, o método desenvolvido aplicou a combinação de diferentes classificadores como VGG16 e ResNet50, adicionou o LSTM para obter mais precisão no processamento dos dados, ter mais camadas para armazenar e processar os dados para classificação e, consequentemente, agregar na acurácia do método. Ao realizar a análise em cima dos cenários de aplicação do método na base RAVDESS, a acurácia obtida pelo método para toda a base foi 2% maior que os trabalhos relacionados que aplicaram essa base de dados sem nenhuma divisão de informação. Porém, ao analisar o comportamento do método desenvolvido por este trabalho e o comportamento do trabalho relacionado que aplicou a divisão da base RAVDESS por intensidade, alta e baixa, o presente trabalho atingiu uma acurácia 8% maior na intensidade baixa em relação ao trabalho relacionado, ou seja, uma diferença considerável. Já para a intensidade alta o método proposto atingiu uma acurácia 3% menor que o trabalho relacionado, porém, essa divisão de intensidade e mostrar diferentes cenários de validação do método agregam para desmontrar seu comportamento na classificação das emoções.

Analisando o método proposto por este trabalho em relação aos trabalhos relacionados, o método proposto obteve uma acurácia 4% maior que o trabalho (RAJAK; MALL, 2019), comparando com o trabalho (MUSTAQEEM; KWON, 2020) a acurácia do método proposto foi 13% maior, e, em relação ao trabalho (SLIMI et al., 2020) a acurácia foi 2% maior. Essas diferenças tem como objetivo agregar ao método sobre o seu comportamento e valor a área de estudo. Visualizando os resultados por cenários de validações, como, a divisão por intensidade baixa e alta, que, é possível a partir das informações da base RAVDESS, o método proposto teve sua acurácia 11% maior que o trabalho (SLIMI et al., 2020) para a intensidade baixa, e em contra partida, obteve uma acurácia 4% menor para a intensidade alta. Isso significa que o cenário de intensidade alta deve ser explorada pelo autor, porém, não afeta no valor agregado alcançado pelo método já que o mesmo se comporta em vários cenários de maneira melhor que os trabalhos relacionados.

Revisitando as comparações com os trabalhos relacionados de maneira geral, a comparação com o trabalho (MUSTAQEEM; KWON, 2020) agregou muito ao todo o processo de desenvolvimento técnico e teórico, pois, o mesmo possui uma abordagem bem parecida e, consequentemente, cada um com suas diferenças na arquitetura, aplicação da base de dados, característica do áudio e sua representação, e com isso, facilitou a validar e a fundamentar todos os pontos que foram implementados.

# 6 CONCLUSÃO

Após a apresentação de todos os conceitos, etapas, ferramentas e resultados do método proposto por este trabalho, este capítulo tem como objetivo revisitar todos os objetivos, questões de pesquisas e outros pontos que podem ser mencionados para concluir este trabalho, e, com isso, analisar a necessidade de melhorias, ou, até mesmo, ampliar a visão para os próximos passos envolvendo o tópico desenvolvido por este trabalho.

Iniciando a discussão pelas questões de pesquisas, a primeira questão envolvia a análise se o pré-processamento dos sinais de áudio tinham influencia nos resultados do método, e, foi possível visualizar tanto nos trabalhos relacionados e no método proposto por este trabalho, que analisou diferentes parâmetros e etapas de pré processamento, é possível concluir que existe sim influencia nos resultados, pois, a diferença entre os resultados possuíram influências dessa questão.

Uma outra questão envolvia a detecção das emoções em formato de imagem, e, ao comparar com os trabalhos relacionados é possível observar que ao utilizar diferentes características como espectrograma e o mel espectrograma os resultados obtidos são diferentes, e também, são relevantes para a área de pesquisa, já que cada característica se comporta de uma maneira diferente para cada emoção.

Uma das questões envolvia a utilização de técnica de aprendizagem profunda e, a partir dos resultados obtidos pelo método proposto comparados ao resultados dos trabalhos relacionados, é possível concluir que os resultados são melhores analisado por diferentes perspectivas, porém, nada impede a realização de comparações com métodos de aprendizagem de máquina, por exemplo.

Abrangendo a questão anterior sobre aprendizagem profunda,uma questão que podemos levantar é a combinação dessa técnica com a aprendizagem profunda com rede neural, pois, são duas técnicas diferentes e combinadas por este trabalho. Sendo assim, se faz necessário avaliar se os resultados obtidos foram relevantes, e então, é possível concluir a partir do método desenvolvido com a combinação de quatro algoritmos diferentes aplicados envolvendo a combinação de parâmetros, os resultados obtidos são relevantes e interessantes para a área de estudo.

