UNIVERSIDADE FEDERAL DO TOCANTINS
CÂMPUS UNIVERSITÁRIO DE PALMAS
CURSO DE CIÊNCIA DA COMPUTAÇÃO

RECONHECIMENTO AUTOMÁTICO DE EMOÇÕES PELO DISCURSO

MATHEUS ALMEIDA FARIAS DA SILVA

PALMAS (TO)
2020

MATHEUS ALMEIDA FARIAS DA SILVA

RECONHECIMENTO AUTOMÁTICO DE EMOÇÕES PELO DISCURSO

Trabalho de Conclusão de Curso II apresentado à Universidade Federal do Tocantins para obtenção do título de Bacharel em Ciência da Computação, sob a orientação do(a) Prof.(a) Dr. Rafael Lima de Carvalho.

Orientador: Dr. Rafael Lima de Carvalho

PALMAS (TO)

2020

MATHEUS ALMEIDA FARIAS DA SILVA

RECONHECIMENTO AUTOMÁTICO DE EMOÇÕES PELO DISCURSO

Trabalho de Conclusão de Curso II apresentado à UFT – Universidade Federal do Tocantins – Câmpus Universitário de Palmas, Curso de Ciência da Computação foi avaliado para a obtenção do título de Bacharel e aprovada em sua forma final pelo Orientador e pela Banca Examinadora.

Data de aprovação: 11 / 12 / 2020

Banca Examinadora:

Prof. Dr. Nome do Primeiro Examinador Sobrenome

Profa. Dra. Nome do Segundo Examinador Sobrenome

Profa. Ma. Nome do Terceiro Examinador Sobrenome

Dados Internacionais de Catalogação na Publicação (CIP)
Sistema de Bibliotecas da Universidade Federal do Tocantins

S586r Silva, Matheus Almeida Farias da .

Reconhecimento automático de emoções pelo discurso. / Matheus Almeida Farias da Silva. – Palmas, TO, 2020.

54 f.

Monografia Graduação - Universidade Federal do Tocantins – Câmpus Universitário de Palmas - Curso de Ciências da Computação, 2020.

Orientador: Rafael Lima de Carvalho

1. Reconhecimento de emoções pelo discurso. 2. Teorias das emoções. 3. Aprendizado de máquina. 4. Processamento de voz. I. Título

CDD 004

Elaborado pelo sistema de geração automática de ficha catalográfica da UFT com os dados fornecidos pelo(a) autor(a).

A alguém cujo valor é digno

desta dedicatória.

AGRADECIMENTOS

Gostaria de agradecer a todos.

RESUMO

Apesar de ser um ramo subjetivo de estudo, a análise de emoções pode ser simulada por algoritmos de Aprendizado de Máquina que são treinados para que, através de bases de dados de áudio e vídeo catalogados, reconheçam padrões nessas mídias que possam estar relacionados a sua emoção. Algorítimos de Redes Neurais (ramo do Aprendizado de Máquina) vêm sendo desenvolvidos para atuar no reconhecimento dessas emoções, com o foco apenas no áudio, e são conhecidos como *Speech Emotion Recognition* (SER), geralmente obtendo médias desiguais de resultados referentes as reconhecimento de emoções entre as bases. Esse trabalho de pesquisa tem como foco implementar um algoritmo SER, baseado em trabalhos atuais que obtiveram bons desempenhos, aplicando um método de *Data Augmentation* chamado de Janela Deslizante, proposto por este trabalho, que tem como objetivo aumentar as médias de reconhecimento nas bases de dados selecionadas. Como resultado, este trabalho conseguiu alcançar uma evolução de reconhecimento de 11.95% na base EMO-DB, 22.76% na SAVEE e 18.82% na RAVEDESS.

Palavra-chave: Reconhecimento de emoção pelo Discurso. Teorias das emoções. Aprendizado de máquina. Processamento de voz.

ABSTRACT

Despite of being a subjective branch of study, emotion analysis can be simulated by Machine Learning algorithms that are trained so that, through cataloged audio and video databases, they can recognize patterns in these media that could be related to their emotion . Neural Network Algorithms (branch of Machine Learning) provided to act on the recognition of these emotions, with a focus only on audio, and are known as Speech Emotion Recognition (SER), generally obtaining unequal averages of referring results such as recognition of emotions between the bases. This research work focuses on implementing an SER algorithm, based on current works that have achieved good performances, applying a Data Augmentation method called Slide Window, proposed by this work, which aims to increase the averages of recognition in selected databases. As a result, this work managed to achieve an evolution of recognition of 11.95% on the EMO-DB base, 22.76% on SAVEE and 18.82% on RAVEDESS.

Keywords: Speech Emotion Recognition. Theories of emotions. Machine Learning. Voice processing.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

O reconhecimento de emoções pelo discurso, tradução para *Speech Emotion Recognition* (SER), consiste em simular o processo de percepção da emoção das pessoas usando uma máquina para extrair as informações nela contida Jiang Hongliang Fu (2019). Neste contexto, uma das áreas de investigação consiste em propor algoritmos de extração de atributos da voz de indivíduos, tais como características prosódicas, espectrais e de qualidade de voz Xu (2018). Ao analisar essas características, o próximo passo para a automatização é encontrar um modelo capaz de identificar e separar padrões na dimensão destes atributos.

Ordeiramente, as tecnologias de reconhecimento das emoções pela fala evoluíram o suficiente para serem aplicadas em vários cenários da vida real, como *call centers*, diagnóstico auxiliar de doenças, educação remota, segurança na condução e jogos de computador Jing Xia Mao (2018). Na área da robótica, protótipos que utilizam a SER como forma de melhor reconhecer os sentimentos podem ser muito úteis no diálogo e assistência para pessoas mais velhas ou com algum tipo de deficiência Imani (2019).

As características extraídas da fala, como já citadas anteriormente, carregam uma riqueza de dados nos quais podem ser naturalmente relacionados. Como forma de fazer essa análise, um ramo do Aprendizado de Máquina chamado Aprendizado Profundo, se faz proficiente. O aprendizado profundo representa uma série de estruturas de Redes Neurais artificiais multicamadas com valores de peso sintonizáveis aprendidos com os dados de treinamento subjacentes Hanab Ying Lia (2019).

Trabalhos atuais relacionados ao SER, como o Zhao, Mao e Chen (2019), o Jiang Hongliang Fu (2019), ou o Lee et al. (2019), arquitetaram ótimas estruturas de Redes Neurais com os métodos mais utilizados que compõem o estado da arte, atingindo excelentes resultados no treinamento. Entretanto, mesmo com tantas variações nos modelos, os algoritmos evoluem lentamente a acurácia quando testados com dados não incluídos nos treinamentos, ou seja, o algoritmo reconhece muito bem os dados de treinamento mas não com a mesma precisão os de teste, essa ocorrência é chamada de *Overfitting*. Um dos motivos nos quais podem explicar esse comportamento, é pela baixa quantidade e falta de variação nos dados das bases de dados utilizadas. Esse percalço pode ser revertido ao aplicar técnicas chamadas de *Data Augmentation* no intuito de gerar mais dados a partir dos já existentes.

Diante do exposto, percebe-se que a detecção automática de emoções é um importante problema e que existem soluções compondo o estado da arte com bons resultados, mas que ainda precisam evoluir em alguns aspectos. Dessa maneira, este trabalho teve como objetivo construir uma Rede Neural baseada no Trabalho Zhao, Mao e Chen (2019)

e demonstrar a efetividade de um método de *Data Augmentation* chamado de Janela Deslizante, proposto por esse trabalho, nas bases de dados escolhidas.

O método de Janela Deslizante consiste na criação de novos arquivos de áudios a partir de arquivos originais de cada base. Uma Janela de 2 segundos é aplicada ao áudio, gerando um áudio novo com esse tempo em segundos. A Janela deslizará até que chegue ao fim do áudio, gerando no mínimo dois novos áudios, ou seja, quando maior o áudio, maior será a quantidade de novos áudios gerados.

O algoritmo base para a execução do SER, foi escolhido por ter obtido bons resultados em suas execuções. Ao final do trabalho é possível ter conhecimento sobre o impacto que o método proposto gera.

1.1 Motivação

Uma das motivações que tornaram a concepção desse trabalho relevante, é o fato de que o SER, apesar de ter sido bastante explorado ultimamente, não atingiu ainda um nível de acurácia satisfatório o suficiente para que possa ser usada com asseveração. A maior parte dos bancos de dados utilizados para esse ramo da pesquisa, como o RAVDESS, EMO-DB e SAVEE, lidam com até 8 emoções, tendo algumas em comum entre eles, que são elas: Raiva, Nojo, Medo, Felicidade, Tristeza e Neutra. Alcançar um notável reconhecimento nessas emoções por meio pela fala, poderá ser bastante proveitoso em tarefas do cotidiano.

Uma aplicabilidade do SER que pode trazer benefícios, é integra-lo aos alto-falantes inteligentes (*Smart Speakers*), como por exemplo o Echo da Amazon ou o Google Home, do Google. Eles permitem que os usuários consigam usar comandos de voz para interagir com os assistentes pessoais da empresa fornecedora, podendo inclusive gerenciar todos os dispositivos conectados à internet dentro da casa em que está instalado. Segundo Haeb-Umbach et al. (2019), o sucesso desses auto-falantes pode ser atribuído aos avanços em todas as tecnologias constituintes que incluem um assistente digital, processamento do sinal digital envolvido na captação da voz do utilizador, dentre outros. Se aplicado à essa tecnologia, o SER pode ser útil para analisar o estado emocional do usuário, possibilitando uma mudança de comportamento do assistente, tal qual usar palavras reconfortantes ou sugerir uma música para relaxar.

1.2 Objetivo Geral

Este trabalho visou comparar os resultados obtidos nas bases de dados EMO-DB, RAVDESS e SAVEE gerados por um algoritmo de Rede Neural baseado na implementação produzida no Trabalho Zhao, Mao e Chen (2019), tendo a aplicação de um processo de *Data Augmentation* chamado Janela Deslizante. Ao final deste trabalho é possível

visualizar as diferenças entre as precisões medidas em diferentes funções de avaliações, em que evidenciam a desigualdade de resultados gerados com e sem a Janela Deslizante.

1.3 Objetivos Específicos

Os objetivos específicos deste trabalho tem o propósito de expor o caminho seguido para a construção de um algoritmo SER de alto desempenho, que utilize os dados provenientes do processo de Janela Deslizante. São eles:

- Selecionar trabalhos mais recentes da área do reconhecimento de emoções pelo discurso;
- Analisar a eficácia das técnicas de extração das características do áudio e classificadores utilizados nos trabalhos recentes;
- Elencar bases de dados contendo clipes de áudio utilizados em trabalhos relacionados;
- Estudar técnicas de classificação de padrões tais como redes com aprendizado profundo observando a aplicabilidade no problema considerado;
- Implementar um algoritmo de Rede Neural seguindo a base do que foi desenvolvido no Trabalho Zhao, Mao e Chen (2019);
- Aplicar o processo de Janela Deslizante nos áudios das bases selecionadas;
- Gerar os resultados de treinamento das bases de dados selecionadas com e sem o processo de Janela deslizante;
- Reportar o desempenho da implementação através de métricas de avaliação, comparando a diferença de resultados com abordagem que este trabalho propõe e com trabalhos relacionados;

1.4 Estrutura da monografia

Este trabalho foi divido em seis capítulos. O Capítulo 2 aborda conceitos essenciais, terminologias e classificadores de forma a subsidiar a compreensão do trabalho. No Capítulo 3, tem a descrição de alguns trabalhos relacionados, com modelos de implementações mais atuais na área do SER. O Capítulo 4 tem objetivo de descrever quais são os métodos e parâmetros de construção do algoritmo implementado, assim como descreve as etapas seguidas para tal. Dessa forma, o Capítulo 5 tem o propósito de apresentar os resultados obtidos conforme a implementação deste trabalho. E finalmente, no Capítulo

6 são feitas as considerações finais sobre o que a abordagem desse trabalho conseguiu resultar e como ela pode ser importante para futuros trabalhos na área do reconhecimento automático de emoções pelo discurso.

2 REFERENCIAL TEÓRICO

No presente capítulo, são expostos conceitos essenciais para a compreensão do reconhecimento emocional pela voz, bem como, métodos de classificação do problema, as características do arquivo de áudio usualmente manipuladas, e, uma abordagem fundamental sobre as emoções e como se constituem.

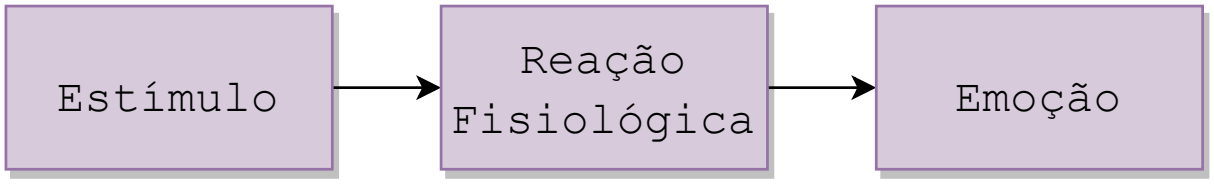
2.1 Emoções

As emoções descrevem estados fisiológicos e são geradas de forma subconsciente. Normalmente, são respostas corporais autônomas a certos eventos externos ou internos. Sabendo disso, a análise sobre como esses estados fisiológicos foram gerados fica bastante subjetiva, tendo algumas teorias de como as emoções são propriamente geradas. Nessa seção são citadas algumas teorias sobre as emoções, assim como a descrição do que pode ser chamado de emoções básicas.

