UNIVERSIDADE ESTADUAL PAULISTA - UNESP

Instituto de Biociências, Letras e Exatas- campus de São José do Rio Preto

GUSTAVO PEREIRA PAZZINI

Verificação de Locutor baseada em Aprendizado Profundo a partir de Espectrogramas Mel Multirresolução

São José do Rio Preto

2026

Gustavo Pereira Pazzini

Verificação de Locutor baseada em Aprendizado Profundo a partir de Espectrogramas Mel Multirresolução

Dissertação, apresentada à Universidade Estadual Paulista (UNESP), Instituto de Biociências, Letras e Exatas, São José do Rio Preto, para obtenção do título de Mestreem Ciência da Computação.

Área de Concentração: Computação Aplicada

Orientador: Profº Dr. Arnaldo Cândido Júnior

São José do Rio Preto

2026

P348v

Pazzini, Gustavo Pereira
Verificação de locutor baseada em aprendizado profundo a partir de espectrogramas mel multirresolução / Gustavo Pereira Pazzini. -- São José do Rio Preto, 2026
95 f.

Dissertação (mestrado) - Universidade Estadual Paulista (UNESP), Instituto de Biociências Letras e Ciências Exatas, São José do Rio Preto
Orientador: Arnaldo Cândido Júnior

1. Ciência da computação. 2. Aprendizagem profunda. 3. Inteligência artificial. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Dados fornecidos pelo autor(a).

IMPACTO POTENCIAL DESTA PESQUISA

Este trabalho propõe uma abordagem baseada em múltiplas resoluções espectrais e aprendizado profundo para verificação de locutores, visando modelos mais robustos e eficientes para autenticação por voz. A pesquisa contribui para o avanço da biometria da voz, com potencial de aplicação em sistemas de segurança, análises forenses e serviços digitais, além de promover o desenvolvimento científico e tecnológico na área de inteligência artificial.

POTENTIAL IMPACT OF THIS RESEARCH

This work proposes an approach based on multiple spectral resolutions and deep learning for speaker verification, aiming to develop more robust and efficient models for voice authentication. The research contributes to advances in voice biometrics, with potential applications in security systems, forensic analysis, and digital services, while promoting scientific and technological development in the field of artificial intelligence.

GUSTAVO PEREIRA PAZZINI

VERIFICAÇÃO DE LOCUTOR BASEADA EM APRENDIZADO PROFUNDO A PARTIR DE ESPECTROGRAMAS MEL MULTIRRESOLUÇÃO

Dissertação apresentada à Universidade Estadual Paulista (UNESP), Instituto de Biociências, Letras e Exatas, São José do Rio Preto, para obtenção do título de Mestre em Ciência da Computação.

Área de Concentração: Computação Aplicada

Data de defesa: 07/08/2026

BANCA EXAMINADORA

Profº Dr. Arnaldo Cândido Júnior
UNESP – Instituto de Biociências, Letras e Exatas – Campus de São José do Rio Preto

---

Profº Dr. Marcelo Matheus Gauy
UNESP – Instituto de Ciências e Engenharia – Campus de Itapeva

---

Profº Dr. Moacir Antonelli Ponti

ICMC – Instituto de Ciências Matemáticas e de Computação – Universidade de São Paulo

Dedico este trabalho aos meus avós, que neste ano completam 50 anos de casados, José Luiz Pazzini e Neide Bortollotto Pazzini, que tiveram um papel fundamental como alicerce da família e sempre me apoiaram desde o início. E por fim, dedico à minha irmã Julia de Oliveira Pazzini, lembre-se sempre: Tenha Esperança, Acredite, Mantenha-se forte e Ouse.

Agradecimentos

Agradeço aos meus pais, Filomeno e Márcia, pelo apoio dado para a realização deste trabalho. Aos meus amigos de faculdade, que me proporcionaram muitos momentos de felicidade durante esta árdua caminhada, e, em especial, ao meu melhor amigo João, uma pessoa brilhante que a vida me proporcionou e que esteve presente desde a quinta série em minha vida e me incentivou a realizar este mestrado. Agradeço imensamente ao Professor Arnaldo Cândido Júnior por aceitar me orientar e por todos os ensinamentos que recebi durante este período. Também agradeço aos amigos que trabalham na UNESP e na FUNDUNESP. Por fim, agradeço à nutricionista e futura doutora Thayane Carla Rodrigues Costa Caobianco, uma pessoa excepcional, que sempre me ajudou e incentivou a alcançar os grandes objetivos que conquistei, meu sincero muito obrigado. O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 88887.976050/2024-00.

“Às vezes, dar um passo à frente significa deixar algumas coisas para trás.”
(Powder, Arcane)

Resumo

A verificação de locutor é uma área fundamental da biometria da voz, visto que a autenticação automática de indivíduos por meio da comparação entre pares de áudios faz-se necessária principalmente em contextos de segurança e em técnicas forenses, a fim de validar se ambos os áudios foram produzidos pela mesma pessoa. Nos últimos anos, os avanços obtidos na área foram, em grande parte, devidos à chegada da era do grande volume de dados, com bases que superam um milhão de trechos de áudio. Entretanto, fatores externos e internos, como ruído e o próprio estado emocional do indivíduo, podem interferir no processo de verificação e dificultar a criação de modelos com uma maior taxa de acerto. Neste contexto, este trabalho tem como objetivo investigar o uso de múltiplas resoluções de áudio, comparado à utilização de apenas duas resoluções. Para isso, propõe-se um modelo baseado no uso de múltiplas resoluções espectrais para extração de características em um espectrograma Mel, com a utilização de uma arquitetura de redes neurais pré-treinada (PANN) e ECAPA2, integradas a uma rede prototípica. No modelo baseado em PANN, a configuração de três resoluções alcançou uma Equal Error Rate (EER) de $10.10\,\%$, representando reduções de aproximadamente $30.9\,\%$ e $38.6\,\%$ em comparação com as configurações de duas resoluções ($14.62\,\%$) e resolução única ($16.45\,\%$), respectivamente. Além disso, experimentos que utilizaram o modelo da arquitetura ECAPA2 também demonstraram melhorias ao empregar a abordagem multirresolução, alcançando um EER de $6.97\,\%$ para o modelo de três resoluções, o que corresponde a reduções EER de aproximadamente $10.2\,\%$ e $15.6\,\%$ em relação às configurações de duas resoluções ($7.76\,\%$) e resolução única ($8.25\,\%$), respectivamente. Como contribuição científica, desenvolveram-se adaptações de modelos capazes de priorizar as características fundamentais da voz, a fim de aumentar o desempenho na comparação de locutores entre amostras de áudios distintas e avaliar o impacto de diferentes quantidades de resoluções espectrais.

Palavras-chave: verificação de locutor; espectrograma multirresolução; aprendizado profundo.

