Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Classificação de Intensidade das Emoções na Fala em Português Brasileiro por meio de Deep Learning

Henrique Tibério Brandão V. Augusto

Dissertação apresentada como requisito parcial para
conclusão do Mestrado em Informática

Orientador
Prof. Dr. Geraldo Pereira Rocha Filho

Brasília
2024

Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Classificação de Intensidade das Emoções na Fala em Português Brasileiro por meio de Deep Learning

Henrique Tibério Brandão V. Augusto

Dissertação apresentada como requisito parcial para
conclusão do Mestrado em Informática

Prof. Dr. Geraldo Pereira Rocha Filho (Orientador)
DCET/UESB - CiC/UnB

Prof. Dr. Marcelo Antonio Marotta
Universidade de Brasília

Dr. Rafael Duarte Coelho dos Santos
Instituto Nacional de Pesquisas Espaciais

Prof. Dr. Rodrigo Bonifácio Almeida
Coordenador do Programa de Pós-graduação em Informática

Brasília, 18 de Dezembro de 2024

Dedicatória

Ao leitor.

Agradecimentos

Agradeço a todas as pessoas que fizeram parte da minha vida, próximas ou distantes, na presença ou ausência, e que contribuíram inexoravelmente para minha formação enquanto indivíduo. Agradeço também ao acaso.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES), por meio do Acesso ao Portal de Periódicos.

Resumo

A fala costuma ser a nossa primeira forma de comunicação e de expressão de emoções. O Reconhecimento de Emoção na Fala é um problema complexo, pois a expressão emocional depende da linguagem falada, do dialeto, do sotaque e do histórico cultural dos indivíduos. A intensidade dessa emoção pode afetar nossa percepção e nos induzir a interpretar a informação de maneira inadequada, havendo perspectiva de aplicabilidade em diversas áreas, como: monitoramento de pacientes, segurança, sistemas comerciais e entretenimento. Este trabalho realizou uma tarefa de Aprendizado de Máquina utilizando Aprendizado Profundo para inferir a intensidade das emoções na voz em português, através da Fusão de Domínios com duas bases de dados distintas. Para tal, foi criado um *Autoencoder* para realizar a extração de características e, posteriormente, um modelo supervisionado para efetuar a classificação das intensidades entre quatro classes: (i) Fraca; (ii) Moderada; (iii) Alta; (iv) Pico de intensidade. Os resultados indicam a possibilidade de inferir a intensidade, embora o conjunto de dados seja reduzido, mesmo ao combinarmos dois *datasets*. Foram realizados dois cenários experimentais com arquiteturas análogas, variando apenas a quantidade de características representativas utilizadas como dado de entrada para os modelos. Além disso, observando as métricas de desempenho em ambos experimentos, foi possível notar reincidência da mesma classe (forte) com a menor variação, enquanto as classes mais distantes (fraca e pico) tiveram os melhores desempenhos, o que levanta questionamentos para estudos posteriores.

Palavras-chave: aprendizado de máquina, aprendizado profundo, reconhecimento de emoção na voz, intensidade da emoção, português brasileiro, verbo, vivae

Abstract

Speech is often our first form of communication and expression of emotions. Speech Emotion Recognition is a complex problem, as emotional expression depends on spoken language, dialect, accent, and the cultural background of individuals. The intensity of this emotion can affect our perception and lead us to interpret information inappropriately, with potential applications in various fields such as: Patient monitoring, security, commercial systems, and entertainment. This work performed a Machine Learning task using Deep Learning to infer the intensity of emotions in Portuguese speech, employing Domain Fusion with two distinct databases. To do so, an Autoencoder was created to extract features, and then a supervised model to classify intensities into four classes: (i) Weak; (ii) Moderate; (iii) High; and (iv) Peak intensity. The results indicate the possibility of inferring intensity, although the final dataset is limited, even when combining two datasets. Two experimental scenarios were carried out, with analogous architectures, varying only the quantity of representative features used as input for the models. Additionally, observing the performance metrics on both experiments, it was possible to note the recurrence of the same class (strong) with the lowest variation while the most distant classes (weak and peak) had the best performance, which raises questions for further studies.

Keywords: machine learning, deep learning, speech emotion recognition, emotion intensity, brazilian portuguese, verbo, vivae

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

ADAM *Adaptative Moment Estimation.*

AE *Autoencoder.*

AI *Artificial Intelligence.*

CNN *Convolutional Neural Network.*

DL *Deep Learning.*

DNN *Deep Neural Networks.*

DT *Decision Tree.*

EDA *Exploratory Data Analysis.*

FN *False Negative.*

FP *False Positive.*

GAN *Generative Adversarial Network.*

GRU *Gated Recurrent Unit.*

IA Inteligência Artificial.

KNN *K-Nearest Neighbors.*

LSTM *Long Short-term memory.*

MFC *Mel Frequency Cepstrum.*

MFCC *Mel Frequency Cepstral Coefficients.*

ML *Machine Learning.*

MLP *Multilayer Perceptron.*

MSE *Mean Squared Error.*

NN *Neural Networks.*

PANAS *Positive and Negative Affect Scale.*

PCA *Principal Component Analysis.*

ReLU *Rectified Linear Unit.*

RF *Random Forest.*

RNN *Recurrent Neural Network.*

SER *Speech Emotion Recognition.*

SVM *Support Vector Machine.*

TF *Transformada de Fourier.*

TN *True Negative.*

TP *True Positive.*

TRF *Transformada Rápida de Fourier.*

VAE *Variational Autoencoder.*

VIVAE *Variably Intense Vocalizations of Affect and Emotion Corpus.*

Capítulo 1

Introdução

A fala costuma ser nossa primeira forma de comunicação e de expressão de emoções [4]. Desde a infância, antes mesmo de utilizarmos nosso idioma, expressamos emoções através de sons não verbais que possuem significado para o emissor. Não obstante, existem estudos que investigam a emoção de bebês por meio do choro [5].

Continuamos a expressar emoções dessa maneira ao longo da vida e, em um mundo moderno, estamos interagindo cada vez mais com, e por meio de, ferramentas tecnológicas (*e.g.*: assistentes virtuais como Alexa e Siri).

A emoção é um estado psicológico relacionado com o sistema sensorial, provocada por alterações hormonais que podem estar relacionadas a observações, sentimentos, interações sociais ou algum nível de satisfação ou frustração, e que causam uma alteração distinguível na fala [6, 7]. A intensidade dessa emoção pode afetar nossa compreensão [8] e nos induzir a interpretá-la de forma inadequada.

A análise de emoções na voz se tornou uma área de pesquisa proeminente, graças ao aumento da capacidade computacional e à eficiência de algoritmos [9, 10]. Desse modo, a classificação das emoções e de sua intensidade ganha um papel importante [11] no desenvolvimento da ciência e da tecnologia, uma vez que a mensagem transmitida pode ter sua semântica alterada pelas emoções [12].

O Reconhecimento de Emoção na Fala (*Speech Emotion Recognition*, SER) é um problema complexo [13], pois a expressão emocional depende da linguagem falada, do dialeto, do sotaque e do histórico cultural dos indivíduos [14]. O reconhecimento e a avaliação das emoções apresentam dificuldades por sua natureza interdisciplinar: o reconhecimento de emoções e a avaliação da intensidade são objetos das ciências da Psicologia; a aferição e a avaliação de dados fisiológicos estão relacionadas às ciências médicas; a análise e a solução de dados de sensores são objetos da mecatrônica [15].

Inferir a intensidade da emoção encontra aplicações potenciais em diversas áreas, como saúde [16], segurança [17], entretenimento (através de *smart environments* [18] e *smart*

assistant [19]) e comercial [20]. Trabalhos como [21] buscam entender a intensidade da emoção para melhorar o desempenho da síntese de uma vocalização emocional oriunda de um mecanismo de *speech to text*.

Desenvolver um método para classificar a intensidade da emoção poderia colaborar com o monitoramento do estado de saúde de pacientes, por meio de um modelo embarcado em assistentes virtuais inteligentes; auxiliar no atendimento ao público, encaminhando chamadas de *call centers* compreendidas como urgentes para fora do fluxo automatizado e para o atendimento humano; melhorar o desempenho de sistemas de segurança baseados em reconhecimento de voz, reduzindo a alteração na voz causada por variações do estado emocional, bem como soluções que reduzem ruído; ser utilizado para fins comerciais no sistema financeiro, como uma empresa que alega possuir uma solução *From Voice to Revenue* que aumentou em 20% o sucesso da renegociação para recuperação de crédito de liquidação duvidosa [22].

Esta dissertação foi desenvolvida com o intuito de explorar possibilidades para inferência da intensidade da emoção na voz em nosso idioma nativo. Problemas de SER são problemas complexos, uma vez que a expressão emocional é afetada por diversos fatores que podem influenciar sua compreensão.

1.1 Objetivos

Este trabalho visa propor, desenvolver e avaliar uma arquitetura de classificação da intensidade da emoção na voz Português Brasileiro. A arquitetura será composta por dois modelos de *Deep Learning*, sendo o primeiro para a redução de dimensionalidade e extração de características representativas, e o segundo para realizar a predição da intensidade. Os objetivos específicos incluem: (i) coletar e preparar um conjunto de dados de fala em português brasileiro com anotações emocionais; (ii) desenvolver um modelo de *Deep Learning* capaz de reconhecer e classificar a intensidade da emoção na fala em nosso idioma nativo; (iii) interpretar os resultados obtidos para avaliar o trabalho realizado; e (iv) identificar desafios e propor melhorias para futuras pesquisas na área.

1.2 Estrutura deste Trabalho

Esta dissertação está organizada como se segue. Os demais capítulos, enumerados de 2 a 6, apresentam, respectivamente: Fundamentação Teórica, Trabalhos Relacionados, Pesquisa, Resultados e Conclusões. No capítulo 2, é feita uma fundamentação para oferecer os conceitos necessários para a compreensão do texto; no capítulo 3, são discutidas as comparações com base na literatura relacionada; no capítulo 4, é apresentada a estratégia

empregada nesta pesquisa; no capítulo 5, são trazidos os detalhes da implementação, bem como a discussão dos resultados; por fim, no capítulo 6, são apresentadas as conclusões desta dissertação.

Capítulo 2

Fundamentação Teórica

Neste capítulo serão expostos os conceitos necessários para a compreensão desta dissertação, iniciando-se pelo processo de formação da voz e conceituando formas de como catalogar as emoções (Seção 2.1). Em seguida, é feita uma exposição de conceitos relativos à Aprendizagem de Máquina (Seção 2.2), ao Aprendizado Profundo (Seção 2.3) e às abordagens (Seção 2.4) aplicadas em tarefas que utilizam esse tipo de tecnologia. Por fim, são elicitadas técnicas para aferir a eficiência (Seção 2.5) de tarefas de Aprendizagem de Máquina.

2.1 Intensidade da Emoção na Voz

A voz humana é produzida na laringe. Com a passagem do ar oriundo dos pulmões pelas pregas vocais, estas vibram e geram um som. Com o auxílio de outras estruturas fisiológicas como a língua, a boca e os lábios, esse som é transformado e nossa voz é produzida [23]. Nossa fala não é somente um ato de expressão de ideias e emoções por meio da vocalização [24], como também é um componente indispensável para a comunicação entre os indivíduos de uma sociedade. Enquanto humanos, somos especialistas em voz, e conseguimos extrair uma gama de informações socialmente relevantes [25] dessas ondas sonoras.

Sobre as emoções, temos três modelos bastante consolidados na literatura: Ekman, Russel, e Plutchik. O modelo de Ekman [26] afirma que existem seis emoções básicas: Neutra, raiva, medo, surpresa, alegria e tristeza. O autor afirma que são reconhecidas independentemente do idioma, da cultura ou dos meios de expressão (*i.e.*: fala, expressões faciais, etc.).

O modelo de Russel [27] (Figura 2.1), por sua vez, sugere que as emoções podem ser representadas em um espaço bidimensional, onde o eixo horizontal representa a valência

(positiva ou negativa) e o eixo vertical representa a ativação (alta ou baixa). Já o modelo de Plutchik [2] (Figura 2.2) combina os dois modelos anteriores, criando emoções internas (básicas ou primárias) e externas (compostas ou secundárias).

Plutchik estendeu [2] o modelo de Russel e seu modelo, em formato de cone, dispõe de oito emoções básicas com cores distintas. Neste modelo, a intensidade da emoção fica denotada pela intensidade da cor naquela região, indo do mais intenso (centro) para o menos intenso (borda), por exemplo: em relação ao medo, o terror é mais intenso que a apreensão. As emoções estão dispostas de acordo com seu grau de similaridade: as mais similares estão próximas e as mais antagônicas estão diametralmente opostas. No modelo de Plutchick, as emoções compostas são aquelas formadas por duas emoções básicas.

Uma vez que a intensidade da emoção pode afetar nossa percepção da mesma [8] (*e.g.*: Felicidade pode ser confundida com euforia, que são semelhantes em qualidade de voz mas distintas quanto à intensidade [28]), correlacionar a intensidade da emoção com o volume da voz é demasiada simplificação. A intensidade da emoção não pode ser inferida apenas pela energia na fala [29]. As diferenças entre características acústicas da voz podem ser maiores entre intensidades distintas de uma mesma emoção do que entre emoções diferentes [8].

A intensidade emocional deve ser vista como uma combinação de traço de extroversão e traço de neuroticismo. Assim, o reconhecimento de uma função integradora da intensidade da experiência emocional coloca dificuldades do ponto de vista da medida e da metodologia, uma vez que indivíduos que experimentam emoções positivas intensas também tendem a experimentar emoções negativas intensas [30].

Numa tentativa de resumir a natureza múltipla da intensidade emocional global [31] nos dizem que um dos aspectos mais perceptíveis de uma emoção é a sua intensidade. Quando alguém descreve uma experiência emocional, esta pessoa quase sempre se referirá à sua intensidade. Portanto, seria intrigante que esse aspecto da emoção fosse quase completamente ignorado como um objeto específico de pesquisa. Os autores também propuseram a seguinte fórmula descritiva: $I_E = f(C, E, A_p, A, P, R)$, na qual, I_E denota a intensidade emocional sentida, C a importância dos objetivos ou interesses envolvidos, E a magnitude do acontecimento, A_p uma componente de avaliação, A o potencial de ação, P componentes relevantes de personalidade e R uma componente de contenção, inibição ou regulação.

Encontramos em [32] uma publicação com experimento realizado no Brasil que busca construir um instrumento de estudo de avaliação psicométrica para mensurar afetos positivos e negativos e que confronta seu experimento com outros testes de avaliação de afetos, como a Escala de Afetos Positivos e Negativos (*Positive and Negative Affect Scale*, PANAS) que também foi observada e validada para fins psicométricos no Brasil, em [33]

e [34]. Do ponto de vista do comportamento social, [35] diz que uma representação da ativação e da intensidade do estado emocional parece essencial, mesmo quando a valência não pode ser determinada.

Figura 2.1: Modelo de Russel [1]

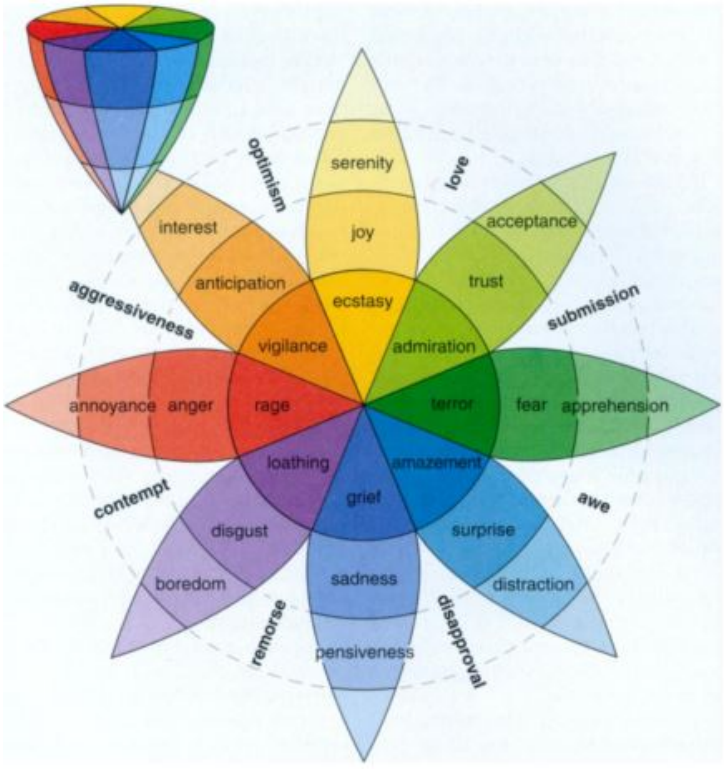

Figura 2.2: Modelo de Plutchik [2]

2.2 Aprendizagem de Máquina

Aprendizagem de Máquina é uma subárea da Inteligência Artificial (*Artificial Intelligence*, AI). A Inteligência Artificial, nome este estabelecido [36] para uma área denominada Inteligência Computacional, consiste no estudo de agentes inteligentes. Um agente é algo capaz de atuar num ambiente, enquanto um agente inteligente é aquele que atua no ambiente de forma inteligente, apropriando-se de circunstâncias para alcançar um objetivo, possivelmente, influenciando o ambiente, com capacidade para alterar esse objetivo, aprendendo com sua experiência e fazendo escolhas adequadas de acordo com suas limitações.