2.1.1 Teorias das emoções

Desde o começo do século passado, vários estudos da área da psicologia objetivaram explicar o funcionamento das emoções. Uma das teorias mais famosas, é a *James-Lang Theory of emotion* JAMES (1884), escrita pelo renomado psicólogo William James. Essa teoria sugere que quando se vê um estímulo externo, que sua reação fisiológica acontecerá.

Figura 1 – Representação do funcionamento da *James-Lang Theory of Emotion*



Fonte: Autoria própria

A sua reação emocional depende da forma como interpreta essas reações físicas. Por exemplo, se uma pessoa ver um animal selvagem perigoso na floresta, ela vai começar a tremer e as batidas do coração acelerarão. A teoria de James-Lange propõe que é possível interpretar as suas reações físicas e concluir que está assustado ("Estou tremendo, por isso, tenho medo"). De acordo com esta teoria da emoção, você não está tremendo porque está assustado. Em vez disso, sente-se assustado porque está a tremer.

Outra teoria bem famosa, é a *Evolutionary Theory of Emotion*. Proposta por Charles Darwin, em 1872, é explorada até os dias atuais, ela descreve que as emoções existem porque servem como um papel adaptativo. Elas motivam as pessoas a responder rapidamente aos estímulos do ambiente, o que ajuda a melhorar as perspectiva de sucesso

e sobrevivência. Em Al-Shawaf e Lewis (2017) os autores trabalham nessa mesma perspectiva da psicológica evolutiva, e sugerem que as emoções coordenam mecanismos cuja função evolutiva é sistematizar uma variedade de programas na mente e no corpo a serviço da resolução de um problema adaptativo específico. Por exemplo, o medo coordena programas ao serviço da prevenção ou da fuga ao perigo, o desgosto regula mecanismos de prevenção de infecções.

Ter conhecimento das diferentes teorias sobre o funcionamento das emoções no ser humano, ajuda a entender o porque as emoções são tão importantes no dia a dia, sendo relacionada até no processo evolutivo.

2.1.2 Emoções básicas

Saber todas as variações das emoções é algo praticamente impossível, já que elas estão relacionadas não só à cada pessoa, mas também à cultura na qual faz parte. Por isso, estudos da psicologia relacionados à essa área, tendem a trabalhar com emoções básicas. Talvez a razão mais comum para propor emoções básicas seja fornecer uma explicação de algumas observações de rotina sobre as emoções Ortony e Turner (1990). Estas observações incluem o fato de algumas emoções parecerem existir em todas as culturas e também em alguns animais, e de algumas emoções parecerem gerar funções biológicas identificáveis relacionadas com as necessidades de sobrevivência do indivíduo e da espécie. Vários pesquisadores apontam, segundo seus estudos, quais são as emoções base, ou seja, que dão origem às demais vertentes de emoções conhecidas ou ainda desconhecidas. Na Tabela 1 é possível visualizar as emoções básicas que cada pesquisador sugere.

Há muitas razões para a falta de um consenso que parece rodear a noção de emoções básicas. Uma razão é que os pesquisadores nem sempre concordam sobre o que são as emoções. O resultado é que as listas de emoções básicas de alguns pesquisadores contêm afirmações que outros não consideram de todo como emoções. Mas dentre as emoções citadas na Tabela 1, e vendo outras abordagens sobre emoções, seis delas se destacam no que diz respeito a frequência, são elas: Felicidade, Tristeza, Medo, Nojo, Raiva, e Surpresa. Essas seis emoções básicas descritas acima são apenas uma parte dos muitos diferentes tipos de emoções que as pessoas são capazes de experimentar.

2.1.3 Como as emoções impactam o comportamento humano

Como foi explicado anteriormente, as emoções estão ligadas inclusive ao processo de evolução humana, isso faz com que, sobretudo, sejam uma parte essencial ao comportamento humano, possibilitando que atue nas atitudes mais primitivas, como a fuga ou até mesmo violência. Os comportamentos são diferentes dos pensamentos e emoções, porque são sobre o que fazemos no mundo, mas as emoções estão presentes diariamente, interferindo para o bem ou para o mal nesse comportamento. Um exemplo contemporâneo que

Tabela 1 – Relação de emoções básicas segundo pesquisadores da psicologia.

Autores	Emoções
Arnold (1960)	Raiva, aversão, coragem, dejeição, desejo, desespero, medo, ódio, esperança, amor e tristeza.
Ekman (1982)	Raiva, nojo, medo, alegria, tristeza e surpresa.
Frijda (1986)	Desejo, felicidade, interesse, surpresa, dúvida e mágoa.
Gray (1982))	Fúria, horror, ansiedade e alegria.
Izard (1971)	Raiva, desprezo, nojo, angústia, medo, culpa, interesse, alegria, vergonha e surpresa.
JAMES (1884)	Medo, sofrimento, amor e fúria.
McDougall (1926)	Raiva, nojo, elação, medo, subjeção, emoção terna e dúvida.
Mowrer (1960)	Dor e prazer.
Oatley e Johnson-laird (1987)	Raiva, nojo, ansiedade, felicidade e tristeza.
Panksepp (1982)	Expectativa, medo, fúria e pânico.
PLUTCHIK (1980)	Aceitação, raiva, antecipação, nojo, alegria, medo, tristeza e surpresa.
Tomkins (1984)	Raiva, interesse, desprezo, nojo, angústia, medo, alegria, vergonha e surpresa.
Watson (1930)	Medo, amor e fúria.
Weiner (1985)	Felicidade e Tristeza.

Fonte: Ortony e Turner (1990)

pode ser citado é no esporte. Se o atleta vem de resultados ruins, ele vai começar a ficar triste consigo mesmo, isso fará ele ter menos vontade de treinar e consequentemente, vai manter os resultados, ruins. Emoções com características positivas, como felicidade ou amor, tendem a dar mais ânimo para a realização de tarefas ou criação de ideias.

2.2 Características do Áudio

Qualquer som, seja ele qual for, é causado por alguma vibração. Essas vibrações geram ondas sonoras que são detectadas pelo tímpano com frequência e amplitude definidas. Algumas características inerentes à essas ondas podem ser identificadas e catalogadas, de forma que seja possível definir sua estrutura geral. Neste capítulo, são exibidas algumas técnicas de extração de características dos sinais sonoros que foram aplicadas à execução do SER deste trabalho, e algumas definições sobre os sinais sonoros que são relevantes para a compreensão deste trabalho.

2.2.1 *Mel-Frequency Cepstral Coefficients*

Estudos psicofísicos demonstraram que a percepção humana da frequência sonora da fala não segue uma escala linear Hossan, Memon e Gregory (2010). A escala Mel é um método de reproduzir essa percepção. Ela recolhe parâmetros da fala semelhantes aos que são utilizados pelo ser humano para ouvir a fala, ao mesmo tempo que considera todas as outras informações Dave (2013). Dessa forma, para cada tom sonoro de frequência real, f, medido em Hz, é medido um tom concorrente numa escala chamada escala de Mel f_{mel} que é uma escala que visa imitar as características únicas perceptíveis pelo ouvido humano. Como pode ser notado na seguinte equação:

$$f_{mel} = 2595 \log_{10} \left(t1 + \frac{f}{700} \right) \tag{1}$$

Para normalizar as frequências na escala Mel, primeiro os *frames* do áudio (pequeno segmento de áudio) são filtrados no processo chamado de *Fast Fourier transform* (FFT), em que são gerados filtros calculados com a média do espectro em torno da frequência central, e têm diferentes larguras de janelas, essas larguras denominam a quantidade de amostras (do áudio bruto) que devem ser calculadas. É claro que, quanto maior a janela, menos filtros serão gerados. Esse processo de transformação é chamado de *filter bank*, que como o nome diz, gera um banco de filtros. Depois desse processo ser efetuado as frequências resultantes do filtro são normalizadas para a escala Mel.

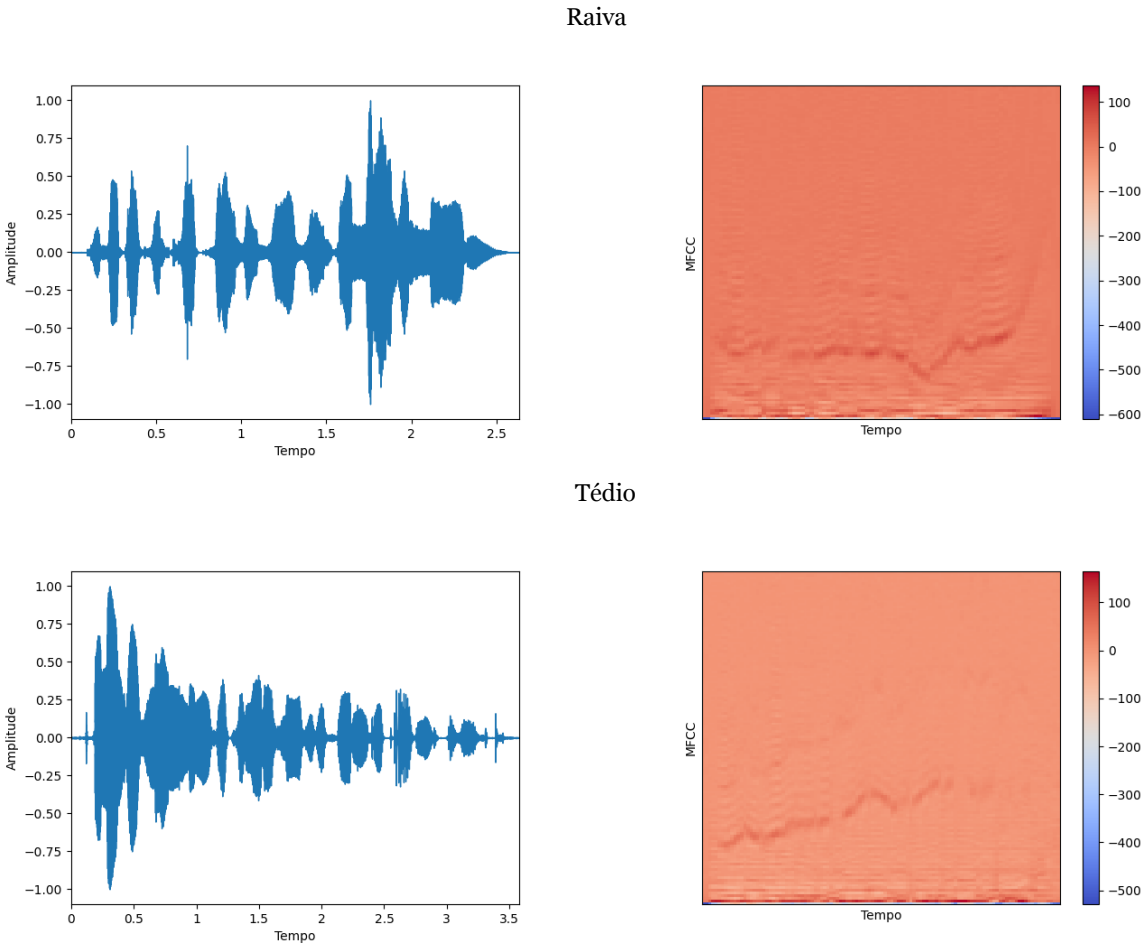
Os *Mel-Frequency Cepstral Coefficients* (MFCC), chamado de Coeficientes Cepstrais de Frequência Mel, é uma das técnicas de extração de características mais populares utilizadas no reconhecimento da fala com base no domínio da frequência utilizando a escala Mel. Para gerar o MFCC, o primeiro passo é dividir o sinal de áudio em *frames* , após isso, vão ser aplicados ao FFT. Posteriormente, o processamento dos *filter bank* é efetuado nos espectros de potência, utilizando a escala Mel Alim e Rashid (2018). A obtenção dos coeficientes MFCC pode ser descrito na seguinte equação:

$$\hat{C}_n = \sum_{n=1}^{k} \left(\log \hat{S}_k \right) cos \left[n \left(k - \frac{1}{2} \right) \frac{\pi}{k} \right] \tag{2}$$

onde k é o número de coeficientes mel, $\hat{S}_k$ é a saída do *filter bank* e $\hat{C}_n$ é o coeficiente final MFCC. Na Figura 2, pode-se observar o gráfico das ondas sonoras de um áudio (esquerda) e o mesmo com coeficientes cepstrais de frequência mel (direita).

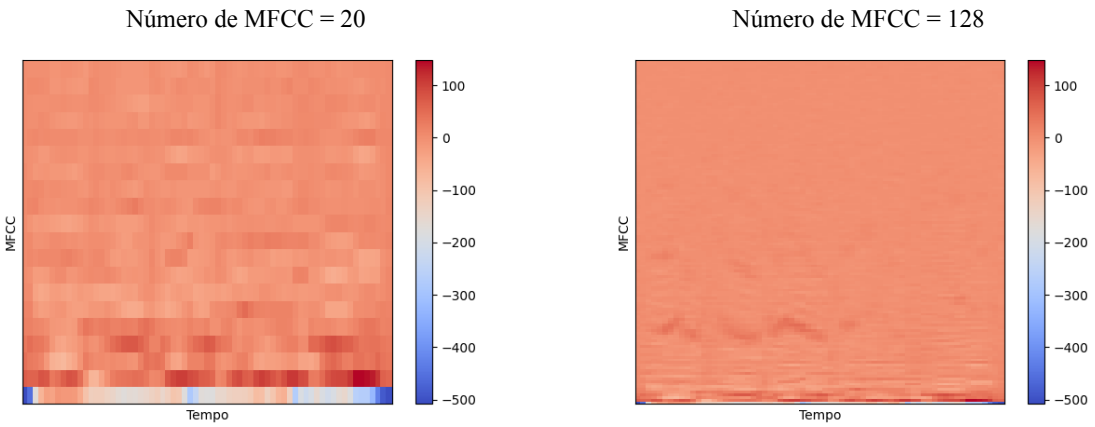
O número de MFCC gerados interfere diretamente no resultado final, pois quanto maior o número de MFCC, maior vai ser o detalhamento das informações contidas sobre cada intervalo de frequência. Como pode ser observado na Figura 3, no gráfico da esquerda, com um número de 20 coeficientes, o resultado gerado será bem mais resumido do que para o gráfico da direita, com 128 coeficientes.