Abstract

Speaker verification is a fundamental area of voice biometrics, since the automatic authentication of individuals through the comparison of audio pairs is necessary primarily in security contexts and forensic techniques, in order to validate whether both audios were produced by the same person. In recent years, the advances achieved in the area were largely due to the arrival of the big data era, with databases that exceed one million audio excerpts. However, external and internal factors, such as noise and the individual's own emotional state, can interfere in the verification process and hinder the creation of models with a higher accuracy rate. In this context, this work aims to investigate the use of multiple audio resolutions, compared to the use of only two resolutions. For this, a model based on the use of multiple spectral resolutions for feature extraction in a Mel spectrogram is proposed, using a pretrained neural network architecture (PANN) and ECAPA2, integrated into a prototypical network. In the PANN-based model, the three-resolution configuration achieved an Equal Error Rate (EER) of $10.10\,\%$*, representing reductions of approximately* $30.9\,\%$ *and* $38.6\,\%$ *compared to the two-resolution (*$14.62\,\%$*) and single-resolution (*$16.45\,\%$*) configurations, respectively. Furthermore, experiments that used the ECAPA2 architecture model also demonstrated improvements when employing the multiresolution approach, achieving an EER of* $6.97\,\%$ *for the three-resolution model, which corresponds to EER reductions of approximately* $10.2\,\%$ *and* $15.6\,\%$ *compared to the two-resolution (*$7.76\,\%$*) and single-resolution (*$8.25\,\%$*) configurations, respectively. As a scientific contribution, model adaptations capable of prioritizing fundamental voice characteristics were developed in order to increase performance in speaker comparison between distinct audio samples and to evaluate the impact of different amounts of spectral resolutions.*

Keywords*: speaker verification; multiresolution spectrogram; deep learning.*

Lista de figuras

Lista de tabelas

Lista de abreviaturas e siglas

ADAM	*Adaptive Moment Estimation (Estimativa Adaptativa de Momento)*
AUC	*Area Under the Curve (Área Sob a Curva)*
CGRU	*Convolutional Gated Recurrent Unit (Unidade Recorrente com Portões e Camada Convolucional)*
CNN	*Convolutional Neural Network (Rede Neural Convolucional)*
CRNN	*Convolutional Recurrent Neural Network (Rede Neural Recorrente Convolucional)*
EER	*Equal Error Rate (Taxa de Erro Igualitária)*
FAR	*False Acceptance Rate (Taxa de Aceitação Falsa)*
FFT	*Fast Fourier Transform (Transformada Rápida de Fourier)*
FIR	*Finite Impulse Response (Resposta de Impulso Finita)*
FRR	*False Rejection Rate (Taxa de Rejeição Falsa)*
GPU	*Graphics Processing Unit (Unidades de Processamento Gráfico)*
GRU	*Gated Recurrent Unit (Unidade Recorrente com Portões)*
ICBEN	*International Commission on Biological Effects of Noise (Comissão Internacional sobre os Efeitos Biológicos do Ruído)*
LSTM	*Long Short-Term Memory (Memória de Longo Prazo e Curto Prazo)*
ROC	*Receiver Operating Characteristic (Curva Característica de Operação do Receptor)*
RNNs	*Recurrent Neural Networks (Redes Neurais Recorrentes)*
SAP	*Self-Attentive Pooling (Pooling Auto-Atencional)*
TAR	*True Acceptance Rate (Taxa de Aceitação Verdadeira)*

Sumário

1 INTRODUÇÃO

O reconhecimento de falantes é essencial para a capacidade humana de identificar padrões, como ao escutar uma música desconhecida e o cérebro tentar associar quem é o cantor baseado nas vozes conhecidas. Esta habilidade de identificação tem grande destaque nas áreas de perícia criminal e execução das leis, principalmente para definir um suposto criminoso pelas suas características vocais únicas, que funcionam como a biometria da voz (HANSEN; HASAN, 2015).

O processamento de fala é uma das áreas na inteligência artificial, com análise de sinais de áudio vindos da voz humana, desde a filtragem até a extração de características fundamentais. Dentro dessa área, destaca-se o reconhecimento de locutor, que tem como base analisar as características intrínsecas da sua voz.

O reconhecimento de locutor comumente é dividido em duas grandes áreas, sendo elas a identificação de locutor e a verificação de locutor. A identificação é responsável por comparar se as características vocais da pessoa em questão pertencem a uma das vozes previamente salvas no sistema; esse caso pode ser considerado um problema de conjunto de dados fechado, pois é necessário um conjunto conhecido de locutores. Por outro lado, a verificação funciona como um processo de autenticação que, após a comparação entre quaisquer pares de áudio, retorna uma pontuação de semelhança entre eles. Com isso, este é um problema de conjunto de dados aberto, no qual é verificado se ambos os áudios, não necessariamente conhecidos previamente, pertencem ao mesmo locutor (JAKUBEC et al., 2024).

O primeiro desafio é a representação da voz contínua em uma representação discreta que seja possível à análise por um modelo. Os trabalhos recentes indicam que a representação com o maior desempenho foi a transformação dos dados brutos em gráficos de tempo-frequência chamados de espectrogramas, principalmente os que atuam em escala de frequência logarítmica que simula o ouvido humano, como nos espectrogramas Mel. Essas representações atuam com grande precisão em dados que não contêm ruídos; entretanto, o nível de ruído no ambiente real pode degradar o aprendizado da fala (LAMBAMO; SRINIVASAGAN; JIFARA, 2023).

Ao definir o escopo do reconhecimento, é necessário um grande volume de dados em diversas situações para treinar o modelo. Com o passar dos anos, foram criadas inúmeras bases de áudio em situações específicas, por exemplo, restrição a apenas um idioma, captação de áudio feita em um ambiente perfeito e sem nenhum ruído, além de poucas amostras e locutores disponíveis. Dentre as bases criadas, destaca-se a sucessora da *VoxCeleb1* (NAGRANI; CHUNG; ZISSERMAN, 2017), de-

nominada *VoxCeleb2*. Esta base é composta por trechos de áudios de entrevistas no ambiente real, com fatores externos atrelados ao áudio do entrevistado, que suprem as principais situações específicas citadas, com o armazenamento de mais de 1 milhão de amostras de aproximadamente 6.000 locutores de 145 países diferentes (CHUNG; NAGRANI; ZISSERMAN, 2018).

1.1 MOTIVAÇÃO E ESCOPO

Apesar de bons resultados de desempenho conquistados pelo avanço na área, os sistemas enfrentam grandes desafios relacionados a ruídos, principalmente em ambientes reais com elevadas taxas ruidosas, como é o caso dos sons produzidos por carros, ventiladores, cafeteiras, entre outros, que resultam em menos acertos (JUNEJA, 2022).

Uma baixa taxa de acerto também pode instigar buscas de falhas na subárea de segurança da biometria da voz, mais especificamente com ataques a modelos de reconhecimento de falantes. Um dos projetos pioneiros neste assunto é o trabalho que apresenta o *Fakebob*, um ataque adversário que não necessita conhecer o código do sistema que irá atacar, utiliza-se da adição de perturbação imperceptível aos humanos nos áudios e algoritmos de otimização para adaptar o som; desta forma, obteve taxas de sucesso de $99\,\%$ em sistemas comerciais e de código aberto (CHEN et al., 2021).

Os estudos da verificação da voz relacionados ao uso doméstico também se tornaram relevantes, com a possibilidade de relacionar reconhecimento facial e de voz como segundo fator de autenticação, com o destaque principal para a utilização em segurança residencial inteligente, como integração dos fatores ao sistema de câmeras, que pode considerar, além da imagem captada, a possível degradação devido à baixa iluminação, uma frase dita ao sistema para considerar a autenticação válida ou inválida (SAXENA; VARSHNEY, 2021).

