UNIVERSIDADE ESTADUAL DO SUDOESTE DA BAHIA
DEPARTAMENTO DE CIÊNCIAS EXATAS E TECNOLÓGICAS - DCET
CURSO DE CIÊNCIA DA COMPUTAÇÃO

IAN SILVA ANTUNES RAMOS

RECONHECIMENTO DE EMOÇÕES NO PROCESSAMENTO DE VOZES EM LÍNGUA PORTUGUESA COM REDES NEURAIS SEM PESOS

VITÓRIA DA CONQUISTA
2024

IAN SILVA ANTUNES RAMOS

RECONHECIMENTO DE EMOÇÕES EM PROCESSAMENTO DE VOZES UTILIZANDO REDES NEURAIS SEM PESOS

Trabalho de Conclusão de Curso (TCC) apresentado ao Curso de Ciência da Computação da Universidade Estadual do Sudoeste da Bahia (UESB) como requisito parcial para a conclusão do Curso de Ciências da Computação.

Orientador(a): Prof. Dr. José Carlos Martins Oliveira

VITÓRIA DA CONQUISTA

2024

DEDICATÓRIA

Dedico este trabalho aos meus pais, que sempre me apoiaram incondicionalmente e me ensinaram o valor do conhecimento. Aos meus amigos, pela parceria e incentivo em cada etapa desta jornada, e ao meu orientador, pelo apoio e pelas valiosas orientações que tornaram este trabalho possível.

RESUMO

Um dos desafios da inteligência artificial (IA) e da interação humano computador (IHC) é a evolução dos sistemas de reconhecimento de emoção na fala. As arquiteturas presentes atualmente para o *Speech Emotion Recognition (SER)* em fala são em sua maioria baseados em sistemas de *Deep Learning (DL),* principalmente as redes neurais convolucionais*,* e sistemas multimodais, onde é combinado diferentes entradas como texto ao processamento em busca de melhor detectar a emoção, mas estes modelos ainda estão longe de serem perfeitos, além de serem complexos estes também possuem uma alta demanda computacional e requerem grandes quantidades de dados. A Fim de evoluir os sistemas de SER, este trabalho tem como objetivo principal utilizar Redes Neurais Sem Pesos (RNSP), redes que utilizam neurônios baseados em dispositivos de memória RAM para treinar modelos, para identificar e classificar espectros das emoções em faixas de áudio de falantes da língua portuguesa, e depois compará los á um modelo similar produzido usando redes neurais convolucionais (CNN). Para treinar o modelo é utilizado um banco de dados de voz da língua portuguesa chamado VERBO. A fim de avaliar os resultados de precisão obtidos com a rede neural sem peso em relação às redes neurais convolucionais é realizado uma comparação dos resultados do modelo sem peso aqui apresentado com os resultados de um modelo convolucional chamado DEEP (*DEtection of voice Emotion in Portuguese language*) produzido com o mesmo banco de dados VERBO.

Palavras-chave: Reconhecimento de emoção na voz, Redes Neurais Sem Pesos, RNSP, língua portuguesa, VERBO

ABSTRACT

One of the challenges in artificial intelligence (AI) and human-computer interaction (HCI) is the advancement of speech emotion recognition (SER) systems. Current SER architectures are mostly based on deep learning (DL) models, particularly convolutional neural networks (CNNs) and multimodal systems that combine different inputs, such as text, to enhance emotion detection. However, these models remain far from perfect, as they are computationally complex, require large amounts of data, and demand high processing power. To contribute to the evolution of SER systems, this study aims to utilize Weightless Neural Networks (WNN), a type of network that employs memory-based neurons to train models, for the identification and classification of emotional spectrums in speech samples from Portuguese-speaking individuals. The performance of this approach is then compared to a similar model based on convolutional neural networks. The training process employs the Portuguese-language speech dataset called VERBO. To assess the accuracy of the weightless neural network, the results obtained are compared to those of a convolutional model named DEEP (Detection of Voice Emotion in Portuguese language), which was trained using the same VERBO dataset.

Keywords: Speech Emotion Recognition, Weightless Neural Networks, WNN, Portuguese language, VERBO.

SUMÁRIO

1. Introdução

As tecnologias já estão em um nível evolucional onde o ser humano é cada vez mais dependente de seu uso e evolução, fazendo-se necessário que a forma como esta interação humano computador (IHC) seja cada vez mais facilitada. Com esta evolução novas formas de interação com a máquina, além da interação física, estão ganhando cada vez mais uso, como por exemplo o uso de câmeras e microfones como principais dispositivos para a interação, que permitem que pessoas com deficiência motora consigam interagir com a máquina além de novos tipos de processos possíveis como o reconhecimento de emoção do usuário.

Segundo Picard (2000) a computação afetiva é uma computação que se relaciona com, surge de ou influência emoções. Os computadores estão progressivamente ganhando a habilidade de expressar e reconhecer emoções, e em breve poderão desenvolver a capacidade de "sentir". Estudos neurológicos recentes destacam a importância crucial das emoções na cognição e percepção humana, sugerindo que computadores afetivos não só irão melhorar seu desempenho na assistência a humanos, mas também aprimoraram suas capacidades de tomada de decisão.

Uma das áreas da computação afetiva que vem sendo mais utilizadas é o reconhecimento de emoções na voz. O SER tem se tornado uma área de pesquisa cada vez mais relevante dentro do campo da inteligência artificial e da interação homem-máquina, promovendo avanços na tecnologia de reconhecimento de voz que podem ser aplicados em diversas áreas, como atendimento ao cliente, assistentes virtuais, e sistemas de saúde.

Dentre os modelos mais utilizados para Reconhecimento de Emoção em Fala (SER), destacam-se as redes neurais convolucionais (CNNs), que extraem características de representações espectrais do áudio, as redes neurais recorrentes (RNNs) e suas variantes, como as LSTMs, que modelam a dinâmica temporal da fala, além das máquinas de vetores de suporte (SVMs), que realizam classificações baseadas em hiperplanos ótimos (SCHULLER et al., 2018). Essas abordagens, no entanto, apresentam desafios como alta demanda computacional, necessidade de grandes volumes de dados rotulados e dificuldade em generalizar para diferentes sotaques e variações de fala (EYBEN et al., 2016).

No contexto da língua portuguesa, o desenvolvimento de sistemas de SER enfrenta desafios adicionais devido à limitada disponibilidade de bases de dados anotadas. Pesquisadores têm explorado redes neurais pré-treinadas e técnicas de aumento de dados para melhorar a robustez dos modelos para o português brasileiro (NOGUEIRA, 2018). Além

disso, esforços como a criação do banco de dados VERBO (TORRES NETO et al., 2018) são fundamentais para o avanço da área, pois possibilitam o treinamento de modelos mais adaptados às particularidades da prosódia e fonética da língua portuguesa.

O uso de Redes Neurais Sem Pesos (RNSP) no SER surge como uma alternativa promissora, pois oferece vantagens como baixo custo computacional, rapidez no treinamento e maior robustez a pequenas variações no sinal de áudio (ALEKSANDER, 1967). Diferente dos modelos tradicionais baseados em deep learning, as RNSP utilizam dispositivos de memória RAM para armazenar e recuperar padrões, tornando a inferência mais eficiente.

1.1 Motivação e contextualização do problema

A interação entre humanos e máquinas tem evoluído significativamente nos últimos anos, impulsionada por avanços na inteligência artificial e na computação afetiva. À medida que os dispositivos tecnológicos se tornam cada vez mais integrados ao cotidiano, torna-se essencial que sua interação com os usuários seja mais intuitiva e natural. Dentre as diversas formas de interação humano-computador (IHC), o reconhecimento de emoções na voz tem se destacado como uma ferramenta promissora para tornar as máquinas mais responsivas ao estado emocional do usuário.

O reconhecimento de emoções na fala pode beneficiar diversas aplicações, desde assistentes virtuais e atendimento ao cliente até sistemas de suporte emocional e monitoramento da saúde mental. No entanto, os modelos atuais de SER são predominantemente baseados em redes neurais profundas, como as redes neurais convolucionais (CNNs) e redes recorrentes (RNNs). Embora essas abordagens tenham demonstrado bons resultados, elas apresentam desafios significativos, como alta demanda computacional, necessidade de grandes volumes de dados rotulados para treinamento e dificuldades na generalização para diferentes idiomas e sotaques.

No contexto da língua portuguesa, o desenvolvimento de modelos de reconhecimento de emoções enfrenta desafios adicionais. A maior parte dos modelos existentes foi treinada com bases de dados em inglês, o que limita sua eficácia quando aplicados a idiomas com estruturas fonéticas e prosódicas distintas. Além disso, o número de bancos de dados públicos e anotados para o português ainda é reduzido, dificultando a evolução da pesquisa na área. Um dos poucos conjuntos de dados disponíveis é o VERBO, criado especificamente para o

reconhecimento de emoções na voz em português, sendo utilizado como referência para o presente estudo.

Diante desses desafios, este trabalho busca investigar uma abordagem alternativa para o reconhecimento de emoções na voz, utilizando Redes Neurais Sem Pesos (RNSP). Diferente das redes neurais tradicionais, as RNSPs não dependem de operações matemáticas complexas para aprendizado e classificação, baseando-se no uso de memórias RAM para armazenar padrões e realizar inferências. Essa característica pode tornar o treinamento significativamente mais rápido e reduzir o custo computacional, permitindo a implementação do SER em dispositivos com recursos limitados, como sistemas embarcados e dispositivos móveis.

1.2 O Estado da arte

O reconhecimento de emoções na fala (*Speech Emotion Recognition - SER*) é uma área que tem evoluído de forma significativa, refletindo o avanço das tecnologias de inteligência artificial. Nos seus primórdios, o SER era baseado em métodos tradicionais de extração de características, onde a emoção era identificada a partir de parâmetros acústicos como pitch, intensidade e ritmo (Schuller et al., 2018). Inicialmente, esses métodos utilizam classificadores como máquinas de vetores de suporte (SVMs) e redes neurais simples, que, apesar de eficazes, apresentavam limitações significativas em relação à adaptação a diferentes sotaques e ruídos de fundo.

Com o avanço das redes neurais profundas e da computação, modelos mais sofisticados começaram a ser aplicados, como as redes neurais convolucionais (CNNs), redes neurais recorrentes (RNNs) e, mais recentemente, os transformers, que são capazes de capturar tanto as características espectrais quanto a dinâmica temporal da fala. A chegada dos sistemas híbridos multimodais, que combinam dados de áudio, texto e vídeo, trouxe uma nova dimensão ao SER, ampliando sua precisão e robustez ao permitir a consideração de diferentes fontes de informação (Eyben et al., 2016).

Apesar dos avanços, o SER enfrenta alguns desafios persistentes. A variabilidade da fala, como sotaques, entonação e variações de pronúncia, dificulta a criação de modelos que possam generalizar bem para diferentes populações de falantes. Além disso, o reconhecimento de emoções é sensível ao ruído de fundo e a diferentes condições acústicas, o que pode reduzir a precisão do modelo em ambientes do mundo real. A necessidade de grandes volumes de dados rotulados para treinamento é outra dificuldade crítica, pois esses dados são

caros e demorados para serem anotados, limitando a capacidade de treinar modelos mais robustos.

No contexto da língua portuguesa, os desafios se ampliam devido à falta de bases de dados robustas e anotadas especificamente para esse idioma. A diversidade linguística presente no português, com variações significativas entre os diferentes sotaques e dialetos, também representa uma barreira. Além disso, as diferenças fonéticas entre o português e outros idiomas mais amplamente estudados, como o inglês, exigem modelos específicos que considerem essas particularidades. Poucos estudos focados especificamente no português brasileiro, como o banco de dados VERBO (Torres Neto et al., 2018), são fundamentais para suprir essa lacuna e permitir o avanço da área nesse contexto específico.