Outra análise para a conclusão deste trabalho é a partir das comparações realizadas no capítulo de resultados. Essas comparações esclareceram diferentes pontos sobre o método proposto, e, o primeiro deles que a ideia inicial de utilizar todas as emoções da base de dados escolhida foi bem desenvolvida e os resultados são relevantes e interessantes, porém, quando comparado com um dos trabalhos relacionados, das oito emoções alvo seis delas possuem sua classificação mais efetiva pelo método desenvolvido deste trabalho. Um

valor considerável de emoções classificadas quando comparado aos trabalhos relacionados apresentados nesse trabalho. A partir disso, uma conclusão é que futuramente o método pode ser revisado, de maneira teórica e prática, para analisar diferentes melhorias nas etapas e com isso chegar na classificação mais eficiente de todas as emoções, de maneira isolada e comparativa. Outra conclusão é em relação a base de dados, os trabalhos relacionados apresentam uma ou mais bases de dados que foram utilizadas para o desenvolvimento dos mesmos, e sendo assim, um dos futuros passos é adicionar outra bases no método, verificar seu comportamento, e provavelmente desenvolver melhorias para classificar as emoções em duas bases diferentes.

Identificamos alguns pontos importantes que podem ser explorados para pesquisas futuras:

- Agregar ao método mais bases de dados para avaliar o seu comportamento.
- Avaliar melhorias técnicas para o método obter uma acurácia média melhor para as emoções.
- Agregar as etapas de pré processamento com mais características e representações, podendo ser combinações ou não, por exemplo.

# REFERÊNCIAS

AGNIHOTRI, PIYUSH. *EmoDB Dataset*. 2020. Disponível em: <https://www.kaggle.com/datasets/piyushagni5/berlin-database-of-emotional-speech-emodb>. Citado na página 27.

AYADI, Souha; LACHIRI, Zied. A combined cnn-lstm network for audio emotion recognition using speech and song attributs. In: *2022 6th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)*. [S.l.: s.n.], 2022. p. 1–6. Citado 5 vezes nas páginas 36, 37, 38, 60 e 61.

AYADI, Souha; LACHIRI, Zied. Deep neural network for visual emotion recognition based on resnet50 using song-speech characteristics. In: *2022 5th International Conference on Advanced Systems and Emergent Technologies (IC_ASET)*. [S.l.: s.n.], 2022. p. 363–368. Citado na página 61.

BADSHAH, Abdul Malik; AHMAD, Jamil; RAHIM, Nasir; BAIK, Sung Wook. Speech emotion recognition from spectrograms with deep convolutional neural network. In: *2017 International Conference on Platform Technology and Service (PlatCon)*. [S.l.: s.n.], 2017. p. 1–5. Citado na página 18.

DARWIN, Charles. *The Expression Of The Emotions In Man And Animals*. 1872. Disponível em: <http://darwin-online.org.uk/content/frameset?pageseq=1&itemID=F1142&viewtype=text>. Citado na página 13.

DINIZ, Eduardo A. B. Da Silva e Sergio L. Netto Paulo S. R. *Processamento Digital de Sinais: Projeto e Análise de Sistemas*. [S.l.]: Grupo A, 2014. ISBN 9788582601235. Citado na página 21.

EDU, Princeton. *FFT and Spectrogram*. 2021. Disponível em: <https://www.princeton.edu/~cuff/ele201/files/spectrogram.pdf>. Citado 3 vezes nas páginas 8, 20 e 22.

EKMAN, Paul. Universals and cultural differences in facial expressions of emotion. In: UNIVERSITY OF NEBRASKA PRESS. *Nebraska symposium on motivation*. [S.l.], 1971. Citado 2 vezes nas páginas 13 e 14.

ELECTRICAL, CS; ELECTRONICS. *Difference Between CNN And RNN Architecture In Deep Learning*. 2022. Disponível em: <https://cselectricalandelectronics.com/difference-between-cnn-and-rnn-architecture-in-deep-learning/>. Citado 2 vezes nas páginas 8 e 31.

FERRARI, Leandro Nunes De Castro Silva e Daniel Gomes. *Introdução a mineração de dados*. [S.l.]: Saraiva, 2017. ISBN B076C18GM4. Citado na página 33.

GUPTA, Manas; CHANDRA, Satish. Speech emotion recognition using mfcc and wide residual network. In: *2021 Thirteenth International Conference on Contemporary Computing (IC3-2021)*. [S.l.]: Association for Computing Machinery, 2021. p. 320–327. ISBN 9781450389204. Citado 4 vezes nas páginas 36, 37, 38 e 60.

JACKSON, Philip; HAQ, Sanaul. *Surrey Audio-Visual Expressed Emotion (SAVEE) Database*. 2015. Disponível em: <http://kahlan.eps.surrey.ac.uk/savee/Database.html>. Citado na página 27.

KABAL, Peter. *Audio File Format Specifications*. 2022. Disponível em: <https://www.mmsp.ece.mcgill.ca/Documents/AudioFormats/WAVE/WAVE.html>. Citado na página 17.

KERAS. *Dense layer*. 2022. Disponível em: <https://keras.io/api/layers/core_layers/dense/>. Citado na página 47.

KERAS. *Dropout layer*. 2022. Disponível em: <https://keras.io/api/layers/regularization_layers/dropout/>. Citado na página 52.

KERAS. *Flatten layer*. 2022. Disponível em: <https://keras.io/api/layers/reshaping_layers/flatten/>. Citado na página 47.