1.2 OBJETIVOS E METODOLOGIA

Ao observar a importância da verificação de locutor nos dias atuais, pode-se afirmar que a voz é considerada um dos fatores-chave de autenticação, juntamente com outros métodos, como o reconhecimento facial e impressão digital. A combinação destes fatores contribui para tornar a verificação mais fácil, rápida e precisa em diferentes tipos de aplicações, desde a produção de provas periciais em processos jurídicos até a autenticação para efetuar transações bancárias. Diante desse cenário, o objetivo deste estudo é avaliar e construir um modelo baseado em rede neural eficaz em verificar se dois áudios a serem avaliados foram produzidos pelo mesmo locutor.

Para atingir o objetivo geral, é proposto os seguintes objetivos específicos descritos abaixo:

- Implementar e treinar arquiteturas de redes neurais para validação de desempenho em diferentes configurações;
- Analisar a influência do tamanho da janela temporal nos resultados;
- Investigar se o uso de múltiplas resoluções apresenta desempenho superior comparado à utilização de apenas duas resoluções.

A metodologia empregada tem como base a otimização do modelo proposto por Virgilli et al. (2024), que é constituído pela extração de características por meio de duas janelas temporais de tamanho fixo. O foco deste projeto foi explorar novas abordagens e permitir que o modelo desenvolvido aceite múltiplas resoluções de entrada, com diversas resoluções espectrais e que utilize diferentes estruturas de redes neurais profundas para extração de características, denominadas *backbones*.

Nos testes realizados, a metodologia empregada foi a utilização de três tamanhos de janela temporal de 8 ms, 16 ms e 32 ms, com tamanho de salto referente a razão de 4,8, que resulta em saltos com aproximadamente 1,66 ms, 3,33 ms e 6,66 ms. A partir dessas configurações, foram gerados espectrogramas Mel utilizados como representação de entrada para os modelos de aprendizado profundo. A utilização de técnicas de otimização e mecanismos de atenção foram incorporadas ao fluxo de aprendizado, de acordo com o modelo base de extração de características do espectrograma.

A avaliação dos resultados do modelo foi feita a partir das métricas de Equal Error Rate (EER) e minimum Detection Cost Function (minDCF), amplamente utilizadas na literatura para verificação de locutor. A comparação do desempenho foi realizada entre as diferentes configurações de resoluções espectrais, arquiteturas de extração de características e trabalhos correlatos.

A utilização de espectrogramas multiescala é justificada pela capacidade de contornar o trade-off tradicional entre as resoluções de tempo e frequência. A eficácia em integrar múltiplas representações espectrais para treinar arquiteturas de redes neurais tem se mostrado uma estratégia robusta para a captura de padrões acústicos complexos, o que traz ganhos de desempenho não apenas no reconhecimento de locutores, mas também em outras áreas de fronteira do processamento de áudio, como na classificação de vocalizações na fauna (DIAS; PONTI; MINGHIM, 2025).

1.3 CONTRIBUIÇÕES

Dessa forma, o trabalho teve como contribuição científica a construção de um modelo capaz de obter melhorias significativas na verificação de locutor comparado aos outros modelos propostos, baseado em otimizações da entrada de múltiplos espectrogramas com diferentes janelamentos de tamanhos e intervalos distintos, ao demostrar a superioridade da utilização de múltiplas janelas com tamanho de salto variado ao invés de apenas duas janelas com tamanho de salto fixo.

1.4 ORGANIZAÇÃO DA MONOGRAFIA

Esta monografia está organizada de forma que o leitor compreenda a base teórica do projeto, os trabalhos correlatos, a proposta desenvolvida e as conclusões. No capítulo 2, é exposto todo o conhecimento necessário para a projeção do trabalho, com referências ao processamento de áudio, representações do som e redes neurais. No capítulo 3, são citados os principais trabalhos correlatos na área de processamento de fala, com foco no reconhecimento e verificação de locutores. No capítulo 4, são mencionados os detalhes da proposta escolhida para este projeto, os materiais e métodos que foram utilizados, e a exposição dos resultados com a análise detalhada de cada experimento. Ao final, no capítulo 5, são apresentadas a conclusão e as considerações finais deste estudo.

2 FUNDAMENTAÇÃO TEÓRICA

No capítulo atual, são apresentados os tópicos essenciais para o entendimento do trabalho proposto, organizados na seguinte sequência: Abordagem sobre o som e suas características, representação no meio digital e transformadas para conversão do sinal em um espectrograma (Seção 2.1). Seguido de tópicos relacionados ao aprendizado de máquina e aprendizado profundo (Seção 2.2). Por fim, são apresentadas as formas de verificação de locutor, desafios e métricas do projeto (Seção 2.3).

2.1 SINAIS DE ÁUDIO E PROCESSAMENTO DE FALA

O som é um princípio físico que produz perturbação de acordo com o deslocamento do meio em que é propagado. Este meio pode ser a água, ar e diversos objetos com diferentes propriedades, com exceção do vácuo. As perturbações baixas são denominadas infrassom e as altas são denominadas ultrassom. Apesar de inaudíveis ao ser humano, estas ondas também são consideradas som. Dentre as áreas em que o som é necessário, pode-se citar a música, a acústica, a telefonia e a gravação de voz (PIERCE, 2019).

2.1.1 Propriedades do som

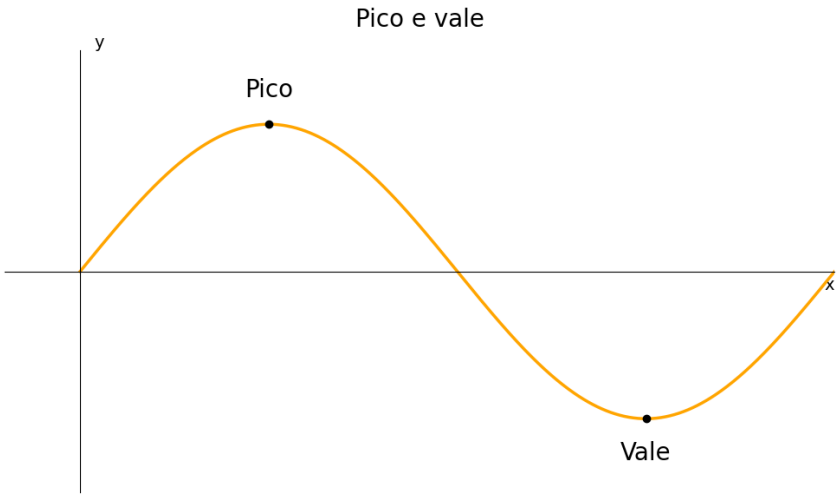
Ao representar o som, o formato mais comum é de onda em um plano cartesiano, em que o eixo x representa o tempo e o eixo y a pressão do ar ou amplitude da onda. Sons periódicos, como uma nota musical, apresentam um padrão de repetição regular. A voz humana, por sua vez, alterna entre trechos aproximadamente periódicos, como as vogais, e trechos aperiódicos, como diversos sons consonantais. Entretanto, ambos os sons periódicos e sons não periódicos podem ser representados de forma contínua em uma função de onda ou de forma discreta em um áudio digital ao ser armazenado em um computador. Dada a forma não contínua, é possível extrair as características da fala baseadas nos fatores de amplitude, frequência e tempo (CONSTANTINESCU; BRAD, 2023).