Figura 2 – Visualização gráfica de uma onda sonora (esquerda) e seu gráfico de MFCC gerados (direita)



Fonte: Autoria própria

Figura 3 – Diferenças de resultados em um áudio da emoção nojo com variações de MFCC da base de dados EMO-DB



Fonte: Autoria própria

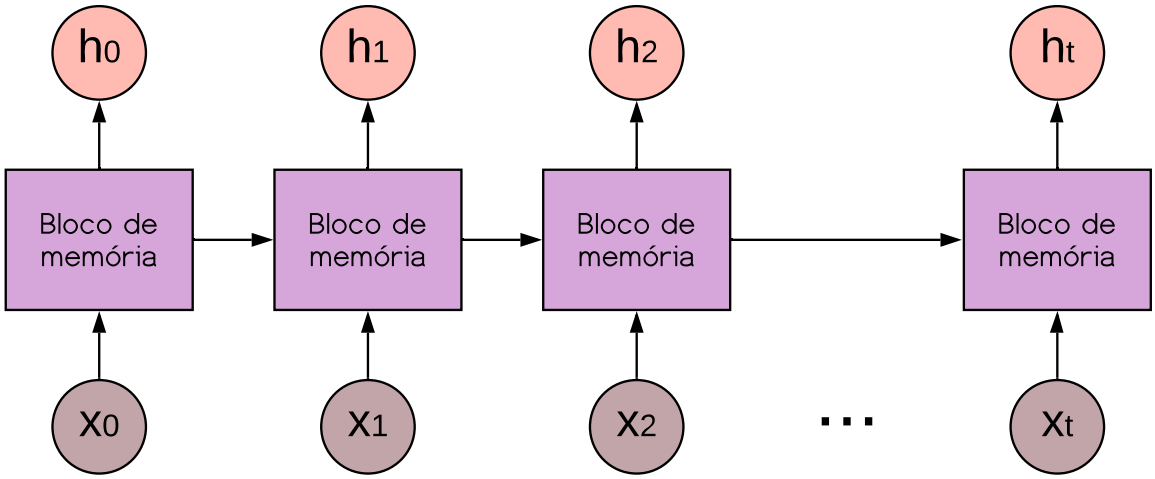
2.3 Classificadores

Os classificadores são itens essenciais para o trabalho abordado nesta pesquisa. Neste sentido, essa seção apresenta os principais classificadores encontrados nos trabalhos relacionados. A saber, a Subseção 2.3.1 aborda a *Long Short-Term Memory*, uma rede neural que possui conexões de *feedback* para analisar a recorrência dos dados. A Subseção 2.3.2, explica sobre a *Convolucional Neural Network*, um tipo de rede neural focada em mapear características marcantes em um vetor matricial.

2.3.1 *Long Short-Term Memory*

As *Recurrent neural networks* (RNN) tem como característica a retenção informações das camadas anteriores para serem consultadas por camadas posteriores, e com esse embasamento, calcular seus valores de saída Heaton (2015). Contudo, quanto maior a distância entre as camadas, mais a informação retida irá perder relevância.

Figura 4 – Esquema de funcionamento de uma *Recurrent Neural Network*



Fonte: Autoria própria

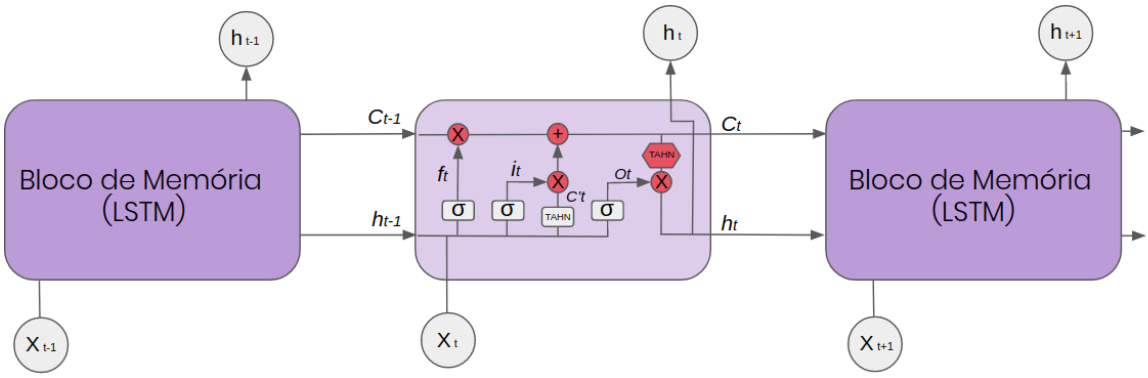
Uma RNN pode ser considerada múltiplas cópias de uma mesma rede (Figura 4), em que, cada cópia calcula os pesos referentes aos dados, e passa esses pesos para as cópias seguintes. Isso serve para a última cópia de RNN sofrer interferência das anteriores, o que possibilita o aprendizado temporal.

É então que em 1997 em Hochreiter e Schmidhuber (1997) os autores propuseram uma de rede neural recorrente chamada *Long Short-Term Memory* (LSTM) que foi projetada para lidar com os problemas de memória encontrados nas RNNs da época. Isso foi possível por meio de um algoritmo eficiente, baseado em gradientes que efetuam cálculo de relevância dentro das camadas. O LSTM contém unidades especiais chamadas blocos de

memória na camada oculta recorrente. Os blocos de memória contêm células de memória com auto-conexões que armazenam o estado temporal da rede, além de unidades multiplicativas especiais chamadas portões para controlar o fluxo de informação, atribuindo prioridades Sak, Senior e Beaufays (2014).

Na Figura 5 é elucidado como a estrutura de uma célula de memória LSTM se constituem.

Figura 5 – Esquema de funcionamento da célula ou bloco de memória de uma LSTM



Fonte: Autoria própria

A célula atual recebe o estado da célula passada C_{t-1}, junto com a camada camada oculta h_{t-1}. Esses dados, em conjunto com a entrada da célula atual X_t, irão gerar o C e h da célula atual.

O *Forget gate* (f_t), decide qual informação é permitida passar a diante e qual deve ser suprimida.

$$f_t = \sigma(W_{xf}x^{(t)} + W_{hf}h^{(t-1)} + b_f) \tag{3}$$

O *input gate* (i_t) e o *input node* (g_t) serão inseridos na equação $(C^{(t)})$ para gerar um valor no qual representará o estado daquela célula de memória.

$$i_t = \sigma(W_{xi}x^{(t)} + W_{hi}h^{(t-1)} + b_i) \tag{4}$$

$$g_t = \sigma(W_{xg}x^{(t)} + W_{hg}h^{(t-1)} + b_g) \tag{5}$$

$$C^{(t)} = (C^{(t-1)} * f_t) + (i_t * g_t) \tag{6}$$

O *output gate* (C'_t) decide como atualizar os valores das unidades ocultas h

$$C'_t = \sigma(W_{xc'}x^{(t)} + W_{hc'}h^{(t-1)} + b_{c'}) \tag{7}$$

Dessa forma será possível gerar as unidades da ocultas da célula atual

$$h^{(t)} = C'_t * tanh(C^{(t)}) \tag{8}$$

Seguindo esse padrão de funcionamento, as células posteriores dificilmente sofrerão escassez de informações prévias, pois o cálculo de prioridade irá garantir que só informações que possivelmente serão relevantes se sobreponham Raschka e Mirjalili (2017).

2.3.2 *Convolutional neural network*

A *Convolutional neural network* (CNN) é uma vertente do aprendizado profundo que tem como objetivo o cálculo matricial. Tem sido usada para tarefas de reconhecimento de padrões, tais como reconhecimento facial e reconhecimento numérico escrito à mão Matsugu et al. (2003). A CNN aprende a mapear uma determinada imagem, no caso uma matriz de valores, para sua categoria correspondente, detectando uma série de características específicas de cada entrada Khan et al. (2018). Uma CNN pode ter dezenas ou centenas de camadas, onde, cada uma aprende a detectar diferentes características de uma matriz. As CNNs realizam a identificação e classificação de imagens, texto, som e vídeo. Antes dos dados serem passados para a camadas da rede, eles são pré-processados, para então serem treinados. É importante mencionar que esse pré-processamento varia de acordo com o tipo de dado de entrada, por isso existem diversas formas de fazê-lo. Após ter efetuado esse pré-processamento, os dados poderão ser elencados às camadas. Comumente, as CNNs são compostas por várias camadas convolutivas como já citado, e de subamostragem (também conhecidas como *Pooling* (P)) que são seguidas por uma ou mais camadas Completamente Conectadas, em inglês *Fully Connected* (FC), no final Raschka e Mirjalili (2017).

As camadas convolutivas são compostas por múltiplos mapas de características, eles podem ser representados como:

$$x_j^l = f\left(\sum_{i \epsilon M_j} x_i^{l-1} * k_{ij}^l + b_j^l\right) \tag{9}$$

Nessa função o x_j^l irá representar o j-ésimo mapa na camada convolucional l, e vai ser ativada pela convolução de todos os mapas de características da camada superior e pela adição de uma máscara. k_{ij}^l e b_j^l são valores dos pesos e da máscara usada, respectivamente.

Na camada de subamostragem ou *pooling*, é usada para fazer uma redução do mapa de características. Cada saída do mapa de características são descritas:

$$x_j^l = f(u_j^l) \tag{10}$$

$$u_j^l = down(x_j^{l-1}) \tag{11}$$

A *down* é a função que representará o mapa de caraterísticas de x_j^{l-1} com um determinado tamanho de amostragem reduzido, como por exemplo uma matriz 2x2 ou 4x4.

A camada completamente conectada está geralmente localizada no final da rede. Cada neurônio está totalmente conectado a todos os neurônios da camada superior; ele pode integrar informações locais com discriminação de categoria nas camadas de convolução ou de agrupamento. O x_j^l representa o valor de saída da camada FC, e é obtido da seguinte forma:

$$u^l = w^l x^{l-1} + b^l \tag{12}$$

$$x_j^l = f(u_j^l) \tag{13}$$

A conexão local, a partilha de peso e a amostragem na CNN podem reduzir eficazmente a complexidade do modelo de rede e o número de parâmetros de formação. Os dados de entrada são abstraídos em representação de características de alto nível através de operações algorítmicas entre camadas Jiang Hongliang Fu (2019).

As CNNs podem ser classificadas em diferentes dimensões, sendo as mais utilizadas a CNN 1D, unidimensional, e a CNN 2D, bidimensional. A escolha da dimensão da arquitetura de CNN usada é relativa aos dados que serão utilizados no treinamento, sendo a CNN 2D melhor para o processamento de fotos ou vídeos e a CNN 1D para sinais sonoros unidimensional.

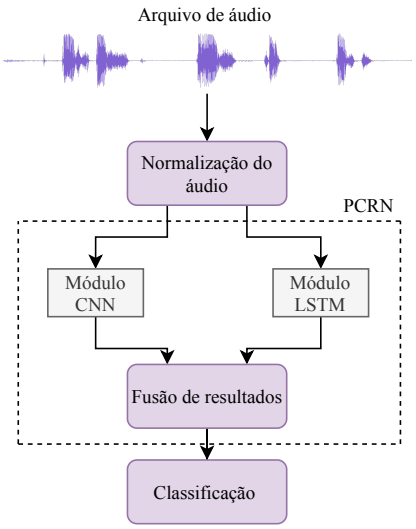
3 TRABALHOS RELACIONADOS

Seguir uma linha de raciocínio mais atual, é de suma importância, visto que, o problema do reconhecimento das emoções pela voz não chegou à perfeição e vem sendo testado com diferentes tipos de algoritmos como tentativa de aumentar sua precisão. Desse modo, nesta seção são expostos alguns trabalhos na área do SER que estão relacionados ao presente sistema desenvolvido. Os motivos da escolha dos seguintes trabalhos foram, sua modernidade tal qual seu desempenho nos principais bancos de dados exploradas nesse tipo de pesquisa.

3.1 *Parallelized Convolutional Recurrent Neural Network With Spectral Features for Speech Emotion Recognition*

No trabalho de Jiang Hongliang Fu (2019), foi construído um algoritmo de reconhecimento de emoções pelo discurso chamado de *Parallelized Convolutional Recurrent Neural Network* (PCRN), que é divido em dois módulos de execução. Um módulo com uma rede CNN e outro com uma rede LSTM. Cada áudio analisado, passa pela execução dos dois módulos de rede neurais, e ao fim das execuções, os resultados gerados por cada uma são mesclados e normalizados para enfim, proporcionar um único resultado. O objetivo do trabalho é gerar um resultado no qual contenha as características do CNN para aprender detalhes da emoção pelo tempo e frequência da fala e do LSTM para identificar as mudanças temporais dos detalhes das emoções, como citado pelos autores.