Tarefas de ML costumam ser descritas em termos de como o sistema de Aprendizagem de Máquina deve processar um exemplo [37]. Um exemplo consiste numa coleção de recursos, de um objeto ou evento, que foi aferida quantitativamente e que desejamos que seja processada por esse sistema. Costuma-se representar um exemplo como um vetor $x \in R^n$, onde cada coordenada x_i do vetor x é uma característica (*feature*) desse vetor. Por exemplo, se x representa uma imagem, cada x_i pode ser o valor de um pixel dessa imagem.

Dentre as tarefas de ML, duas tarefas comuns são: a classificação e a regressão. A classificação busca descobrir a qual de k classes possíveis um vetor x pertence, produzindo uma função $f : R^n \rightarrow \{1, ..., k\}$, de modo que quando $y = f(x)$, o modelo atribui a uma entrada (*input*) x uma saída (*output*) numérica de valor y que representa uma categoria (classe). Na regressão, o pensamento é análogo, porém, ao final, o modelo não tenta encontrar a qual classe x pertence, mas predizer um valor (contínuo) para $f(x)$.

Dentre os algoritmos de ML utilizados em trabalhos de SER, encontramos casos de uso de: Floresta Aleatória (*Random Forest*, RF) em [38], Árvore de Decisão (*Decision Tree*, DT) [39], Máquinas de Vetores de Suporte (*Support Vector Machine*, SVM) [40] e K-vizinhos mais próximos (*K-Nearest Neighbors*, KNN) [41].

2.3 Aprendizado Profundo

Aprendizado Profundo é um tipo específico de Aprendizado de Máquina [37]. Algoritmos de ML citados anteriormente - conhecidos como algoritmos tradicionais - costumam funcionar bem em uma grande variedade de problemas importantes. Entretanto, não costumam ter desempenho tão bom em problemas que envolvem reconhecimento de fala ou de objetos. O desenvolvimento do Aprendizado Profundo foi motivado, em parte, pela falha desses algoritmos tradicionais em generalizar bem para essas tarefas de Inteligência Artificial. Dada essa necessidade de explorar modelos mais robustos, alguns trabalhos

estudaram o impacto de algoritmos de *Deep Learning* no reconhecimento de emoção na voz [42, 43].

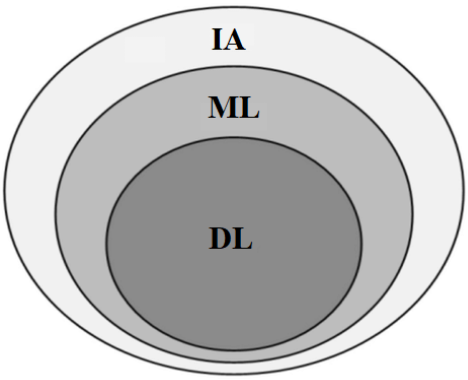

Figura 2.3: Relação entre IA, ML e DL [3]

Neste universo, Redes Multicamadas de Perceptrons (*Multilayer Perceptron*, MLP) são predecessoras da construção de modelos de DL, conhecidos como Redes Neurais (*Neural Networks*, NN). O Perceptron é uma unidade (Figura 2.4) composta por valores (pesos) wi e uma função de ativação f, que recebe as *features*, realiza uma operação matemática entre wi, xi, aplica a função de ativação $y = f(x, w)$ e emite esse resultado como *output*. Podemos realizar o mapeamento de vários Perceptrons, recebendo as *features* e produzindo *outputs*, ao longo de camadas, onde cada camada atua como *input* para a próxima camada e, sucessivamente, até chegar a um *output* final, dessa maneira, formando uma MLP (Figura 2.5).

A camada inicial é chamada de camada de entrada (*input layer*), as camadas intermediárias são chamadas de camadas ocultas (*hidden layers*) e a camada final é chamada de camada de saída (*output layer*). Compreendendo uma rede neural como um conjunto de nós e arestas, cada nó da rede será chamado de neurônio.

2.3.1 Redes Neurais Profundas

Nosso cérebro tem uma grande capacidade de generalização, o que nos ajuda a raciocinar de forma indutiva, sendo o primeiro passo do nosso aprendizado [44]. Redes neurais são capazes de aprender relações não lineares complexas e criar relações entre *input* e *output*, formando sistemas utilizados em várias áreas de ML, e SER não é uma exceção [45].

Com base no conceito e na organização de uma rede neural, o conjunto formado pela disposição dos neurônios, pesos e funções de ativação pode ser organizado de forma a criar diferentes arquiteturas que, ao longo do tempo, se mostraram eficientes para generalizar bem em certas áreas de conhecimento. Diz-se que uma rede neural se torna uma Rede Neural Profunda quando possui grande quantidade de neurônios e camadas ocultas em

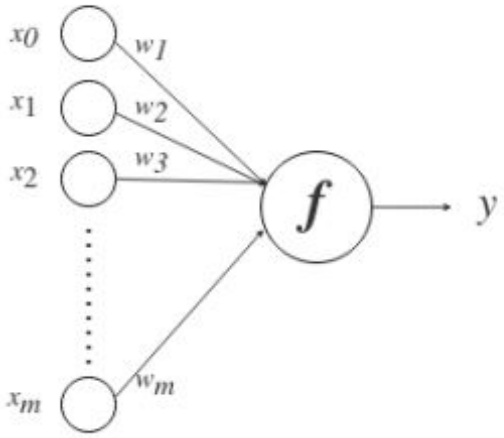

Figura 2.4: Exemplo de Perceptron

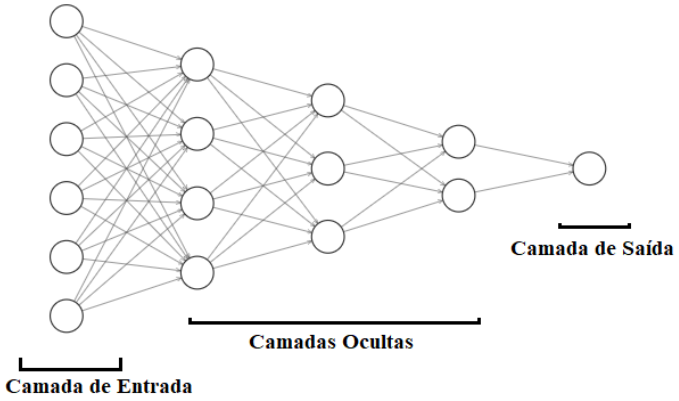

Figura 2.5: Exemplo de arquitetura MLP

sua arquitetura, embora não haja um valor específico que a habilite a se tornar profunda ao superá-lo.

2.4 Abordagens Supervisionada e Não-Supervisionada

No processo de criação de modelos de ML, uma vez definida a arquitetura, a rede irá começar a aprender com os dados de forma iterativa. A etapa na qual o modelo começa a ver os *inputs* e tenta inferir a classe é chamada de treinamento (*training*). Duas abordagens tradicionais para esse momento são: a Aprendizagem Supervisionada (*Supervised Learning*) e a Aprendizagem Não Supervisionada *(Unsupervised Learning*).

2.4.1 Abordagem Supervisionada

Para essa abordagem, faz-se necessário que os dados já estejam rotulados com a classe (*label*) à qual cada registro pertence. Por exemplo, ao utilizar uma base de dados que envolve imagens, pode ser trivial etiquetar os dados com base no objeto que aparece naquela imagem (*e.g.*: ou gato ou cachorro).

Ao lidar com emoções, essa atividade irá demandar pessoas especializadas e capacitadas. Uma pesquisa [44] realizada no ano de 2021 aponta essa dificuldade na área de SER, pois mesmo quando há variedade de conjuntos de dados (*datasets*) diferentes, estes podem apresentar poucas classes, poucos registros por classe ou ambos.

Sejam X, Y, conjuntos de *inputs* e *outputs*, respectivamente, $X = \{x_i, ..., x_n\}$ e $Y = \{y_i, ..., y_n\}$, onde o y_i é a classe de x_i, podemos generalizar a etapa de treinamento como o seguinte fluxo $\forall x_i \in X' \subset X$:

Seja $f : X \to Y$, tal que $f(x) = y$, então,

1. O modelo f é apresentado a um dado x_i e produz um resultado $f(x_i) = y'_i$;
2. É calculado um erro $d_i = d(y_i, y'_i)$ entre o resultado apresentado e o resultado esperado;
3. d_i é utilizado para atualizar os parâmetros de f;
4. O processo é repetido para o próximo exemplo x_{i+1}.

A etapa de treinamento costuma ser seguida pela etapa de testes (*tests*), na qual o modelo é exposto aos $x_j \in (X \setminus X')$ e, a partir daí, são calculadas métricas para aferir o desempenho de f. Cabe observar que f não precisa ser injetora, uma vez que mais de um x_i pode pertencer a mesma classe y_i.

2.4.2 Abordagem Não-Supervisionada

Para uma abordagem não supervisionada, os modelos são apresentados a um conjunto de dados sem rótulos e tentam aprender características importantes da sua estrutura. A distinção entre supervisionado e não supervisionado costuma se dar pela presença, ou não, da classe (*target*) daqueles dados. É importante notar a possibilidade de aplicar uma abordagem não supervisionada a um *dataset* mesmo quando a classe é conhecida, basta excluí-la do processo de treinamento do modelo.

Alguns casos de uso de *Unsupervised Learning* incluem: (1) Formação de conjuntos (*clustering*); (2) Aprendizado de Características (*Feature Learning*); (3) Redução de dimensionalidade; (4) Automatizar a rotulação de amostras; e (5) Aprendizado sobre o *dataset* através de análise exploratória (*Exploratory Data Analysis*, EDA).

Trabalhos envolvendo aprendizado não supervisionado apontam as capacidades de Autocodificadores (*Autoencoder*, AE) tanto para aprendizagem de características (*Feature Learning*) [46, 47] quanto para redução de dimensionalidade [48, 49].

2.5 Métricas de Avaliação de Modelos

Modelos de aprendizagem de máquina precisam ser validados para que possamos avaliar o seu desempenho e gerar inferências a respeito do seu comportamento. Existem diversas formas de avaliar seus resultados, sendo esta mais ou menos adequadas à natureza do modelo. Na lista a seguir temos as variáveis utilizadas para calcular métricas de avaliação do treinamento e teste de modelos de ML:

- Verdadeiros Positivos (VP) ou *True Positive* (TP): Classificação correta da classes positivas;
- Verdadeiros Negativos (VN) ou *True Negative* (TN): Classificação correta da classes negativas;
- Falsos Positivos (FP) ou *False Positive* (FP): Erro onde o modelo previu uma classe positiva, quando o valor real pertencia a classe negativa;
- Falsos Negativos (FN) ou *False Negative* (FN): Erro onde o modelo previu uma classe negativa, quando o valor real pertencia a classe positiva.

Com essas variáveis, podemos calcular as seguintes métricas:

- Acurácia (*Accuracy*): Indica o desempenho geral do modelo.

$$\frac{VP + VN}{VP + FP + FN + FN} \tag{2.1}$$

- Precisão (*Precision*): Dentre as classificações positivas que o modelo fez, quais foram corretas.

$$\frac{VP}{VP + FP} \tag{2.2}$$

- Sensibilidade (*Recall*): Dentre todas as classificações positivas esperadas, quantas foram corretas.

$$\frac{VP}{VP + FN} \tag{2.3}$$

- *F1-Score*: Média harmônica entre precisão e sensibilidade.

$$2 * \frac{precision * recall}{precision + recall} \tag{2.4}$$

2.6 Considerações

Neste capítulo, são abordados os principais conceitos para fornecer subsídios a esta Dissertação. A partir dos conceitos apresentados, foram citados trabalhos relacionados que os utilizam, demonstrando a aderência desse trabalho às práticas utilizadas nesta área de pesquisa. Mais detalhes da estratégia deste trabalho, como aquisição dos dados e arquitetura e dos modelos serão apresentados no Capítulo 4. No capítulo a seguir (3), serão apresentados trabalhos relacionados a este com o intuito de identificar e validar as lacunas deste tema na literatura.

Capítulo 3

Trabalhos Relacionados

Neste capítulo, serão discutidos trabalhos relacionados a esta dissertação. Serão apresentadas suas propostas, características, metodologias, possíveis limitações e possibilidades de melhoria. Posteriormente, será feita uma análise comparativa com esta dissertação, evidenciando pontos em comum e possíveis divergências e, por fim, será apresentado como este trabalho se propõe a colaborar com a produção científica e o estado da arte na área de SER.

No âmbito das Ciências da Computação, o processamento de voz é uma área de pesquisa ativa, com publicações datando desde o final do século XX, podemos citar citar [50], no ano de 1991 e [51], em 1995.

Donn Morrison, Ruili Wang e Liyanage C. De Silva [52], lembram da necessidade de aprimorar a naturalidade das interações humano-computador e da importância de interpretar com precisão informações emocionais, dada a ubiquidade de sistemas automatizados. Além de um *dataset* tradicional (ESMBS), utilizaram uma base de dados fornecida por uma companhia elétrica, composta por ligações de clientes para o *call-center* da empresa, o que significa que não se tratava de um *dataset* simulado ou seminatural, mas, sim, de uma massa de dados com amostras naturais e espontâneas. Afirmam que um modelo de SER pode colaborar com o atendimento aos clientes em chamadas de acordo com a urgência percebida. O que corrobora este trabalho na importância, não só do diagnóstico da emoção, como também da sua intensidade. Caso fosse detectada determinada emoção, o sistema transferiria a ligação para um atendente humano fornecer assistência. Entretanto, mesmo com a combinação de duas massas de dados, o *dataset* final contava com pouco mais de 1000 amostras (1100), sendo 93% pertencente à classe neutra.

Mayank Bhargava e Tim Polzehl [53], se propõem a melhorar o desempenho em tarefas de SER, incorporando mais *features* aos dados de entrada. Apontam que pesquisas da época dependiam fortemente de MFCCs, da tonalidade e da amplitude. Decidiram

utilizar um algoritmo (*Voice Activity Detection*) para separar, em nível de amostra, os momentos com linguagem presente e performar a extração de características separadamente. Validaram sua abordagem com dois modelos, um DNN e um SVM, ao longo de sete classes. Apesar dos modelos, o escopo do trabalho seria o de relacionar a melhora no desempenho dos modelos com a adição ou a combinação de mais *features*, entretanto, observaram que os melhores resultados de índice de acerto foram obtidos em dois cenários: (1) 74.02% utilizando MFCCs e características rítmicas, embora o resultado utilizando apenas *features* rítmicas fosse de apenas 34.6%; (2) 71.93% utilizando apenas MFCCs. Sobre as características rítmicas, ressaltaram que sua baixa contribuição dá-se, possivelmente, em virtude da baixa dimensionalidade, uma vez que é mais de dez vezes menor que a dos MFCCs. Os resultados de [52] foram, de certa forma, contrários à sua hipótese, e reasseguraram a eficiência de utilizar *features* espectrográficas para tarefas de SER.

Xiaoming Zhao e Shiqing Zhang [54], propõem um método para classificação de emoções capaz de aprender com representações de coeficientes mais esparsas, utilizando características prosódicas, espectrais e de qualidade da voz. Uma vez que a localidade dos dados tem sido amplamente utilizada em muitos problemas de reconhecimento de padrões, como *clustering*, redução de dimensionalidade e classificação de imagens. Combinaram técnicas para tentar compor sobre o problema de perda de captura da estrutura dos dados e apontam para o desafio de produzir soluções de SER que atuem em tempo real.

Zhang, S. et al. [55], apontam o distanciamento entre as emoções subjetivas e as características de baixo nível. Observam o bom desempenho de Redes Neurais Convolucionais (CNN) em tarefas de reconhecimento de imagem e detecção de objeto e exploram como utilizar essas redes para tentar encurtar essa distância. De maneira análoga ao que é feito em tarefas que envolvem imagens, extraindo os três canais correspondentes às cores vermelho, verde e azul, este trabalho extrai três canais a partir do *Mel Spectrogram* da amostra. Para realizar a extração de características, utilizaram a AlexNET, uma rede neural pré-treinada para reconhecimento de imagens, para tentar aprender características de alto nível a respeito das vocalizações. Propuseram, então, uma *Discriminant Temporal Pyramid Matching* para agrupar as características aprendidas em uma unidade maior e aplicando um modelo SVM linear a esse resultado. Embora tenham utilizado mais de uma base de dados em sua tarefa, apontam a quantidade limitada de amostras disponíveis para treinamento e afirmam que o modelo AlexNET apresenta um bom desempenho na tarefa de extração de características.