A amplitude está relacionada à intensidade do som, ou seja, quanto maior a amplitude, mais alto o som é percebido, e quanto mais baixa, menor será a percepção do som. Pode-se calcular a amplitude de um trecho de áudio ao observar o topo, comumente chamado de “pico” e a parte inferior da onda, chamada de “vale” em relação ao deslocamento do eixo central. O domínio do tempo no plano cartesiano é

observado à medida que o som avança pelo eixo *x*. Desta forma, é possível verificar a evolução do áudio ao longo do tempo e definir sua duração (CONSTANTINESCU; BRAD, 2023).

Na Figura 1 é exibida a representação de um pico no ponto mais alto e um vale no ponto mais baixo de uma onda senoidal no domínio do tempo.

Figura 1 – Representação do pico e do vale de uma onda.

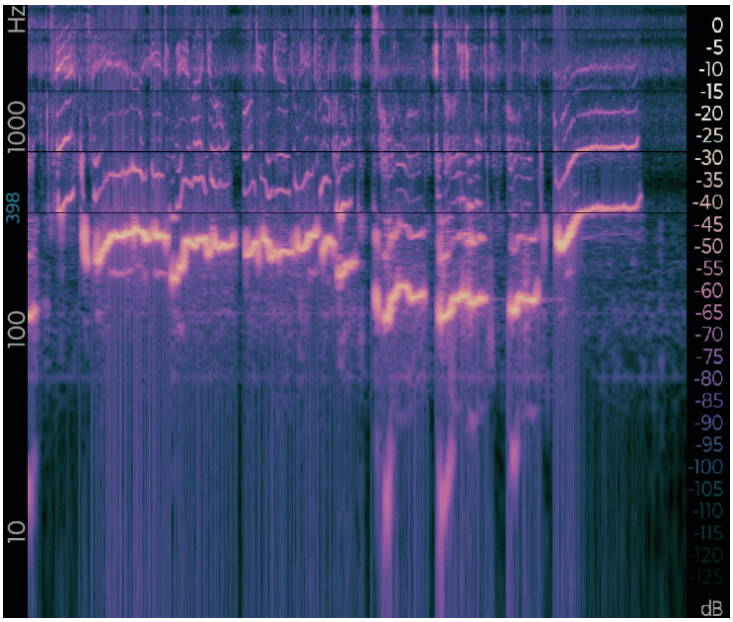


Fonte: Adaptado de Serway e Jewett (2008).

O domínio do tempo não é a única representação de uma onda; dentre as diversas representações, é possível utilizar também o domínio da frequência, que é demonstrado no gráfico quais frequências compõem a amostra em um determinado tempo *t*. Para decompor uma onda e construir o domínio da frequência, é necessário aplicar o Teorema de Fourier, que permite representar a onda como uma soma de funções seno e cosseno com diferentes frequências (FOURIER, 1883; AMBAYE, 2020).

É apresentado na Figura 2 um trecho de áudio da voz do autor em formato de espectrograma com um total de 3 dimensões, sendo o eixo *x* o tempo (ms), o eixo *y* a frequência (Hz) e o eixo *z* a amplitude (dB). O eixo do tempo representa uma variável contínua que registra a história da onda ao longo do tempo, apresentada da esquerda para a direita. A frequência é exibida em escala logarítmica, que proporciona ênfase nas frequências abaixo de 1.000 Hz, relacionando-se à capacidade de produção de som pela voz humana. A amplitude difere-se dos outros dois eixos, pois o eixo *z* é apresentado com a variação da cor, que é indicada pelos valores na escala de decibel ao lado direito da imagem.

Figura 2 – Espectrograma de um trecho de áudio feito por voz humana.



Fonte: Elaborado pelo autor.

2.1.2 Áudio digital

O som é uma onda contínua com infinitos valores de amplitude. Por conter essas propriedades, é impossível armazenar o sinal bruto do áudio em um computador com memória finita. Para a resolução deste impasse, é necessário transformar o som analógico em áudio digital, o que deteriora a qualidade do som. Para transformar o áudio contínuo em discreto, é necessário estabelecer duas técnicas: a primeira é a amostragem, que é utilizada para definir intervalos no tempo em que serão coletadas as amostras para formar a onda sonora no eixo *x* e a segunda é a quantização, que define valores finitos na amplitude em relação ao eixo *y* (YAROSLAVSKY, 1985).

A amostragem é a transformação do tempo contínuo em tempo discreto com a extração de amostras em um período regular. Esta operação resulta em uma sequência discreta, como expressa a Equação (1), onde o sinal amostrado $x(n)$ é obtido pelo sinal contínuo $f(t)$ nos instantes $t = n \cdot T$.

$$x(n) := f(n \cdot T) \tag{1}$$

Os termos da Equação (1) são:

- $x(n)$: sinal no tempo discreto;
- $f(t)$: sinal original no tempo contínuo;
- T: intervalo de tempo entre duas amostras consecutivas (em segundos);

- n: índice inteiro que representa o número da amostra, com $n \in \mathbb{Z}$.

Na Equação (2), T é representado o período de amostragem, e a taxa de amostragem F_s está relacionada ao inverso do período, chamada de frequência de amostragem, que é medida em Hertz.

$$F_s := \frac{1}{T} \tag{2}$$

Ao definir o tempo T da equação, é necessária atenção, pois, caso o espaço de tempo seja pequeno, pode-se ocupar mais memória do que o previsto. Caso o espaço de tempo entre as amostras seja grande, é possível acontecer o efeito chamado de *aliasing*, que acarreta a distorção do som na reconstrução da onda por falta de amostras (MÜLLER, 2015).

Este impasse é resolvido com o Teorema de Nyquist, o qual estabelece que, para que um sinal seja digitalizado, é necessário amostrar com uma frequência de pelo menos duas vezes maior que a frequência mais alta do sinal, ou seja, somente sinais em que a frequência máxima seja menor que a metade da frequência de amostragem podem ser corretamente convertidos para o domínio digital (NYQUIST, 1928; AUSTERLITZ, 2003).