Figura 6 – Estrutura resumida do modelo do algoritmo PCRN



Fonte: Autria própria.

Este trabalho utilizou quatro diferentes bases de dados de áudio, para fazer os testes

experimentais. Foram elas: EMO-DB, CASIA, ABC e SAVEE. As funções de avaliações escolhidas foram a média ponderada (WA) e a média aritmética (UA) dos resultados da matriz de confusão. Também é utilizado um método de validação cruzadas chamado de *Leave-One-Speaker-Out* (LOSO). Nessa estratégia, todo os áudios de um ator ficam de fora do treinamento, sendo usados apenas no teste do algoritmo, isso acontece até que todos os atores passem por essa exclusão. Finalmente, a média dos vários experimentos é calculado como resultado.

Tabela 2 – Comparação da média (%) de resultados da matrizes de confusão de diferentes arquiteturas do modelo PCRN

Banco de dados	Modelo	WA	UA
CASIA	PCRN_CNN	48.66	48.66
	PCRN_LSTM	49.67	49.67
	PCRN	58.25	58.25
EMO-DB	PCRN_CNN	76.64	72.85
	PCRN_LSTM	80.01	78.37
	PCRN	86.44	84.53
ABC	PCRN_CNN	50.97	44.90
	PCRN_LSTM	58.77	54.88
	PCRN	61.63	57.59
SAVEE	PCRN_CNN	55.62	49.64
	PCRN_LSTM	54.79	49.76
	PCRN	62.49	59.40

Fonte: Jiang Hongliang Fu (2019).

Para fim de verificação de resultados, além de terem comparado com outros trabalhos, foi feita uma observação com os resultados obtidos com as variações de arquitetura do PCRN. Como pode ser observado na Tabela 2, o modelo "PCRN_CNN" seria o algoritmo desenvolvido pelos autores, porém utilizando apenas os resultados obtidos via o módulo CNN, assim também se constitui o "PCRN_LSTM". E o "PCRN" sendo o resultado do algoritmo final, com os dois módulos analisados.

3.2 *A CNN-Assisted Enhanced Audio Signal Processing for Speech Emotion Recognition*

Em Mustaqeem (2019) é desenvolvido um algoritmo de reconhecimento de emoções com uma arquitetura de CNN 2D em que se objetiva na otimização do tempo de treinamento e melhor seleção de dados chave, tendo como principal foco o pré-processamento dos áudios, que concentra seu processo em eliminar os ruídos e intervalos sem nenhuma fala. Seus experimentos foram efetuados nas bases IEMOCAP e RAVDESS. No trabalho é citado que foi centralizado o foco de execução no pré-processamento pelo fato de que as arquiteturas atuais de CNN não revelaram nenhuma melhora significativa em termos de

precisão e complexidade de custo no processamento de sinais de voz, e o uso da LSTM é útil para treinar dados sequenciais, mas são difíceis de treinar com eficácia e são mais complexos computacionalmente.

Para o pré-processamento, os arquivos de áudio passam por uma filtragem de ruídos e informações irrelevantes que é feita por uma função que tem como entrada os dados de energia e amplitude, e como resultado um único valor para o segmento de áudio analisado. Após essa etapa, é gerado o espectrograma desses áudios, que são suas informações em representação bidimensional, representando a força do sinal do áudio pelas diferentes frequências.

O algoritmo de treinamento é composto por 7 camadas de CNN, sendo o seu diferencial a exclusão das camadas *Poolings*. Isso foi possível pela quantidade de camadas de CNN que são estruturadas junto com o tamanho de filtragem fixa para cada.

Na Tabela 3 é possível analisar os resultados obtidos pelo trabalho. Nele são comparadas precisões de reconhecimento do modelo de algoritmo em questão com a filtragem dos áudios (*Clean spec*) e sem a filtragem (*Raw spec*).

Tabela 3 – Comparação das precisões com diferentes bases de dados e entrada de áudios gerados pelo algoritmo presente em Mustaqeem (2019)

Modelo	Entrada	Base de dados	Média ponderada de precisão	Média aritimética de precisão	F1 Score
Proposto	Raw spec	IEMOCAP	76%	72%	77%
Proposto	Clean spec	IEMOCAP	84%	82%	84%
Proposto	Raw spec	RAVDESS	68%	61%	70%
Proposto	Clean spec	RAVDESS	80%	79%	81%

Fonte: Releitura de Mustaqeem (2019).

Ao examinar a Tabela 3, fica claro o impacto da filtragem dos áudios, além disso, denota que, ter uma maior atenção no pré-processamento dos dados que serão inseridos no algoritmo de rede neural (Como a CNN) pode implicar em um efeito positivo maior do que focar exclusivamente na arquitetura das suas camadas.

3.3 *Speech emotion recognition using deep 1D & 2D CNN LSTM networks*

No trabalho de Zhao, Mao e Chen (2019), são implementados dois modelos de arquiteturas de um algoritmo para a efetuação do reconhecimento das emoções pela voz, chamados de 1D CNN LSTM e 2D CNN LSTM. Com o objetivo de aumentar a precisão do reconhecimento emocional, os autores propuseram um bloco de aprendizado chamado *Local Feature Learning Block* (LFLB), que, é responsável pelo processo convolucional do aprendizado. Esse bloco é composto por uma camada convolucional, uma camada de normalização, uma camada de unidade linear exponencial (ELU) e uma camada *Max-Pooling*, que serve para reduzir as escalas da matriz de características. O algoritmo contém três LFLB, que geram um resultado parcial para ser inserido em uma rede LSTM.

O trabalho cita as arquiteturas como CNN LSTM, pelo fato de usar a camada de CNN e LSTM em sequência. Na 1D CNN LSTM, é utilizada uma CNN unidimensional, e a arquitetura utiliza como dado de treino o áudio bruto, já na 2D CNN LSTM, é utilizada uma camada de CNN bidimensional, tendo como entrada os logaritmos da escala Mel.

Foram consumidas duas base de dados, a alemã EMO-DB e a americana IEMO-CAP. Os autores usam dois diferentes métodos de organização dos dados para avaliarem os resultados gerados pelos algoritmos. Um é chamado de *Speaker-Dependent*, no qual faz todo o procedimento de reconhecimento emocional utilizando apenas os áudios de uma só pessoa. Sendo que o resultado final é a média aritmética da precisão obtida das execuções de cada indivíduo. E ou outro método é o *Speaker-Independet*, que faz o procedimento com os áudios de todos os indivíduos da base de dados e uma única execução.

Tabela 4 – Comparação da precisão média de reconhecimento da rede 2D CNN LSTM realizada na base de dados EMO-DB. Melhores resultados em negrito.

Trabalho de pesquisa	**Precisão (Speaker-Dependent)**	**Precisão (Speaker-Independent)**
Wu, Falk e Chan (2011)	91.6	85.8
Huang et al. (2014)	88.3	85.2
Huang et al. (2015)	75.5	-
Demircan e Kahramanli (2016)	-	92.9
Zhao, Mao e Chen (2019)	**95.33**	**95.89**

Fonte: Zhao, Mao e Chen (2019).

4 METODOLOGIA

Até este momento, foram abordados os principais conceitos referentes à estruturação deste trabalho e terminologias utilizadas ao longo do mesmo. Neste Capítulo são apresentadas as técnicas e métodos de construção que foram bases para a obtenção dos resultados do estudo proposto neste trabalho.

4.1 Bases de dados

Um sistema de reconhecimento de emoção pela voz pode ser modelado através de um sistema de classificação supervisionado. Para tanto, é preciso ter uma base de dados catalogada para que se possa conhecer os rótulos (neste caso as emoções) de maneira que as medidas de desempenho possam ser aplicadas. Dessa forma, nesta seção, são descritas algumas da bases de dados coletadas através da leitura bibliográfica, direcionadas para o estudo do SER.

4.1.1 *Surrey Audio-Visual Expressed Emotion*

Surrey Audio-Visual Expressed Emotion (SAVEE) é uma base de dados audiovisual que foi projetada com o propósito de ser aplicada no desenvolvimento de sistemas de reconhecimento automático de emoções Jackson e haq (2011). Ela consiste em gravações de 4 atores masculinos, estudantes de pós-graduação e pesquisadores da Universidade de Surrey com idades entre os 27 e 31 anos, com total de 7 emoções. As emoções são descritas psicologicamente em categorias discretas: raiva, repugnância, medo, felicidade, tristeza e surpresa. Também foi acrescentada a emoção neutra para fornecer as 7 categorias desejadas. O material de texto, no qual os atores leram, consistia de 15 frases por emoção: 3 frases comuns, 2 específicas de emoção e 10 genéricas que eram diferentes para cada emoção e equilibradas foneticamente. As 3 frases comuns e $2 \times 6 = 12$ frases específicas de emoções foram gravadas como neutras para dar 30 frases neutras. Isto resultou em um total de 120 sentenças por ator. Os dados foram gravados em um laboratório de mídia visual com equipamentos audiovisuais de alta qualidade, processados e etiquetados.

4.1.2 *Ryerson Audio-Visual Database of Emotional Speech and Song*

A *Ryerson Audio-Visual Database of Emotional Speech and Song*, chamada de RAVDESS, é uma base de de dados canadense multimodal de fala e música com o foco na análise emocional que foi inteiramente gravada em inglês. A RAVDESS consiste de 24 atores profissionais, cada um realizando 104 vocalizações únicas com emoções que incluem: felicidade, tristeza, raiva, medo, surpresa, nojo, calmo e neutra Livingstone e

Russo (2018). Por ser multimodal, elea se divide em três módulos. No primeiro, *audio-visual* (AV), cada arquivo contém tanto a gravação do vídeo quanto do áudio. No segundo, *video-only* (VO), é constituído apenas pelo vídeo das expressões faciais dos atores. E por fim, o *audio-only* (AO), contendo apenas o áudio dos atores. No módulo *audio-only*, sendo especificamente focado na fala (não na música), é composto por 60 arquivos de áudio de cada ator, totalizando 1440 arquivos. É importante ressaltar que apenas o módulo *audio-only* dessa base de dados foi utilizado.

Nessa base de dados, as emoções foram gravadas com dois níveis de intensidade, a normal e a forte. Esse tipo de definição foi escolhida com intuito de tentar captar ao máximo a emoção por completo, já que uma emoção pode conter muitos níveis de intensidade.

4.1.3 *Berlin Database of Emotional Speech*

Com falas de 10 atores, sendo 5 mulheres e 5 homens, a *Berlin Database of Emotional Speech* (EMO-DB) Burkhardt et al. (2005) é uma base de dados alemã de áudios para análise da emoção. Os atores geraram 10 frases cada, sendo que metade foram frases longas, a outra metade, frases curtas e algumas frases extras, que podem ser utilizadas na comunicação diária e são interpretáveis em todas as emoções aplicadas. As emoções são divididas em 7, que são elas: neutra, raiva, medo, alegria, tristeza, repugnância e tédio. As gravações foram feitas numa câmara anecoica com equipamento de gravação de alta qualidade. O material contido na base de dados compreende cerca de 800 frases. Para sua construção, não foi preferível o uso de atores profissionais para efetuarem as gravações, dessa forma, a procura foi feita por meio de um anúncio no jornal. Cerca de 40 pessoas se inscreveram e foram convidadas para uma sessão de pré-seleção. Tiveram de fazer uma apresentação em cada uma das emoções pretendidas. Das 40 pessoas, três especialistas selecionaram as 10 pessoas que constituíram o corpo de atores, representando igualmente os sexos, julgando a naturalidade e o reconhecimento da atuação. Ao efetuarem todas as gravações, o material gerado foi avaliado num teste de audição automatizada e cada frase foi julgada por 20 ouvintes no que diz respeito ao reconhecimento e naturalidade da emoção manifestada.

4.2 Dados e Pré-Processamento

Nesta seção são apresentados os métodos de extração das características dos dados utilizados no treinamento do algoritmo de Rede Neural (explicado na Seção 4.5), bem como o pré-processamento intitulado de Janela Deslizante para aumentar a quantidade de amostras de áudios que foram dividas para treinamento e teste.

4.2.1 MFCC

Escolher de forma mais apropriada o método de extração das características dos áudio a serem processadas é altamente importante, uma vez que para diferenciar as emoções, é preciso observar pequenas variações que ocorrem com a frequência do áudio. Tendo isso em vista, o método escolhido para a obtenção dos dados são os MFCC. Como explicado na Subseção 2.2.1 os MFCC são coeficientes cepstrais extraídos a partir da frequência do áudio na escala Mel, que contêm um alto detalhamento de informações ajustadas a simular o ouvido humano.

A biblioteca Librosa McFee et al. (2015), dispõe de métodos de extração das características do áudio, dentre elas os MFCC, podendo escolher o número de coeficientes a serem gerados. Considerando que, quanto maior o número de coeficientes, mais detalhado serão os dados retornados. Assim, neste trabalho foram utilizados 128 coeficientes, com o de FFT no tamanho de 2048 *frames* por filtro, e um pulo de 512 *frames* por filtragem. Com o intuito de manter todos os áudios do mesmo tamanho, todos os áudios menores que 8 segundos são preenchidos por eles mesmo até atingir o valor necessário e áudios maiores que 8 segundos são reduzidos para este valor. Por consequência, como todos os áudios têm 8 segundos, a matriz de resultante será de 128x251 para todos os áudios, sendo o primeiro valor o número de coeficientes e o segundo o número de valores resultantes.