Sefik Emre Eskimez, Zhiyao Duan e Wendi Heinzelman [56], propõem a utilização de modelos não supervisionados do tipo *Autoencoder*, separadamente, para tentar remediar a escassez de dados para tarefas de SER, questionando a viabilidade de aprender características de *datasets* de outros domínios de voz e utilizá-los para treinar modelos

de classificação de emoções. Os modelos não supervisionados incluem *Denoising Autoencoder*, *Variational Autoencoder*, *Adversarial Autoencoder* e *Adversarial Variational Bayes* com adição de ruído. Utilizaram um SVM e uma Rede Neural Convolucional como modelos de base para avaliar o desempenho dos *Autoencoders*. Os modelos receberam *Mel Spectrograms* para efetuar a classificação das emoções e os autores chegaram à conclusão de que os modelos inferenciais (*Variational Autoencoder*, *Adversarial Autoencoder* e *Adversarial Variational Bayes*) obtiveram um desempenho superior na tarefa de aprendizado de características.

Yuanchao Li, Tianyu Zhao e Tatsuya Kawahar [57], propõem um modelo multitarefa, utilizando uma camada de autoatenção (*self attention layer*) para classificar tanto a emoção quanto o sexo da pessoa, apesar das dificuldades em virtude da variabilidade em dados de fala e emoção. Utilizaram os espectrogramas como *input* para um modelo composto por uma Rede Neural Convolucional que alimenta uma rede *Bidirectional Long Short-Term Memory*, seguida pela camada de autoatenção para agregar informações da camada anterior ao longo do tempo.

M. S. Akhtar et al. [58], propõem uma combinação de modelos para a multitarefa de classificar emoção e intensidade em texto. Utilizaram três modelos para extração de características (CNN, *Long Short-term memory* e *Gated Recurrent Unit*) e processamento de linguagem natural para adicionar mais uma dimensão ao vetor de entrada. Observaram que modelos multitarefa costumam ter desempenhos superiores em matéria de generalização. As bases de dados utilizadas apresentavam as emoções como classe, e a intensidade foi calculada a partir da técnica de Vader [59]. Akhtar avaliou tanto o desempenho dos modelos individuais quanto combinações (*ensembles*) para concluir que as combinações propostas apresentaram desempenho superior aos modelos individuais.

Zhu et al. [21], pretendem melhorar o desempenho da síntese de uma vocalização emocional oriunda de um mecanismo de conversão de texto para fala (*speech to text*). Focaram apenas no controle sutil da intensidade da emoção, simplificando o controle da emoção e da intensidade através de um vetor e de um escalar, respectivamente, para que consigam modular com facilidade a intensidade de uma amostra. A base de dados utilizada é composta por várias sentenças que são repetidas, tendo sua classe (emoção) alterada. Embora não possua rótulos (*labels*) para a intensidade, busca calcular uma função de ranqueamento de atributo (*attribute ranking function*) entre pares formados por classes distintas de uma mesma amostra e utilizar esse atributo em combinação com uma amostra de outra classe para produzir a amostra pertencente a esta, de acordo com o atributo. Uma vez calculada a *ranking function*, um algoritmo não supervisionado de redução de dimensionalidade (*Principal Component Analysis*, *PCA*) foi aplicado em uma amostra da base de dados, onde foi observada a coerência entre as classes de dados de um

mesmo conjunto (*cluster*).

Aggelina Chatziagapi et al. [60], propõem um modelo de GAN para atacar o problema de conjuntos de dados desbalanceados em tarefas de SER. Fato observado em [52] que se soma a um ponto descrito nesse trabalho, sobre a discrepância em volume de dados para tarefas de SER com relação a outras tarefas, como reconhecimento de imagens, e mais ainda no contexto de intensidade das emoções, buscando produzir novas amostras de espectrogramas para classes menos representadas. Embora não seja propriamente uma proposta para reconhecimento de emoções ou de intensidade das emoções, [60] utilizou duas bases de dados, IEMOCAP e FEEL-25k, para validar o resultado do seu modelo. Uma vez que os espectrogramas gerados artificialmente eram incluídos ao conjunto dos dados, calculou a distribuição das classes e, então, passou a retirar amostras aleatoriamente, observando que as porcentagens das classes se mantinham equilibradas ao longo das emoções.

Campos e Moutinho [61], propõem a criação de um modelo mais robusto, com uma implementação híbrida. Observaram que a utilização de múltiplas redes neurais de maneira sequencial pode ocasionar a propagação de erros entre os modelos. Assim, o modelo é composto por vários modelos especialistas, que combinam redes neurais convolucionais e redes neurais profundas, treinados de forma supervisionada. Campos e Moutinho treinaram um modelo especialista para cada uma das sete emoções presentes no *dataset* VERBO. Além de reiterar a usabilidade da base de dados, tiveram ganhos em sua taxa média de acerto acima de 10% quando comparado com sua implementação com uma CNN simples.

Neelakshi Josh [62], se propôs a explorar diferentes atributos de um banco de dados em português brasileiro. Procurou explorar características espectrais, prosódicas e temporais numa tarefa de SER, afirmando que a utilização conjunta dessas características pode melhorar o seu índice de acerto. Totalizando 38 atributos em um vetor, treinou quatro algoritmos, utilizando apenas a abordagem supervisionada: SVM, MLP, RF e KNN. Também utilizou o *dataset* VERBO, análogo a este trabalho, colaborando com a evidência de que, apesar de ser o primeiro *dataset* para SER em Português Brasileiro, pode ser utilizado para trabalhos de SER e suas amostras possuem características suficientes para desenvolver tarefas de ML. Embora tenha utilizado quatro modelos distintos, Neelakshi não se aprofundou em demais arquiteturas de DL.

Abbaschian, Sierra-Sosa e Elmaghraby [44], revisaram publicações que envolvem trabalhos de DL para SER, bem como as bases de dados utilizadas. Realizaram um comparativo entre 11 bases de dados amplamente utilizadas em trabalhos dessa natureza, dispostas em três categorias: Simuladas, seminaturais e naturais. Realizaram um cruzamento entre 25 bases de dados e trabalhos que as utilizaram, organizando essa relação

ao longo do tempo, disposta num eixo horizontal que compreende os anos entre 2005 e 2020. Nessa relação, podemos perceber a predominância de certos entes em determinados períodos do tempo, tanto com relação a *datasets* quanto a modelos. Quanto às bases de dados, a partir de 2018, percebemos a predominância do IEMOCAP, sendo utilizado o dobro de vezes em relação ao EMO-DB, o segundo mais utilizado. Quanto aos modelos, percebemos a presença constante de CNNs, a partir de 2016, e de variações de *Autoencoders*, a partir de 2018, ambos permanecendo presentes até 2020, ano final da cobertura da pesquisa.

Kun Zhou et al. [63], comentaram a parca presença de estudos relativos à intensidade da emoção, uma vez que têm um grande potencial para a conversão de voz. O trabalho consistiu em conseguir controlar a intensidade da emoção em um modelo de sequência para sequência que converte texto para voz. De maneira análoga a [21], Zhou et al. também decidiram encontrar uma *attribute ranking function* para operar com o dado de entrada e alterar a intensidade da emoção da maneira desejada. Kun [63] e Zhu [21] diferem na sua arquitetura: enquanto [21] utilizaram um *Autoencoder* que recebia o escalar relativo a intensidade após a fase de *encoding*, [63] utilizaram três modelos com funções distintas: (1) Extrair as *features* do *input*; (2) Adicionar a emoção desejada para a saída; (3) Adicionar a alteração na intensidade.

Gonçalves et. al. [64], descrevem um desafio voltado ao reconhecimento de emoções em discursos espontâneos, utilizando um *dataset* com 237 horas de dados anotados. O desafio inclui duas tarefas: reconhecimento de emoções categóricas (oito classes emocionais) e predição de atributos emocionais (excitação, valência e dominância). Os resultados apresentados referem-se ao desempenho de modelos de referência (*baseline*), que servem como ponto de partida para participantes desenvolverem metodologias inovadoras. Entre os resultados apresentados, o melhor desempenho para detecção de excitação (*arousal*) em dados de teste foi obtido pelo seu próprio modelo *baseline*, atingindo 53% da métrica de desempenho adotada (*concordance correlation coefficient*).

Também ressaltaram os desafios relativos a tarefas de SER em que mesmo os modelos de melhor desempenho ainda têm dificuldades para lidar com a complexidade inerente à fala espontânea, na qual as expressões emocionais são sutis e dependentes do contexto, sugerindo que os modelos são mais bem ajustados para capturar expressões emocionais médias e superiores em vez das extremas e mais baixas, sinalizando uma direção para o refinamento futuro nesta área, assim como, Sonia Xylina Mashal e Kavita Asnani [65], quando observam que, embora sejam desenvolvidos trabalhos na área de reconhecimento de emoções, o próximo passo seria determinar a intensidade dessa emoção.

3.1 Considerações

Com base nos trabalhos relacionados, a Tabela 3.1 apresenta um comparativo entre esta dissertação e os demais encontrados na literatura e citados anteriormente, apontando a abordagem empregada em cada um dos trabalhos, o objeto de classificação (emoção ou intensidade) e a presença do nosso idioma.

Nos trabalhos citados foi observado que as dificuldades relativas à massa de dados não é um problema exclusivo de trabalhos em SER, uma vez que os trabalhos que lidam com dados textuais também reiteram essa queixa, sejam os que lidam puramente com textos ou os que realizam tarefas de *text-to-speach* com emoção.

No ano de 2007, [52] apontaram para a ubiquidade de sistemas automatizados, que se mostra ainda mais presente contemporaneamente. Ao efetuar um comparativo (Tabela 3.1) entre trabalhos com tarefas de reconhecimento de emoções na fala, notamos que, apesar de [65] afirmar que a atividade de classificar emoções poderia ter um papel adicional como catalisador para trabalhos envolvendo a intensidade da emoção, no entanto, a incidência de trabalhos nessa área específica parece baixa.

A subjetividade das emoções demanda cuidado e atuação profissional na composição de bancos de dados para esse tipo de tarefa. Ao reduzir o escopo da pesquisa para trabalhos com dados em português, a literatura se torna mais escassa, haja vista a publicação do VERBO apenas no ano de 2018.

Características espectrais aparentam ser uma característica importante para os modelos, carregando grande valor informacional sobre a amostra. Observa-se, em [53], que a ausência de MFCCs acarretou uma queda superior a 50% no desempenho da classificação, enquanto que [56] os utilizou como *target* para redes generativas em sua busca para aumentar a quantidade de amostras disponíveis para treinamento de modelos.

Outro ponto a ser observado é que, embora existam trabalhos utilizando mais de um *dataset*, essa prática ainda não parece estar totalmente difundida entre as publicações de SER, uma vez que os escopos e naturezas (simulada, seminatural ou natural) das bases de dados costumam ser distintas.

Analisando as informações expostas na Tabela 3.1, nota-se que este trabalho se aproxima dos demais por utilizar características espectrais e envolver uma abordagem supervisionada. Também tem em comum o fato de utilizar técnicas consolidadas ("clássicas") de ML, ao mesmo tempo que investiga arquiteturas de DL, como DNNs e AEs.

Há um distanciamento deste trabalho frente aos demais quanto a trabalhar com áudios para a língua portuguesa. Além disso, a abordagem não supervisionada também foi utilizada, o que não é tão comum em uma mesma tarefa de SER. Outro ponto de inovação se dá uma vez que os trabalhos que encontramos utilizando o *dataset* VERBO utilizaram apenas Aprendizagem Supervisionada e não lidam com a intensidade da emoção. Ade-

Tabela 3.1: Comparativo entre este trabalho e a literatura correlata

Trabalho	Aprendizagem Supervisionada	Aprendizagem Não Supervisionada	*Machine Learning*	*Deep Learning*	Português	Emoção	Intensidade
[55]	x		x			x	
[56]	x	x		x		x	
[57]	x			x		x	
[61]	x				x	x	
[62]	x		x	x	x	x	
[66]	x	x	x	x		x	
[64]	x			x		x	x
Dissertação	x	x	x	x	x		x

mais, não foram encontrados trabalhos utilizando o *dataset* VIVAE (2020), possivelmente tornando este trabalho pioneiro.

Capítulo 4

Classificação de Intensidade das Emoções na Fala em Português Brasileiro por meio de Deep Learning

Este capítulo apresenta uma arquitetura para classificação da intensidade da emoção na voz em Português. Para tal, foram criados dois modelos de DL, a saber: (i) *Autoencoder*, responsável pela redução de dimensionalidade e extração de características representativas dos dados; (ii) Rede Neural, para efetuar a predição da classe relativa à intensidade da vocalização.

4.1 Visão Geral

Na Figura 4.1, é apresentada uma visão geral da arquitetura. Conforme a imagem, três etapas principais serão necessárias para o reconhecimento da intensidade das emoções: (A) Aquisição de informações; (B) Extração de características; (C) Classificação da intensidade.

A primeira etapa (A) lida com a obtenção dos dados que serão utilizados na dissertação e com a sua conversão para uma interpretação possível de ser utilizada por modelos de aprendizagem de máquina. Podemos descrever os dados como um conjunto de registros rotulados que serão utilizados para treinamento e teste dos modelos implementados na proposta.

A segunda etapa (B) lida com a extração de características dos dados convertidos. Essas características serão obtidas através de um modelo não supervisionado para a re-

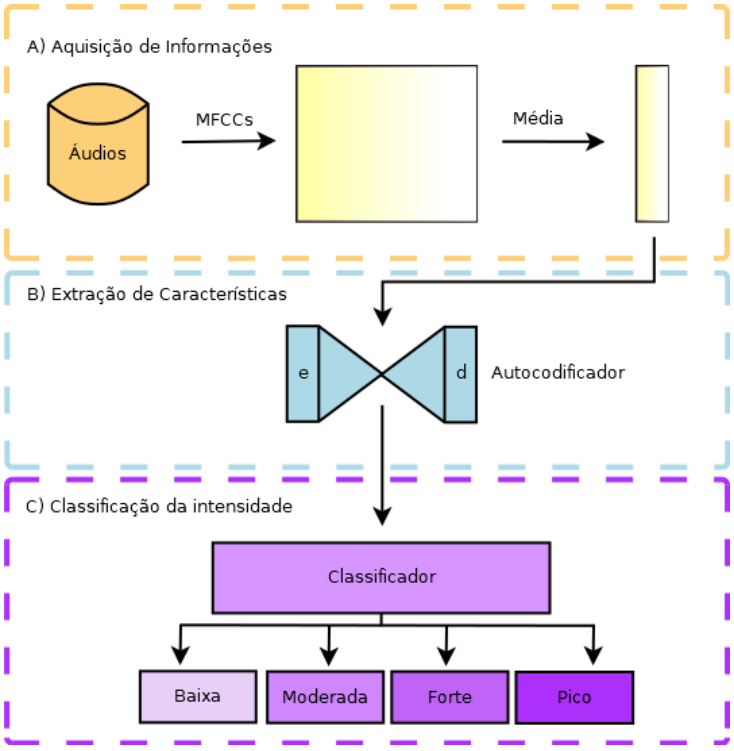

Figura 4.1: Visão geral da arquitetura

dução de dimensionalidade e, posteriormente, utilizadas como entrada de um modelo supervisionado de classificação da intensidade da emoção.

A terceira e última etapa (C) é responsável pela inferência da intensidade. O modelo recebe as características obtidas na etapa anterior e realiza o treinamento e testagem do modelo de classificação de acordo com quatro classes possíveis: (i) Baixa; (ii) Moderada; (iii) Forte; (iv) Pico de intensidade.

Este trabalho aplicará esta arquitetura em dois cenários modelados a serem detalhados no capítulo posterior (5), juntamente com os resultados obtidos.

4.2 Aquisição de informações da base de conhecimento

Para modelar a intensidade da emoção, uma das dificuldades é a falta de dados rotulados [63]. Tradicionalmente, em áreas como visão computacional ou reconhecimento de voz, os *datasets* chegam a ter milhões de registros, como, por exemplo: ImageNet (imagem) com mais de 14 milhões e Google AudioSet (áudio) com mais de 2 milhões de amostras. Observa-se um comparativo na Tabela 4.1 entre *datasets* populares [44] em trabalhos de SER: AudioSet, *Berlin Database of Emotional Speech* (EMO-DB) [67], *Danish emotional speech database* (DES) [68], *The Ryerson Audio-Visual Database of Emotional Speech*

Tabela 4.1: Comparativos entre *datasets* populares para SER

Base de dados	Quantidade de amostras	Duração Média (s)	Português?
DES	210	2,7	Não
EMO-DB	700	2,8	Não
RAVDESS	2496	3,7	Não
TESS	2800	2,1	Não
CREMA-D	7442	2,5	Não

and Song (RAVDESS) [69], *Toronto Emotional Speech Set* (TESS) [70] e *Crowd-Sourced Emotional Multimodal Actors Dataset* (CREMA-D) [71].