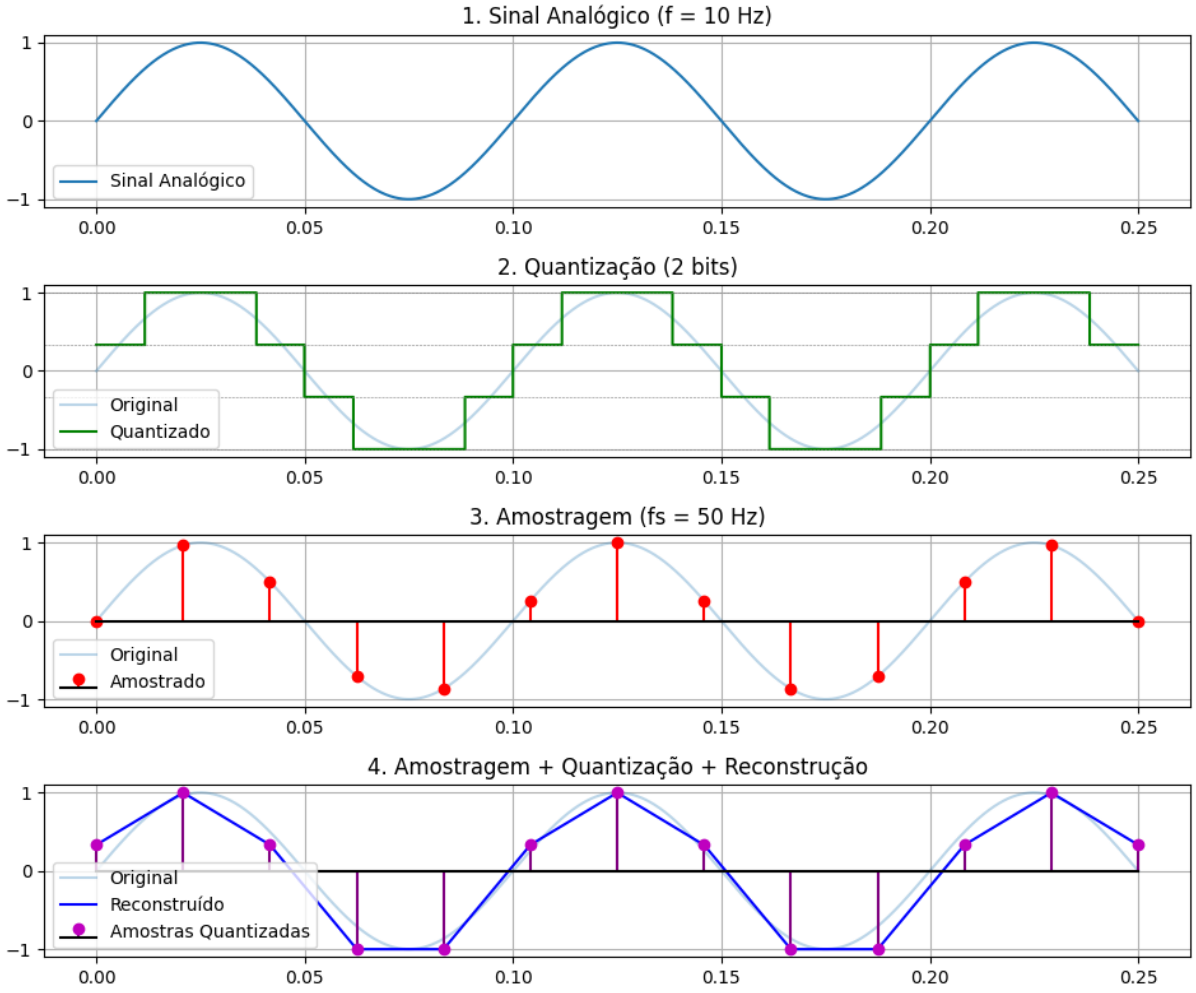
Em relação à quantização, esta é representada por um sistema não linear que possui o propósito de representar os valores da amplitude em um conjunto finito. Esta representação deve-se ao fato de aproximar os valores ao nível de quantização mais próximo, que rotineiramente é exibido como 2^x, sendo x o número de bits em que o valor da amplitude pode assumir (OPPENHEIM; SCHAFER; BUCK, 1999).

Na Figura 3 é representada a conversão de um sinal analógico senoidal para digital. A imagem contém quatro gráficos que ilustram cada etapa da conversão:

a) Sinal Analógico
O primeiro gráfico apresenta um sinal analógico puro, descrito por uma função seno $s(t) = \sin(2\pi f t)$, com a frequência do sinal escolhida de $f = 10$ Hz. Este sinal é contínuo no tempo e nos valores de amplitude, e servirá de base para a transformação do áudio analógico para digital.

b) Quantização
O segundo gráfico realiza a quantização, que consiste em mapear os valores analógicos para um conjunto finito de níveis discretos. A resolução utilizada no exemplo foi de 2 bits, que corresponde a $2^2 = 4$ níveis de quantização de tamanhos iguais entre -1 e 1. As linhas horizontais representam os níveis de quantização disponíveis, e o sinal quantizado tende a se aproximar o máximo possível do valor analógico com base nos níveis.

Figura 3 – Gráfico da conversão de sinal analógico para digital.



Fonte: Elaborado pelo autor.

c) Amostragem

O terceiro gráfico realiza a amostragem temporal, onde o sinal contínuo é discretizado no tempo. A frequência de amostragem utilizada foi de 50 Hz, o que satisfaz o Teorema de Nyquist, dado que a frequência do sinal é 10 Hz. Para satisfazer o teorema, a taxa de amostragem f_s deve ser, no mínimo, o dobro da maior frequência presente no sinal ($f_s \geq 2f$). Como $f_s = 50$ Hz $> 2 \times 10 = 20$ Hz, não há problemas de *aliasing*.

d) Amostragem, Quantização e Reconstrução

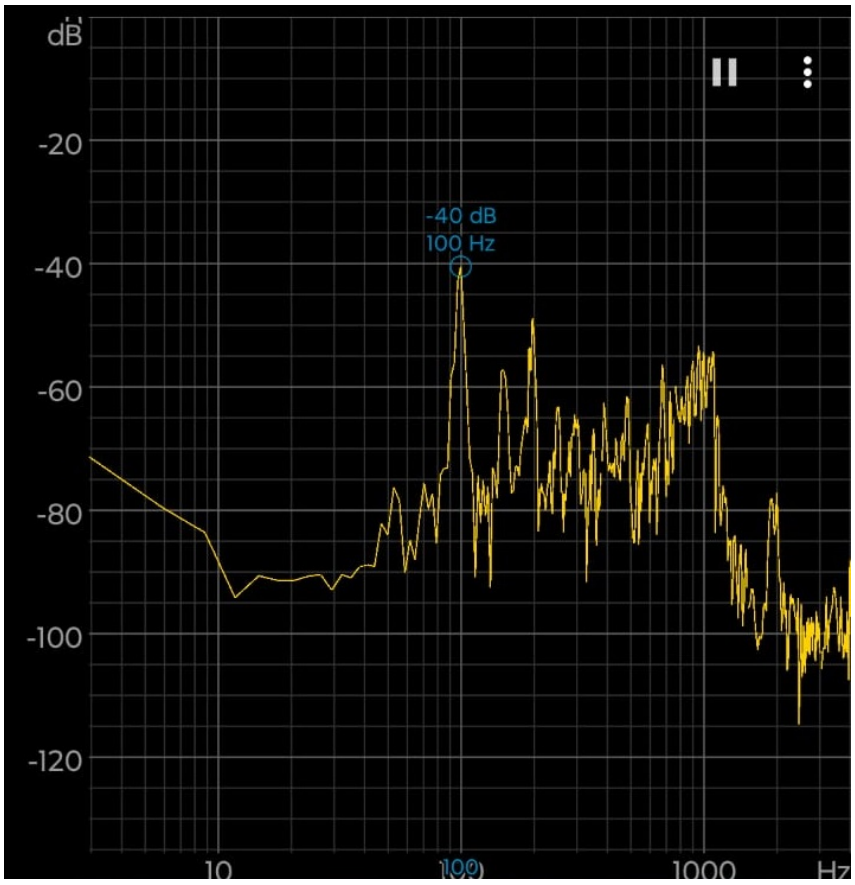
Por fim, o quarto gráfico mostra o resultado da amostragem e da quantização, com a reconstrução do sinal digital. Após a discretização temporal (amostragem) e de amplitude (quantização), o sinal é reconstruído e se aproxima da forma original do sinal analógico definido inicialmente.