4.2.2 Janela Deslizante

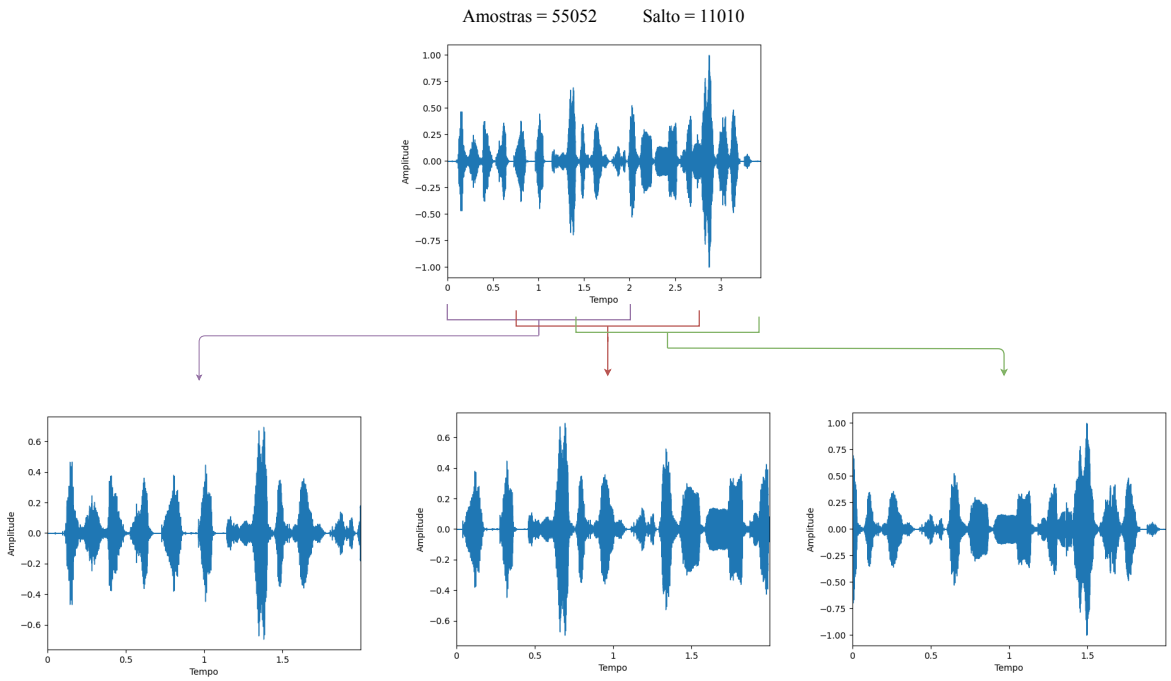
Um problema bem recorrente encontrado nos treinamentos das redes neurais, o *Overfitting* Rice, Wong e Kolter (2020), faz com que o algoritmo de aprendizado de máquina fique muito bom em reconhecer os padrões dos dados de treino, mas não obtenha desempenho tão bom no reconhecimento dos dados de teste, resumindo, o *Overfitting* limita a generalização do algoritmo. Algumas técnicas usadas durante a construção do algoritmo de rede neural podem evitar esse problema, uma delas é a *Data Augmentation*, que é relativa a geração de novas amostras a partir dos dados existentes, para assim aumentar o tamanho da base de dados Lashgari, Liang e Maoz (2020). Com isso o algoritmo de rede neural receberá uma quantidade maior para dados, tanto para treino, quanto para teste, e consequentemente sua capacidade de generalização será maior, uma vez que ele estará apto para reconhecer mais amostras de dados.

Para este trabalho, foi utilizado um método chamado Janela Deslizante, proposto por este trabalho, como forma de efetuar a técnica de *Data Augmentation*, na qual consiste em definir uma janela fixa de tempo, que, irá pegar todo conteúdo de áudio no determinado intervalo da amostra para transformá-lo em uma nova amostra independente, e um salto para janela, ou seja, a quantidade de dados que essa janela irá se descolar (deslizar) para então gerar um novo dado.

4.2.2.1 Estruturação da janela

Como a análise que o algoritmo faz sobre os áudios para determinar a emoção é de forma temporal, ou seja, investiga e relaciona os dados pelo seu comportamento durante toda extensão do áudio, faz com o que seja importante, além de tudo, ter um nível mínimo da margem de extração de uma nova amostra, para não perder a essência da emoção contida no áudio. Por isso, foi definido que o a janela deve ser fixa de 2 segundos para todos os áudios, assim, as novas amostras geradas preservam sua identidade. Para fazer o processamento dos áudios, é delimitada uma *sample rate* (taxa de amostragem), que representa a quantidade de amostras de frequência devem ser recolhidas por segundo. Neste trabalho foi definida uma *sample rate* de 16000 (Dezesseis mil), dessa forma, como a Janela Deslizante foi definida em 2 segundos, seu número de amostras será de 32000 frequências. Sabendo disso, foi estabelecido que para um áudio ser processado pela Janela Deslizante, ele deve ter um número mínimo de 40000 amostras de frequências (2,3 segundos). Isso significa que os áudios menores de 2,3 segundos não sofreram a multiplicação, para que não perdessem a riqueza nos dados.

Figura 7 – Multiplicação de um áudio em três novos áudios através do processo de Janela Deslizante



Fonte: Autoria própria

A Figura 7 ilustra como ocorre o processo de Janela Deslizante na prática. No exemplo em questão, foram gerados 3 novos áudios de 2 segundos, a partir de um áudio de 3.44 segundos. É importante citar que, caso o próximo áudio a ser gerado pela Janela Deslizante exceda o tamanho máximo do áudio de origem, o processo será interrompido e

só serão utilizados os áudios gerados previamente.

4.2.2.2 Cálculo do tamanho do Salto

Em virtude de fazer a Janela Deslizante se descolar pelos dados do áudio, é calculado um tamanho de pulo para cada áudio, tendo como embasamento seu tamanho. Essa determinação foi imposta para que a quantidade de novos dados não ficasse desbalanceada para áudios que forem grandes. A Equação seguinte mostra como foi feito o cálculo, onde S é a quantidade de dados de frequência que a Janela Deslizante se movimenta e tam o tamanho total de dados de frequência do áudio de entrada:

$$S = \frac{tam}{Df} \tag{14}$$

onde Df é o coeficiente de divisão, no qual determina o tamanho do pulo da Janela Deslizante ao dividir a quantidade de amostras de frequência do áudio original. Para este trabalho foi considerado $Df = 5$ devido ao tamanho mínimo de áudio aceito (2,3 segundos) no processo de Janela Deslizante e o tamanho dos áudios resultantes (2 segundos). O $Df = 5$ garante que serão gerados dois áudios de 2 segundos para um áudio com o tamanho mínimo aceitável.

4.2.2.3 Transformação da base de dados

A quantidade de áudios gerados provenientes do processo de Janela Deslizante está ligado diretamente à duração de cada áudio base, por isso, o crescimento foi diferente tanto para emoções, quanto para bases de dados diferente, pois, os áudios não possuem tamanho fixo, tendo maior tamanho de duração 8 segundos e o menor com cerca de 1,12 segundos, ambos da base EMO-DB. Observando a Tabela 5, vê-se que a base de dados RAVDESS possui praticamente todos o seus dados com mais de 3 segundos, isso implica em um crescimento maior para a base. Confirmando essa afirmação, na Tabela 6 é possível perceber que a RAVDESS foi a que obteve o maior percentual de crescimento, já a EMO-DB, obteve o menor percentual por ter áudios mais curtos que as demais bases.

Tabela 5 – Análise sobre a distribuição da quantidade de áudios por variação de duração nas base de dados utilizadas.

Quantidade de áudios por segundos nas bases de dados			
Segundos	**EMO-DB**	**RAVDESS**	**SAVEE**
1 a 2 segundos	126	0	13
2 a 3 segundos	224	1	92
3 a 4 segundos	136	997	191
Maiores de 4 segundos	49	250	184
TOTAL	**535**	**1248**	**480**

Fonte: Autoria própria

Ainda observando a Tabela 6, nota-se que que há uma desigualdade de crescimento por emoção, como já citado. Esse comportamento se justifica pelo fato das bases possuírem diferentes distribuições de áudios e tamanhos para os mesmo. Nas bases RAVDESS e SAVEE, em que além de terem quantidades iguais de áudios para cada emoção (exceto a Neutra), possuem uma média de 3,42 e 3,49 segundos de duração dos seus áudios, isso implicou em um crescimento que permaneceu acima dos 200%. Em contraste, a base EMO-DB, com 2,46 segundos de média, obteve um crescimento inferior, tendo praticamento só metade dos seus áudios aplicados a Janela Deslizante.

Tabela 6 – Análise do crescimento de dados após o processo de Janela Deslizante em cada uma das bases de dados utilizadas.

Crescimento das emoções com a Janela Deslizante - EMO-DB			
Emoção	**Original**	**Com Janela**	**Crescimento**
Raiva	127	287	126%
Tédio	81	195	141%
Nojo	46	141	207%
Medo	69	119	72%
Felicidade	71	150	111%
Neutra	79	143	81%
Tristeza	62	223	260%
TOTAL	**535**	**1258**	**135%**

Crescimento das emoções com a Janela Deslizante - RAVDESS			
Emoção	**Original**	**Com Janela**	**Crescimento**
Neutra	96	364	279%
Calma	192	759	295%
Felicidade	192	748	290%
Tristeza	192	755	293%
Raiva	192	762	297%
Medo	192	728	279%
Nojo	192	764	298%
TOTAL	**1248**	**4880**	**291%**

Crescimento das emoções com a Janela Deslizante - SAVEE			
Emoção	**Original**	**Com Janela**	**Crescimento**
Raiva	60	209	248%
Nojo	60	224	273%
Medo	60	211	252%
Felicidade	60	221	268%
Neutra	120	422	252%
Tristeza	60	245	308%
Surpresa	60	218	263%
TOTAL	**480**	**1750**	**264,5%**

Fonte: Autoria própria

4.3 Métodos de Treino

Alguns métodos de treinos devem ser levados em conta quando é planejada a construção de um algoritmo de aprendizado de máquina, dentre eles está a separação dos dados que serão usados para o treinamento do algoritmo e aqueles que servirão para testar a eficiência do treino. Por essa razão, foi escolhida uma abordagem bastante presente em estudos relacionados ao Reconhecimento de Emoções Pelo Discurso, a *Speaker-Dependent* Dependência do Locutor.

A *Speaker-Dependent* (SD), é a maneira mais popular de separação dos dados para a execução do algoritmo. Neste método a base é divida em duas partes, uma onde vão os dados de treino e a outra com dados de teste. Essa divisão é feita por porcentagem, na qual é preciso informar quantos por cento cada parte terá de dados. Ao separar os dados, o algoritmo só efetuará o treinamento com dados de treino e o teste com os de teste. Apesar de ser um método já conhecido, ele leva esse nome pelo fato do algoritmo depender dos dados de todos os Atores contidos na base de dados, tanto na etapa de teste quando no treino. Neste trabalho, o uso do método *Speaker-Dependent* foi definido com 80% de divisão dos dados para treino e os 20% restantes para testes.

4.4 Funções de avaliação

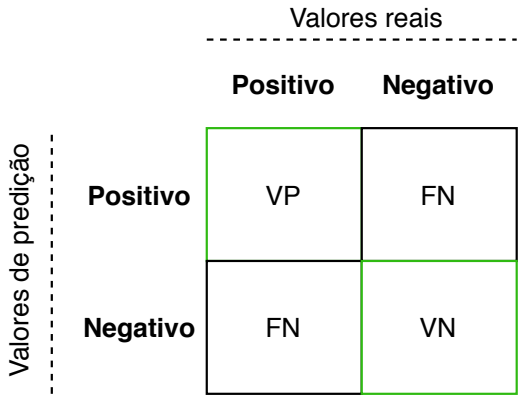
Assim que treinada, a Rede Neural precisa passar por uma avaliação de qualidade, na qual aponta a eficácia do treinamento da rede. Essa avaliação pode ser feita de várias formas, com tanto que seja com os resultados de classificação gerados pela inserção dos dados de teste no modelo da Rede Neural treinada.

4.4.1 Matriz de confusão

Uma das técnicas mais utilizadas é a matriz de confusão. Uma matriz de confusão é um tipo de medição de desempenho de classificação usada em Redes Neurais, assim como em outras técnicas do Aprendizado de Máquina. É uma espécie de tabela que ajuda a conhecer o desempenho do modelo de classificação num conjunto de dados de teste para que os verdadeiros valores sejam conhecidos. Ela é divida em valores reais, dispostos nas colunas, e valores de predição, dispostos na horizontal.

Como pode ser observado na Figura 8, no índice da primeira coluna, na primeira linha, “V” significa verdadeiro, naquela posição será indiciado os valores que a rede neural indicou corretamente a predição dos dados testados, seja ele verdadeiro negativo (VN), ou verdadeiro positivo (VP) . Já na mesma primeira linha, porém em uma coluna diferente, é possível identificar o “FN”, falso negativo, ele seria a predição da rede neural de um valor positivo, mas que na verdade é um valor negativo, nesse caso, seria uma predição errada. Portanto, os valores indicados corretamente estão na coluna e linha de mesmo índice.