A pesquisa contínua sobre modelos mais adaptados ao português brasileiro e a criação de novos datasets são passos importantes para superar os obstáculos atuais e permitir o desenvolvimento de sistemas de SER mais precisos e amplamente aplicáveis.

Diante das dificuldades enfrentadas pelos modelos tradicionais de redes neurais no campo do reconhecimento de emoções na fala (SER), surgem alternativas como as Redes Neurais Sem Pesos (RNSP), que oferecem vantagens como menor custo computacional e maior rapidez no treinamento. No entanto, a adoção de modelos como o WiSARD, uma RNSP, ainda é limitada. Isso ocorre principalmente porque modelos baseados em deep learning, como Redes Neurais Convolucionais (CNNs) e Redes Neurais Recorrentes (RNNs), já estão amplamente consolidados na indústria e na pesquisa. Essas abordagens têm forte suporte de ferramentas e bibliotecas populares, como TensorFlow, PyTorch e Keras, que facilitam o desenvolvimento e a experimentação (ABADI et al., 2015; PASZKE et al., 2019). Além disso, o uso de hardware avançado, como GPUs e TPUs, otimizou o processamento desses modelos, tornando-os mais eficientes e acessíveis.

Esses modelos profundos são predominantes em benchmarks e competições internacionais de SER, como as promovidas por conferências como INTERSPEECH e IEEE ICASSP, que consolidam as redes profundas como o padrão da área. Como resultado, modelos como as RNSP, apesar de suas vantagens, não são amplamente explorados no contexto do SER (SCHULLER et al., 2018).

Ademais, as RNSP apresentam algumas limitações que dificultam sua adoção. Um desses desafios é a dependência da qualidade da base de treinamento. Modelos como o WiSARD têm dificuldades em generalizar para variações sutis da fala, como sotaques e entonações, o que pode comprometer a eficácia do modelo em cenários mais diversos. Além

disso, a fala é um sinal contínuo e dinâmico, o que exige modelos que possam capturar variações temporais. As RNSP, por sua vez, funcionam melhor com padrões discretos e categorizáveis, o que as torna menos eficazes para esse tipo de processamento.

Outro ponto crítico é a escassez de ferramentas e bibliotecas robustas para a implementação de redes sem pesos, o que limita sua disseminação e adoção no mercado. Além disso, a pesquisa aplicada a RNSP no contexto de processamento de áudio é escassa, sendo a maior parte da pesquisa focada em áreas como reconhecimento de padrões visuais e classificação de dados estruturados, em vez de na análise de sinais de áudio (ALEKSANDER, 1967). Essas dificuldades, juntamente com os avanços em redes profundas, fazem com que as RNSP permaneçam uma alternativa pouco explorada no campo do reconhecimento de emoção na fala.

1.2.1 Trabalhos relacionados

O reconhecimento de emoções a partir da fala (SER) é uma área de pesquisa em constante evolução, com abordagens baseadas em diferentes técnicas de aprendizado de máquina, incluindo deep learning e Redes Neurais Sem Pesos. A seguir, são apresentados alguns dos trabalhos mais relevantes que serviram como base para o desenvolvimento deste estudo destacando seus objetivos, metodologias e limitações. Ao final, será feita uma análise comparativa entre os trabalhos existentes e a proposta deste estudo.

Em 2020, Campos e Moutinho realizaram um estudo inovador ao aplicar o banco de dados VERBO, composto por áudios em português brasileiro, para o reconhecimento de emoções a partir da fala. O modelo DEEP (Detection of Emotion in Portuguese Language) foi desenvolvido utilizando redes neurais convolucionais (CNNs) com o objetivo de especializar a detecção de emoções na língua portuguesa, levando em consideração as particularidades fonéticas e prosódicas do idioma. O modelo DEEP, que utiliza deep learning para classificar as emoções na fala, foi um marco importante na adaptação dos modelos de SER para o português e serve como base para comparações neste estudo. No entanto, o modelo DEEP apresenta limitações como a alta demanda computacional e a necessidade de grandes volumes de dados rotulados para treinamento (CAMPO & MOUTINHO, 2020).

Em contraste, Oliveira (2018) propôs a utilização das Redes Neurais Sem Pesos (RNSPs) para a resolução de problemas de diagnóstico e identificação de falhas em sistemas dinâmicos, com uma abordagem baseada em neurônios que utilizam memórias RAM para

aprender as características dos dados de treinamento. O modelo WiSARD (Wilkie, Stonham e Aleksander 's Recognition Device), uma rede neural sem pesos, foi utilizado para identificar padrões com rapidez, precisão e consistência, sem a necessidade de retreinamento. Apesar de não ser aplicado diretamente ao reconhecimento de emoções, o trabalho de Oliveira demonstrou o potencial das RNSPs em tarefas de classificação e reconhecimento de padrões. As vantagens das RNSPs incluem a redução no custo computacional e maior flexibilidade em comparação com modelos baseados em deep learning, o que torna essa abordagem promissora para o SER, especialmente em cenários com recursos limitados (OLIVEIRA, 2018).

Além desses, o estudo "Pantheon: Classificação de emoções faciais utilizando a rede neural sem pesos WiSARD" (2018) também explora o uso das RNSPs, mas focando no reconhecimento de emoções faciais. Embora o trabalho seja voltado para um tipo diferente de sinal, ele reforça a viabilidade do uso das RNSPs em tarefas de reconhecimento de emoções, sugerindo que esse modelo pode ser adaptado para o domínio da fala, dado o seu desempenho em outras áreas do reconhecimento de padrões.

A principal diferença entre os trabalhos existentes e a proposta deste estudo reside na escolha do modelo. Enquanto os trabalhos de Campos e Moutinho (2020) e o modelo DEEP utilizam redes neurais convolucionais (CNNs) baseadas em deep learning para reconhecimento de emoções na fala, o presente estudo investiga a aplicação de Redes Neurais Sem Pesos (RNSPs), mais especificamente o modelo WiSARD, para a mesma tarefa. A principal vantagem das RNSPs sobre as abordagens tradicionais de deep learning é a redução no custo computacional e a maior rapidez no treinamento, fatores que tornam essa abordagem interessante, especialmente em ambientes com recursos limitados.

1.3 Objetivos geral e específicos

1.3.1 Objetivo geral

Este trabalho tem como objetivo comparar dois modelos de rede neural, um convolucional (CNN) e outro com modelo RNSP (Rede Neural Sem Pesos), ambos treinados no dataset VERBO (banco de dados produzido com diferentes falantes da língua portuguesa e separado por emoções) com o objetivo de classificar emoções em áudio. O modelo convolucional chamado DEEP, terá seus resultados de acurácia na classificação das emoções comparado com os resultados do modelo de RNSP proposto com a ideia de provar que um

modelo RNSP pode apresentar melhores resultados de reconhecimento e classificação de padrões de áudio do que o modelo convencional.

1.3.2 Objetivos específicos

Os objetivos específicos deste trabalho visam detalhar as etapas e abordagens adotadas para atingir o objetivo geral mencionado anteriormente. Estes objetivos são direcionados à análise comparativa entre os dois modelos de redes neurais (CNN e RNSP), além de aspectos técnicos e implementação. Os Objetivos Específicos São:

- **Revisar os fundamentos teóricos:** Revisar o conteúdo disponível sobre a estruturação de voz e emoção para máquinas e os modelos computacionais do problema, como foco em RNSPs.
- **Analisar o banco de dados VERBO**: Examinar as características e os dados presentes no banco VERBO, com foco nas emoções representadas e nas particularidades do áudio em português brasileiro, garantindo que o dataset seja adequado para a tarefa de reconhecimento de emoções.
- **Desenvolver o modelo RNSP**: Projetar e implementar um modelo de Rede Neural Sem Pesos (RNSP), especificamente o WiSARD, para classificar as emoções a partir dos áudios, utilizando características acústicas extraídas dos sinais de áudio, como Mel-frequency cepstral coefficients (MFCCs).
- **Realizar o treinamento do modelo**: Treinar o modelo RNSP utilizando os dados do banco de dados VERBO, realizando ajustes de parâmetros necessários para otimizar a acurácia dos modelos nas tarefas de classificação emocional.
- **Comparar os resultados de desempenho**: Avaliar o desempenho de ambos os modelos (CNN e RNSP) utilizando métricas como acurácia, precisão, recall e F1-Score, comparando os resultados obtidos em termos de eficácia e eficiência.
- **Analisar a aplicabilidade do modelo RNSP**: Investigar a aplicabilidade e as vantagens do modelo RNSP em relação ao modelo convolucional, considerando aspectos como tempo de treinamento, complexidade computacional e desempenho em diferentes cenários de dados.
- **Identificar os benefícios e limitações de cada abordagem**: Identificar as forças e limitações de cada abordagem (CNN e RNSP), contribuindo para a literatura científica

ao mostrar que o modelo RNSP pode oferecer uma alternativa eficiente para o reconhecimento de emoções em áudio, especialmente em situações com recursos computacionais limitados.

1.4 Organização do trabalho

Além do capítulo de introdução, este trabalho possui mais 5 capítulos apresentados de forma discursiva e técnica. Os conteúdos distribuídos em capítulos estão organizados como apresentado a seguir.

O “capítulo 2” apresenta toda a base teórica e referencial necessária ao entendimento do trabalho. É apresentada uma descrição dos processos de identificação de emoções em relação a espectros de áudio e principalmente uma revisão bibliográfica sobre machine learning e redes neurais, principalmente as sem pesos.

O “capítulo 3” é apresentado de forma detalhada como foi o processo de processamento dos dados disponíveis, a RNSP utilizada e o fluxo de desenvolvimento do trabalho. Em respeito ao modelo sem pesos utilizado, irei somente o explicar superficialmente pois eles já estão detalhados no trabalho de Oliveira [4].

O “capítulo 4” os resultados obtidos dos modelos são analisados, comparados e jogados sobre diferentes perspectivas e então é avaliado se o objetivo principal do trabalho se concretizou

No “capítulo 5” são postas as conclusões e considerações obtidas com o trabalho realizado e possíveis sugestões para trabalhos futuros.

2. Fundamentação teórica

O objetivo deste capítulo é fornecer os fundamentos teóricos e bibliográficos essenciais para o desenvolvimento deste trabalho. Na seção 2.1, são discutidos os conceitos centrais de *machine learning*, abordando as técnicas mais utilizadas para reconhecimento de padrões. Na seção 2.2, será explorada a relação entre voz e emoção, com foco em como as características acústicas da fala podem refletir diferentes estados emocionais. A seção 2.3 faz uma introdução às redes neurais, explicando sua estrutura e funcionamento. Em seguida, na seção 2.4, são descritas as redes neurais com peso, com destaque para as Redes Neurais

Convolucionais (CNN), amplamente utilizadas em tarefas de reconhecimento de emoções. Finalmente, a seção 2.5 apresenta as Redes Neurais Sem Peso (RNSP), discutindo seus princípios, benefícios e como elas se comparam às redes tradicionais.

2.1 Reconhecimento de voz e emoção

A relação entre a voz humana e as emoções é um tema central na área de interação humano-computador, especialmente no contexto de sistemas de reconhecimento de emoções a partir da fala. A fala humana não é apenas um meio de transmitir palavras e frases, mas também um veículo carregado de elementos paralinguísticos que comunicam o estado emocional do falante. Tais elementos incluem a entonação, ritmo, timbre, intensidade e modulação da voz, que são profundamente influenciados pelo estado emocional do indivíduo. Isso significa que a voz não só transmite o conteúdo verbal, mas também carrega informações emocionais que podem ser decodificadas por meio da análise do sinal de áudio.