Não obstante, nenhum destes apresenta a intensidade da emoção ou é em português - língua falada pela sexta[1] maior população e nona[2] maior economia do mundo - como também estão distantes do AudioSet, tanto em quantidade de amostras ($> 2.000.000$) quanto em duração média ($\approx 10s$).

Neste trabalho, serão utilizados dois conjuntos de dados, VERBO e VIVAE, que satisfazem requisitos de ser em português e apresentar emoções catalogadas (VERBO); além de ter rótulos para emoções e intensidades (VIVAE). A serem detalhados nas subseções 4.2.2 e 4.2.3, respectivamente.

Os *datasets* da Tabela 4.1 são simulados (*simulated*), ou seja, os áudios são gravados a partir de pessoas treinadas, realizando a leitura de um texto e interpretando com emoções diferentes. Existem, também, *datasets* seminaturais (*semi-natural*) compostos tanto por atores como por pessoas comuns lendo um roteiro, assim como os ditos naturais (*natural*), com áudios extraídos de programas de TV, centrais telefônicas, vídeos da internet e outros meios.

Também é comum que as amostras não apresentem ruídos ou alguma poluição sonora, o que as distancia de situações reais. Sistemas treinados nesses *datasets* podem não ser bem sucedidos em situações reais [44]. Há também *datasets* gerados a partir da participação de um usuário ou cliente de algum serviço, entretanto, este cliente é informado da gravação, o que pode comprometer a qualidade do dado.

Outro fator é o efeito da cultura e da linguagem, já que ambos podem afetar a percepção do sentimento na fala [44]. A incerteza na anotação (categorização dos dados) representa mais um desafio para *datasets* de SER, uma vez que, num discurso emocional, um participante pode rotular um enunciado como eufórico e outro como raivoso. Essa subjetividade torna a tarefa mais complexa e pode limitar a possibilidade de combinar os bancos de dados para criar superconjuntos de dados emocionais.

[1]Disponível em `https://brasilescola.uol.com.br/geografia/populacao-mundial.htm`

[2]Disponível em `https://www.gov.br/funag/pt-br/ipri/publicacoes/estatisticas/as-15-maiores-economias-do-mundo`

4.2.1 Fusão de Domínios

Dado esse cenário, pode-se utilizar uma técnica que permite fundir o conhecimento de vários conjuntos de dados organicamente para uma tarefa de aprendizado de máquina. A Fusão de Domínios [25] (*Domain Fusion*) é uma técnica para aproveitar mais bases de dados e produzir informações mais robustas e úteis do que as fornecidas por uma única fonte de dados individualmente. Um exemplo de utilização da Fusão de Domínios pode ser observado em [11] para tentar melhorar a generalização do seu modelo, tendo percebido uma melhora do desempenho em dados inéditos (distintos das amostras de treinamento e teste).

Uma das metodologias da Fusão de Domínios é a Fusão de Dados Baseada em Aprendizado de Transferência (*Transfer Learning-Based Data Fusion*). Uma das possibilidades que essa metodologia aborda compreende a fusão de bases de dados de natureza semelhante (*Transductive Learning*) quando a tarefa é a mesma, mas o domínio (ponto de partida) e o contra domínio (ponto de chegada) são distintos, como no nosso caso, em que partimos da voz para chegar na intensidade da emoção. Por mais que estejam relacionados, são distintos. Por exemplo: em uma tarefa de predição de tráfego urbano, pode-se utilizar os dados da cidade A para tentar uma previsão sobre o tráfego na cidade B, caso os dados sobre B sejam limitados.

Uma vez compreendido o processo de formação da fala, o seu emprego na transmissão de emoções, as formas de categorizá-las e como se dá uma tarefa que envolve algoritmos de DL, faz-se necessário conseguir uma massa de dados que possa ser utilizada para esta atividade. Para isso, 4.2.2 e 4.2.3 satisfazem nossa necessidade.

4.2.2 *VERBO*

Voice Emotion Recognition Database in Portuguese Language[3] [72] é uma base de dados com 1176 arquivos em formato *.wav*, publicada em 2018, criada no Instituto de Matemática e Ciências da Computação da Universidade de São Paulo, ICMC-USP, formada por arquivos de áudio na língua portuguesa do Brasil, rotulados com emoções. É o primeiro [62] *dataset* para SER em português do Brasil.

Os áudios têm duração entre 2 e 5 segundos, gravados por 12 atores brasileiros - seis homens e seis mulheres - de diferentes idades e regiões do país. Compreende 14 enunciados (*utterances*) validados por um profissional linguístico, acomodando todos os fonemas da língua portuguesa. Possui exemplos das seis emoções básicas propostas por Russel: (1) Alegria; (2) Nojo; (3) Medo; (4) Raiva; (5) Surpresa; (6) Tristeza. Por fim, foi adicionado

[3]Disponível em `https://sites.google.com/view/verbodatabase/`

Tabela 4.2: Distribuição por classe das 1167 sentenças do *dataset* VERBO

Classe	Total
Raiva	167
Nojo	167
Medo	166
Alegria	166
Tristeza	167
Surpresa	167
Neutro	167

Tabela 4.3: Distribuição por intensidade das 1085 sentenças do *dataset* VIVAE

Intensidade	Total
Baixa	262
Moderada	269
Forte	272
Pico	282

um sétimo estado emocional, denominado de (7) Neutro. A distribuição das classes pode ser observada na Tabela 4.2.

4.2.3 *VIVAE*

Variably Intense Vocalizations of Affect and Emotion Corpus[4] [35] é uma base de dados com 1085 arquivos em formato *.wav*, publicada em 2020, criada por pesquisadores alemães e estadunidenses, formada por vocalizações não verbais. Os áudios foram gravados por 11 pessoas, compreendendo três sentimentos positivos e três negativos de duração média de aproximadamente um segundo, Os positivos são: satisfação (*achievement/triumph*), prazer sexual (*sexual pleasure*) e surpresa (*positive surprise*). Os negativos são: raiva (*anger*), medo (*fear*) e dor física (*physical pain*).

Todos foram gravados com a intensidade variando entre baixa, moderada, forte e pico de emoção. A distribuição das intensidades pode ser observada na Tabela 4.3 e a das classes na Tabela 4.4.

4.2.4 Comparativo entre as bases de dados

Fazendo uma intersecção entre as classes dos *datasets*, conforme Tabela 4.5, observamos apenas quatro classes em comum: alegria, medo, raiva e surpresa. Realizando uma contagem das classes em comum, em ambas as bases de dados (Tabela 4.6), teremos 1364 amostras, o que representa uma quantidade maior do que alguns dos *datasets* em 4.1.

[4]Disponível em `https://zenodo.org/records/4066235#.YWWe4ZpByUk`

Tabela 4.4: Distribuição por classe das 1085 sentenças do *dataset* VIVAE

Sentimento	Intensidade	Quantidade	Total
Satisfação	*Low*	5	16
	Moderate	3	
	Strong	5	
	Peak	3	
Raiva	Low	4	14
	Moderate	4	
	Strong	2	
	Peak	4	
Medo	*Low*	4	16
	Moderate	4	
	Strong	3	
	Peak	5	
Dor	*Low*	3	17
	Moderate	5	
	Strong	5	
	Peak	4	
Prazer	*Low*	5	19
	Moderate	4	
	Strong	5	
	Peak	5	
Surpresa	*Low*	4	13
	Moderate	3	
	Strong	2	
	Peak	4	

Tabela 4.5: Comparativo das emoções presentes em VERBO e VIVAE

Emoção (Português / Inglês)	VERBO	VIVAE	Comum
- / *Pain*	Não	Sim	
- / *Pleasure*	Não	Sim	
Alegria / *Achievement*	Sim	Sim	**X**
Medo / *Fear*	Sim	Sim	**X**
Neutro / -	Sim	Não	
Nojo / –	Sim	Não	
Raiva / *Anger*	Sim	Sim	**X**
Surpresa / *Surprise*	Sim	Sim	**X**
Tristeza / -	Sim	Não	

Tabela 4.6: Total de amostras por classe em comum utilizando VERBO e VIVAE

Sentimento	Base de dados		Total
	VERBO	VIVAE	
Alegria (*Achievement*)	166	161	327
Medo (*Fear*)	166	176	342
Raiva (*Anger*)	167	174	341
Surpresa (*Surprise*)	167	187	354

Das amostras compreendidas pela combinação das bases de dados, 698 delas pertencem ao VIVAE, o que significa que aproximadamente 51% do nosso *dataset* tem, além das classes para emoções, classes para a intensidade.

4.2.5 Processamento de Dados

Para realizar tarefas de ML a partir de arquivos de áudio, faz-se necessário convertê-los para uma forma passível de ingestão pelo modelo. Os arquivos das bases de dados estão em formato *.wav* (encurtamento de *WAVEform*), que não realiza compressão do som digital, mantendo-o o mais próximo da expressão do som natural. Então, precisamos de uma forma de transformar os dados em uma representação que preserve suas características.

Compreende-se um sinal como a variação de uma quantidade ao longo do tempo, no caso da voz, a variação da pressão do ar. Amostras da pressão do ar são aferidas ao longo do tempo com uma determinada frequência, e tem-se um sinal unidimensional, ou seja, com uma única variável, a amplitude do som distribuída ao longo do tempo (Figura 4.2). No ato da fala, as pregas vocais oscilam um número de ciclos de acordo com o seu comprimento, tamanho da massa de vibração envolvida e tensão. Essa oscilação também pode ser aferida uma quantidade de vezes por segundo, obtendo-se sua frequência.

Para transportar um sinal do domínio do tempo para o domínio da frequência, utiliza-se a Transformada de Fourier (*Transformada de Fourier*, TF), que irá decompor o sinal em seus componentes de frequências, realizada computacionalmente através da Transformada Rápida de Fourier (*Transformada Rápida de Fourier*, TRF).

Com o resultado da TRF de uma amostra, calcula-se o seu espectrograma (*spectrogram*): uma representação da densidade das frequências ao longo do tempo. Entretanto, como humanos não compreendem todo o espectro sonoro [73], aplica-se uma normalização às frequências e calcula-se o Espectrograma de Mel (*Mel-Spectrogram*).

A normalização em questão é a feita através da Escala de Mel (*Mel Scale*), construída para tornar tons equidistantes perceptivelmente equidistantes ao ouvido humano. A escala Mel é uma escala perceptual utilizada para modelar como os humanos percebem diferenças tonais, enfatizando a sensibilidade (não linear) do ouvido humano a diferentes

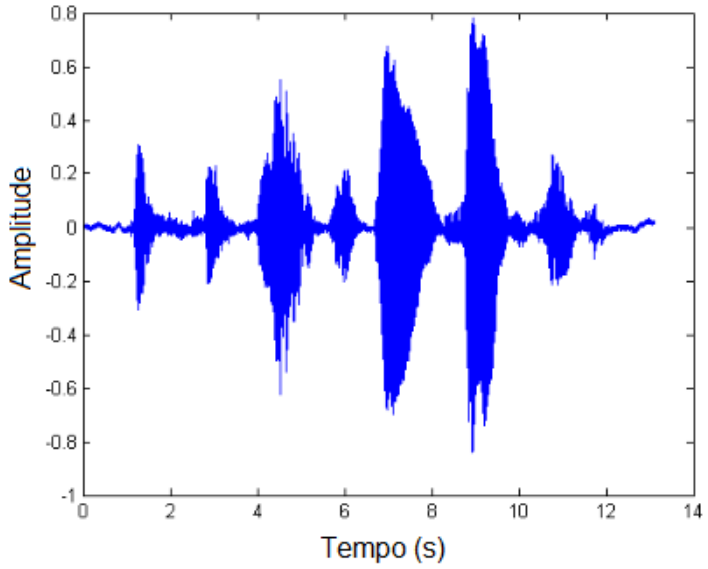

Figura 4.2: Exemplo de visualização de sinal sonoro, medindo a amplitude ao longo do tempo

frequências. Ao lidar com espectrogramas, a escala Mel é aplicada para transformar a frequência em uma escala mais alinhada à percepção auditiva humana. Isso é realizado utilizando um banco de filtros que distribui os filtros de forma mais densa em frequências mais baixas e mais espaçada em frequências mais altas, refletindo a maior sensibilidade do ouvido a mudanças tonais em frequências mais baixas. Ao aplicar a escala Mel aos espectrogramas, as características são extraídas de forma a destacar informações perceptualmente relevantes, tornando-a especialmente útil em tarefas de processamento de áudio. O *Mel-Spectrogram* tem sido amplamente utilizado em problemas de SER, como [55, 74], e em [54] e [52], que também utilizam técnicas de *Autoencoder*. A Escala de Mel é dada por:

$$m(f) = 1127 * \log_e (1 + \frac{f}{700}) \tag{4.1}$$

Outro atributo observado na literatura são os Coeficientes Cepstrais de Frequência Mel (MFCCs). Um MFC é uma representação de curto prazo do espectro de potência de um som, assim, MFCs são os coeficientes que formam os MFCCs coletivamente. Calcular os MFCCs consiste em aplicar a Transformada Discreta do Cosseno ao *Mel-Spectrogram*. MFCCs podem ser compreendidos como uma compressão [75] do *Mel Spectrogram*. São encontramos trabalhos que utilizam MFCCs, [76] e [60], cujas arquiteturas contém uma Rede Neural Convolucional e uma GAN, respectivamente. A relevância desse atributo também é encontrada em trabalhos, como [53], que explorou a combinação de caracterís-

Tabela 4.7: Atributos dos datasets VERBO, VIVAE, ideal e da fusão de domínios

	Datasets			
Atributos	VERBO	VIVAE	Ideal	*Data Fusion*(VERBO,VIVAE)
Idioma	X		X	X
Emoções	X	X	X	X
Intensidade		X	X	X

ticas espectrais, tonais e rítmicas para observar sua contribuição no desempenho de seus modelos de classificação, concluindo que a utilização apenas de MFCCs gerou resultados quase duas vezes superiores à combinação de mais características.

Enquanto o *dataset* VERBO é constituído por pares {amostra, classe}, o *dataset* VIVAE é constituído por pares {amostra, classe, intensidade}. Nestes, as classes são a emoção atribuída àquela amostra e a intensidade é o rótulo da intensidade da classe daquela amostra. Vamos denotar as amostras do VERBO por X_{VERBO} e do VIVAE por X_{VIVAE}, e em virtude da diferença da duração dos áudios entre VERBO e VIVAE, visando reduzir a quantidade de variáveis no problema, estes sinais foram empilhados horizontalmente uma quantidade inteira de vezes para tentar alcançar a maior duração (≈ 5 segundos) entre as amostras e o tempo restante foi preenchido com zeros.

Seja Y_{VERBO} o conjunto das classes (emoções) $y_i \forall x_i \in X_{VERBO}$, analogamente para Y_{VIVAE}, definimos $Y = Y_{VERBO} \bigcap Y_{VIVAE}$ e por Z o conjunto das intensidades. Uma vez que as intensidades estão apenas presentes em VIVAE, serão utilizadas suas quatro classes (baixa, moderada, forte, pico). Sabemos das Tabelas 4.3 e 4.6 que:

- $Y = \{alegria, medo, raiva, surpresa\}$
- $Z = \{baixa, moderada, forte, pico\}$

Vamos redefinir $X_{VERBO} = \{x_i \mid y_i \in Y\}$, analogamente para X_{VIVAE}, e vamos definir nosso domínio $X = X_{VERBO} \bigcap X_{VIVAE}$, assim:

- $\forall x_i \in X, \exists y_i \in Y$ tal que y_i é a classe de x_i
- $\forall x_j \in X_{VIVAE}, \exists z_j \in Z$ tal que z_j é a intensidade de x_j

Ao realizar tarefas de DL, os dados de entrada serão transformados em um formato passível de ingestão pelos modelos. Os arquivos *.wav* de X serão lidos e serão gerados seus respectivos MFCCs, convertendo o sinal de áudio ao longo do tempo para uma representação do sinal no domínio da frequência a ser utilizada pelos modelos. Então, tem-se um mapa $M(x_i) : MFCCs_i$.

4.3 Extração de características

De posse do mapa $M(x_i)$, seguimos para a etapa de extração de características. Uma vez que X é formado por dados de *datasets* diferentes, busca-se uma forma de extrair características relevantes com boa capacidade de generalização ao ser aplicada em ambos X_{VERBO} e X_{VIVAE}.