2.1.3 Transformadas e representações no domínio da frequência

Uma onda periódica pode ser representada por um conjunto de ondas senoidais simples; a somatória dessas ondas compõe o sinal original. Logo, a Transformada de Fourier é o processo que decompõe uma onda complexa em diversas ondas simples no domínio da frequência (FOURIER, 1883; BRACEWELL, 2000).

Na Figura 4, é possível observar a análise do espectro de frequência do som emitido pela voz do autor, representado pela linha amarela. Diante disso, é possível identificar quais ondas simples compõem o trecho de fala, com destaque para a componente de 100 Hz emitida a -40 dB, que é predominante e está próxima à média da frequência da voz humana.

Figura 4 – Espectro da frequência em escala logarítmica.

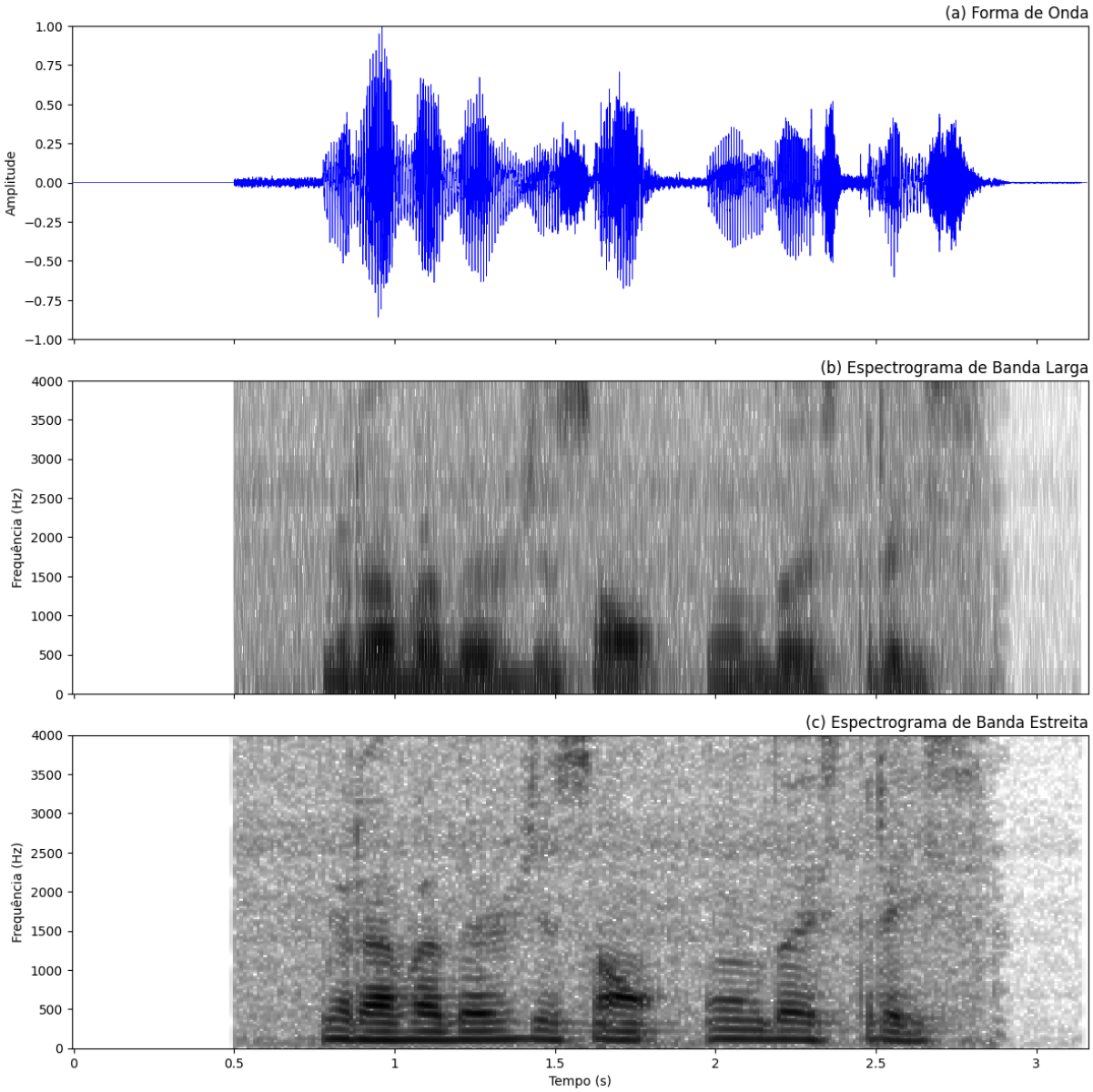


Fonte: Elaborado pelo autor.

Na criação do espectrograma, deve-se definir o tamanho da janela temporal em que será amostrado o áudio. O tamanho da janela temporal influencia inversamente a largura de banda do espectrograma. Logo, o espectrograma de banda larga é feito com janelas curtas e o espectrograma de banda estreita é feito com janelas temporais longas. No trabalho de Virgilli et al. (2024), foram utilizados os dois tipos de espectrogramas com janelas de 30 ms para banda estreita e 5 ms para banda larga,

em que os espectrogramas são representados, respectivamente, em 5(b) e 5(c), na Figura 5.

Figura 5 – Espectrograma de banda larga e banda estreita.



Fonte: Elaborado pelo autor.

As principais diferenças entre os espectrogramas de banda larga e banda estreita estão na forma como representam o tempo e a frequência. O espectrograma de banda larga exibe uma melhor resolução na representação do domínio do tempo, pois utiliza janelas temporais curtas, o que permite identificar com precisão o momento em que um som ocorreu. No entanto, esta configuração dificulta distinguir quais frequências estavam presentes naquele instante. Em relação ao espectrograma de banda estreita, é proporcionada uma visualização mais nítida das frequências, devido ao uso de janelas temporais mais longas, mas com uma menor resolução temporal, o

que prejudica a identificação de quando uma frequência inicia e termina.

2.1.4 Espectrograma mel

Ao determinar a frequência como uma medida física linear e comparar com a percepção humana, percebe-se que o ouvido humano segue a escala logarítmica em vez da escala linear, ou seja, a percepção da frequência do som, comumente chamada de tom, indica que frequências mais altas são percebidas como sons mais agudos, e frequências mais baixas são percebidas como sons mais graves. Por conta do estudo realizado por Stevens (1937), foi criada uma escala empírica para conversão de frequência (Hz) em tom, conhecida como escala Mel, com auxílio de ouvintes.

A representação do som em imagem pode ser feita por um espectrograma padrão, em que a escala de frequência permanece linear, como por exemplo: 100 Hz, 200 Hz, 300 Hz, etc. Mas, como explicado anteriormente, essa representação não corresponde a como o ouvido humano percebe os sons. Para lidar com essa divergência, foi desenvolvido o espectrograma em escala Mel. Essa representação se diferencia por utilizar a Equação 3, que converte o domínio da frequência linear para a escala Mel, que se aproxima da percepção do ouvido humano, em que f representa a frequência que será convertida em escala Mel (LACERDA et al., 2021).

$$\text{Mel}(f) = 2595 \cdot \log_{10}\left(1 + \frac{f}{700}\right) \tag{3}$$

A principal diferença entre o espectrograma e o espectrograma Mel é a escala logarítmica realçar as frequências mais baixas, como é possível observar na Figura 6 (LACERDA et al., 2021).