As emoções humanas afetam a voz de maneira complexa e variada, influenciando diretamente características acústicas que podem ser utilizadas para inferir o estado emocional de uma pessoa. Por exemplo, emoções como raiva e felicidade frequentemente estão associadas a um aumento na frequência fundamental da voz (F0) e na intensidade acústica, tornando a fala mais rápida e enérgica. Por outro lado, emoções como tristeza e medo tendem a reduzir essas variáveis, resultando em uma fala mais baixa, lenta e monótona. Além disso, emoções como surpresa e nojo podem ser expressas por variações rápidas e abruptas na prosódia vocal, enquanto emoções como o desprezo podem ser identificadas por uma combinação específica de características tonais e temporais da fala.

A capacidade dos seres humanos de identificar emoções na fala é altamente desenvolvida e intuitiva, sendo parte fundamental da comunicação social e interpessoal. Porém, a tarefa de fazer essa identificação por parte dos computadores é desafiadora devido à complexidade das emoções e às variações na forma como elas se manifestam na fala. Enquanto os humanos conseguem perceber nuances emocionais de forma quase imediata, os sistemas automatizados enfrentam dificuldades em captar esses sinais, devido à variabilidade de fatores como sotaque, contexto cultural e a interferência de ruídos ambientais (RICHARD et al., 2022). Portanto, o reconhecimento de emoções a partir da fala exige modelos sofisticados que consigam processar e interpretar as variações acústicas, muitas vezes utilizando técnicas de aprendizado de máquina, como redes neurais profundas, para identificar padrões nas características do sinal de áudio.

Atualmente, com o avanço das técnicas de aprendizado profundo (Deep Learning), como redes neurais convolucionais (CNNs) e redes neurais recorrentes (RNNs), os sistemas de reconhecimento de emoções estão se tornando mais precisos e eficientes, embora ainda enfrentam desafios relacionados à diversidade de expressões emocionais e à necessidade de grandes volumes de dados rotulados para treinamento.

2.2.1 Características acústicas relevantes

Para analisar e classificar emoções na fala, algumas características acústicas principais são observadas. As características prosódicas englobam elementos como a entonação, o ritmo e a duração da fala; mudanças na entonação podem sugerir emoções positivas ou negativas, enquanto um ritmo acelerado pode indicar excitação ou raiva. Já as características espectrais referem-se à distribuição de energia em diferentes frequências do sinal de áudio, e sua análise permite identificar padrões específicos associados a diversas emoções. Representações comuns para capturar essas variações incluem espectrogramas e MFCCs (Mel-Frequency Cepstral Coefficients). Além disso, características de intensidade e pitch, como a energia do sinal e a frequência fundamental da voz, possuem forte relação com estados emocionais, sendo que um pitch mais elevado é geralmente associado à felicidade ou raiva, enquanto uma intensidade reduzida pode indicar tristeza. Essas características acústicas são essenciais para identificar estados emocionais, pois capturam nuances que refletem as expressões vocais de diferentes emoções.

2.2.2 Modelo circunflexo das emoções

Para organizar as emoções e entender suas relações, o **Modelo Circunflexo das Emoções**, proposto por Russel (1980), é amplamente utilizado. Este modelo posiciona as emoções em um círculo, onde o eixo horizontal representa o nível de prazer (positivo ou negativo) e o eixo vertical representa o nível de excitação (alto ou baixo), com o centro representando um estado neutro. O modelo sugere que as emoções próximas no círculo possuem características semelhantes, mas que uma completa distinção entre elas não é possível devido à sobreposição de suas características. Essa noção é essencial para o presente trabalho, pois destaca que as emoções compartilham padrões, indicando que um estado emocional pode se aproximar de outros em termos de suas propriedades acústicas.

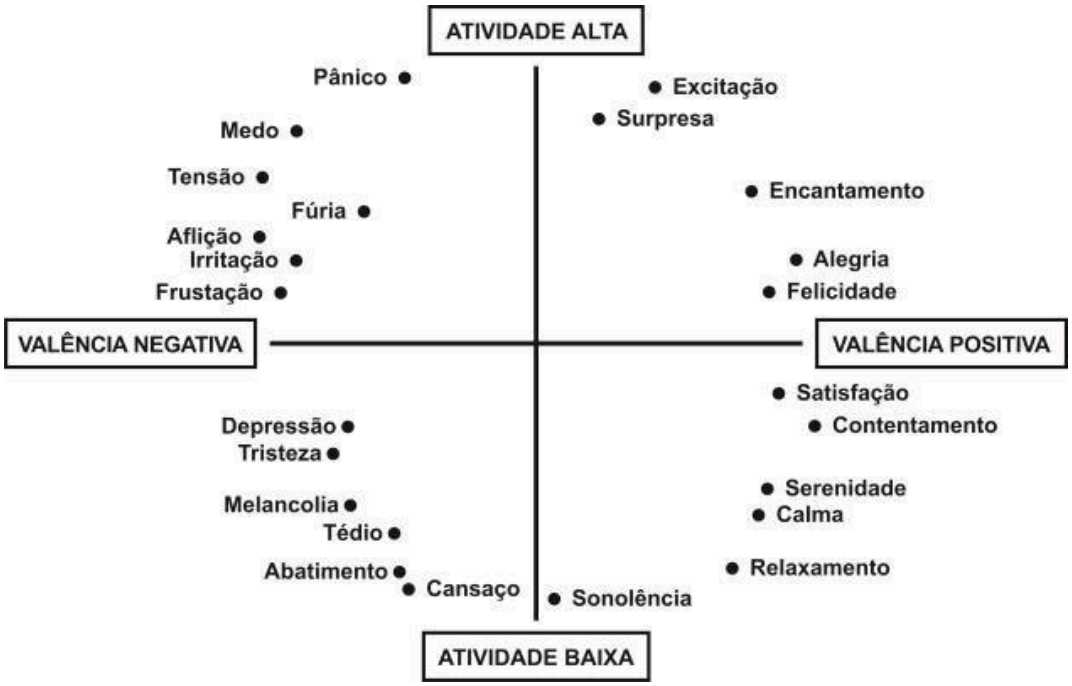

Figura 01: Modelo Circunflexo das Emoções (Adaptado de NOGUEIRA, 2018)

2.2 Aprendizado de máquina

Aprendizagem de máquina é um ramo da inteligência artificial que envolve o desenvolvimento de algoritmos capazes de aprender a partir de dados e tomar decisões ou fazer previsões sem serem explicitamente programados para uma tarefa específica. O aprendizado ocorre por meio da identificação de padrões e relações nos dados, o que permite ao modelo melhorar seu desempenho com o tempo e com mais exemplos.

Algoritmos de aprendizagem de máquina são a base de várias aplicações na área de processamento de linguagem natural, visão computacional e reconhecimento de voz. Em tarefas de reconhecimento emocional na fala, como neste trabalho, o ML permite que os sistemas identifiquem características emocionais de forma automatizada, com base em dados de áudio. Modelos específicos de redes neurais, que são uma das subcategorias mais usadas em ML, contribuem para essa tarefa ao capturar e aprender padrões complexos e não lineares que representam a diversidade de expressões emocionais.

A tarefa de ML utilizada para a realização deste trabalho será a classificação. A tarefa de classificação envolve treinar um modelo para separar dados em categorias específicas. Os classificadores associam um conjunto de entradas de tamanho m* (X1,X2,X3,...,Xm) a

rótulos de tamanho n* (y1,y2,y3,...,yn) usando um algoritmo que vincula cada entrada “Xk” ao seu respectivo rótulo. Neste trabalho, a classificação será usada para relacionar entradas de faixas de áudio de acordo com os rótulos de emoção estabelecidos.

As faixas de áudio de entrada são compostas por várias informações diferentes compensadas, e por isso é essencial uma escolha de quais informações serão necessárias para a resolução do problema e que comporá o conjunto {X1, X2, X3, . . . , Xm}. Essas características escolhidas são chamadas de features. A escolha destas para este trabalho foi feita com a mesma metodologia do trabalho de Campos e Moutinho (2020), a ser explorada em capítulos posteriores.

O processo de aprendizagem em ML é geralmente categorizado em aprendizagem supervisionada, não supervisionada e por reforço, mas para o treinamento deste trabalho, o método supervisionado foi escolhido. Segundo Russell e Norvig (2002), no aprendizado supervisionado o objetivo é encontrar uma função que aproxima a função verdadeira que gera as saídas a partir das entradas em um conjunto de dados de treinamento. Esse conjunto consiste em pares de entrada e saída, (x1,y1),(x2,y2),...,(xN,yN), onde y é o valor de saída, gerado por uma função desconhecida y=f(x). Ou seja, a tarefa de aprendizado envolve então descobrir uma função, chamada hipótese H, que se aproxima bem da função f, de forma a fazer previsões adequadas, inclusive para dados novos.

2.2.1 Principais critérios de avaliação do aprendizado de máquina

O sucesso dos modelos de aprendizagem de máquina depende da qualidade dos dados de treinamento e do processo de extração de características, que envolve transformar os dados brutos em informações relevantes. Esses dados alimentam o modelo, permitindo-lhe reconhecer padrões que podem ser aplicados a novos dados não vistos durante o treinamento. A precisão e a eficácia dos sistemas de ML podem ser avaliadas com métricas como a acurácia, a precisão, o recall e a F1-score, que quantificam o quão bem o modelo está lidando com a tarefa proposta. Estas métricas serão utilizadas para futura comparação do modelo proposto com o DEEP ou futuros modelos. A seguir, são apresentadas as variáveis e fórmulas que serão utilizadas no cálculo dos critérios de avaliação.

- **Verdadeiros Positivos (VP):** São os casos em que o modelo identifica corretamente uma classe positiva, ou seja, quando a previsão e a realidade correspondem a um evento positivo.

- **Verdadeiros Negativos (VN):** Representam os casos em que o modelo identifica corretamente uma classe negativa, acertando ao não prever um evento que realmente não ocorreu.
- **Falsos Positivos (FP):** Ocorrências em que o modelo identifica erroneamente um caso como positivo quando, na realidade, ele é negativo. Esse tipo de erro é conhecido como "alarme falso".
- **Falsos Negativos (FN):** Casos em que o modelo falha ao não identificar uma classe positiva, ou seja, ele prevê um resultado negativo para algo que realmente era positivo, perdendo uma detecção importante.

As fórmulas de avaliação:

- **Acurácia:** mede a proporção de previsões corretas em relação ao total de previsões feitas. É útil para ter uma visão geral da performance do modelo.

$$Acurácia = \frac{VP+VN}{VP+VN+FP+FN} \quad (1.1)$$

- **Precisão:** Avalia a exatidão das previsões positivas, ou seja, a proporção de casos verdadeiros entre todos os casos classificados como positivos pelo modelo.

$$Precisão = \frac{VP}{VP+FP} \quad (1.2)$$

- **Sensibilidade (ou Recall):** mede a capacidade do modelo de identificar corretamente todos os casos positivos reais, mostrando a taxa de acertos entre todos os verdadeiros positivos existentes.

$$Sensibilidade = \frac{VP}{VP+FN} \quad (1.3)$$

- **F1-Score:** Combina Precisão e Sensibilidade em uma única métrica, calculando a média harmônica entre as duas. É útil em cenários com classes desbalanceadas.

$$F1 - Score = 2 * \frac{Precisão \times Sensibilidade}{Precisão + Sensibilidade} \quad (1.4)$$

2.3 Redes neurais

As Redes Neurais Artificiais (RNAs) são sistemas inspirados nas redes de neurônios biológicos, adaptados para resolver problemas computacionais através de unidades de processamento interligadas. Também chamadas de computação neural, elas abrangem aplicações que vão da engenharia às ciências cognitivas, aproveitando a capacidade dessas redes de realizar tarefas complexas a partir de dados incompletos ou ambíguos. Estruturalmente, uma RNA consiste em "neurônios" artificiais, ou nodos, conectados em camadas, que processam e transmitem sinais por meio de interconexões, formando um modelo distribuído de processamento de informações.