4.3.1 Codificador Automático

Um Codificador Automático (AE) é uma rede neural criada para tentar reproduzir o seu *input* no seu *output*. Pode ser descrito como um conjunto de funções f, f' de modo que, dado um *input* x, queremos $f'(f(x)) = x' \approx x$, onde f realiza a codificação (*encoding*) de x e f' realiza a decodificação (*decoding*) do resultado $f(x)$. Assim, um *Autoencoder* é uma rede neural composta por um *encoder* e um *decoder* que tenta reproduzir uma função de identidade. Modelos do tipo AE são encontrados em trabalhos como [56] que os utilizaram, juntamente com características espectrais, para aprender *features* de *datasets* visando remediar a escassez de dados.

O resultado da etapa de *encoding* costuma ter uma dimensionalidade menor do que a do dado de entrada. O espaço composto por dados codificados (*encoded*) é chamado de Espaço Latente (*Latent Space*), um espaço composto por representações significativas dos dados, contendo informações sobre cada amostra que, possivelmente não estariam visíveis nas representações de alta dimensionalidade [77].

*Autoencoder*s não podem simplesmente aprender a generalizar $f'(f(x)) = x$ para todo tipo de dado. Ao invés disso, são impostas restrições para que consiga esse tipo de generalização apenas para os dados relevantes à sua tarefa, forçando o modelo a priorizar características que devem ser aprendidas.

Tradicionalmente, o *Autoencoder* pode ser utilizado para reduzir a dimensionalidade de dados ou para aprender características (*feature learning*) sobre os dados. Atualmente, AEs são bastante utilizados como peça fundamental em Redes Generativas Adversariais (GAN) e *Autoencoder*s Variacionais (VAE).

Redes neurais constituem uma boa ferramenta para tarefas de SER. AEs, podem criar uma representação de qualidade com dimensionalidade reduzida em seu espaço latente, enquanto DNNs encontram espaço na literatura como bons discriminadores ou classificadores.

Será construído um AE, ilustrado na Figura 4.3, que tenta reproduzir uma função identidade, composto por uma função *encoder* (f_e) e uma função *decoder* (f_d), de modo que $AE : M \rightarrow M'$ faça:

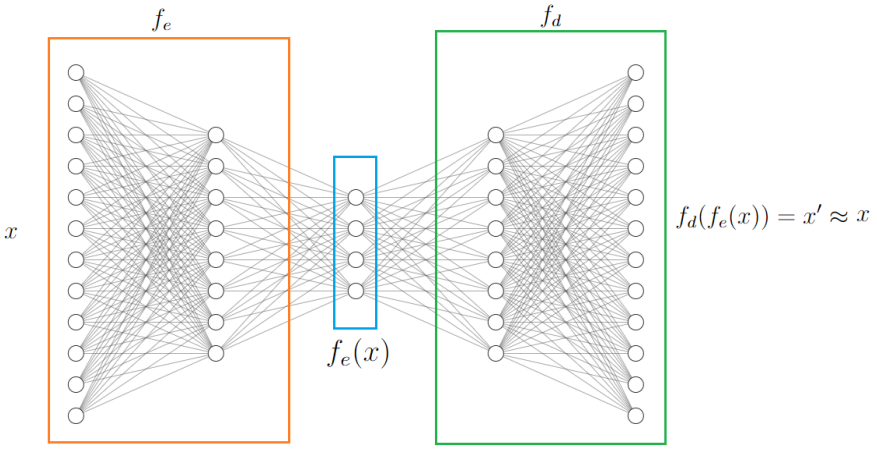

Figura 4.3: Composição do AE

$$AE(x) = f_d(f_e(x)) = x' \approx x \tag{4.2}$$

Para o treinamento do modelo AE com os dados de X, este será dividido em um conjunto de treinamento e outro de teste, contendo 75% e 25% dos registros de X, respectivamente. Em seu espaço latente, a representação do dado de entrada com a dimensionalidade reduzida, preserva suas características de maneira suficiente para que possa ser reconstruído (x') ao aplicar a função de *decoding*.

A estrutura dos *Autoencoder*s é análoga para ambos os experimentos. A Figura 4.4 representa a arquitetura do modelo para 64 MFCCs enquanto a 4.5 representa a arquitetura do modelo para 128 MFCCs. Ambos seguem uma estrutura composta por três camadas, a serem detalhadas abaixo: entrada, *encoder* e *decoder*.

O treinamento foi realizado utilizando o Erro Quadrático Médio (*Mean Squared Error*, MSE) como função de Perda (*Loss*) e a Estimativa de Momento Adaptativo (*Adaptative Moment Estimation*, ADAM) como função de otimização. Para efeitos de legibilidade, será definido dim_{MFCCs} como a quantidade de MFCCs utilizada em cada experimento.

1. Entrada: Responsável por receber os dados com dimensão igual $(1, dim_{MFCCs})$.
2. *Encoder*: Responsável por realizar a redução de dimensionalidade do dado oriundo da camada de entrada. Na camada de *encoder*, o dado entra com dimensão $(1, dim_{MFCCs})$ e é comprimido para uma dimensionalidade $(1, dim_{MFCCs}/2)$. Assumimos que buscamos preservar apenas os atributos com maior relevância, então utilizamos uma função de ativação do tipo Unidade Linear Retificada (*Rectified Linear Unit*, ReLU).

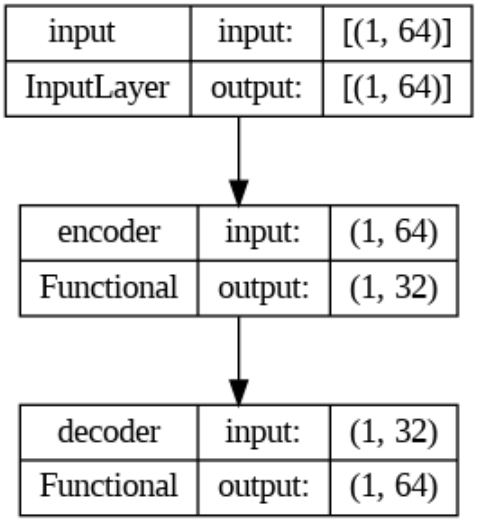

Figura 4.4: Arquitetura do *Autoencoder* para 64 MFCCs

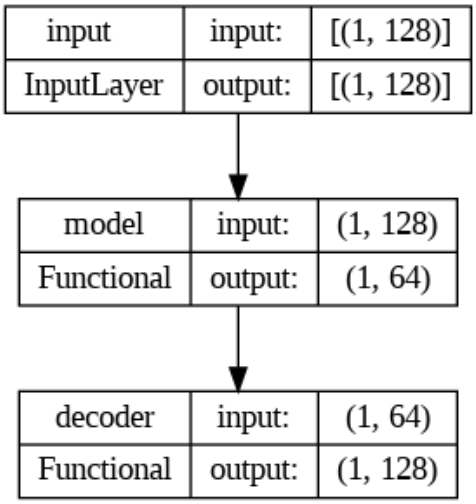

Figura 4.5: Arquitetura do *Autoencoder* para 128 MFCCs

3. *Decoder*: Responsável por receber os dados com dimensão reduzida e expandir-los para a dimensão de entrada $(1, dim_{MFCCs})$, tentando reproduzir os valores originais. Para isso, esta camada utilizou uma função de ativação do tipo Linear, que reproduzirá valores contínuos oriundos das operações matriciais desta camada, não limitados ao intervalo $(0, max(0, x)$ como é o caso da ReLU.

4.4 Classificação da Intensidade

Redes neurais são capazes de aprender relações complexas entre as características dos dados, o que pode ser especialmente benéfico para a tarefa de classificação. Ao contrário de métodos mais tradicionais, as redes neurais não estão limitadas por suposições lineares ou por características específicas pré-definidas, permitindo uma modelagem mais flexível e adaptativa dos dados.

Com o *Autoencoder* treinado, este será aplicado aos registros de X_{VIVAE}, separados em dois conjuntos para treinamento ($X_{VIVAE_{treino}}$) e testes ($X_{VIVAE_{teste}}$), contendo 75% e 25% desses dados, respectivamente. Estes serão os dados utilizados pelo nosso modelo supervisionado.

Este modelo (Figura 4.6) foi construído para classificar a intensidade da emoção de forma direta. Uma vez que os $x_i \in X_{VIVAE_{treino}}$ têm correspondentes em Z, contradomínio das intensidades, então, $j : M \to Z$, é tal que $j(f_e(m)) = z$

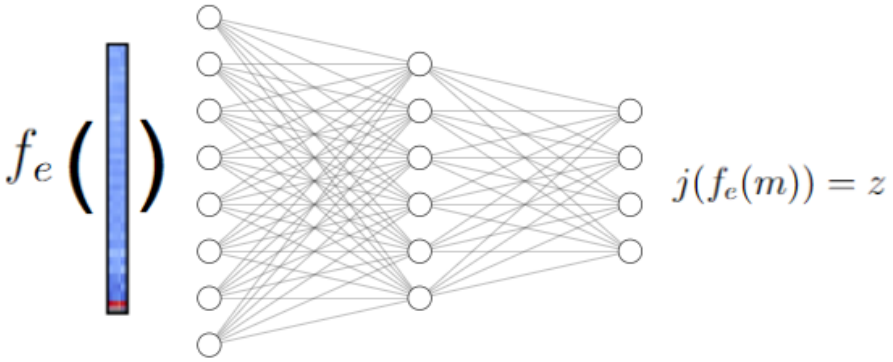

Figura 4.6: Modelo supervisionado j

A estrutura dos Classificadores é análoga para ambos os experimentos. A Figura 4.7 representa a arquitetura do modelo para 64 MFCCs, enquanto a 4.8 representa a arquitetura do modelo para 128 MFCCs. Ambos seguem uma estrutura composta por cinco camadas, a serem detalhadas a seguir: entrada, três camadas densas intermediárias e uma camada densa de saída.

O treinamento foi realizado utilizando *Categorical Cross Entropy* como função de *Loss* e ADAM como função de otimização. Uma vez que o AE realiza a função de reduzir a dimensionalidade dos dados que serão utilizados como *input* para os classificadores, para efeitos de legibilidade, será definido dim_{encode} como a dimensão do dado comprimido obtido através do *Autoencoder* da etapa anterior, sendo estas, 32 e 64, respectivamente.

Neste trabalho, perceberemos um erro de falso negativo de maneira equivalente a um erro de falso positivo. Portanto, não buscamos otimizar as métricas como *Precision* ou

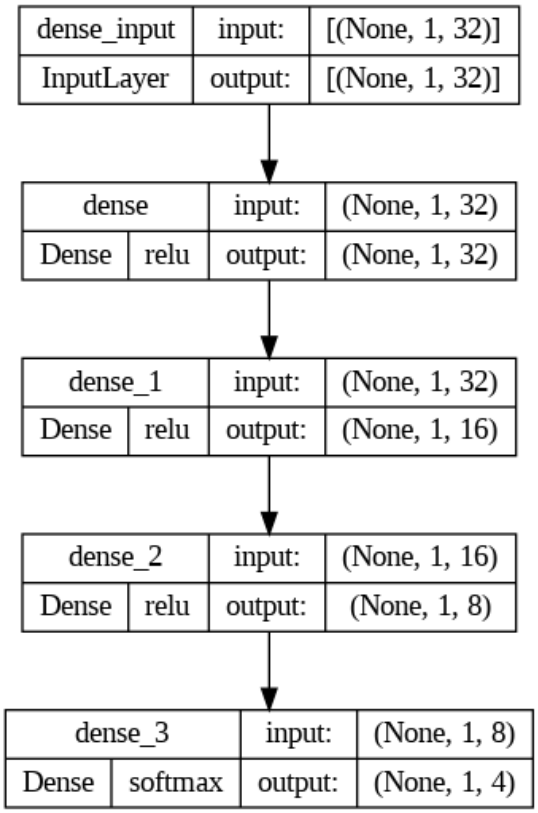

Figura 4.7: Arquitetura do classificador do experimento com 64 MFCCs

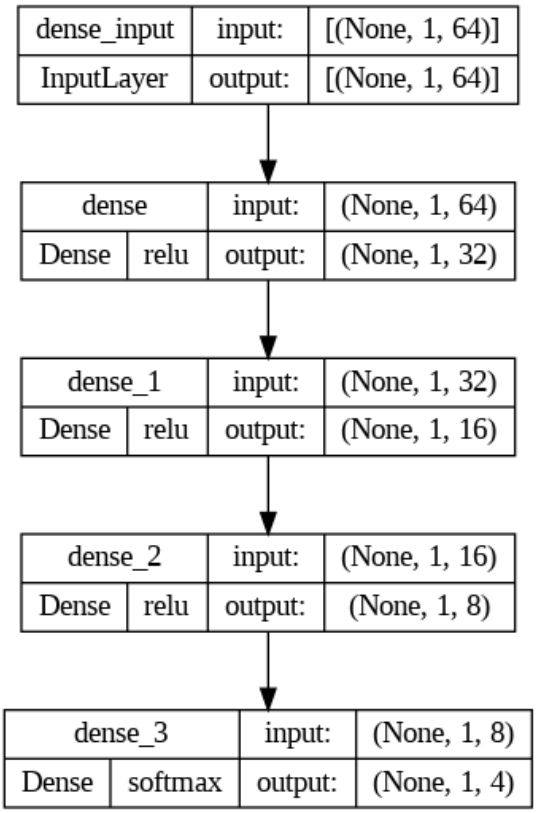

Figura 4.8: Arquitetura do classificador do experimento com 128 *MFCC*s

Recall. Assim, utilizaremos *F1-Score* como parâmetro de desempenho para o modelo supervisionado.

1. Entrada: Responsável por receber os dados com dimensão igual $(1, dim_{encode})$.
2. Densas intermediárias: As camadas posteriores (*dense*, *dense_1* e *dense_2*) são camadas totalmente conectadas que utilizam a função de ativação ReLU e cujo resultado é um vetor com dimensão $(1, 8)$.
3. Densa de saída: A camada *dense_3* é responsável por efetuar a ativação final da rede neural e entrega um vetor com dimensão $(1, 4)$. Utiliza uma função de ativação *Softmax*, cujo resultado será interpretado para classificar a intensidade da emoção em uma de quatro categorias: Fraca, moderada, forte ou pico de intensidade.

4.5 Considerações

Como considerações deste capítulo, destacamos que a descrição detalhada do projeto permitiu compreender os objetivos, as técnicas utilizadas e os desafios enfrentados ao

longo do processo de desenvolvimento. Essa visão geral proporciona uma base firme para a análise que será realizada no próximo capítulo (5), onde os resultados obtidos serão expostos e discutidos à luz dos objetivos estabelecidos.

Capítulo 5

Resultados

Neste capítulo serão apresentados os detalhes da implementação, os resultados, as métricas de desempenho e a metodologia utilizada na avaliação deste trabalho.

5.1 Cenários Modelados

Esta seção detalha nossa implementação para efetuar o reconhecimento da intensidade das emoções na fala em nosso idioma nativo, utilizando aprendizagem supervisionada e não supervisionada para uma tarefa de reconhecimento da intensidade da emoção expressa vocalmente em português.

Para tanto, em virtude do desafio deste trabalho, ratificado pela escassez de dados, nos aproveitamos da Fusão de Dados para criar, primeiramente, uma solução não supervisionada que consiga extrair características que sejam representativas o suficiente para que, mesmo com a dimensionalidade reduzida em comparação com a original, consigamos reconstruir o dado de entrada; e, posteriormente, utilizar esses dados comprimidos, constituídos de atributos suficientemente relevantes, para desenvolver um modelo supervisionado de inferência da intensidade.

Definida a arquitetura composta por um *Autoencoder* e um classificador, tomamos apenas as emoções comuns entre as duas bases de dados (alegria, medo, raiva e surpresa), totalizando 1364 registros, dentre os quais 51% apresentam o rótulo relativo à intensidade.

Foram realizados dois cenários experimentais, utilizando 64 e 128 MFCCs, respectivamente. Em ambos, o *Autoencoder* foi treinado e validado com dados de X e, em seguida, treinamos uma rede neural densa multicamada para classificar a intensidade, treinada e validada apenas na porção dos dados resultantes da aplicação de f em X_{VIVAE}, uma vez que X_{VERBO} não tem correspondência com o contradomínio (Z) das intensidades.

Uma vez que não há rótulo correspondente às intensidades para X_{VERBO}, devemos investigar se há algum sentido nos resultados quando aplicarmos a classificação a esses dados, que, até então, não foram vistos pelo classificador. Assim, faz-se necessária uma forma de analisar os registos de X_{VIVAE} e X_{VERBO} quanto às intensidades e à predição dessas intensidades, respectivamente. Podemos utilizar o PCA para reduzir a dimensionalidade dos registros e observar o comportamento da classificação.