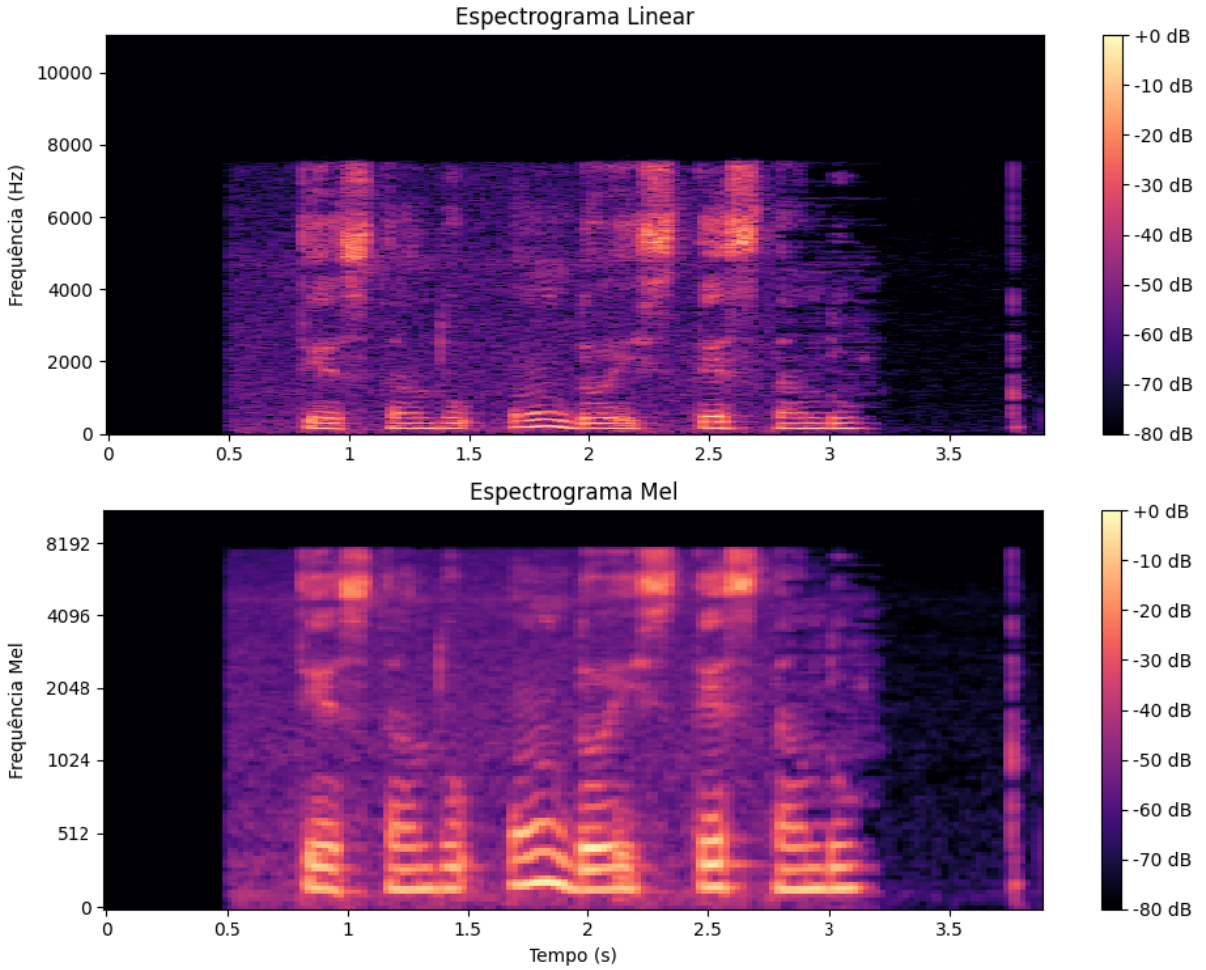
O áudio em questão foi produzido pela voz do autor, nota-se que as faixas de frequências mais fortes estão abaixo de 1.000 Hz, sendo possível perceber que os detalhes da onda no espectrograma Mel são mais visíveis do que no espectrograma linear.

2.1.5 Extração de características

O processo de extração de características de áudio no domínio da frequência, como a escala Mel, é composto por quatro etapas principais: pré-ênfase, segmentação e janelamento, geração do espectro de potência e aplicação do banco de filtros Mel (SIDHU; LATIB; SIDHU, 2025; JOSHI; PAREEK; AMBATKAR, 2023). A seguir, cada etapa é descrita em detalhes:

a) Pré-ênfase

Figura 6 – Comparação do espectrograma linear com o espectrograma Mel.



Fonte: Elaborado pelo autor.

A pré-ênfase é a etapa responsável por equilibrar o espectro do sinal de fala, com o realce das frequências mais altas do som com a utilização de filtros passa-alta[1]. Em relação à verificação de locutor, este processo atua na função de reduzir os componentes de baixa frequência e destacar as informações presentes nos componentes de alta frequência, o que favorece a extração de características relevantes. A subamostragem possui o foco de limitar o sinal da amostra aos níveis audíveis do ser humano, ou seja, reduzir o áudio para 16 kHz, o que, segundo o teorema de Nyquist, possibilita a representação de frequências de até 8 kHz. Para evitar o *aliasing* na subamostragem, é aplicado um filtro passa-baixa para remover componentes acima da nova frequência máxima, e se utiliza um filtro passa-alta com corte de 80 Hz para eliminação de ruídos (SIDHU; LATIB; SIDHU, 2025). A equação discreta que descreve o filtro de resposta de impulso finita passa-altas de primeira ordem, do inglês *Finite Impulse Response* (FIR), é apresentada na Equação 4 (NYQUIST, 1928).

[1] Filtro passa-alta é um filtro que permite a passagem de sinais com frequências acima de um valor determinado e atenua frequências mais baixas

$$y[n] = x[n] - \alpha x[n-1] \tag{4}$$

Os termos da Equação (4) são:

- $y[n]$: Valor da amostra do sinal de saída no instante discreto n, após o processamento pelo filtro;
- $x[n]$: Valor da amostra do sinal de entrada no instante discreto n;
- $x[n-1]$: Valor da amostra do sinal de entrada no instante discreto anterior ($(n-1)$), representando um atraso de uma amostra (dependente de f_s);
- α: Constante que determina a intensidade da influência da amostra anterior e as características de realce. O valor utilizado na literatura geralmente está entre 0 e 1, sendo comum $\alpha \approx 0.95$.

b) Segmentação e janelamento do sinal

Após a pré-ênfase, o áudio é dividido em janelas temporais menores. Entre as janelas mais utilizadas, destaca-se a janela de Hamming, conforme apontado por Ayvaz et al. (2022) e Kabir et al. (2021). Além dos tamanhos variáveis, é possível obter resultados promissores com múltiplas janelas (VIRGILLI et al., 2024).