Essa flexibilidade e a capacidade de tratar dados não estruturados fazem das redes neurais artificiais ferramentas muito eficazes em diversas tarefas de reconhecimento de padrões, onde o conhecimento humano é limitado ou as regras são difíceis de definir, como no reconhecimento de emoções em fala, visão computacional e processamento de linguagem natural.

Com base em Hecht-Nielsen (1990) [9] é possível definir um RNA como uma estrutura de processamento distribuída e paralela que pode ser representada por um grafo direcionado. Nessa estrutura, os nós representam unidades de processamento, também chamadas de neurônios artificiais, e as arestas, chamadas de conexões ou sinapses, transmitem sinais em apenas uma direção. Cada unidade de processamento pode tanto receber várias conexões de entrada quanto ter múltiplas saídas, desde que os sinais transmitidos pelas conexões de saída permaneçam idênticos. Essas unidades também podem manter uma memória local. Cada unidade de processamento executa uma função de transferência, combinando os dados da memória local com os sinais recebidos para gerar uma resposta que será transmitida. Os sinais de entrada podem vir de fontes externas ou de circuitos internos de realimentação, e as saídas, que são as respostas da rede, são transmitidas para fora dela.

As RNAs operam em duas fases principais: o aprendizado e o uso. Na fase de aprendizado, a rede ajusta seus parâmetros para produzir uma saída específica, enquanto na fase de uso, ela aplica o conhecimento adquirido, generalizando para responder corretamente a novas entradas. Essa habilidade de generalização é fundamental para o reconhecimento de

padrões, permitindo à RNA identificar e classificar entradas nunca vistas, com base em exemplos de treinamento. Os exemplos de treinamento apresentados ao RNA são chamados de “padrões". Esses padrões podem ser visuais, ou auditivos,. Padrões podem ser estáticos, sem necessidade de considerar o tempo (como pontos em uma imagem) ou temporais, em que a ordem e velocidade dos elementos são importantes (como uma sequência de fonemas em uma faixa de áudio). Mesmo partes de um padrão, captadas por unidades da rede, são interpretadas como representações completas do padrão.

A capacidade de aprender com padrões e generalizar o conhecimento adquirido é, segundo OLIVEIRA [4], a principal característica das redes neurais. Para isso, as RNAs são treinadas utilizando algoritmos de aprendizado. Um algoritmo de aprendizado pode ser definido como um conjunto de regras e processos matemáticos que guiam a rede neural a ajustar seus parâmetros, como os pesos das conexões entre neurônios, com o objetivo de aprender padrões a partir dos dados de entrada. Algoritmos de aprendizagem podem variar de diferentes modelos de acordo com seus paradigmas, arquiteturas e formas de neurônio apresentadas.

As RNAs podem operar em três paradigmas de aprendizado: supervisionado, não supervisionado e por reforço. No aprendizado supervisionado, a rede recebe pares de entrada e saída, e um algoritmo ajusta as conexões internas para minimizar o erro. No aprendizado não supervisionado, sem rótulos de saída, a rede organiza os dados em classes semelhantes de forma autônoma. Já no aprendizado por reforço, o ajuste ocorre com base no feedback de acertos ou erros, mesmo sem uma saída específica para cada entrada.

Existe atualmente uma variedade de formas de neurônios artificiais existentes, mas, essencialmente todos podem ser reduzidos a dois tipos fundamentais: o modelo de McCulloch-Pitts (MCP), que representa uma simplificação do neurônio biológico e é conhecido como "neurônio com pesos" devido à escolha de “valor” nas entradas, e o modelo digital, mais próximo do funcionamento de uma máquina do que o natural, é chamado de "neurônio sem pesos." Ambos os tipos de neurônios serão discutidos com mais detalhes nas seções 2.4 e 2.5, respectivamente, adiante.

Os tipos mais comuns de RNAs incluem as Redes Neurais Convolucionais (CNNs) e as Redes Neurais Recorrentes (RNNs). As CNNs são amplamente aplicadas em visão computacional por capturarem padrões espaciais em imagens, enquanto as RNNs são ideais para séries temporais e processamento de linguagem, devido à capacidade de manter informações de estados anteriores. Além disso, as redes sem peso, como o modelo WiSARD,

abordam o reconhecimento de padrões de uma forma inovadora, empregando memórias RAM para representar e processar informações de maneira eficaz e com baixo custo computacional.

2.4 Redes neurais com peso

Dando continuidade ao estudo das Redes Neurais Artificiais (RNA), é essencial explorar as Redes Neurais Com-Peso, que se destacam pelo uso de conexões ponderadas para ajustar seu aprendizado. Essa abordagem permite maior flexibilidade na resolução de problemas complexos, sendo amplamente utilizada em diversas aplicações, como reconhecimento de padrões, visão computacional e predição.

O desenvolvimento das Redes Neurais com Pesos teve como base o modelo proposto por McCulloch e Pitts (1943) descrito no trabalho "A logical calculus of the ideas immanent in nervous activity". Esse modelo, conhecido como neurônio McCulloch-Pitts (MCP), oferece uma representação matemática simplificada do funcionamento do neurônio biológico. O MCP é um dispositivo binário que processa entradas e gera uma saída booleana (0 ou 1), dependendo da soma ponderada das entradas. Nesse contexto, as entradas simulam os estímulos recebidos pelos dendritos, os pesos representam as sinapses, o somatório emula o corpo celular e a saída corresponde ao disparo do axônio (McCulloch e Pitts, 1943).

Matematicamente, o funcionamento do neurônio MCP pode ser descrito pela soma ponderada das entradas e pela aplicação de uma função de ativação. A equação que descreve essa operação é:

$$S = \sum_{i=1}^{n} w_i x_i + \mathrm{b} \tag{2.1}$$

$$y = f(s) \tag{2.2}$$

Nessa formulação, S é o somatório ponderado com o viés (b), $f(s)$ é a função de ativação, w_i são os pesos associados às entradas x_i, e y é a saída do neurônio. A função de ativação pode variar entre linear e não linear, sendo as funções sigmóide e tangente hiperbólica amplamente usadas devido à continuidade e diferenciabilidade, o que facilita a aplicação de métodos baseados em gradiente descendente.

O modelo de McCulloch-Pitts foi posteriormente expandido por Rosenblatt em 1958, levando ao desenvolvimento do perceptron, um modelo capaz de resolver problemas

linearmente separáveis. Esse modelo baseia-se em um processamento em que as entradas são ponderadas pelos pesos, e a saída é ativada com base em uma função específica. No entanto, o perceptron simples apresenta limitações em problemas não linearmente separáveis, como o problema XOR, o que motivou o surgimento de estruturas mais complexas.

Visualmente, um neurônio típico com pesos pode ser representado conforme a imagem abaixo:

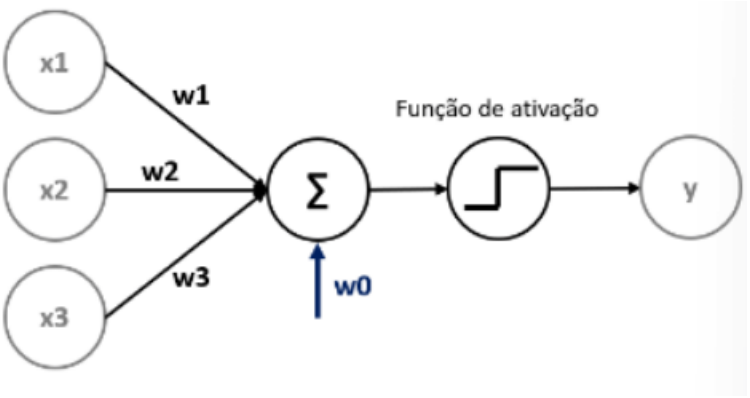

Figura 2: Representação do Perceptron (adaptado por ARAÚJO,2023)

O Perceptron Multicamadas (MLP), ou "MultiLayer Perceptron", surgiu como uma evolução significativa das redes com pesos. Essa rede é composta por múltiplas camadas, incluindo uma camada de entrada, uma ou mais camadas intermediárias (ou ocultas) e uma camada de saída. A camada de entrada realiza um mapeamento inicial dos dados, enquanto as camadas intermediárias introduzem funções de ativação não lineares, permitindo que a rede modele relações complexas. Já a camada de saída pode ser composta por neurônios lineares ou não, dependendo da aplicação.

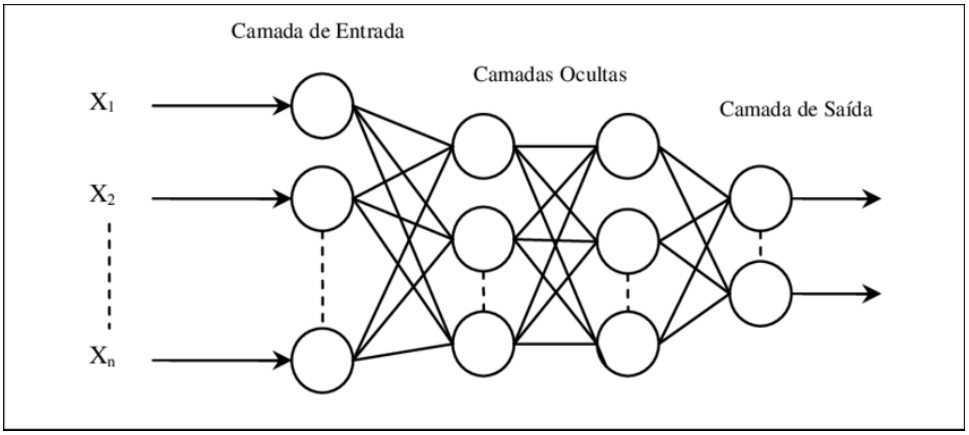

Figura 3: Perceptron Multicamadas (Sobreiro et al., 2008)

As redes MLP possuem a capacidade de aproximar qualquer função matemática, o que as torna úteis na solução de problemas que envolvem classificação, regressão e reconhecimento de padrões. Essa característica foi formalizada por Cybenko (1988), que demonstrou que uma rede com pelo menos uma camada intermediária é suficiente para aproximar funções arbitrárias. Além disso, a inclusão de camadas intermediárias possibilitou a resolução de problemas não linearmente separáveis, ampliando consideravelmente o campo de aplicações dessas redes.

O treinamento das redes MLP geralmente é realizado de forma supervisionada, utilizando algoritmos como o backpropagation, introduzido por Rumelhart et al. em 1986. Esse método ajusta os pesos da rede para minimizar o erro entre as saídas previstas e os valores reais, utilizando o gradiente descendente para guiar o processo de aprendizado. A capacidade de ajustar os pesos com eficiência permitiu que as redes MLP se consolidassem como uma das principais ferramentas em inteligência artificial.

As Redes Neurais Com-Peso têm demonstrado grande relevância na resolução de problemas complexos. A sua flexibilidade, aliada à capacidade de aprendizado e generalização, continua impulsionando avanços no campo da Inteligência Artificial, com aplicações em áreas como visão computacional, processamento de linguagem natural e predição de séries temporais.

2.5 Redes neurais sem-peso

As Redes Neurais Sem-Peso (RNSP) surgiram como uma abordagem inovadora ao paradigma tradicional das redes neurais artificiais, sendo introduzidas por Aleksander (1967). Diferentemente das redes com peso, as RNSP foram projetadas com base em dispositivos digitais de memória, como as memórias de acesso aleatório (RAM). O modelo inicial, chamado SLAM (Stored Logic Adaptive Microcircuit), foi posteriormente refinado para o modelo RAM, resultando em neurônios digitais conhecidos como "neurônios sem peso" ou "neurônios baseados em RAM" (Aleksander, 1966). Esses neurônios foram fundamentais para a criação de sistemas como o WiSARD (Wilkie, Stonham e Aleksander’s Recognition Device), amplamente utilizado para reconhecimento de padrões desde os anos 1980.