A utilização de PCA encontra registros na literatura, tanto de ML e DL aplicados à voz e às emoções. Temos [78] que utilizou o PCA para gerar *features* de um modelo, observando que modelos treinados com as *features* obtidas a partir do PCA mantiveram seu desempenho quando comparados aos treinados com outros atributos; [79] que apesar de fazer uma *feature selection* prévia, optou por utilizar o PCA para reduzir em quase três vezes a dimensionalidade do dado para conseguir uma visualização bidimensional dos registros; [80] que buscou desenvolver um sistema de reconhecimento de emoções para áudios com ruídos, utilizando dados 64-dimensionais que seriam reduzidos para um espaço 6-dimensional, utilizando e comparando técnicas distintas, sendo uma delas o PCA; e [81] que aproxima o grau de *encoding* de um *Autoencoder* linear em seu espaço latente n-dimensional ao de um PCA com n componentes.

O PCA pode ser utilizado tanto para permitir a visualização de dados através da redução de dimensionalidade quanto para gerar os atributos utilizados em modelos de reconhecimento de emoções, o que nos diz que o PCA seria uma solução adequada para reduzir a dimensionalidade dos dados enquanto preserva características relevantes das amostras.

5.2 Resultados dos experimentos

5.2.1 Primeiro experimento: 64 *MFCC*s

A Figura 4.4 apresenta a arquitetura do modelo *Autoencoder*, onde podemos observar que os vetores do espaço latente, portanto, o vetor comprimido para ser posteriormente reconstruído, terão 32 posições.

Para este modelo, o valor final da *Loss* média ao aplicá-lo ao seu conjunto de teste foi de $6,40$. A Figura 4.7 apresenta a arquitetura do modelo de classificação de intensidade, enquanto a Tabela 5.1 apresenta métricas de desempenho para sua aplicação no conjunto de teste. Podemos observar que o melhor resultado para o *F1-Score* deu-se para a intensidade Fraca, com o valor de $0,68$, ficando abaixo de $0,5$ apenas para a classe Forte.

	Métricas		
Intensidade	Precision	Recall	F1-Score
Fraca	**0,73**	0,64	**0,68**
Moderada	0,53	0,52	0,53
Forte	0,46	0,50	0,48
Pico	0,64	**0,68**	0,66

Tabela 5.1: Métricas de avaliação para o classificador do experimento para 64 MFCCs

Aplicamos o classificador aos dados oriundos do X_{VERBO} e anotamos o resultado. Então efetuamos um PCA com 2 componentes ao resultado do *encoding* de X (Figura 5.1), onde as cores representam a classe - original para os dados do $VIVAE$ e predição do classificador para os dados do $VERBO$ - e podemos observar a formação de 2 grandes *clusters*, um à esquerda e outro à direita.

Percebemos através dos rótulos que estes conjuntos conseguiram distinguir bem as duas bases de dados, bem como um movimento descendente de aumento da intensidade da emoção em ambos os *clusters*. Temos uma visualização mais simples na Figura 5.2, onde agrupamos as classes Fraca e Moderada em uma nova classe denominada Baixa e as classes Forte e Pico em uma classe denominada Alta.

5.2.2 Segundo experimento: 128 *MFCC*s

A Figura 4.5 apresenta a arquitetura do modelo *Autoencoder*, onde podemos observar que os vetores do espaço latente, portanto, o vetor comprimido para ser posteriormente reconstruído, terão 64 posições.

Para este modelo, o valor final da *Loss* média ao aplicá-lo ao seu conjunto de teste foi de $0,84$. A Figura 4.8 apresenta a arquitetura do modelo de classificação de intensidade, enquanto a Tabela 5.2 apresenta métricas de desempenho para sua aplicação no conjunto de teste. Podemos observar que o melhor resultado para o *F1-Score* deu-se para a intensidade Pico, com o valor de $0,64$, ficando abaixo de $0,5$ apenas para a classe Forte.

Aplicamos o classificador aos dados oriundos do X_{VERBO} e anotamos o resultado. Então efetuamos um PCA com 2 componentes ao resultado do *encoding* de X (Figura 5.3), onde as cores representam a classe - original para os dados do $VIVAE$ e predição do classificador para os dados do $VERBO$ - e podemos observar a formação de 2 grandes *clusters*, um à esquerda e outro à direita.

Percebemos através dos rótulos que estes *conjuntos* conseguiram distinguir bem as duas bases de dados, bem como um movimento descendente de aumento da intensidade da emoção em ambos os *clusters*. Temos uma visualização mais simples na Figura 5.4,

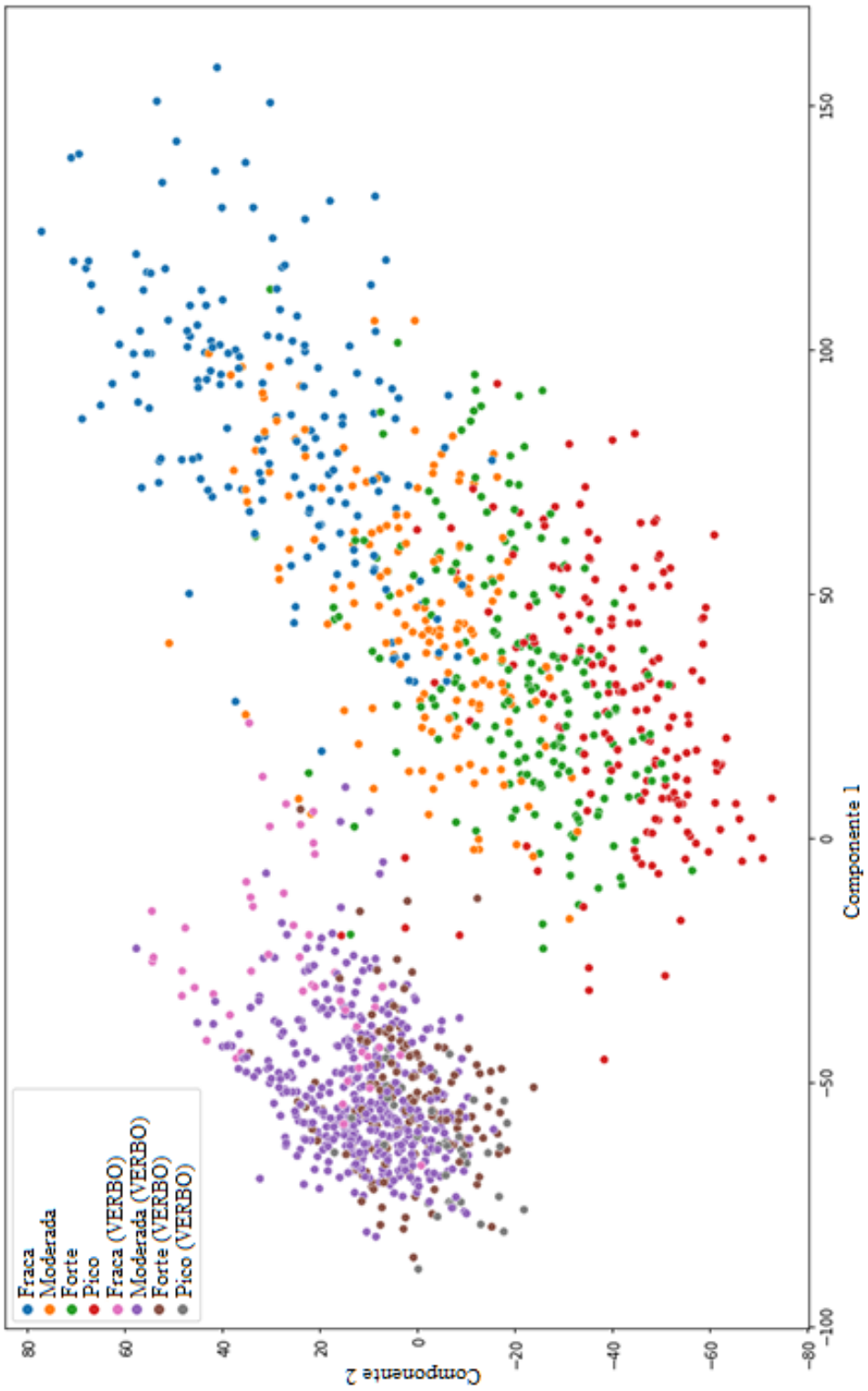

Figura 5.1: PCA com 2 componentes aplicado ao resultado do *encoding* do primeiro experimento

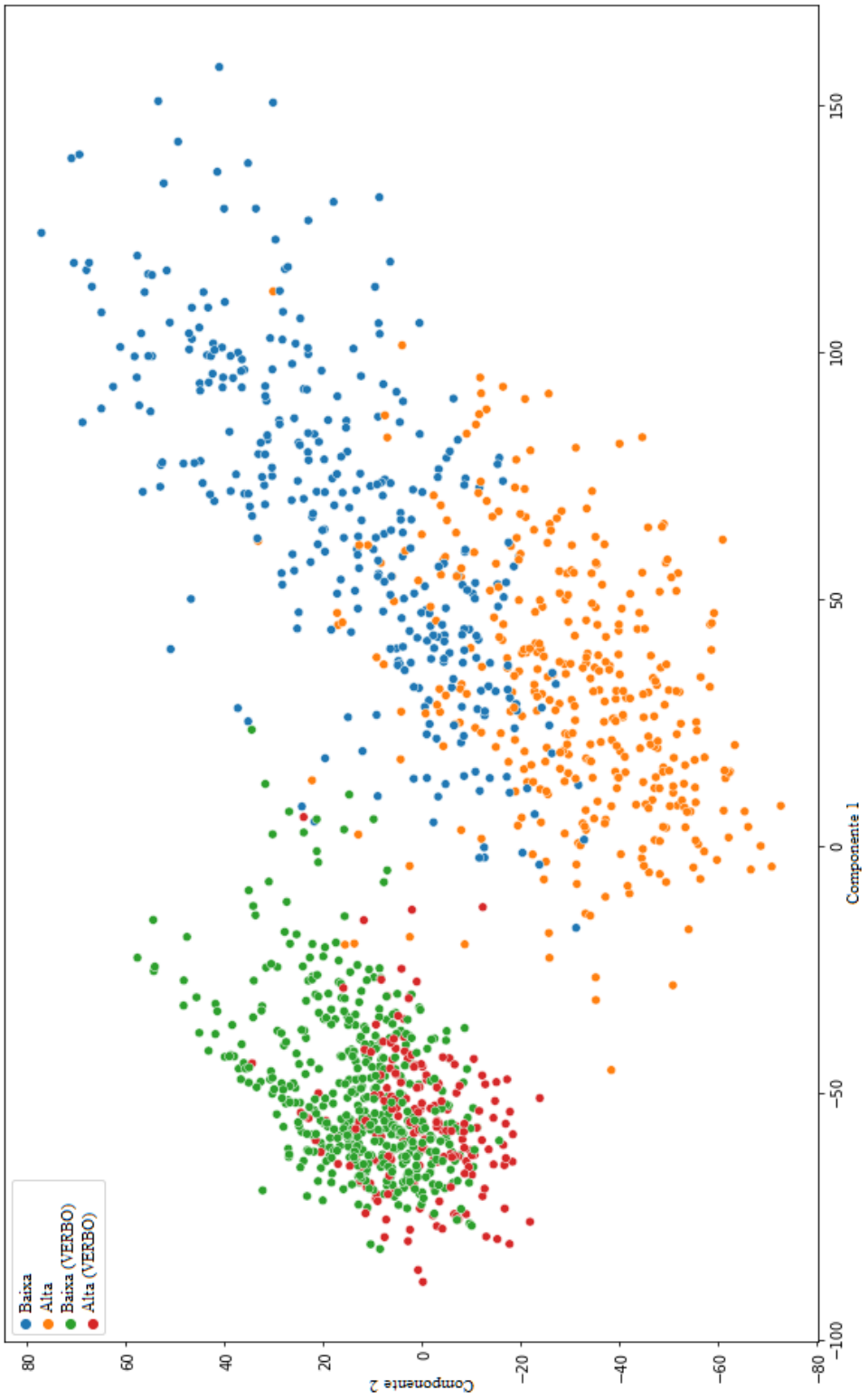

Figura 5.2: Agrupamento em 2 classes para PCA com 2 componentes aplicado ao resultado do *encoding* do primeiro experimento

	Métricas		
Intensidade	Precision	Recall	F1-Score
Fraca	**0,62**	0,55	0,58
Moderada	0,55	0,48	0,51
Forte	0,49	0,50	0,49
Pico	0,57	**0,73**	**0,64**

Tabela 5.2: Métricas de avaliação para o classificador do experimento para 128 MFCCs

onde agrupamos as classes Fraca e Moderada em uma nova classe denominada Baixa e as classes Forte e Pico em uma classe denominada Alta.

5.3 Discussão dos resultados

Nesta seção iremos iniciar uma discussão sobre os resultados alcançados e apresentaremos as dificuldades encontradas, além das limitações gerais do projeto.

Com base nos resultados da Tabela 5.3, verificamos que o primeiro experimento obteve desempenho superior no que tange à métrica selecionada para o modelo de classificação de intensidade, tendo um *F1-Score* superior aos do primeiro experimento em três das quatro classes. Em ambos os experimentos, a classe com pior resultado do classificador foi a classe Forte, enquanto as classes com melhor desempenho são Fraca e Pico de intensidade, respectivamente.

Dadas as Figuras 5.1 e 5.3, conseguimos observar que as classes Fraca e Pico ocupam os extremos dos conjuntos, o que pode tornar sua separação mais fácil frente às amostras das classes Moderada e Forte, as quais podemos notar mais amalgamadas na região central. Esta distribuição parece estar de acordo com nossos resultados, uma vez que os dois melhores desempenhos - em ambos os cenários - foram das classes Fraca e Pico, e os piores das classes Moderada e Forte.

Na Tabela 5.4, vemos um ganho de desempenho superior a sete vezes para o valor da *Loss* no segundo experimento, no qual utilizamos um número maior de MFCCs. Embora a *Loss* tenha apresentado uma queda significativa no segundo experimento, o que significa que o *Autoencoder* apresenta um desempenho melhor para reproduzir o dado de entrada, preservando melhor as características e atributos do dado, essa melhora no desempenho de uma função identidade se mostra contraditória frente ao desempenho do segundo classificador. Então, podemos supor que uma quantidade estritamente maior de MFCCs colaborou para a reconstrução da amostra, enquanto não demonstrou ganhos semelhantes na sua utilização para classificar a intensidade da emoção.