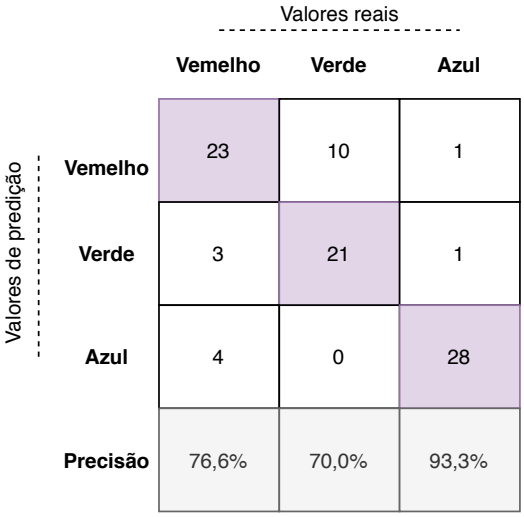
Figura 8 – Exemplo de uma matriz de confusão



Fonte: Autoria própria

Uma forma de observar os dados da matriz de confusão, é tirando médias sobre os seus verdadeiros positivos em relação as predições incorretas, sendo o cálculo de precisão por classe um ótimo método quando os dados são desbalanceados. Para calcular a precisão de cada classe, é preciso dividir o valor no qual a predição da rede neural indicou corretamente a classe, pela soma de toda aquela coluna. Na Figura 9 é possível observar como é representada a precisão de cada classe, reproduzida por coluna na matriz de confusão.

Figura 9 – Exemplo de calculo da precisão de cada classe



Fonte: Autoria própria

A porcentagem da precisão média, portanto, é a soma das precisões resultantes de cada classe , dividia pelo número de classes analisadas.

$$\frac{76,6 + 70 + 93,3}{3} = 79,9\% \tag{15}$$

No exemplo da Figura 9, nota-se que a classe “azul” teve uma precisão de 93,3%

em sua predição, ainda assim, pelo fato das demais classes obterem precisões inferiores, a precisão, como pode ser observada na Equação 15, foi de 79,9%, sendo cerca de 14% à menos da precisão da classe melhor reconhecida.

4.4.2 *Unweighted Average*

Como as bases de dados deste trabalhos foram balanceadas com o método de treino SD, uma outra função de avaliação baseada nos resultados gerados na matriz de confusão encontrou-se viável. A *Unweighted Average* (UA), trata-se de uma média aritmética fundamentada apenas nos verdadeiros positivos, desconsiderando os outros possíveis resultados obtidos pelas classes. Esse modo de avaliar o desempenho do algoritmo com a Matriz de Confusão é bastante presente em trabalhos relacionados ao SER, como em Zhao, Mao e Chen (2019) e em Jiang Hongliang Fu (2019). Este trabalho utiliza esse método de avaliação como principal referência, porque encontra-se como um dos mais utilizados, além de ser uma ótima métrica quando os dados não são desbalanceados.

4.4.3 ROC-AUC

Para fazer uma avaliação sobre a sensibilidade do algoritmo em classificar os dados é utilizado o ROC-AUC. Em que ROC (*Receiver Operating Characteristic*) é representado por uma curva gráfica gerada a partir de uma função em que calcula os verdadeiro positivos com os falsos positivos, podendo ser representada pela AUC (*Area Under the Curve*), que exibe a área que a curva cobre. Quanto maior ROC-AUC, maior é a capacidade do algoritmo de destingir as classes.

4.4.4 F1-Score

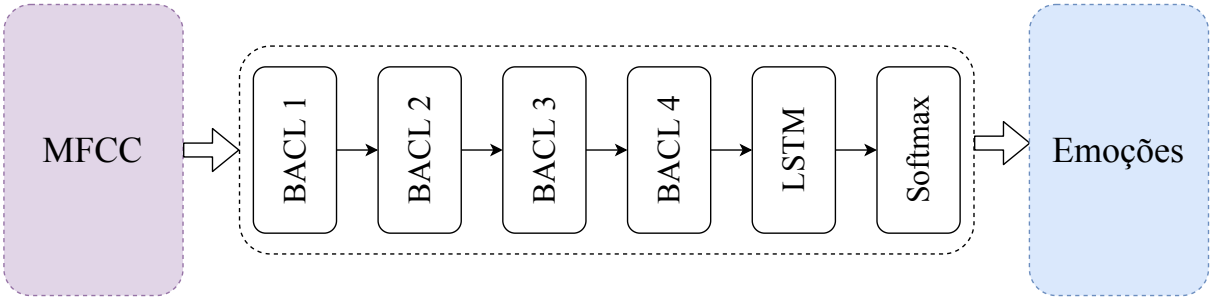
Dividido em Macro e Micro, o F1-Score é uma métrica de avaliação sobre os dados de teste que tem como objetivo exibir o grau de precisão do algoritmo levando em consideração o balanço de cada classe. Neste sentido, quando mais parecido com a precisão do algoritmo (no caso deste trabalho, a UA), mais os dados estão balanceados, comprovando a veracidade em sua precisão. Para o F1-Score com a abordagem Micro, o F1-Score mede um único valor levando em conta o balanceamento das classes como um todo, já o Macro, é originário a partir do cálculo com cada classe.

4.5 Algoritmo de Classificação

Fazer um computador reconhecer uma emoção, exige sobretudo, uma análise profunda e ampla em relação as características dos áudios, tanto se especificando em como cada valor se consiste, quanto verificando a relação desses valores como um todo ao decorrer do tempo. Tendo isso em vista, neste trabalho foi utilizado o mesmo modelo de

algoritmo presente em Zhao, Mao e Chen (2019), no qual se constitui de quatro Blocos de Aprendizagem de Características Locais (BACL), que são blocos compostos por camadas de CNN 2D focados em abstrair o máximo de informações relevantes do áudio, uma camada de rede LSTM, na qual foi selecionada com o objetivo de analisar dependências entre uma sequência dos dados, seu papel é identificar quais são as dependências de contexto temporal a partir dos dados resultantes dos BACL, e por fim uma camada *Softmax*, que é uma generalização da regressão logística para problemas de classificação com várias classes, que vai determinar a emoção do áudio de acordo com as características aprendidas. A Figura 10 mostra o arranjo do algoritmo usado para este trabalho

Figura 10 – Diagrama da arquitetura do algoritmo para a classificação das emoções a partir dos MFCC extraído dos áudios



Fonte: Zhao, Mao e Chen (2019)

4.5.1 CNN 2D

Para fim de obter uma maior abstração de quais características dos áudios resultarão na classificação de determinada emoção, foi utilizada para a composição das BACLs, as camadas de rede neural CNN 2D. Esses tipos de camadas contam com a filtragem dos dados que mais se destacam, que é feita através de um *Kernel*. Para todas as camada de convolução, foi definido um *Kernel* de tamanho 3x3, ou seja, 9 unidades da matriz de características passaram pela filtragem por vez, tendo uma *Stride* de 1x1, que informa quantas unidades o *Kernel* deslocou-se até sua próxima filtragem. Foi definido um número pequeno de *Stride* para que não fosse perdida muitas informações sobre o áudio.

4.5.2 LSTM

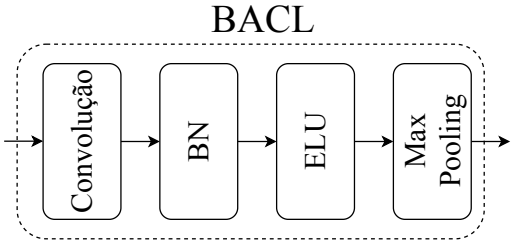
A camada de LSTM neste algoritmo, teve o propósito de, através de uma análise temporal, examinar a composição e a variação dos dados do áudio provenientes das camadas de CNN, para então gerar os últimos dados de classificação da rede neural. Cada célula (Bloco de memória) que compõe a camada de LSTM efetua um cálculo de análise sobre o dado inserido, dessa forma, foi utilizado um número total de 256 células de LSTM,

o dobro do valor de saída da última camada de CNN. Isso foi definido para aumentar a confiabilidade do resultado gerado.

4.5.3 Blocos de Aprendizagem de Características Locais

Cada BACL é composto por uma camada Convolucional, uma *Batch Normalization* (BN), uma camada ELU (*exponential linear unit*), e uma camada *Max Pooling*, sendo as camadas Convolucional e *Max Pooling* as fundamentais desse conjunto Zhao, Mao e Chen (2019). Na Figura 11 é exibida a composição de um BACL e sua ordem de operações.

Figura 11 – Composição de um Bloco de Aprendizagem de Características Locais



Fonte: Zhao, Mao e Chen (2019)

Como citado na Subseção 2.3.2, as camadas Convolucionais filtram os dados de entrada como forma de mapear as características marcantes do mesmo, essas características são também chamadas de ativações. A camada BN normaliza os resultados da camada Convolucional de cada lote de áudios treinados, isso implica em uma melhora de performance e estabilidade das camadas de aprendizado. A transformação aplicada pela BN mantém a ativação média próxima a 0 e o desvio padrão da ativação próximo a 1. Após a camada BN, encontra-se a ELU, que define a saída da BN. Ela força a média dos valores de ativações a ficarem mais próximos de zero para diminuir o tempo de aprendizagem e aumentar o nível de reconhecimento. Por fim, a camada *Max Pooling*, tipo de camada *Pooling*, que para esta abordagem com áudios, faz com quem as características mais importantes se sobressaíam sobre as distorções e barulhos contidos nos áudios. Ela consegue fazer essa abstração dividindo a entrada em regiões e dando como saídas os valores máximos de cada região.

4.5.4 Parâmetros das camadas

Para que o algoritmo de classificação execute de forma correta, é importante dar atenção em alguns detalhes de implementação. Um desses detalhes é o controle de dados resultante de cada camada. Para cada BACL, o que define o tamanho da sua saída é o tamanho do *Kernel Size*, região que filtra as ativações da camada, e o *Stride* que informa quantas unidades o *Kernel Size* deve se deslocar após a ativação. E para a camada de

LSTM, o tamanho da sua saída é definido por quantas unidades de LSTM a camada se compõe. A Tabela 7 apresenta os dados utilizados neste trabalho.

Tabela 7 – Parâmetros utilizados nas camadas do algoritmo utilizado neste trabalho.

Camada	**Tamanho da saída**	***Kernel Size***	***Stride***
BACL 1 - Convolução	128 x 256 x 64	3 x 3	1 x 1
BACL 1 - Max Pooling	64 x 125 x 64	2x2	2x2
BACL 2 - Convolução	64 x 125 x 64	3 x 3	1 x 1
BACL 2 - Max Pooling	16 x 31 x 64	4 x 4	4 x 4
BACL 3 - Convolução	16 x 31 x 128	3 x 3	1 x 1
BACL 3 - Max Pooling	4 x 7 x 128	4 x 4	4 x 4
BACL 4 - Convolução	4 x 7 x 128	3 x 3	1 x 1
BACL 4 - Max Pooling	1 x 1 x 128	4 x 4	4 x 4
LSTM	256	-	-
Softmax - Dense	7	-	-

Fonte: Autoria própria

5 RESULTADOS

Apresentar qual é o desempenho do algoritmo de Rede Neural adjunto ao pré-processamento com a Janela Deslizante exige a realização de repetidos testes com o designo de capturar a média dos resultados. Assim, a estabilidade do treinamento pode ser comprovada. Outrossim, ter a visualização da evolução de resultados dependendo do crescimento da base após o *Data augmentation*, suporta a implementação do método proposto por este trabalho, uma vez que o algoritmo de treinamento é o mesmo para cada base. Neste Capítulo serão expostas as configurações utilizadas na obtenção dos resultados do trabalho proposto, bem como sua comparação, tanto internamente, quanto com os resultados gerados pelos trabalhos relacionados, ambicionando uma contextualização de efetividade da implementação do que fora apresentado até então neste trabalho.

5.1 Configuração dos testes

Os áudios utilizados para a execução do algoritmo foram divididos em 80% para treinamento e 20% para teste. Para certificar que a divisão seria feita igualitariamente em termos de porcentagem de cada classe de emoção, não só para a base como um todo, foi utilizado um parâmetro chamado *stratify* pertencente à função de separação de dados de treino e teste da biblioteca *sklearn*. Nesse parâmetro foi colocado um vetor contendo as emoções dos áudios, dessa forma a porcentagem da separação tanto por classe de emoção, quanto pra base de modo geral, foram as mesmas.

Um algoritmo de Rede Neural possui alguns critérios de parada, que o fazem interromper seu treinamento para que seus resultados sejam reportados. Um deles é definir o limite máximo de épocas que ele pode alcançar. Uma época é como é chamado todo o processo de interação dos dados desde a entrada no algoritmo, até o recálculo dos pesos da Rede Neural pós-predição da classe que o algoritmo sugere que o conjunto de dados de entrada pertence, gerando um modelo contendo as configurações de pesos das camadas da época em questão. Parar o treinamento do algoritmo baseando-se apenas na quantidade de épocas pode gerar um *Overfitting*, já que ele ficará cada vez mais específico no processamento dos dados de treino. Por esse motivo, para a obtenção do modelo que trará melhores resultados tanto com os dados de treino quanto com dados de teste, este trabalho utilizou o método de parada baseado na maior acurácia obtida com os dados de validação por 15 épocas, isso quer dizer que, se após 15 épocas o algoritmo não obter uma acurácia maior de acerto na predição dos dados de validação, o treinamento se encerrará, salvando como modelo definitivo, o que obteve a maior acurácia de predição nos dados de validação.