A principal diferença entre as RNSP e as redes neurais com peso (RNA tradicionais) reside na forma como armazenam e processam informações. Nas redes com peso, os dados aprendidos durante o treinamento são armazenados nos pesos ajustáveis associados às conexões entre os neurônios. Já nas RNSP, o aprendizado ocorre diretamente nas memórias dos neurônios, em forma de tabelas-verdade. Essas tabelas registram padrões binários, eliminando a necessidade de ajustes contínuos de parâmetros, como ocorre nos modelos baseados em gradiente descendente.

Além disso, as RNSP têm conectividade parcial, ou seja, cada neurônio é conectado apenas a uma fração das entradas do sistema. Essa característica é consequência da natureza exponencial do espaço de memória necessário para lidar com grandes números de entradas. A conectividade parcial reduz a complexidade computacional, tornando as RNSP mais eficientes em termos de processamento e consumo de recursos. Contudo, essa característica também limita sua capacidade de representar relações complexas entre os dados, uma vez que o aprendizado e a generalização ocorrem predominantemente no nível da rede, e não nos neurônios individuais.

Embora menos flexíveis que as redes com peso para resolver problemas altamente não lineares, as RNSP demonstram diversas vantagens. Sua robustez a ruídos, eficiência na classificação de padrões e capacidade de aprendizado rápido tornam-nas ideais para aplicações práticas, como sistemas de autenticação, análise de dados binários e reconhecimento de padrões. Estudos apontam que, mesmo com suas limitações, as RNSP podem ser equiparadas em poder computacional a Máquinas de Turing, devido à capacidade de reescrever e atualizar informações em sua memória (de Oliveira, 1992).

Entre os modelos mais conhecidos de neurônios sem peso, além do WiSARD, estão o Probabilistic Logic Neuron (PLN), o Multiple-valued Probabilistic Logic Neuron (MPLN) e o Probabilistic RAM (pRAM). Esses modelos expandiram as possibilidades de aplicação das RNSP, permitindo que lidasse com entradas contínuas e aprendizado por reforço, características que as tornam úteis para aproximação de funções e aprendizado em sistemas adaptativos.

Em resumo, as Redes Neurais Sem-Peso representam uma alternativa poderosa e eficiente para aplicações específicas, destacando-se por sua simplicidade, velocidade de processamento e adaptabilidade. Sua origem, marcada pelos avanços de Aleksander e seus colaboradores, continua a influenciar a pesquisa em computação neural e sistemas digitais modernos.

3. Reconhecimento de emoções no processamento de voz em língua portuguesa com redes neurais sem pesos

Este capítulo apresenta a descrição detalhada dos métodos e algoritmos utilizados para o reconhecimento de padrões que identificam emoção na voz humana utilizando redes neurais sem peso. Os passos são descritos desde a forma como o dataset foi organizado e criado até a forma como o sistema funciona, de forma superficial, e ao final a metodologia utilizada para realizar os testes é explicada.

3.1 Introdução

Diferentes áreas da tecnologia que utilizam fortemente da interação humano-computador estão em demanda da identificação de emoções em faixas de áudio humana. Como por exemplo assistentes virtuais que buscam sempre um melhor atendimento e empatia com o usuário, sistemas de monitoramento emocional que precisam detectar possíveis estados de depressão ou ansiedade, aplicações de entretenimento que se adaptam na hora ao estado emocional da pessoa para uma melhor experiência, sistemas de segurança que identificam sinais de estresse medo ou raiva alertando os profissionais de segurança entre outros.

O objetivo de utilizar RNSP nas situações relatadas tem como objetivo aumentar a eficiência computacional e a simplicidade estrutural destes sistemas, podendo ser utilizado amplamente e com uma maior precisão. O trabalho de OLIVEIRA [4] que inspirou este projeto testa resolver o problema de detecção de falhas em sistemas industriais dinâmicos e aplicam diversas abordagens diferentes de RNSPs para a resolução, inclusive uma abordagem não dinâmica que utiliza “bleaching” em seu processo de aprendizado denominada “WiSARD Contador”, que foi escolhida como algoritmo principal a ser utilizado neste projeto.

3.2 Dados utilizados

O banco de dados utilizado e features extraídas seguiram o padrão do trabalho DEEP de [3] Campos e Moutinho, para que uma comparação nas mesmas condições seja realizada no final do trabalho, porém adaptado para a entrada do algoritmo da rede WiSARD. O banco

de dados utilizado o “VERBO: Voice Emotion Recognition dataBase in Portuguese language”, sendo ele composto por 1176 arquivos de áudio, de 2 à 5 segundos, com diferentes emoções na língua portuguesa do Brasil, realizadas por 12 atores brasileiros de diferentes idades, regiões e gêneros. As faixas de áudio são divididas com base no modelo de Russel [7], sendo estas “alegria”, “nojo”, “medo”, “raiva”, “surpresa” e “tristeza” adicionando ao final uma sétima classificação denominada “neutro”.

3.2.1 Extração de features

Para a realização deste trabalho foram importadas do trabalho DEEP [3] três arquivos .csv (MFCC.csv, Chroma.csv e Prosody.csv) que representam as features já extraídas do banco de dados VERBO, porém como caracter de entendimento será explicado mais adiante como foi realizado o processo de extração destes sinais.

A extração de features realizada foi o processo da retirada de características de trechos da faixa de áudio representadas de forma bidimensional, ou seja o processo retira partes úteis da faixa de áudio em formato numérico que vão ser utilizadas no processo de aprendizagem do sistema. As features selecionadas foram as **MFCC**, **Cromáticas** e **Prosódicas**. Cada uma delas foi escolhida pois a **MFCC** é ideal para capturar características espectrais de sinais de fala e música, as **Cromáticas** são úteis para tarefas relacionadas à tonalidade e as **Prosódicas** fornecem informações emocionais e são essenciais para analisar sentimentos ou estados emocionais na fala.

Para analisar cada unidade de frequência individualmente sem precisar analisar o áudio como um todo é realizado o processo de framing, que divide a faixa de áudio em diversos frames ou quadros que vão ter suas features extraídas individualmente, deixando a classificação mais precisa. Esses quadros possuem sobreposição, possuem 50ms de comprimento e intervalos de captura de 10ms como representado na **figura 4** a seguir.

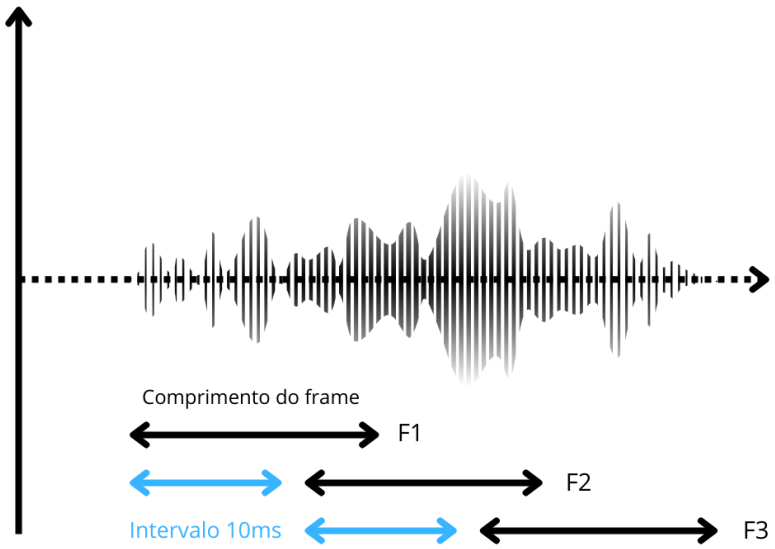

Figura 4: Representação do processo de enquadramento. Fonte: *Elaboração própria*.

3.2.2 Pré-processamento dos dados

Os dados importados ainda precisam passar por um processo de adequação ao sistema tornando-os consistentes, limpos e adequados para análise ou treinamento de modelos chamado de pré-processamento. Primeiramente os 3 datasets gerados tinham o problema de cada faixa de audio ser variável em seu tempo, por isso possuíam diferentes quantidades de frames gerados por áudio, então um processo de "Padding" foi realizado adicionando valores nulos em todas as faixas de áudio até que todas possuíam a mesma quantidade de frames igual ao áudio com o maior valor de frames, padronizando todos os áudios em 540 frames.

3.3 Rede WiSARD

O sistema utilizado é a rede RNSP mais famosa, à rede WiSARD (Wilkie, Stonham e Aleksander's Recognition Device, dispositivo de reconhecimento de Wilkie, Stonham e Aleksander) proposta por Aleksander et al. (1984) [14]. A rede WiSARD é composta por dois ou mais discriminadores, cada um destes diretamente responsável por uma classe do sistema, armazenando seus padrões. A quantidade de RAM em cada discriminador depende

do modelo de mapeamento dos padrões de entrada, onde cada sub padrão é representado por uma n-tupla. Durante a fase de teste, todos os discriminadores recebem a mesma entrada, e cada RAM identifica e processa o padrão correspondente. A saída de um discriminador é a soma das RAMs que retornam o valor 1, sendo a classe com maior soma a reconhecida.

No sistema que irei utilizar serão usados 7 discriminadores, cada um deles representando uma emoção do sistema. Representado na **figura 5** está um exemplo de funcionamento de um discriminador, onde ele divide a entrada em n-tuplas e a fornece para cada RAM do discriminador, e o resultado da memória de cada RAM é somada e devolvida para identificação da classe.

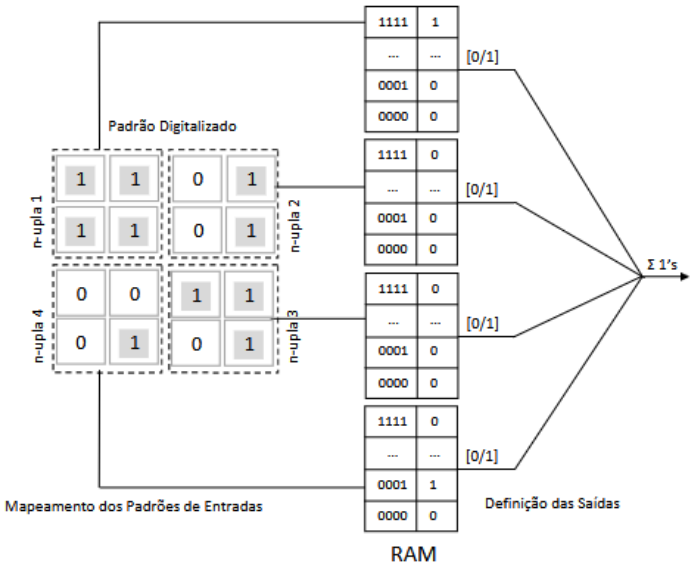

Figura 5: "Representação Esquemática de um Discriminador". (OLIVEIRA, 2018).

A rede também calcula um nível de confiança no reconhecimento (Cw), baseado na diferença entre a maior e a segunda maior pontuação como mostrado no caçulo a seguir:

$$C_w = \frac{P_{max} - P_{2,max}}{P_{max}} \quad (3)$$

Caso ocorra empate, a confiança é nula e é utilizado o critério de desempate aleatório para escolher um vencedor. Este critério aleatório pode reduzir a acurácia do sistema, uma vez que existe a chance do sistema ser classificado erroneamente. Outra limitação da WiSARD é a saturação dos endereços de memória das RAMs e mediante a estes problemas

técnicas têm sido propostas para superar as limitações da aleatoriedade e da saturação dos endereços.