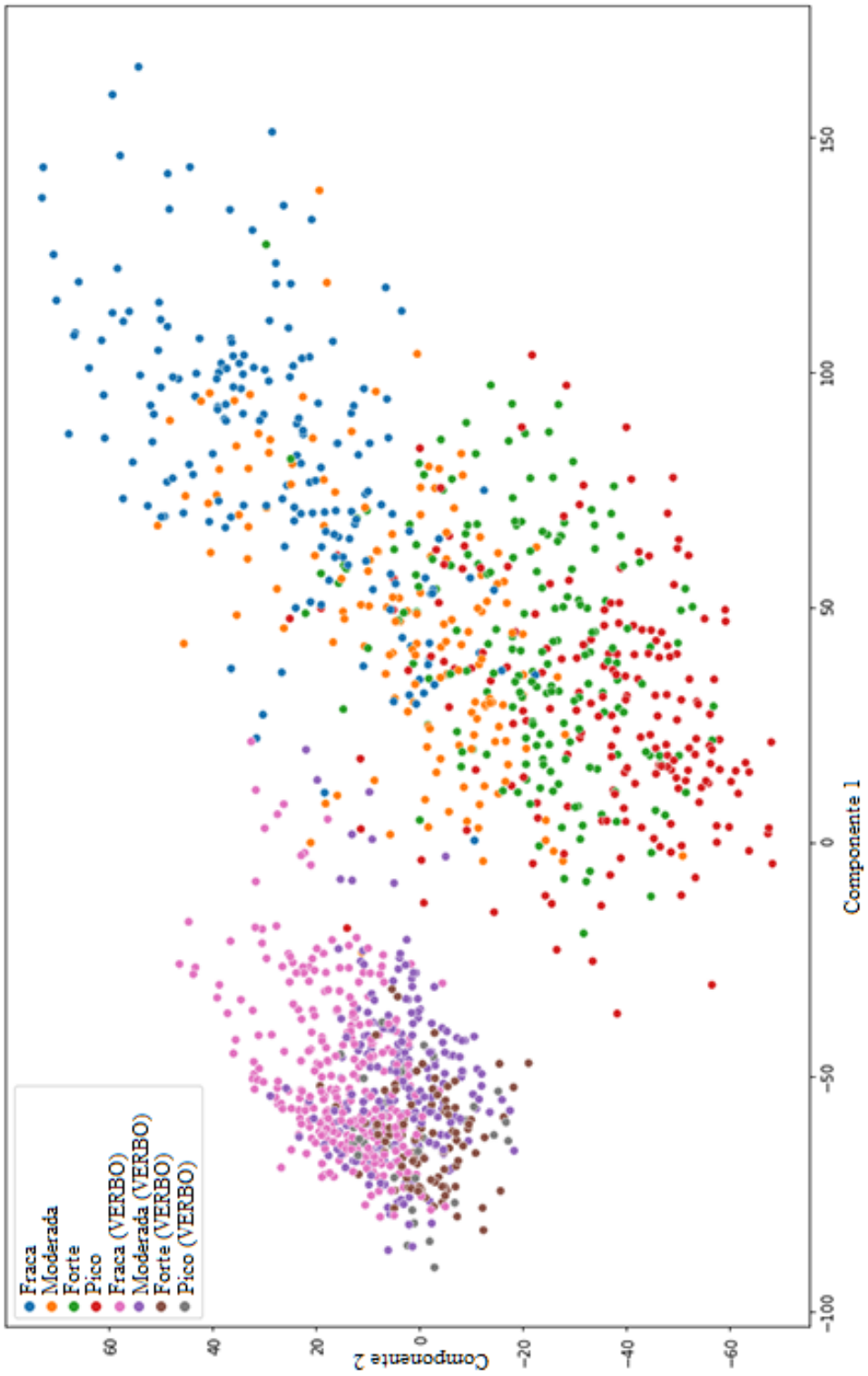

Figura 5.3: PCA com 2 componentes aplicado ao resultado do *encoding* do segundo experimento

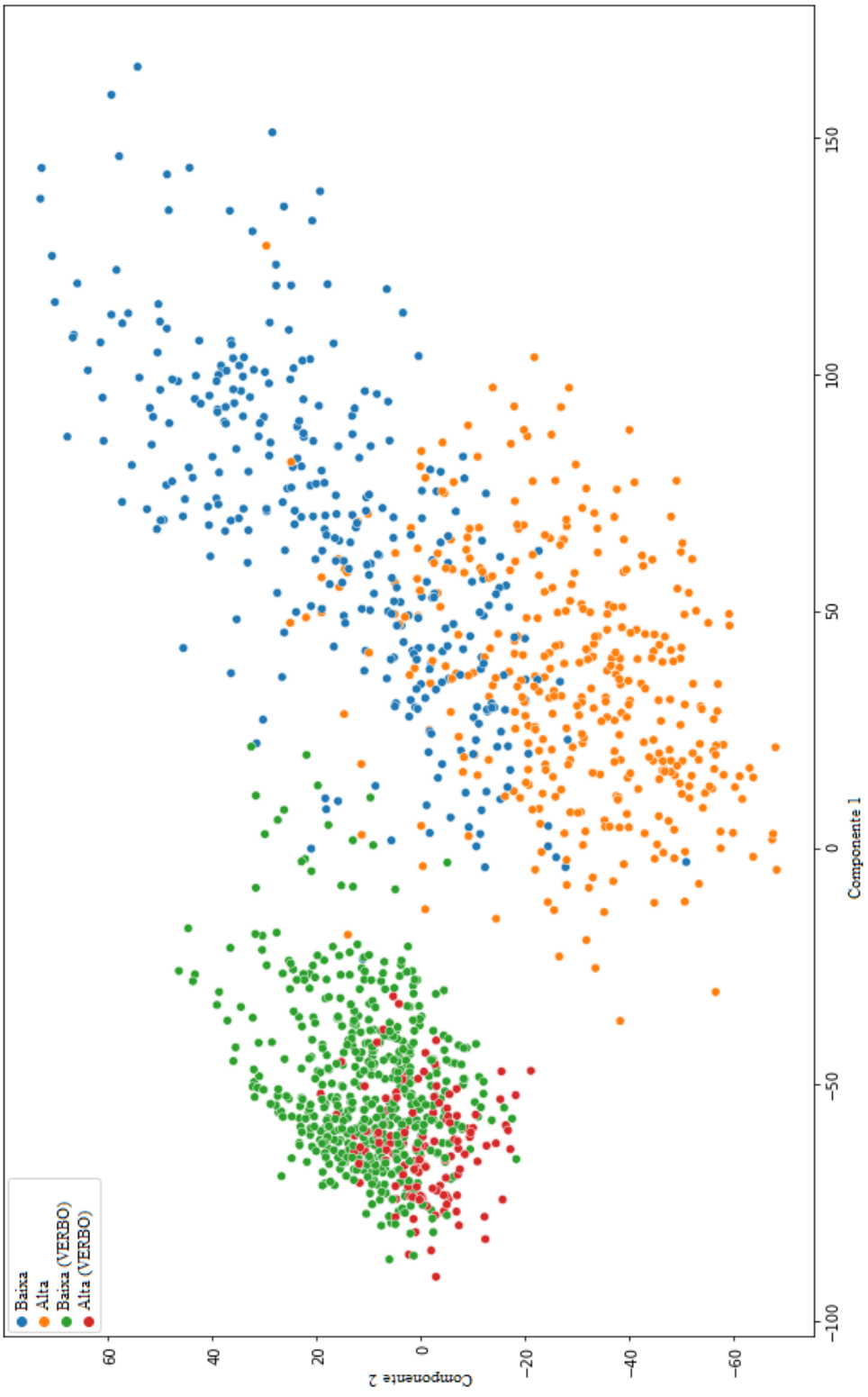

Figura 5.4: Agrupamento em 2 classes para PCA com 2 componentes aplicado ao resultado do *encoding* do segundo experimento

Intensidade	F1-Score para 64 MFCCs	F1-Score para 128 MFCCs
Fraca	**0,68**	0,58
Moderada	**0,53**	0,51
Forte	0,48	**0,49**
Pico	**0,66**	0,64

Tabela 5.3: Comparativo de *F1-Score* entre os experimentos

Resultado	MFCCs	
	64	128
Maior *F1-Score*	**0,68**	0,64
Classe com maior *F1-Score*	Fraca	Pico
Menor *F1-Score*	0,48	**0,49**
Classe com menor *F1-Score*	Forte	Forte

Tabela 5.4: Comparativo de atributos de desempenho dos experimentos

De posse dos resultados dos experimentos, conseguimos observar que os vetores do espaço latente dos *Autoencoder*, mesmo com uma redução de dimensionalidade à metade, aparentam manter características relativas à intensidade da emoção presentes em ambos os experimentos.

Quando agrupamos as classes Fraca e Moderada no conjunto denominado Baixo e as classes Forte e Pico no conjunto denominado Alto, conseguimos observar, nas Figuras 5.2 e 5.4, que mesmo com o desempenho superior do classificador do primeiro experimento, ainda há uma correspondência entre as intensidades dos registros. Uma vez que os rótulos de X_{VIVAE} são originais, observamos que as classes atribuídas aos dados de X_{VERBO} parecem ser condizentes, vide o comportamento descendente do aumento da intensidade. Assim, o classificador aprendeu com X_{VIVAE} a classificar a intensidade e aplicou essa lógica aos dados em X_{VERBO}.

Dados os rótulos corretos para os dados do VIVAE, essa visualização também fornece a ideia de que seríamos capazes de traçar uma linha em $Component\ 2 = 0$ de forma que os dados para ambos os experimentos possam ser divididos em duas categorias macro: Baixa (Fraco e Moderada) e Alta (Forte e Pico). Assim, para ambos os experimentos, dado um ponto de dados $x_i = (i_{component_2}, i_{component_2})$, se $i_{component_2} >= 0$ ele pertenceria à classe Baixa, enquanto se $i_{component_2} < 0$ ele pertenceria à classe Alta. Essa interpretação ingênua levanta a observação de que o $Component\ 2$ seria responsável pela intensidade de uma dada declaração de maneira quase que exclusiva.

Capítulo 6

Conclusões

Neste capítulo, serão abordadas as considerações finais sobre o projeto. Elicitaremos quais objetivos foram alcançados e as contribuições que o projeto traz. Também serão levantadas as dificuldades encontradas e as limitações desta pesquisa. Por fim, serão elucidados os possíveis trabalhos futuros.

Ao longo desta dissertação, nós delimitamos e contextualizamos um problema de pesquisa e vimos um breve panorama da literatura relacionada. Em seguida, formalizamos a metodologia e os componentes do trabalho proposto para, finalmente, realizar os experimentos com base nessa proposta.

Os resultados indicam que parece ser possível inferir a intensidade. Porém, o conjunto de dados ainda é bastante escasso. Também não temos conhecimento de uma base de dados em português que apresente tanto emoções quanto suas intensidades. Embora o português seja uma língua falada pela sexta maior população e pela nona maior economia do mundo, quando comparamos o VERBO com conjuntos de dados como o AudioSet, percebemos a enorme distância tanto em número de amostras ($\approx$ 2.000.000) quanto em duração média ($\approx 10s$). É de esperar que uma base de dados mais robusta melhore o desempenho da investigação.

Quando comparamos as métricas obtidas por trabalhos correlatos a este, podemos notar, por exemplo, que [53] obteve o valor de 71,93% para sua acurácia utilizando apenas MFCCs como atributo de entrada, enquanto [62] obteve um *F1-Score* médio de 74,6% utilizando *Random Forest* para classificar as emoções dos registros do *dataset* VERBO. Embora o objetivo não seja o mesmo, uma vez que estes trabalhos lidam com a emoção e não com a intensidade, percebemos que nossos resultados obtidos revelaram um desempenho subótimo para o classificador, indicando desafios persistentes na tarefa proposta.

Ao observarmos os resultados obtidos em [64], que realizou um desafio de *Speech Emotion Recognition* para detectar atributos emocionais (excitação, valência e dominância),

vemos que o melhor resultado para excitação (*arousal*) em dados de teste, frente à métrica escolhida, foi obtido pelo seu modelo de referência, com valor de 0.53, o que fica abaixo do *F1-Score* ponderado de 0.58 obtido no primeiro experimento deste trabalho. Ainda não se tratando de trabalhos coincidentes, esses valores reiteram a dificuldade presente na tarefa desenvolvida por este trabalho e apontam o resultado superior obtido por este trabalho em comparação a uma das tarefas realizadas por seus pares.

A natureza dos resultados sugere a necessidade contínua de aprimoramentos nas metodologias adotadas, bem como a exploração de novas abordagens e dados para potencializar a qualidade das predições. Além disso, este estudo enfatiza a importância de considerar as nuances linguísticas e culturais específicas do português ao desenvolver modelos de inferência emocional na voz.

O conhecimento adquirido durante esta pesquisa pode orientar futuras investigações, incentivando o desenvolvimento de estratégias mais robustas e sensíveis às particularidades do idioma, visando aprimorar a eficácia da inferência de intensidade emocional na voz. Dessa forma, a presente dissertação proporciona uma base para futuros estudos que busquem aprimorar a compreensão e a aplicação prática dessa importante área no campo do processamento de linguagem natural e reconhecimento de emoções.

Como conclusão desta dissertação de mestrado, foi possível explorar e analisar a inferência de intensidade da emoção na voz em língua portuguesa, utilizando diversas técnicas e abordagens. Apesar das limitações encontradas, esta pesquisa contribuiu para o avanço do conhecimento nesse campo específico, destacando a complexidade da inferência de intensidade emocional em um contexto linguístico diversificado. Como desdobramento da pesquisa realizada nesse trabalho, foi publicado um artigo [82] no *Symposium on Knowledge Discovery, Mining and Learning (KDMiLe)*, ocorrido em Novembro de 2024.

Para trabalhos futuros, pretendemos implementar uma Rede Neural Recorrente (*Recurrent Neural Network*, RNN) para avaliar os dados ao longo do eixo do tempo, tentando melhorar o desempenho do modelo de classificação. Pretendemos, também, realizar análises exploratórias para entender se o modelo está apresentando algum tipo de viés (como a classe Forte tendo o pior desempenho em ambos os experimentos) e, em caso afirmativo, entender como mitigá-lo. Uma outra abordagem seria dar um passo atrás e enfrentar um grande desafio em nossa tarefa e criar nosso próprio conjunto de dados com emoções e intensidade, sendo o primeiro em nossa língua nativa.

Referências

[1] Nogueira, Kennyo: *Estudo de respostas emocionais às cores no contexto de cartazes de cinema*. Design e Tecnologia, 8:1, junho 2018. x, 6

[2] Plutchik, Robert: *The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice*. American Scientist, 89(4):344–350, 2001. http://www.jstor.org/stable/27857503. x, 5, 6

[3] Bhatt, Chandradeep, Indrajeet Kumar, · Vijayakumar, · Kamred, Kamred Singh e Abhishek Kumar: *The state of the art of deep learning models in medical science and their challenges*. Multimedia Systems, 27, agosto 2021. x, 8

[4] Lieberman, Phillip: *The evolution of human speech: Its anatomical and neural bases*. Current anthropology, 48(1):39–66, 2007. 1

[5] Yamamoto, Shota, Yasunari Yoshitomi, Masayoshi Tabuse, Kou Kushida e Taro Asada: *Recognition of a baby's emotional cry towards robotics baby caregiver*. International Journal of Advanced Robotic Systems, 10(2):86, 2013. https://doi.org/10.5772/55406. 1

[6] Sarkar, Uddalok, Sayan Nag, Chirayata Bhattacharya, Shankha Sanyal, Archi Banerjee, Ranjan Sengupta e Dipak Ghosh: *Language independent emotion quantification using non linear modelling of speech*, 2021. https://arxiv.org/abs/2102.06003. 1

[7] Filho, Geraldo P Rocha, Rodolfo I Meneguette, Fábio Lúcio Lopes de Mendonça, Liriam Enamoto, Gustavo Pessin e Vinícius P Gonçalves: *Toward an emotion efficient architecture based on the sound spectrum from the voice of portuguese speakers*. Neural Computing and Applications, 36(32):19939–19950, 2024. 1

[8] Juslin, N Patrik, Laukka e Petri: *Impact of intended emotion intensity on cue utilization and decoding accuracy in vocal expression of emotion.* Emotion, 1(4):381, 2001. 1, 5

[9] Krishnan, P.T., Joseph A. N. Raj e V. Rajangam: *Emotion classification from speech signal based on empirical mode decomposition and non-linear features*. Complex Intell. Syst., 7:1919–1934, 2021. https://link.springer.com/article/10.1007/s40747-021-00295-z. 1

[10] Islam, Md. Riadul, M. A. H. Akhand, Md Abdus Samad Kamal e Kou Yamada: *Recognition of emotion with intensity from speech signal using 3d transformed feature and deep learning*. Electronics, 11(15), 2022, ISSN 2079-9292. https://www.mdpi.com/2079-9292/11/15/2362. 1

[11] Liu, Rui, Berrak Sisman, Björn Schuller, Guanglai Gao e Haizhou Li: *Accurate Emotion Strength Assessment for Seen and Unseen Speech Based on Data-Driven Deep Learning*. Em *Proc. Interspeech 2022*, páginas 5493–5497, 2022. 1, 23

[12] Koolagudi, S. G. e K. S. Rao: *Emotion recognition from speech: a review*. Int J Speech Technol, 15:99—117, 2012. https://link.springer.com/article/10.1007/s10772-011-9125-1. 1

[13] Sonmez, YeSim Ülgen e Asaf Varol: *New trends in speech emotion recognition*. Em *2019 7th International Symposium on Digital Forensics and Security (ISDFS)*, páginas 1–7, 2019. 1

[14] Manoret, Pongpak, Punnatorn Chotipurk, Sompoom Sunpaweravong, Chanati Jantrachotechatchawan e Kobchai Duangrattanalert: *Automatic detection of depression from stratified samples of audio data*, 2021. https://arxiv.org/abs/2111.10783. 1

[15] Dzedzickis, Andrius, Artūras Kaklauskas e Vytautas Bucinskas: *Human emotion recognition: Review of sensors and methods*. Sensors, 20(3), 2020, ISSN 1424-8220. https://www.mdpi.com/1424-8220/20/3/592. 1

[16] Elsayed, Nelly, Zag ElSayed, Navid Asadizanjani, Murat Ozer, Ahmed Abdelgawad e Magdy Bayoumi: *Speech emotion recognition using supervised deep recurrent system for mental health monitoring*, 2022. https://arxiv.org/abs/2208.12812. 1

[17] Nassif, Ali Bou, Ismail Shahin, Ashraf Elnagar, Divya Velayudhan, Adi Alhudhaif e Kemal Polat: *Emotional speaker identification using a novel capsule nets model*. Expert Systems with Applications, 193:116469, 2022, ISSN 0957-4174. https://www.sciencedirect.com/science/article/pii/S0957417421017498. 1