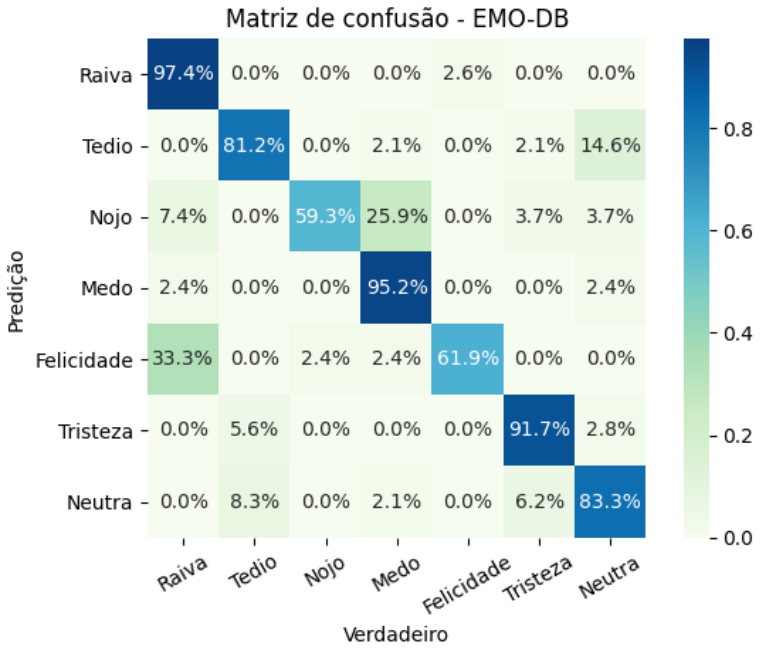
Visando ter uma análise convincente sobre o real desempenho de treinamento do

algoritmo, cada base de dados foi treinada 3 vezes, tendo seu desempenho recolhido ao fim de cada execução. Ao final das 3 execuções, foram tiradas as médias aritméticas dos dados de matriz de confusão, UA, ROC-AUC, F1-Score Micro e F1- Score Macro, gerando um único valor para tais métricas de avaliação que representam o desempenho do algoritmo em cada base de dados.

5.2 Resultados sem a Janela Deslizante

Os resultados gerados sem o processamento da Janela Deslizante, servem para denotar a atual capacidade que o algoritmo de Rede Neural implementado neste trabalho, utilizando os MFCC como entrada, pode alcançar.

Figura 12 – Matriz de confusão resultante do treinamento da base de dados EMO-DB sem a Janela Deslizante

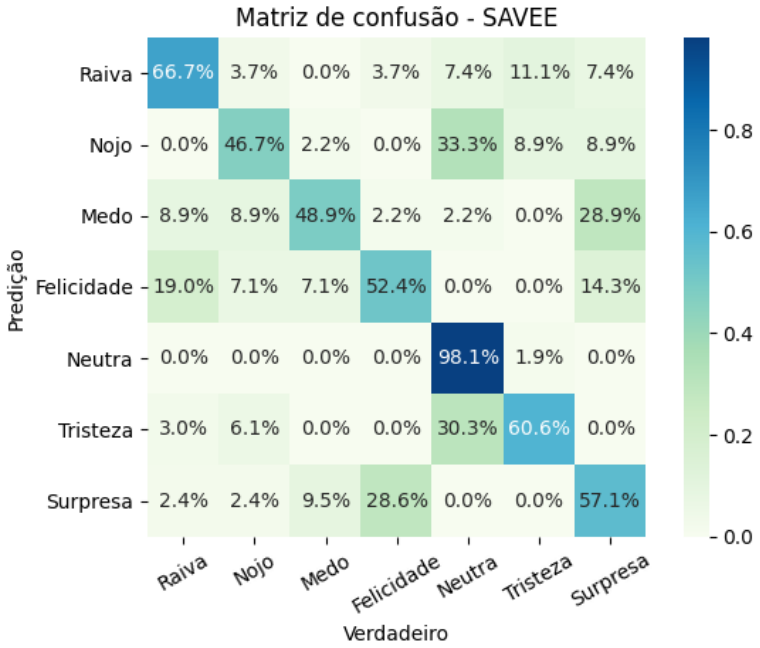


Fonte: Autoria própria

Ao observar a Figura 12, percebe-se que algumas emoções como Medo e Tristeza alcançaram uma alta porcentagem de reconhecimento, mesmo com uma quantidade inferior de dados analisados que as demais pertencentes a base EMO-DB, já outras como Nojo e Felicidade obtiveram resultados bastante inferiores, esse desbalanceamento pode ser explicado por não ter uma quantidade suficiente de dados para o algoritmo conseguir diferenciar precisamente as emoções, o que pode fazer ele reconhecer muito bem certas emoções e outras não. A emoção Felicidade mantém uma baixa porcentagem de reconhecimento em todos as base, como pode ser visto nas Figuras 12, 13 e 14. Além da quantidade de áudios, o que acarreta essa ocorrência é o fato do algoritmo geralmente

confundir com emoções mais intensas, como Raiva e Surpresa por exemplo.

Figura 13 – Matriz de confusão resultante do treinamento da base de dados SAVEE sem a Janela Deslizante



Fonte: Autoria própria

Outra característica presente nas execuções dos resultados sem a Janela Deslizante que é importante esclarecer, são as diferenças de porcentagem no reconhecimento das emoções por base de dados. Por exemplo, a emoção Nojo, visualizando as Figuras 13 e 14. Mesmo que o número de amostras para essa emoção seja igual as das outras emoções em cada uma das bases SAVEE e RAVDESS em questão (como pode ser visto na Tabela 6), a desigualdade dos resultados ainda é grande. Esses acontecimentos são consequentes do método de gravação dos áudios de cada base, em que podem tratar as emoções de forma um pouco mais variada, fazendo então seus dados diferirem de outras bases, e por isso, serem classificados com algumas variações.

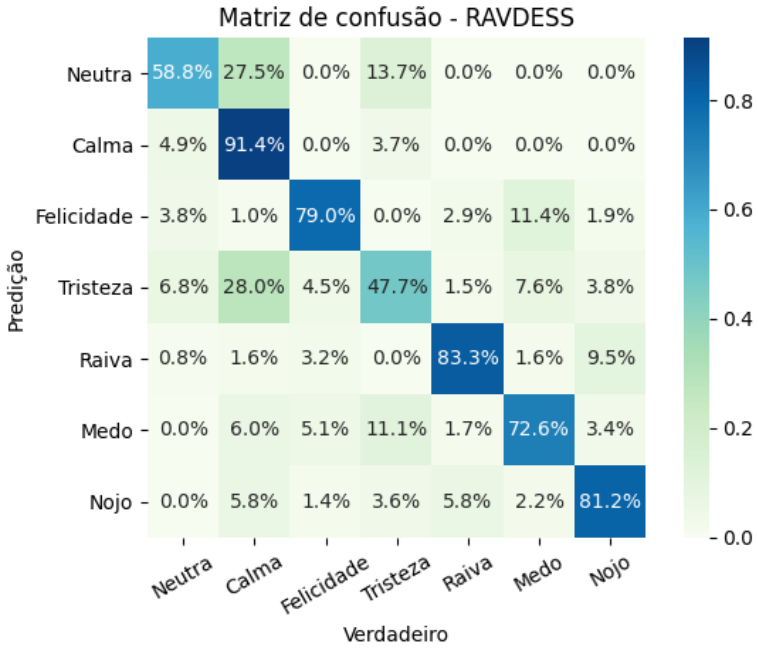
Tabela 8 – Funções de avaliação geradas a partir da execução das bases de dados sem a Janela Deslizante.

Modelo	Entrada	Base de dados	ROC-AUC	F1-Score Macro	F1-Score Micro	UA
Sem JD	MFCC	EMO-DB	98.19%	82.50%	84.50%	**81.44%**
Sem JD	MFCC	SAVEE	90.34%	61.1%	62.5%	**61.50%**
Sem JD	MFCC	RAVDESS	95.29%	72.21%	73.60%	**73.44%**

JD = Janela Deslizante.
Fonte: Autoria própria.

A Tabela 8 revela os resultados finais de execução do treinamento de cada base,

Figura 14 – Matriz de confusão resultante do treinamento da base de dados RAVDESS sem a Janela Deslizante



Fonte: Autoria própria

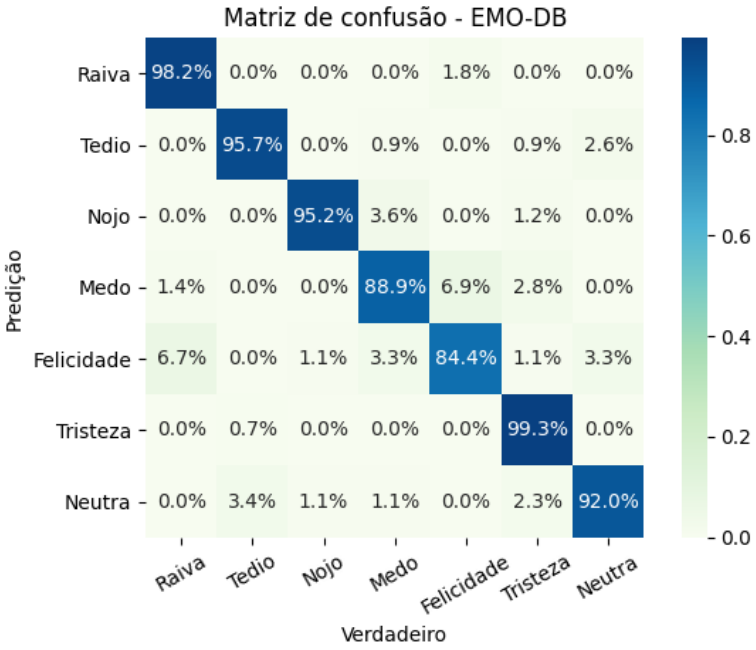
apresentando uma diferença de praticamente 20% de reconhecimento entre as bases EMO-DB e SAVEE, em relação a função UA. A base SAVEE obteve os piores resultados nos treinamentos, e não coincidentemente, possui o menor número de amostras de áudios.

5.3 Resultados com a Janela Deslizante

As Figuras 15, 16 e 17, apresentam as devidas matrizes de confusão geradas pelas execuções em cada base de dados. Contrastando com as matrizes de confusão dispostas na Subseção 5.2, verifica-se que as houve uma redução de disparidade de resultados entre as emoções, como por exemplo as emoções Nojo e Neutra da base SAVEE em que na Figura 13 encontra-se uma diferença de reconhecimento de 51,4%, no entanto , os resultados de sua execução com a Janela Deslizante na Figura 16 mostram uma diferença bem menor, com uma porcentagem de 17,7%. Essa evolução é em decorrência ao aumento de amostras de áudio que o algoritmo teve, isso o fez calcular melhor os pesos de suas camadas, obtendo assim, uma maior capacidade de classificação.

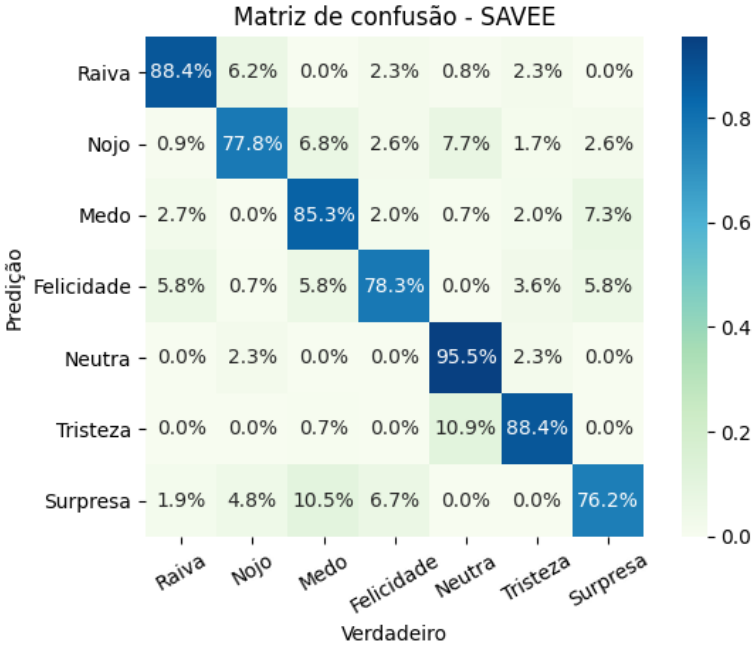
Devido ao aumento das amostras de dados terem sido diferente em cada base de dados, implicou em diferentes margens de crescimento no reconhecimento final das emoções. A Tabela 9 mostra os resultados obtidos com a Janela Deslizante na diversas funções de avaliação, tendo como principal parâmetro de observação a UA, e na Tabela 6 fica perceptível o aumento de reconhecimento em cada base. Na EMO-DB foram gerados

Figura 15 – Matriz de confusão resultante do treinamento da base de dados EMO-DB com a Janela Deslizante



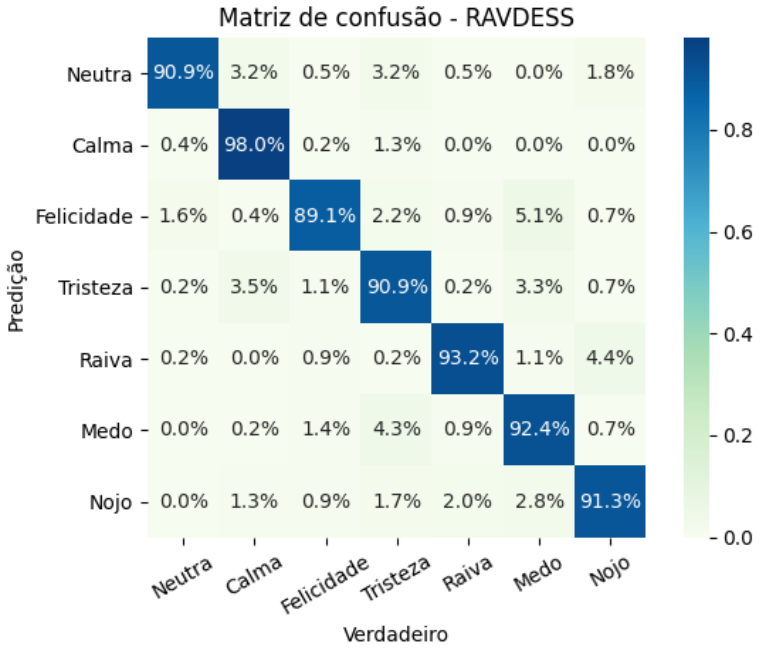
Fonte: Autoria própria

Figura 16 – Matriz de confusão resultante do treinamento da base de dados SAVEE com a Janela Deslizante



Fonte: Autoria própria

Figura 17 – Matriz de confusão resultante do treinamento da base de dados RAVDESS com a Janela Deslizante



Fonte: Autoria própria

menos áudios que as demais bases (como pode ser visto na Tabela 5), isso cominou em um aumento menor no seu desempenho final.