3.3.1 Técnica de “Bleaching”

A técnica de **bleaching** (ou refinamento) é usada em redes WiSARD para reduzir o impacto do critério de aleatoriedade e melhorar a confiabilidade na resolução de empates entre discriminadores. Nesse processo, os endereços de memória das RAM armazenam valores inteiros, representando o número de vezes que foram acessados durante o treinamento. Na fase de teste, a saída de uma RAM é igual a 1 se o valor armazenado for igual ou maior que um limiar pré-definido.

Esse limiar pode ser configurado de duas maneiras principais: em sua forma “convencional” ou em uma versão “percentual”. Para este trabalho o modelo convencional será utilizado. No bleaching convencional o mesmo limiar é usado para todos os discriminadores. Inicialmente, o limiar é zero, funcionando como se a técnica não estivesse ativa. Se dois ou mais discriminadores tiverem a mesma pontuação, o limiar é incrementado gradualmente, reduzindo o número de RAMs com saída igual a 1 até que um discriminador seja escolhido ou todas as pontuações zerem, caso em que a decisão é feita de forma aleatória.

Além de resolver empates, o bleaching ajuda a mitigar a saturação de memória, eliminando conteúdos com pouca relevância ou representatividade. Nesse caso, um limiar de frequência mínima é usado para apagar conteúdos pouco utilizados. Também é possível implementar o bleaching temporal, que apaga conteúdos de endereços que não foram acessados recentemente, liberando memória e descartando discriminadores sem dados relevantes. Essa variação é mais adequada para algoritmos não supervisionados, onde a quantidade de discriminadores é ajustada durante o aprendizado.

3.4 Estrutura do sistema

Este trabalho propõe um sistema capaz de categorizar emoções em faixas de áudio da língua portuguesa, e possui como padrão de entrada do sistema x = [a1,..., an,b1,..., bn,c1,..., cn], onde x equivale aos atributos retirados de uma faixa de áudio a ser classificada. Os

atributos representados por ‘a’ ,‘b’ e ‘c’ correspondem às diferentes features retiradas de n-frames da faixa de áudio, sendo agrupados lado a lado para a etapa de seleção de atributos.

O sistema pode ser dividido em 3 partes que atuam em conjunto, como na figura 3.4, sendo estas a “Seleção de padrões”, o “modelo RNSP WiSARD” e o “Cálculo *Accuracy* e *F1-Score*”.

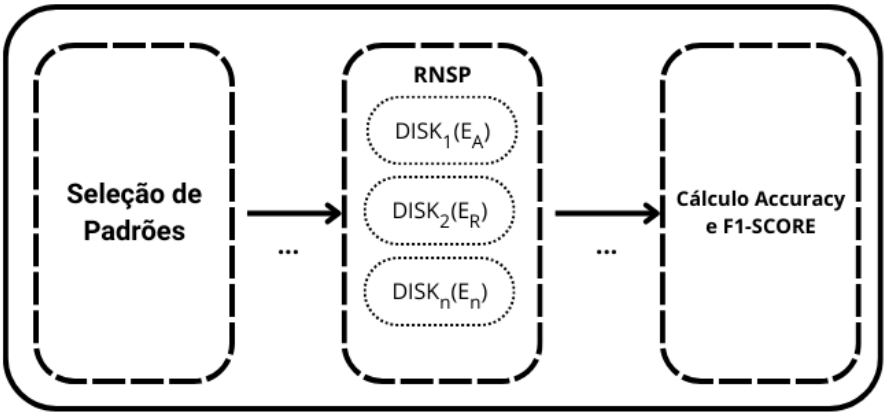

Figura 6: Estrutura do sistema RNSP WiSARD utilizado.

Na representação do sistema acima, Figura 6, o $DISK_n$ se refere aos discriminadores utilizados, cada um referente a uma emoção específica, representada por “E” . Na primeira etapa é uma espécie de pré-processamento, onde é aplicado um algoritmo de seleção dos atributos para a formação de padrões de identificação das emoções. Além disso, é realizado o processamento temporal dos dados, seguindo modelos de mapeamento específicos. Na segunda etapa utiliza-se uma Rede Neural Sem Pesos (RNSP) do tipo WiSARD, composta por 7 discriminadores, correspondentes às emoções, treinados de forma independente para reconhecer somente sua classe de padrões. Na terceira etapa são aplicadas métricas famosas para a avaliação do desempenho e comparação com outros modelos, a *Accuracy* e o *F1-Score*. Todas estas etapas serão descritas com mais detalhes nas seções seguintes.

3.4.1 Processamento dos padrões de entrada

3.4.1.1 Seleção de atributos

A eficácia de um sistema de classificação de padrões está em sua capacidade de dividir o espaço de estados de forma ideal para o problema em análise. Uma solução satisfatória se aproxima dessa partição ideal, exigindo classificadores distintos em suas generalizações para

obter precisão e eficácia na classificação. Em sistemas com múltiplos classificadores (sendo estes no projeto os discriminadores) a seleção de atributos desempenha um papel importante, pois diferentes subconjuntos de atributos permitem maior distinção nas generalizações, reduzindo erros baseados nos limites das classes.

Neste trabalho, foi utilizado o algoritmo **RecPun (Recompensa/Punição)**, proposto por Vale et al. (2010), para realizar a seleção de atributos por classe. Esse algoritmo ordena os atributos com base em sua importância para cada classe, assegurando que os atributos escolhidos sejam relevantes apenas para a classe específica. A rede WiSARD se adapta bem ao RecPun porque seus discriminadores são treinados individualmente para representar classes específicas.

O algoritmo RecPun realiza a seleção em duas etapas. Na primeira, utiliza variância ou correlação de Spearman para ordenar e selecionar as features (atributos) por classe. Na segunda, aplica o parâmetro **RP (Recompensa/Punição),** com a variância fornecida, para refinar a seleção. Segundo Vale et al. (2010) o cálculo do RP pode ser expressado da seguinte forma:

1) $$RP_i = Rec_i + Pun_i$$

2) $$Rec_i = V_{i,c} + \frac{NA}{NA+R_{i,c}}$$

3) $$Pun_i = \frac{1}{NC-1} \cdot \sum_{c=1}^{NC,\, C \neq i} \left(V_{i,c} + \frac{NA}{NA+R_{i,c}} \right)$$

Para evitar a repetição excessiva de atributos entre classes, foi adicionado um passo adicional ao RecPun. Esse passo limita a quantidade de atributos comuns entre classes a, no máximo, 50%. Caso essa quantidade seja excedida, os atributos excedentes são substituídos por outros de menor importância, seguindo a ordenação definida pelo algoritmo. Isso promove maior diferenciação entre os atributos selecionados para as classes e melhora a eficácia do sistema.

3.4.1.2 Mapeamentos dos padrões de entrada

Os modelos de mapeamento são responsáveis por transformar os padrões de entrada em endereços de memória para a RNSP. São divididos em dois tipos principais: **mapeamento simples** e **mapeamento temporal**.

- **Mapeamento simples**: Processa cada padrão de entrada de forma individual, associando-o diretamente ao discriminador da classe correspondente.

- **Mapeamento temporal**: Trabalha com séries temporais, analisando os padrões em um horizonte h>1h > 1h>1, utilizando métodos como **média móvel** e **janela deslizante**. Esse tipo de mapeamento gera padrões comportamentais que alimentam os discriminadores da RNSP, tanto para aplicações uni variáveis quanto multivariáveis. No caso de múltiplos atributos, o padrão comportamental é formado pela concatenação das médias dos atributos processados.

Esses modelos se distinguem pela forma como calculam a média e escolhem o atributo representativo da série. A escolha do modelo e a parametrização adequada são fundamentais, já que aplicações diferentes exigem pré-processamentos específicos. Para o projeto proposto foi utilizado o mapeamento simples pois o problema proposto não possui uma dependência temporal, ou seja os dados são independentes e não têm influência do passado.

3.4.2 O WiSARD contador

A Rede Neural Sem Pesos (RNSP) utilizada neste estudo tem como objetivo classificar emoções presentes em faixas de áudio fornecidas. Para isso, a rede é composta por uma única camada contendo oito discriminadores, cada um responsável por identificar uma emoção específica, como descrito nas seções anteriores.

A estrutura básica de uma rede RAM tradicional exige que todas as memórias associadas a um determinado padrão apresentem um valor positivo para que a entrada seja reconhecida como pertencente a uma classe específica. Dessa forma, esse tipo de rede só consegue identificar padrões já apresentados durante a fase de treinamento, não sendo capaz de generalizar para novos casos.

Para lidar com esse problema e permitir maior flexibilidade na classificação das emoções, foi utilizada a rede WiSARD. Diferente da RAM básica, a WiSARD não exige que todas as memórias concordem para classificar um padrão. Em vez disso, ela atribui a entrada ao discriminador que apresenta a maior quantidade de ativações positivas. Esse critério permite que a rede reconheça emoções que podem não ter sido vistas exatamente da mesma forma durante o treinamento, desde que possuam características semelhantes às emoções

previamente aprendidas. A implementação pode ser realizada utilizando dois modelos distintos da rede WiSARD:

- **WiSARD Padrão (WP)**: Esse modelo consiste em uma única camada com oito discriminadores, onde cada um é treinado separadamente para reconhecer uma das emoções. Durante a fase de teste, o mesmo padrão de entrada é apresentado a todos os discriminadores ao mesmo tempo. Cada unidade RAM verifica os endereços de memória correspondentes ao padrão apresentado e retorna o valor armazenado (0 ou 1). O resultado final da rede é determinado pelo discriminador que apresenta o maior número de respostas positivas. No caso de empate entre dois ou mais discriminadores, a decisão sobre a emoção identificada ocorre de maneira aleatória.
- **WiSARD Contador (WC)**: A estrutura desse modelo é semelhante à WiSARD Padrão, mas com uma diferença fundamental: em vez de armazenar apenas valores binários (0 ou 1), os endereços de memória registram valores numéricos escaláveis. Esse modelo possibilita o uso de um mecanismo chamado bleaching, que reduz a aleatoriedade na tomada de decisão em situações de empate. O bleaching funciona ajustando progressivamente o critério de ativação, permitindo que padrões menos expressivos sejam descartados até que um único discriminador seja escolhido.

Para o sistema proposto foi escolhido o modelo WC do WiSARD, pois lida melhor com as variações emocionais e faz uso da técnica de bleaching mencionada para minimizar decisões aleatórias. Da escolha entre os dois modelos de rede, diferentes parâmetros foram ajustados para otimizar o desempenho do sistema. Alguns dos principais fatores analisados incluem:

- **Forma de representar os padrões de entrada**: Foram testadas diferentes configurações de bits (8 ou 32 bits) para representar os atributos acústicos e calcular as distâncias de Hamming.
- **Número de bits por entrada na rede**: Foram exploradas quantidades de **8, 12, 16 ou 32 bits** multiplicadas pelo número de atributos utilizados.
- **Quantidade de entradas em cada RAM**: Diferentes tamanhos de entradas foram testados, variando entre **2, 3, 4, 6 e 8 bits**.
- **Número de RAMs por discriminador**: Esse número foi determinado dividindo a quantidade total de bits de entrada pelo tamanho da entrada de cada RAM.

- **Método de seleção de atributos**: Foram comparados dois métodos: RecPun e RecPun modificado, que diferem na forma de escolher os atributos mais relevantes para a classificação das emoções.
- **Uso do bleaching**: Foram analisadas duas variações, o bleaching simples e o percentual, para verificar seu impacto na redução de empates e na estabilidade do modelo.