KERAS. *Layer activation functions*. 2022. Disponível em: <https://keras.io/api/layers/activations/>. Citado na página 52.

KERAS. *The Sequential model*. 2022. Disponível em: <https://keras.io/guides/sequential_model/>. Citado na página 46.

KERAS. *TimeDistributed layer*. 2022. Disponível em: <https://keras.io/api/layers/recurrent_layers/time_distributed/>. Citado na página 51.

LIBROSA. *librosa.feature.melspectrogram*. 2022. Disponível em: <https://librosa.org/doc/main/generated/librosa.feature.melspectrogram.html>. Citado 3 vezes nas páginas 9, 40 e 41.

LIBROSA. *librosa.feature.mfcc*. 2022. Disponível em: <https://librosa.org/doc/main/generated/librosa.feature.mfcc.html>. Citado 2 vezes nas páginas 8 e 18.

LIBROSA. *librosa.feature.rms*. 2022. Disponível em: <https://librosa.org/doc/main/generated/librosa.feature.rms.html>. Citado 2 vezes nas páginas 8 e 19.

LIVINGSTONE, Steven R.; RUSSO, Frank A. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PLOS ONE*, Public Library of Science, v. 13, n. 5, p. 1–35, 05 2018. Disponível em: <https://doi.org/10.1371/journal.pone.0196391>. Citado 4 vezes nas páginas 8, 14, 27 e 29.

MADHAVAN, Samaya; JONES, M. Tim. *Deep learning architectures: The rise of artificial intelligence*. 2021. Disponível em: <https://developer.ibm.com/articles/cc-machine-learning-deep-learning-architectures/>. Citado 3 vezes nas páginas 8, 30 e 31.

MUSTAQEEM; KWON, Soonil. A cnn-assisted enhanced audio signal processing for speech emotion recognition. *Sensors*, v. 20, n. 1, 2020. ISSN 1424-8220. Citado 8 vezes nas páginas 36, 37, 38, 60, 61, 62, 63 e 64.

NG, SAMUEL SAMSUDIN. *IEMOCAP Emotion Speech Database*. 2020. Disponível em: <https://www.kaggle.com/datasets/samuelsamsudinng/iemocap-emotion-speech-database>. Citado na página 27.

NOGARE, Diego. *Performance de Machine Learning – Matriz de Confusão*. 2020. Disponível em: <https://diegonogare.net/2020/04/performance-de-machine-learning-matriz-de-confusao/>. Citado 2 vezes nas páginas 8 e 34.

O'SHAUGHNESSY, Douglas. *Speech Communications: Human and Machine*. [S.l.]: Wiley-IEEE Press, 1987. ISBN 9780780334496. Citado na página 24.

RAJAK, Rohan; MALL, Rajib. Emotion recognition from audio, dimensional and discrete categorization using cnns. In: *TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON)*. [S.l.: s.n.], 2019. p. 301–305. Citado 6 vezes nas páginas 36, 37, 38, 60, 61 e 64.

RANJAN, Sandeep; SOOD, Sumesh; VERMA, Vikas. Twitter sentiment analysis of real-time customer experience feedback for predicting growth of indian telecom companies. In: *2018 4th International Conference on Computing Sciences (ICCS)*. [S.l.: s.n.], 2018. p. 166–174. Citado na página 13.

RUSSELL, James; LEWICKA, Maria; NIIT, Toomas. A cross-cultural study of a circumplex model of affect. *Journal of Personality and Social Psychology*, v. 57, p. 848–856, 11 1989. Citado 2 vezes nas páginas 14 e 60.

RUSSELL, James; WEISS, Anna; MENDELSOHN, G. Affect grid: A single-item scale of pleasure and arousal. *Journal of Personality and Social Psychology*, v. 57, p. 493–502, 09 1989. Citado 2 vezes nas páginas 8 e 36.

SLIMI, Anwer; HAMROUN, Mohamed; ZRIGUI, Mounir; NICOLAS, Henri. Emotion recognition from speech using spectrograms and shallow neural networks. In: *Proceedings of the 18th International Conference on Advances in Mobile Computing Multimedia*. [S.l.]: Association for Computing Machinery, 2020. p. 35–39. ISBN 9781450389242. Citado 9 vezes nas páginas 9, 36, 37, 38, 60, 61, 62, 63 e 64.

STEVENS, J. Volkmann S. S.; NEWMAN, E. B. A scale for the measurement of the psychological magnitude pitch. *The Journal of the Acoustical Society of America*, v. 8, n. 185, p. 38, August 1934. Citado 2 vezes nas páginas 18 e 24.

TRENTIN, Bruno Zucuni Prina e Romario. Gmc: Geração de matriz de confusão a partir de uma classificação digital de imagem do arcgis. In: *Anais XVII Simpósio Brasileiro de Sensoriamento Remoto - SBSR*. [S.l.: s.n.], 2015. Citado na página 33.

ZHANG, Shuai Wang Lei; LIU, Bing. Deep learning for sentiment analysis: A survey. *National Science Foundation (NSF), Huawei Technologies Co. Ltd.*, n. 2017, p. 2017, 2017. Citado na página 29.