Os excelentes resultados obtidos com a inclusão da Janela Deslizante, reforçam a ideia de que com uma maior quantidade de amostras de áudios, maior será a capacidade do algoritmo em classificar as emoções de entrada.

Tabela 9 – Funções de avaliação geradas a partir da execução das bases de dados com a Janela Deslizante.

Modelo	Entrada	Base de dados	ROC-AUC	F1-Score Macro	F1-Score Micro	UA
Com JD	MFCC	EMO-DB	99,70%	93,64%	94,44%	**93,39%**
Com JD	MFCC	SAVEE	98,21%	84,61%	86,00%	**84,26%**
Com JD	MFCC	RAVDESS	99,52%	92,39%	92,38%	**92,26%**

JD = Janela Deslizante.
Fonte: Autoria própria.

Tabela 10 – Crescimento no reconhecimento de cada base utilizada.

Base de dados	Crescimento de amostras	UA sem JD	UA com JD	Aumento no reconhecimento
EMO-DB	135%	81,44%	93,39%	**11,95%**
SAVEE	264,5%	61,50%	84,26%	**22,76%**
RAVDESS	291%	73,44%	92,26%	**18,82%**

JD = Janela Deslizante.
Fonte: Autoria própria.

5.4 Comparação com trabalhos relacionados

Os Trabalhos citados no Capítulo 3 servem para apresentar o estado da arte no que diz respeito ao reconhecimento de emoções pelo discurso, e embasando nessa premissa, é possível avaliar externamente os resultados gerados por este trabalho. As Tabelas 11, 12 e 13, mostram a comparação dos resultados entre eles e este trabalho. O método de treino utilizado em Jiang Hongliang Fu (2019) fora diferente dos demais, por esse motivo foi citado nas tabelas. Ele consiste em fazer o treinamento com os áudios de todos os atores, exceto um. Os áudios desse ator não inserido no treinamento são utilizados como teste.

Tabela 11 – Comparação dos resultados da EMO-DB com trabalhos relacionados

Trabalho	Método de treino	UA
Jiang Hongliang Fu (2019)	LOSO	**84,53%**
Zhao, Mao e Chen (2019)	*Speaker Dependent*	**95,02%**
Proposto	*Speaker Dependent*	**93,39%**

Tabela 12 – Comparação dos resultados da SAVEE com trabalhos relacionados

Trabalho	Método de treino	UA
Jiang Hongliang Fu (2019)	LOSO	**59,40%**
Proposto	*Speaker Dependent*	**84,26%**

Tabela 13 – Comparação dos resultados da RAVDESS com trabalhos relacionados

Trabalho	Método de treino	UA
Mustaqeem (2019)	*Speaker Dependent*	***76,85%**
Proposto	*Speaker Dependent*	**92,26%**

*Esse é o resultado desconsiderando a emoção Surpresa, presente em Mustaqeem (2019) mas não neste trabalho.

Os resultados gerados por este trabalho conseguiram obter um ótimo desempenho em relação os trabalhos citados, exceto ao Zhao, Mao e Chen (2019). Como pode ser observado na Tabela 11, o trabalho em questão conseguiu ainda assim quase 2% à mais na avaliação UA.

6 CONSIDERAÇÕES FINAIS

Simular o processo de reconhecimento de emoções humanas em um computador, como já citado, é um processo muito complexo, pois além de cada emoção variar dependendo do contexto cultural em que está inserida, ela pode abranger diferentes entonações, o que de certa forma dificulta o processo de classificação. Este trabalho apresentou um método de multiplicação das amostras dos áudios chamado Janela Deslizante, objetivando a redução da complexidade dessa classificação. Com base em trabalhos que compõem o estado da arte no processamento do SER, foi possível compreender o nível de precisão em que os estudos envolvendo o SER está, bem como a comparar os resultados obtidos por este trabalho.

Conforme apresentado, o algoritmo utilizado por este trabalho, junto ao método de Janela Deslizante como pré-processamento, conseguiu elevar a precisão do reconhecimento das emoções na base EMO-DB em 11,95%, na SAVEE em 22,76%, e na RAVDESS em 18,82%. Essa consequência se deu pelo fato do algoritmo ter à disposição um número maior de amostras de emoções para efetuar seu treinamento, o que possibilitou uma classificação mais concisa dos áudios.

Ter um algoritmo de reconhecimento de emoções consistente, que consiga discernir qual a emoção de um áudio de entrada, pode ser bastante útil nas tarefas diárias e em aplicações mais específicas. Tendo isso em vista, a aplicação da Janela Deslizante ou método similar, se faz proveitoso na implementação de trabalhos futuros para que eles possam alcançar um desempenho ainda maior. Além disso, o que ainda pode ser feito em relação a Janela Deslizante, seria focar mais seu processo em relação a cada emoção, visando deixar as emoções com o mesmo número de amostras de áudio.

REFERÊNCIAS

AL-SHAWAF, L.; LEWIS, D. **Evolutionary Psychology and the Emotions**. [S.l.: s.n.], 2017.

ALIM, S. A.; RASHID, N. K. A. Some commonly used speech feature extraction algorithms. In: LOPEZ-RUIZ, R. (Ed.). **From Natural to Artificial Intelligence**. Rijeka: IntechOpen, 2018. cap. 1. Disponível em: <https://doi.org/10.5772/intechopen.80419>.

ARNOLD, M. Emotion and personality. 1960.

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

DAVE, N. Feature extraction methods lpc , plp and mfcc in speech recognition. In: . [S.l.: s.n.], 2013.

DEMIRCAN, S.; KAHRAMANLI, H. Application of fuzzy c-means clustering algorithm to spectral features for emotion classification from speech. **Neural Computing and Applications**, 11 2016.

EKMAN, P. What emotion categories or dimensions can observers judge from facial behavior. 1982.

FRIJDA, N. The emotions. 1986.

GRAY, J. The neuropsychology of anxiety: An enquiry into the functions of the septo-hippocampal system. 1982.

Haeb-Umbach, R. et al. Speech processing for digital home assistants: Combining signal processing with deep-learning techniques. **IEEE Signal Processing Magazine**, v. 36, n. 6, p. 111–124, 2019.

HANAB YING LIA, X. Z. H. Convolutional neural network learning for generic data classification. **Information Sciences**, ELSEVIER, v. 477, p. 448–465, March 2019. ISSN 0020-0255. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0020025518308703?via%3Dihub>.

HEATON, J. **Artificial Intelligence for Humans, Volume 3: Deep Learning and Neural Networks**. Createspace Independent Publishing Platform, 2015. (Artificial Intelligence for Humans). ISBN 9781505714340. Disponível em: <https://books.google.com.br/books?id=q9mijgEACAAJ>.

HOCHREITER, S.; SCHMIDHUBER, J. Long short-term memory. **Neural computation**, v. 9, p. 1735–80, 12 1997.

Hossan, M. A.; Memon, S.; Gregory, M. A. A novel approach for mfcc feature extraction. In: **2010 4th International Conference on Signal Processing and Communication Systems**. [S.l.: s.n.], 2010. p. 1–5.

Huang, Y. et al. Extraction of adaptive wavelet packet filter-bank-based acoustic feature for speech emotion recognition. **IET Signal Processing**, v. 9, n. 4, p. 341–348, 2015.

HUANG, Z. et al. Speech emotion recognition using cnn. Association for Computing Machinery, New York, NY, USA, p. 801–804, 2014. Disponível em: <https://doi.org/10.1145/2647868.2654984>.

IMANI, G. A. M. M. A survey of emotion recognition methods with emphasis on e-learning environments. **Journal of Network and Computer Applications**, v. 147, 2019. ISSN 1084-8045. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1084804519302759>.

IZARD, C. E. The face of emotion. 1971.

JACKSON, P.; HAQ, S. ul. Surrey audio-visual expressed emotion (savee) database. 04 2011.

JAMES, W. II.—WHAT IS AN EMOTION ? **Mind**, os-IX, n. 34, p. 188–205, 04 1884. ISSN 0026-4423. Disponível em: <https://doi.org/10.1093/mind/os-IX.34.188>.

JIANG HONGLIANG FU, H. T. P. L. L. Z. P. Parallelized convolutional recurrent neural network with spectral features for speech emotion recognition. **IEEE Access**, IEEE, v. 7, 2019. ISSN 2169-3536. Disponível em: <https://ieeexplore.ieee.org/document/8756261>.

JING XIA MAO, L. C. S. Prominence features: Effective emotional features for speech emotion recognition. **Digital Signal Processing**, ELSEVIER, v. 72, 2018. ISSN 1051-2004. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1051200417302439>.

Khan, S. et al. **A Guide to Convolutional Neural Networks for Computer Vision**. [S.l.: s.n.], 2018.

LASHGARI, E.; LIANG, D.; MAOZ, U. Data augmentation for deep-learning-based electroencephalography. **Journal of Neuroscience Methods**, v. 346, p. 108885, 2020. ISSN 0165-0270. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0165027020303083>.

Lee, K. H. et al. A study on speech emotion recognition using a deep neural network. p. 1162–1165, 2019.

LIVINGSTONE, S. R.; RUSSO, F. 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>.

MATSUGU, M. et al. Subject independent facial expression recognition with robust face detection using a convolutional neural network. **Neural Networks**, v. 16, n. 5, p. 555 – 559, 2003. ISSN 0893-6080. Advances in Neural Networks Research: IJCNN '03. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0893608003001151>.

MCDOUGALL, W. An introduction to social psychology. 1926.

MCFEE Brian et al. librosa: Audio and Music Signal Analysis in Python. p. 18 – 24, 2015.

MOWRER, O. Learning theory and behavior. 1960.

MUSTAQEEM, S. K. A cnn-assisted enhanced audio signal processing for speech emotion recognition. **Sensors**, v. 20, p. 183, 12 2019.

OATLEY, K.; JOHNSON-LAIRD, P. Towards a cognitive theory of emotions. **Cognition and Emotion**, v. 1, p. 29–50, 03 1987.

ORTONY, A.; TURNER, T. What's basic about basic emotions? **Psychological review**, v. 97, p. 315–31, 08 1990.

PANKSEPP, J. Toward a general psychobiological theory of emotions. **Behavioral and Brain Sciences**, Cambridge University Press, v. 5, n. 3, p. 407–422, 1982.

PLUTCHIK, R. Chapter 1 - a general psychoevolutionary theory of emotion. Academic Press, p. 3 – 33, 1980. Disponível em: <http://www.sciencedirect.com/science/article/pii/B9780125587013500077>.

RASCHKA, S.; MIRJALILI, V. **Python Machine Learning, 2nd Ed.** 2. ed. Birmingham, UK: Packt Publishing, 2017. ISBN 978-1787125933.

RICE, L.; WONG, E.; KOLTER, J. Overfitting in adversarially robust deep learning. 02 2020.

SAK, H.; SENIOR, A.; BEAUFAYS, F. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. **Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH**, p. 338–342, 01 2014.

TOMKINS, S. Affect theory. 1984.

WATSON, J. Behaviorism. 1930.

WEINER, B. An attributional theory of achievement motivation and emotion. **Psychological review**, v. 92, p. 548–73, 11 1985.

WU, S.; FALK, T. H.; CHAN, W.-Y. Automatic speech emotion recognition using modulation spectral features. **Speech Communication**, v. 53, n. 5, p. 768 – 785, 2011. ISSN 0167-6393. Perceptual and Statistical Audition. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0167639310001470>.

XU, Z. W. F. Emotion recognition research based on integration of facial expression and voice. **2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)**, IEEE, v. 7, 2018. Disponível em: <https://ieeexplore.ieee.org/document/8633129>.

ZHAO, J.; MAO, X.; CHEN, L. Speech emotion recognition using deep 1d 2d cnn lstm networks. **Biomedical Signal Processing and Control**, v. 47, p. 312 – 323, 2019. ISSN 1746-8094. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1746809418302337>.