A correta configuração desses parâmetros foi essencial para melhorar a eficiência do sistema, permitindo que a rede WiSARD identificasse e classificasse as emoções com maior precisão. O ajuste desses fatores também possibilitou avaliar a robustez da rede diante de variações emocionais sutis ou padrões nunca vistos durante o treinamento, aumentando a capacidade da RNSP de lidar com diferentes tonalidades emocionais nas faixas de áudio analisadas. Os atributos finais escolhidos foram padrões de entrada com 8 bits, 12 bits por entrada na rede, bleaching simples e método RecPun original.

3.4.3 Métricas de avaliação

A avaliação do desempenho do sistema WiSARD Contador na classificação de emoções é fundamental para garantir a eficácia do modelo na identificação correta dos padrões emocionais em áudios. Para isso, são empregadas as métricas Accuracy e *F1-Score*, amplamente utilizadas na literatura para medir a qualidade de modelos de classificação (SOKOLOVA; LAPALME, 2009).

O sistema WiSARD Contador proposto neste trabalho realiza a classificação das emoções e, em seguida, o Cálculo de *Accuracy* e *F1-Score* avalia a precisão e confiabilidade dos resultados. A utilização dessas métricas permite validar a eficácia da abordagem adotada, garantindo que o modelo seja capaz de identificar corretamente diferentes espectros emocionais em amostras de áudio, com base na arquitetura WiSARD Contador e nas técnicas de mapeamento temporal empregadas.

3.4.4 Metodologia para o treinamento e teste do sistema

A **figura 7** a seguir representa a metodologia e o processo utilizado para modelagem do sistema, onde “n” corresponde as features selecionadas e “c” as classes de emoção:

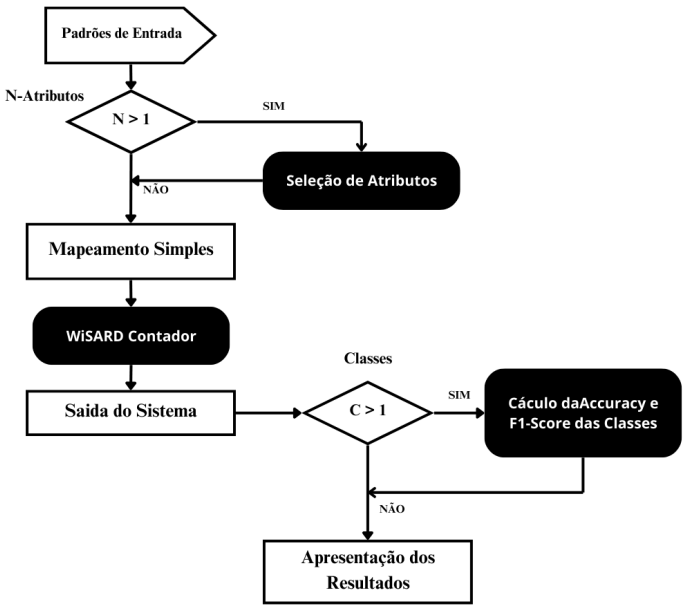

Figura 7: Metodologia para os Treinamento e Testes com o Sistema.

Após o mapeamento dos dados, o padrão comportamental extraído é apresentado à Rede Neural sem Pesos (RNSP) para classificação. Como a rede WiSARD utiliza um discriminador para cada classe emocional, a saída representa a classificação do áudio, indicando o percentual de acertos para cada emoção analisada. Durante a fase de avaliação, os resultados são apresentados com base na quantidade e na taxa de acerto para cada emoção.

4. Apresentação e análise dos resultados

Nesta seção será avaliado o desempenho do módulo de redes neurais sem peso em relação ao projeto DEEP, que utiliza redes convolucionais e Hiper Parametrização, proposto por Campos e Moutinho (2020). Utilizaremos as métricas de avaliação F1-score e Acurácia para comparar o desempenho dos dois modelos e suas respectivas limitações e benefícios.

4.1 Descrição do modelo DEEP e resultados

O modelo DEEP utiliza Redes Neurais Convolucionais (CNNs) para extrair padrões relevantes dos espectrogramas das faixas de áudio. Ele passa por um processo de Hiper Parametrização, ajustando automaticamente parâmetros como número de filtros, tamanho do

kernel e taxa de aprendizado para otimizar o desempenho. Após a extração de características, a rede classifica a emoção presente no áudio com base nos padrões aprendidos. Todo o processo de produção do trabalho e resultados podem ser observados no trabalho referente (CAMPOS; MOUTINHO, 2020, p. 28).

Apesar do processo de processamento dos dados entre a rede sem peso e o DEEP sere, parecidos, estas são somente até certo ponto, especificamente até o processo de preenchimento (*Padding*), pois a redes sem peso possui um pré processamento extra no formato do algoritmo RecPun selecionando os melhores atributos possíveis e a DEEP usa hyper parametrização. Levando em consideração essa diferença entre os processos dos dados é difícil realizar uma comparação direta sobre os diferentes modelos, então os resultados apresentados servirão mais como comparações sobre os diferentes modelos como um todo e não sobre os modelos específicos, impedindo uma comparação direta que mostrasse um modelo superior ao outro.

O processo de treino e teste para obtenção de resultados do modelo sem pesos foi realizado um total de 10 vezes, capturando diferentes entradas de treino e teste, levando em conta o balanceamento de entradas, ou seja, garantindo que a entrada de treino e teste não tivessem quantidade de classes de forma desbalanceada.

A métrica estabelecida para a comparação dos modelos como um todo foi o cálculo do *Accuracy* e *F1-Score* amplamente utilizadas para validação de sistema de classificação. As figuras 8 e 9 a seguir mostram os resultados em média obtidos do módulo sem pesos em relação aos resultados do modelo DEEP em razão das 7 emoções classificadas de acordo com as métricas estabelecidas.

Emoção	Acurácia Sem Peso	F1 Sem Peso	Acurácia DEEP	F1 DEEP
Neutro	1.0000	1.0000	0.7285	0.3378
Nojo	1.0000	1.0000	0.7482	0.3858
Medo	0.9991	0.9969	0.7843	0.3046
Alegria	0.9997	0.9990	0.7637	0.3199
Raiva	0.9994	0.9980	0.6374	0.3371
Surpresa	1.0000	1.0000	0.8556	0.4900
Tristeza	1.0000	1.0000	0.8505	0.3326

Figura 8: Comparação dos resultados do modelo Sem pesos e do modelo DEEP

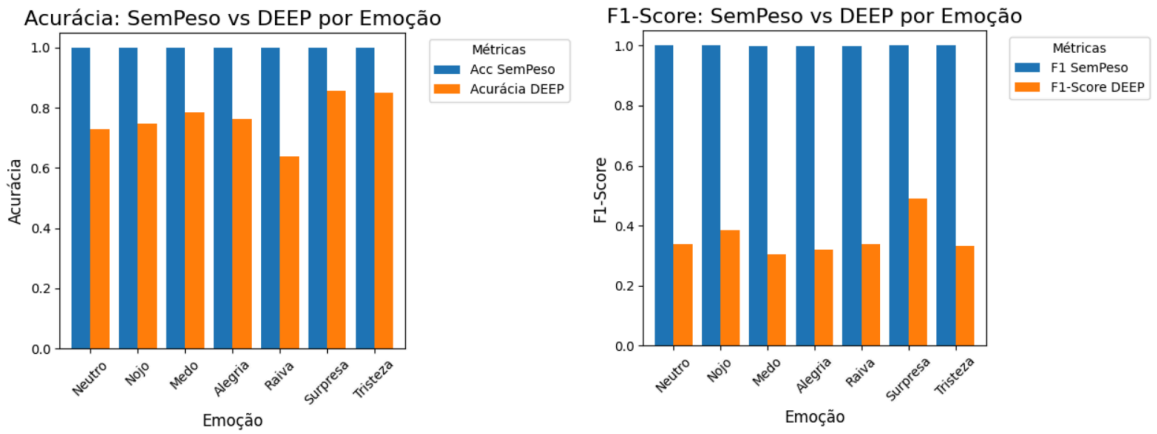

Figura 9: Gráficos de colunas de comparação da Acurácia e Gráfico do F1-Score

Para compreender as variações causadas dos testes realizados individualmente, sem comparação com o modelo DEEP foi gerado uma tabela de acertos por testes realizados na figura 10 como também um mapa de calor na Figura 11. Estes mapas representam as variações da porcentagem de acerto, e o mapa de calor representa através de cores fortes resultados altos e cores fracas para os baixos.

qTestada	qAcertos	%rede	neutro	desgosto	medo	alegria	raiva	surpresa	tristeza	%classificacao
350.0	350.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
350.0	350.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
350.0	349.0	99.71	100.0	100.0	98.0	100.0	100.0	100.0	100.0	99.71
350.0	350.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
350.0	350.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
350.0	350.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
350.0	349.0	99.71	100.0	100.0	98.0	100.0	100.0	100.0	100.0	100.0
350.0	349.0	99.71	100.0	100.0	98.0	100.0	100.0	100.0	100.0	99.71
350.0	350.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
350.0	350.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

Figura 10: Tabela de porcentagem de acertos do Modelo Sem Peso por Amostragem

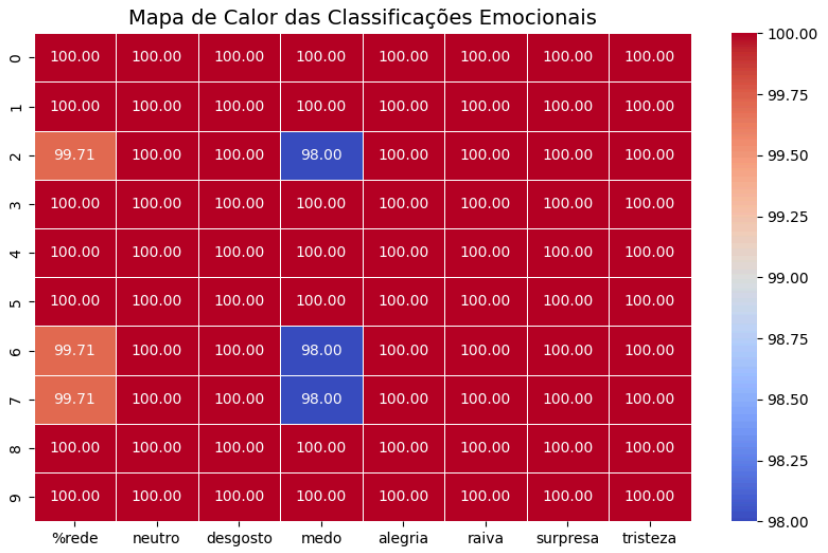

Figura 11: Mapa de calor de acertos do Modelo Sem Peso por Amostragem

4.2 Avaliação dos resultados obtidos

Ao analisar os resultados comparativos entre os modelos Sem Peso e DEEP, observa-se que o modelo Sem Peso apresenta um desempenho superior, principalmente nas métricas de Acurácia e F1-Score. No geral, o modelo Sem Peso demonstrou taxas de acerto de até 100% em quase todas as categorias de emoção, com exceção das emoções alegria, medo e raiva, que apresentaram uma ligeira queda no desempenho. Por exemplo, na identificação de medo, a acurácia foi de 99.71% e o F1-score de 98%, valores que ainda são elevados, mas indicam uma leve perda de desempenho em comparação com as outras emoções, que tiveram 100% de acerto.

A Acurácia Sem Peso foi consistentemente muito alta, variando entre 99.71% e 100% para a maioria das amostras. Além disso, o F1-score Sem Peso também permaneceu próximo de 100% para as emoções mais bem classificadas, como neutro, desgosto, alegria, raiva, surpresa e tristeza. Esses valores indicam uma robustez no modelo Sem Peso ao classificar emoções na língua portuguesa, refletindo sua eficácia na identificação precisa de padrões emocionais na voz.