[18] Cook, Diane e Sajal Kumar Das: *Smart environments: technology, protocols, and applications*, volume 43. John Wiley & Sons, 2004. 1

[19] Purington, Amanda, Jessie G Taft, Shruti Sannon, Natalya N Bazarova e Samuel Hardman Taylor: *" alexa is my new bff" social roles, user satisfaction, and personification of the amazon echo*. Em *Proceedings of the 2017 CHI conference extended abstracts on human factors in computing systems*, páginas 2853–2859, 2017. 2

[20] *Emotionai and customer satisfaction in the financial services*. https://behavioralsignals.com/emotionai-and-customer-satisfaction-in-the-financial-services/. Accessed: 2023-08-18. 2

[21] Zhu, Xiaolian, Shan Yang, Geng Yang e Lei Xie: *Controlling emotion strength with relative attribute for end-to-end speech synthesis*. Em *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, páginas 192–199, 2019. 2, 15, 17

[22] *Ai-mediated conversations (ai-mc) redefining bank's revenues through its call center.* https://behavioralsignals.com/ai-mediated-conversations-case-study/. Accessed: 2023-08-18. 2

[23] Behlau, Mara: *Respostas para perguntasfrequentes na Área de voz*, 2009. https://www.sbfa.org.br/campanhadavoz/FAQs2011.pdf. 4

[24] Deafness, NIH: National Institute on e Other Communication Disorders (NIDCD): *What is voice? what is speech? what is language?*, 2020. https://www.nidcd.nih.gov/health/voice-speech-and-language. 4

[25] Zheng, Yu: *Methodologies for cross-domain data fusion: An overview.* IEEE Transactions on Big Data, 1(1):16–34, 2015. 4, 23

[26] Ekman, Paul: *An argument for basic emotions.* Cognition and Emotion, 6(3-4):169–200, 1992. https://doi.org/10.1080/02699939208411068. 4

[27] Russell, James A: *A circumplex model of affect.* Journal of personality and social psychology, 39(6):1161, 1980. 4

[28] Averill, James R e Thomas A More: *Happiness.* Handbook of emotions, página 617–629, 1993. https://psycnet.apa.org/record/1993-98937-037. 5

[29] Frijda, Nico H, Andrew Ortony, Joep Sonnemans e Gerald L Clore: *The complexity of intensity: Issues concerning the structure of emotion intensity.* 1992. 5

[30] Bachorowski, Jo Anne e Ellen B Braaten: *Emotional intensity: Measurement and theoretical implications.* Personality and individual differences, 17(2):191–199, 1994. https://www.sciencedirect.com/science/article/abs/pii/0191886994900256. 5

[31] Sonnemans, Joep e Nico H Frijda: *The structure of subjective emotional intensity.* Cognition & Emotion, 8(4):329–350, 1994. 5

[32] Zanon, Cristian, Micheline Roat Bastianello, Juliana Cerentini Pacico e Claudio Simon Hutz: *Desenvolvimento e validação de uma escala de afetos positivos e negativos.* Psico-UsF, 18:193–201, 2013. https://www.scielo.br/j/pusf/a/vh7QqFWQLYx5dBptgfQHBJS/?format=pdf&lang=pt. 5

[33] Carvalho, Hudson W de, Sérgio B Andreoli, Diogo R Lara, Christopher J Patrick, Maria Inês Quintana, Rodrigo A Bressan, Marcelo F de Melo, Jair de J Mari e Miguel R Jorge: *Structural validity and reliability of the positive and negative affect schedule (panas): Evidence from a large brazilian community sample.* Brazilian Journal of Psychiatry, 35:169–172, 2013. https://www.scielo.br/j/rbp/a/qLd5P5VfpLRfF5qDmBxTCQS/?lang=en. 5

[34] Otsuka Nunes, Lucas Yukio, Daniel Campos Lopes Lemos, Rodolfo de Castro Ribas Júnior, Cláudia Brandão Behar e Pedro Paulo Pires dos Santos: *Análise psicométrica da panas no brasil.* Ciencias Psicológicas, 13(1):45–55, 2019. https://www.redalyc.org/journal/4595/459559717005/html/. 6

[35] N Holz, P Larrouy Maestri & D Poeppel: *The paradoxical role of emotional intensity in the perception of vocal affect.* Sci Rep, 11(9663), 2021. https://www.nature.com/articles/s41598-021-88431-0. 6, 24

[36] Poole, D.L., D. Poole, A.K. Mackworth, A. Mackworth e R. Goebel: *Computational Intelligence: A Logical Approach.* Oxford University Press, 1998, ISBN 9780195102703. https://books.google.com.br/books?id=3p6KZSHjD4YC. 7

[37] Goodfellow, Ian, Yoshua Bengio e Aaron Courville: *Deep Learning.* MIT Press, 2016. http://www.deeplearningbook.org. 7

[38] Hamsa, Shibani, Ismail Shahin, Youssef Iraqi e Naoufel Werghi: *Emotion recognition from speech using wavelet packet transform cochlear filter bank and random forest classifier.* IEEE Access, 8:96994–97006, 2020. 7

[39] Rong, Jia, Gang Li e Yi Ping Phoebe Chen: *Acoustic feature selection for automatic emotion recognition from speech.* Information Processing & Management, 45(3):315–328, 2009, ISSN 0306-4573. https://www.sciencedirect.com/science/article/pii/S0306457308000885. 7

[40] Milton, A., S. Sharmy Roy e S. Samil Selvi: *Svm scheme for speech emotion recognition using mfcc feature.* International Journal of Computer Applications, 69(9):34–49, 2013. https://research.ijcaonline.org/volume69/number9/pxc3887667.pdf. 7

[41] Lanjewar, Rahul B., Swarup Mathurkar e Nilesh Patel: *Implementation and comparison of speech emotion recognition system using gaussian mixture model (gmm) and k- nearest neighbor (k-nn) techniques.* Procedia Computer Science, 49:50–57, 2015, ISSN 1877-0509. https://www.sciencedirect.com/science/article/pii/S1877050915007358, Proceedings of 4th International Conference on Advances in Computing, Communication and Control (ICAC3'15). 7

[42] Han, Kun, Dong Yu e Ivan Tashev: *Speech emotion recognition using deep neural network and extreme learning machine.* Em *Interspeech 2014*, 2014. 8

[43] Abdel-Hamid, Ossama, Abdel rahman Mohamed, Hui Jiang, Li Deng, Gerald Penn e Dong Yu: *Convolutional neural networks for speech recognition.* IEEE/ACM Transactions on audio, speech, and language processing, 22(10):1533–1545, 2014. 8

[44] Abbaschian, Babak Joze, Daniel Sierra-Sosa e Adel Elmaghraby: *Deep learning techniques for speech emotion recognition, from databases to models.* Sensors, 21(4), 2021, ISSN 1424-8220. https://www.mdpi.com/1424-8220/21/4/1249. 8, 10, 16, 21, 22

[45] Kriesel, D: *A Brief Introduction to Neural Networks.* 2021. http://www.dkriesel.com. 8

[46] Barlow, H.B.: *Unsupervised Learning.* Neural Computation, 1(3):295–311, setembro 1989, ISSN 0899-7667. https://doi.org/10.1162/neco.1989.1.3.295. 11

[47] Hsu, Wei Ning e James Glass: *Extracting domain invariant features by unsupervised learning for robust automatic speech recognition.* Em *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, páginas 5614–5618. IEEE, 2018. 11

[48] Deng, Jun, Zixing Zhang, Erik Marchi e Björn Schuller: *Sparse autoencoder-based feature transfer learning for speech emotion recognition.* Em *2013 humaine association conference on affective computing and intelligent interaction*, páginas 511–516. IEEE, 2013. 11

[49] Deng, Jun, Zixing Zhang, Florian Eyben e Björn Schuller: *Autoencoder-based unsupervised domain adaptation for speech emotion recognition.* IEEE Signal Processing Letters, 21(9):1068–1072, 2014. 11

[50] Juang, Biing Hwang e Laurence R Rabiner: *Hidden markov models for speech recognition.* Technometrics, 33(3):251–272, 1991. 13

[51] al, Robert V Shannon et: *Speech recognition with primarily temporal cues.* Science, 270(5234):303–304, 1995. 13

[52] Morrison, Donn, Ruili Wang e Liyanage C De Silva: *Ensemble methods for spoken emotion recognition in call-centres.* Speech communication, 49(2):98–112, 2007. 13, 16, 18, 27

[53] Bhargava, Mayank e Tim Polzehl: *Improving automatic emotion recognition from speech using rhythm and temporal feature*, 2013. `https://arxiv.org/abs/1303.1761`. 13, 18, 27, 44

[54] Latif, Siddique, Rajib Rana, Junaid Qadir e Julien Epps: *Variational autoencoders for learning latent representations of speech emotion: A preliminary study*, 2017. `https://arxiv.org/abs/1712.08708`. 14, 27

[55] Zhang, Shiqing, Shiliang Zhang, Tiejun Huang e Wen Gao: *Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching.* IEEE Transactions on Multimedia, 20(6):1576–1590, 2018. 14, 19, 27

[56] Eskimez, Sefik Emre, Zhiyao Duan e Wendi Heinzelman: *Unsupervised learning approach to feature analysis for automatic speech emotion recognition.* Em *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, páginas 5099–5103, 2018. 14, 18, 19, 29

[57] Li, Yuanchao, Tianyu Zhao e Tatsuya Kawahara: *Improved end-to-end speech emotion recognition using self attention mechanism and multitask learning.* Em *Interspeech*, páginas 2803–2807, 2019. 15, 19

[58] Akhtar, Md Shad, Deepanway Ghosal, Asif Ekbal, Pushpak Bhattacharyya e Sadao Kurohashi: *All-in-one: Emotion, sentiment and intensity prediction using a multi-task ensemble framework.* IEEE Transactions on Affective Computing, 13(1):285–297, 2022. 15

[59] Hutto, Clayton e Eric Gilbert: *Vader: A parsimonious rule-based model for sentiment analysis of social media text.* Em *Proceedings of the international AAAI conference on web and social media*, volume 8, páginas 216–225, 2014. 15

[60] Chatziagapi, Aggelina, Georgios Paraskevopoulos, Dimitris Sgouropoulos, Georgios Pantazopoulos, Malvina Nikandrou, Theodoros Giannakopoulos, Athanasios Katsamanis, Alexandros Potamianos e Shrikanth Narayanan: *Data augmentation using gans for speech emotion recognition.* Em *Interspeech*, páginas 171–175, 2019. 16, 27

[61] Campos, Gabriel Almeida e Lucas da Silva Moutinho: *Deep: uma arquitetura para reconhecer emoção com base no espectro sonoro da voz de falantes da língua portuguesa*, 2021. `https://bdm.unb.br/handle/10483/27583`. 16, 19

[62] Josh, Neelakshi: *Brazilian portuguese emotional speech corpus analysis.* X Seminário em TI do PCI/CT, 2021. `https://www.gov.br/cti/pt-br/publicacoes/producao-cientifica/seminario-pci/xi_seminario_pci-2021/pdf/seminario-2021_paper_29.pdf`. 16, 19, 23, 44

[63] Zhou, Kun, Berrak Sisman, Rajib Rana, Bjorn W. Schuller e Haizhou Li: *Emotion intensity and its control for emotional voice conversion.* IEEE Transactions on Affective Computing, páginas 1–1, 2022. `https://doi.org/10.1109/TAFFC.2022.3175578`. 17, 21

[64] Goncalves, Lucas, Ali N Salman, Abinay R Naini, Laureano Moro Velazquez, Thomas Thebaud, Leibny Paola Garcia, Najim Dehak, Berrak Sisman e Carlos Busso: *Odyssey 2024-speech emotion recognition challenge: Dataset, baseline framework, and results.* Development, 10(9,290):4–54, 2024. 17, 19, 44

[65] Mashal, Sonia Xylina e Kavita Asnani: *Emotion intensity detection for social media data.* Em *2017 International Conference on Computing Methodologies and Communication (ICCMC)*, páginas 155–158, 2017. 17, 18

[66] Olatinwo, Damilola D., Adnan Abu-Mahfouz, Gerhard Hancke e Hermanus Myburgh: *Iot-enabled wban and machine learning for speech emotion recognition in patients.* Sensors, 23(6), 2023, ISSN 1424-8220. `https://www.mdpi.com/1424-8220/23/6/2948`. 19

[67] Burkhardt, Felix, Astrid Paeschke, M. Rolfes, Walter F. Sendlmeier e Benjamin Weiss: *A database of german emotional speech.* Em *Interspeech*, 2005. 21

[68] Engberg, Inger S., Anya V. Hansen, Ove Andersen e Paul Dalsgaard: *Design, recording and verification of a danish emotional speech database.* 1997. `https://www.isca-speech.org/archive_v0/archive_papers/eurospeech_1997/e97_1695.pdf`. 21

[69] Livingstone, Steven R e Frank A Russo: *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, 13(5):e0196391, 2018. 22

[70] Dupuis, Kate e M Kathleen Pichora-Fuller: *Intelligibility of emotional speech in younger and older adults*. Ear and hearing, 35(6):695–707, 2014. 22

[71] Cao, Houwei, David G. Cooper, Michael K. Keutmann, Ruben C. Gur, Ani Nenkova e Ragini Verma: *Crema-d: Crowd-sourced emotional multimodal actors dataset*. IEEE Transactions on Affective Computing, 5(4):377–390, 2014. 22

[72] Neto, José Torres, Geraldo P.R. Filho, Leandro Y. Mano e João Ueyama: *Verbo: Voice emotion recognition database in portuguese language*. Journal of Computer Science, 14(11):1420–1430, Nov 2018. https://thescipub.com/abstract/jcssp.2018.1420.1430. 23

[73] Purves, D., G. J. Augustine GJ, D. Fitzpatrick D e et al.: *Neuroscience*. Sunderland (MA): Sinauer Associates, 2001. https://www.ncbi.nlm.nih.gov/books/NBK10924. 26

[74] Zhao, Jianfeng, Xia Mao e Lijiang Chen: *Speech emotion recognition using deep 1d & 2d cnn lstm networks*. Biomedical Signal Processing and Control, 47:312–323, 2019, ISSN 1746-8094. https://www.sciencedirect.com/science/article/pii/S1746809418302337. 27

[75] Bui, Khac Hoai Nam, Hyeonjeong Oh e Hongsuk Yi: *Traffic density classification using sound datasets: An empirical study on traffic flow at asymmetric roads*. IEEE Access, 8:125671–125679, 2020. https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9136653. 27

[76] Mekruksavanich, Sakorn, Anuchit Jitpattanakul e Narit Hnoohom: *Negative emotion recognition using deep learning for thai language*. Em *2020 joint international conference on digital arts, media and technology with ECTI northern section conference on electrical, electronics, computer and telecommunications engineering (ECTI DAMT & NCON)*, páginas 71–74. IEEE, 2020. 27

[77] Dillon, Barry M., Tilman Plehn, Christof Sauer e Peter Sorrenson: *Better latent spaces for better autoencoders*. SciPost Phys., 11:061, 2021. https://scipost.org/10.21468/SciPostPhys.11.3.061. 29

[78] Lee, Chul Min, S.S. Narayanan e R. Pieraccini: *Classifying emotions in human-machine spoken dialogs*. Em *Proceedings. IEEE International Conference on Multimedia and Expo*, volume 1, páginas 737–740 vol.1, 2002. https://ieeexplore.ieee.org/document/1035887. 36

[79] Ververidis, D., C. Kotropoulos e I. Pitas: *Automatic emotional speech classification*. Em *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, páginas I–593, 2004. https://ieeexplore.ieee.org/document/1326055. 36

[80] You, Mingyu, Chun Chen, Jiajun Bu, Jia Liu e Jianhua Tao: *Emotion recognition from noisy speech*. Em *2006 IEEE International Conference on Multimedia and Expo*, páginas 1653–1656, 2006. https://ieeexplore.ieee.org/document/4036934. 36

[81] Rovetta, Stefano, Zied Mnasri, Francesco Masulli, Alberto Cabri *et al.*: *Emotion recognition from speech: an unsupervised learning approach.* International Journal of Computational Intelligence Systems, 14(1):23–35, 2020. `https://air.unimi.it/bitstream/2434/955219/1/03-125945494.pdf`. 36

[82] Augusto, Henrique, Vinícius Gonçalves, Edna Canedo, Rodolfo Meneguette, Gustavo Pessin e Geraldo R. Filho: *Unraveling emotional dimensions in brazilian portuguese speech through deep learning.* Em *Anais do XII Symposium on Knowledge Discovery, Mining and Learning*, páginas 33–40, Porto Alegre, RS, Brasil, 2024. SBC. `https://sol.sbc.org.br/index.php/kdmile/article/view/30945`. 45