Por outro lado, o modelo DEEP apresentou um desempenho mais modesto, com a Acurácia DEEP variando entre 63.74% e 85.56%, e o F1-score DEEP apresentando valores entre 30.26% e 49.00%. A diferença mais significativa entre os modelos ocorre nas emoções

medo e raiva, onde o modelo DEEP teve uma acurácia significativamente inferior (em torno de 63% a 73%) e um F1-score consideravelmente mais baixo (aproximadamente 30% a 49%).

Essas comparações revelam que, embora o modelo DEEP seja capaz de capturar certos aspectos das emoções, o modelo Sem Peso parece ser mais eficiente em geral, especialmente em termos de precisão e recall para as categorias de emoção mais comuns.

É importante observar, no entanto, que a análise realizada é preliminar, pois a quantidade de dados testados pode ser insuficiente para uma avaliação conclusiva e de alta confiabilidade. O conjunto de dados utilizado foi composto por 350 amostras em cada categoria, o que pode não ser representativo o suficiente para validar a performance dos modelos em diferentes condições. Essa limitação pode ter contribuído para a baixa variação de erro observada nos testes realizados, que apresentou valores consistentes em todas as amostras.

5. Considerações finais

Este trabalho teve como objetivo principal explorar e aprimorar os sistemas de reconhecimento de emoções na fala (SER), utilizando Redes Neurais Sem Pesos (RNSP) para identificar e classificar emoções em faixas de áudio de falantes da língua portuguesa. Para isso, foi utilizado o modelo WiSARD Contador, que se baseia no armazenamento de padrões diretamente na memória RAM, e seus resultados foram comparados com um modelo similar baseado em Redes Neurais Convolucionais (CNNs).

Os experimentos realizados demonstraram que o modelo sem pesos proposto se mostrou mais eficaz na classificação do conjunto de dados VERBO, obtendo taxas de acerto de até 100% em quase todas as categorias, exceto nas emoções alegria, medo e raiva, onde o desempenho foi inferior. As métricas utilizadas para avaliação, *Accuracy* e *F1-Score*, confirmam que o modelo WiSARD evidencia seu potencial para reconhecimento de emoções na voz com alta precisão.

Durante o desenvolvimento deste estudo, alguns desafios foram enfrentados, como a definição das métricas de avaliação e a escolha das melhores configurações do modelo WiSARD, uma vez que ele possui diversas parametrizações possíveis. Além disso, algumas limitações foram identificadas, como a dependência do banco de dados VERBO, que pode não representar todas as variações linguísticas da língua portuguesa, e o fato de que a

abordagem não foi testada em cenários de fala espontânea, o que pode comprometer sua aplicabilidade em ambientes reais.

Por outro lado, este trabalho trouxe contribuições importantes para o campo do reconhecimento de emoções na fala, demonstrando que modelos sem pesos podem ser uma alternativa eficiente às abordagens tradicionais baseadas em redes neurais profundas. Entre as vantagens do modelo WiSARD Contador, destaca-se o treinamento rápido e simples, pois, ao contrário das redes neurais convencionais que exigem um processo iterativo de ajuste de pesos, o WiSARD armazena padrões diretamente na memória RAM, eliminando cálculos complexos e permitindo um treinamento quase instantâneo. Outro ponto relevante é sua robustez a ruídos e pequenas variações, já que a classificação é baseada na contagem de coincidências entre padrões, o que torna o modelo mais tolerante a variações sutis na voz humana, como pequenas mudanças de tom ou intensidade, tornando-o adequado para ambientes não controlados. Além disso, seu baixo custo computacional se destaca, pois a ausência de cálculos complexos de pesos e gradientes torna o WiSARD altamente eficiente em processamento e consumo de energia, sendo uma alternativa viável para aplicações em dispositivos embarcados e sistemas de hardware limitado.

Entretanto, algumas limitações e desafios do modelo WiSARD também foram identificados ao longo do estudo. O primeiro desafio está na dificuldade em lidar com dados contínuos brutos, já que o modelo requer um pré-processamento dos sinais de áudio para que sejam transformados em padrões binarizados, o que pode levar à perda de informações importantes. Outra questão está na escalabilidade limitada, uma vez que, conforme o número de classes e atributos cresce, o consumo de memória aumenta significativamente, tornando necessário um controle eficiente da estrutura da rede. Além disso, o consumo elevado de memória é um fator limitante, pois o WiSARD memoriza padrões explicitamente, o que pode tornar inviável o armazenamento de grandes conjuntos de dados. A incapacidade de aprendizado incremental também é um desafio, pois, para que o modelo aceitasse novos dados e fosse reajustável, seria necessária a proposição de uma nova arquitetura de projeto.

Diante desses pontos, este estudo abre caminho para futuras pesquisas que busquem minimizar essas limitações e expandir a aplicabilidade do WiSARD. Trabalhos futuros podem explorar a eficácia do modelo em outros conjuntos de dados, testar sua capacidade de generalização para falas espontâneas, adaptá-lo para reconhecimento de emoções em tempo real e investigar os limites físicos da classificação sem pesos para entrada de dados contínuos reais. Além disso, o reconhecimento de emoções na fala possibilita diversas aplicações

práticas que podem tornar a experiência do usuário mais interativa e personalizada. Aplicativos e dispositivos inteligentes podem utilizar o estado emocional do usuário para aprimorar suas funcionalidades, como sistemas de recomendação de músicas e conteúdos audiovisuais que ajustam sugestões conforme o humor identificado. Assistentes virtuais podem melhorar a interação humano-máquina ao adaptar sua comunicação e respostas conforme o estado emocional do usuário. No contexto de e-commerce, lojas online poderiam personalizar ofertas e recomendações de produtos com base nas emoções detectadas. Dispositivos médicos e aplicativos de bem-estar podem ser desenvolvidos para monitorar o estado emocional do usuário, sugerindo técnicas de relaxamento, alertando sobre padrões de estresse e fornecendo suporte terapêutico baseado em inteligência artificial. Outra possibilidade interessante está na aplicação dessa tecnologia em jogos e experiências imersivas, permitindo que os jogos ajustem sua dificuldade ou narrativa conforme a emoção do jogador, tornando a experiência mais dinâmica e envolvente.

Dessa forma, este estudo demonstra o potencial das Redes Neurais Sem Pesos no reconhecimento de emoções na fala, oferecendo uma abordagem eficiente e de baixo custo, com perspectivas promissoras para sua aplicação em dispositivos embarcados, sistemas inteligentes e diversas soluções interativas.

BIBLIOGRAFIA

1. AOUANI, Hadhami; AYED, Yassine Ben. Speech Emotion Recognition with deep learning. *Procedia Computer Science*, v. 176, p. 251-260, 2020. ISSN 1877-0509. DOI: https://doi.org/10.1016/j.procs.2020.08.027. Disponível em: https://www.sciencedirect.com/science/article/pii/S1877050920318512. Acesso em: [data de acesso].
2. EL AYADI, Moataz; KAMEL, Mohamed S.; KARRAY, Fakhri. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, v. 44, n. 3, p. 572-587, 2011. DOI: https://doi.org/10.1016/j.patcog.2010.09.020. Disponível em: https://www.sciencedirect.com/science/article/pii/S0031320310004619. Acesso em: [data de acesso].

3. CAMPOS, Gabriel A.; MOUTINHO, Lucas da S. DEEP: Uma arquitetura para reconhecer emoção com base no espectro sonoro da voz de falantes da língua portuguesa. Universidade de Brasília, Instituto de Ciências Exatas, Departamento de Ciência da Computação, 2020.
4. OLIVEIRA, José Carlos Martins. Detecção e diagnóstico de falhas em processos dinâmicos com redes neurais sem pesos. 2018. Tese (Doutorado em Engenharia Industrial) – Universidade Federal da Bahia, Escola Politécnica, Salvador, 2018.
5. LUSQUINO FILHO, L. A. D. Pantheon: Classificação de emoções faciais utilizando a rede neural sem pesos WiSARD. Dissertação de Mestrado, Universidade Federal do Rio de Janeiro, 2018. Disponível em: (pantheon.ufrj.br)
6. SCHULLER, B. et al. *Speech emotion recognition: Applications and recent developments. IEEE International Conference on Multimedia and Expo (ICME)*, 2018.
7. EYBEN, F.; WÖLLMER, M.; SCHULLER, B. *Opensmile: the munich versatile and fast open-source audio feature extractor.* In: *Proceedings of the 18th ACM International Conference on Multimedia.* ACM, 2016. p. 1459-1462.
8. ABADI, M. et al. TensorFlow: A system for large-scale machine learning. In: *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*. 2016.
9. PASZKE, A. et al. PyTorch: An imperative style, high-performance deep learning library. In: *NeurIPS 2019*. 2019.
10. NETO, José Torres; et al. VERBO: Voice Emotion Recognition database in Portuguese language. *Journal of Computer Science*, v. 14, p. 1420–1430, nov. 2018. DOI: 10.3844/jcssp.2018.1420.1430.
11. NOGUEIRA, Kennyo. Estudo de respostas emocionais às cores no contexto de cartazes de cinema. *Design e Tecnologia*, v. 8, p. 1, 2018. DOI: 10.23972/det2018iss15pp1-11.
12. RUSSELL, James. A Circumplex Model of Affect. *Journal of Personality and Social Psychology*, v. 39, p. 1161-1178, 1980. DOI: 10.1037/h0077714.
13. RUSSELL, Stuart; NORVIG, Peter. *Artificial intelligence: a modern approach*. 2. ed. [S.l.]: [s.n.], 2002.
14. HECHT-NIELSEN, R. *Neurocomputing*. USA: Addison-Wesley Publishing Company, 1990.
15. ARAÚJO, Raquel. O Perceptron – Ciência de Dados. *Hashtag Treinamentos*, 15 maio 2023. Disponível em: https://www.hashtagtreinamentos.com/o-perceptron-ciencia-de-dados/. Acesso em: [20/07/2024].

16. SOBREIRO, Vinicius; SOUSA, Pedro; ARAÚJO, Leão; MENDONÇA, Michelle; NAGANO, Marcelo. *Uma estimação do valor da commodity de açúcar utilizando redes neurais artificiais*. 2008.
17. [AUTOR DESCONHECIDO]. O Perceptron – Parte 1. *Deep Learning Book Brasil*. Disponível em: https://www.deeplearningbook.com.br/o-perceptron-parte-1/. Acesso em: [20/07/2024].
18. ALEKSANDER, I. Adaptive systems of logic networks and binary memories. In: *Proceedings of the Spring Joint Computer Conference*, 30., 1967. p. 707-712.
19. CYBENKO, G. Continuous valued neural networks with two hidden layers are sufficient. *Technical report*, Department of Computer Science, Tufts University, 1988.
20. MCCULLOCH, W. S.; PITTS, W. A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, v. 5, p. 115-137, 1943.
21. ROSENBLATT, F. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, v. 65, n. 6, p. 386-408, 1958.
22. RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J. Learning representations by back-propagation of errors. *Nature*, v. 323, p. 533-536, 1986.
23. VALE, K. M. O.; NETO, A. F.; CANUTO, A. M. P. Using a reinforcement-based feature selection method in classifier ensemble. In: *Proceedings of the 10th International Conference on Hybrid Intelligent Systems – HIS2010*. IEEE, 2010. p. 213-218.
24. GUILFORD, J. P. *Fundamental statistics in psychology and education*. 4. ed. New York: McGraw-Hill Book, 1950.
25. SCHERER, K. R. Vocal communication of emotion: A review of research paradigms. Speech Communication, v. 16, p. 97-110, 2003. DOI: 10.1016/0167-6393(94)00079-2.
26. SOKOLOVA, Marina; LAPALME, Guy**.** A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, v. 45, n. 4, p. 427-437, jul. 2009. Disponível em: https://doi.org/10.1016/j.ipm.2009.03.002. Acesso em: [20/02/2025].