O formato da janela de Hamming é utilizado por sua capacidade de reduzir distorções devido a cortes bruscos no áudio. Diante deste fato, esta janela é a ideal para suavizar os extremos laterais do áudio e gerar a análise do domínio da frequência (BOULAL et al., 2024). A fórmula da janela é descrita na Equação 5.

$$w(n) = 0{,}54 - 0{,}46\cos\left(\frac{2\pi n}{N-1}\right), \quad 0 \leq n \leq N-1 \tag{5}$$

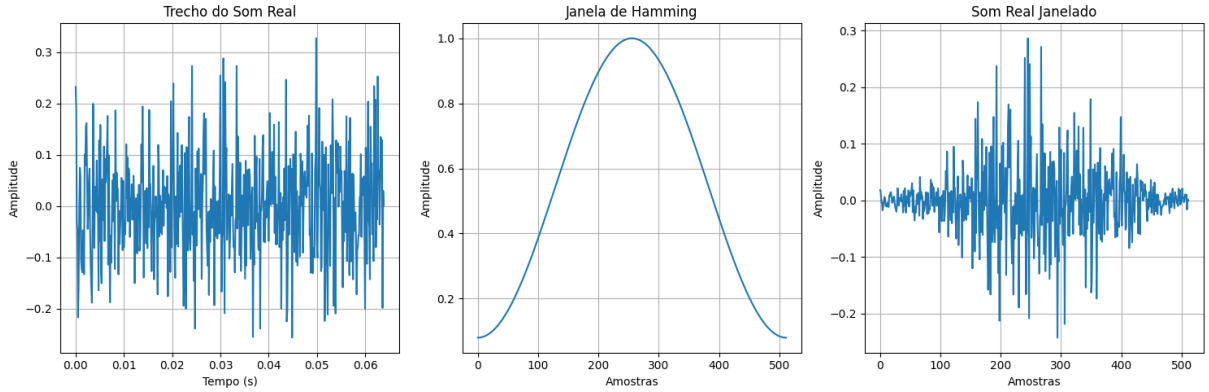
Os termos da Equação (5) são:

- $w(n)$: Valor de peso da janela no índice da amostra n. Ao multiplicar por esses valores, reduz a borda para próximo de zero;
- n: Índice da amostra dentro da janela, que varia de 0 (primeira amostra) até $N-1$ (última amostra da janela);
- N: Comprimento da janela;
- 0.54 e 0.46: Constantes específicas da janela de Hamming para minimizar o vazamento espectral, que reduz a amplitude do sinal das bordas da janela;

- $\cos\left(\frac{2\pi n}{N-1}\right)$: Termo cosseno que cria a forma característica da janela de "montanha".

Para melhor entendimento, foi desenvolvida uma imagem que utiliza a Equação 5 para o janelamento de um trecho de áudio real representado na Figura 7.

Figura 7 – Atuação da janela de Hamming em um trecho de áudio.



Fonte: Elaborado pelo autor.

c) Cálculo do espectro de potência

A geração do espectro de potência é baseada na transformação do sinal de áudio do domínio do tempo para o domínio da frequência. Esta transformação evidencia a quantidade de energia acumulada em cada frequência. O cálculo deve ser feito de acordo com cada trecho da etapa de segmentação e janelamento, pois o fato de que a voz humana não estacionária induz o cálculo a ser de forma localizada. O processo para analisar pequenos segmentos temporais é conhecido como Transformada de Fourier de Curto Período. Para cada janela, é aplicada a Transformada Rápida de Fourier que é um algoritmo eficiente para o cálculo da Transformada Discreta de Fourier. A Transformada Rápida de Fourier, do inglês *Fast Fourier Transform* (FFT), decompõe o sinal em frequências individuais que podem ser descritas como $X[k]$, onde k representa uma determinada frequência. A partir dos coeficientes, calcula-se o módulo ao quadrado de cada componente para gerar o espectro de potência, como demonstrado na equação 6 (FOURIER, 1883).

$$P[k] = |X[k]|^2 \tag{6}$$

O termo $|X[k]|$ denota o módulo do número complexo $X[k]$, e o resultado de $P[k]$ é um número real e positivo que exibe a energia da frequência k (GRAMI, 2016).

d) Aplicação do banco de filtros Mel.

O banco de filtros Mel consiste em aplicar um filtro passa-banda[2] Mel ao espectro de potência. Este filtro foi desenvolvido para extrair escalas não lineares de amostras de áudio relacionadas à voz, com base na percepção de áudio do ouvido humano ser em escala logarítmica (SIDHU; LATIB; SIDHU, 2025). O filtro Mel é composto por filtros triangulares que podem ser representados na Equação 7.

$$H_m(k) = \begin{cases} 0 & \text{se } k < f(m-1) \\ \dfrac{k - f(m-1)}{f(m) - f(m-1)} & \text{se } f(m-1) \leq k < f(m) \\ 1 & \text{se } k = f(m) \\ \dfrac{f(m+1) - k}{f(m+1) - f(m)} & \text{se } f(m) < k \leq f(m+1) \\ 0 & \text{se } k > f(m+1) \end{cases} \tag{7}$$

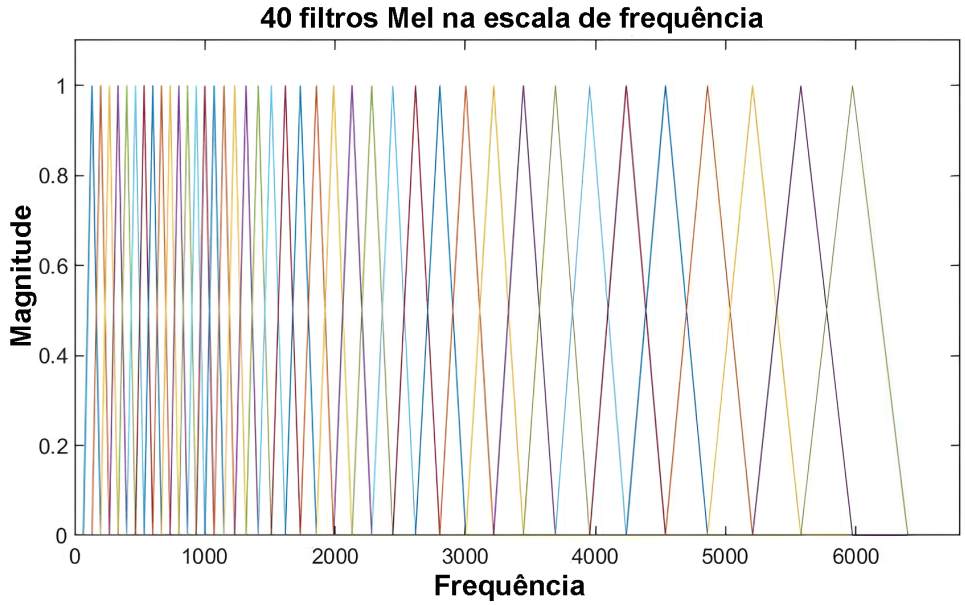
Os termos da Equação (7) são:

- $H_m(k)$: Representa a resposta de frequência do m-ésimo filtro do domínio de k;
- m: Índice que identifica o filtro atual dentro do banco de filtros;
- k: Variável de frequência em Hertz, que foi aplicada ao filtro;
- $f(m-1)$: Frequência de corte inferior do m-ésimo filtro em Hz;
- $f(m)$: Frequência central do m-ésimo filtro, onde a resposta é a máxima (valor 1);
- $f(m+1)$: Frequência de corte superior do m-ésimo filtro em Hz.

Diante da Equação (7), é possível construir um gráfico de 40 filtros triangulares apresentados na Figura 8.

[2] Filtro passa-banda é um filtro que permite a passagem de sinais dentro de uma faixa específica de frequências e atenua as frequências que estão fora dessa faixa

Figura 8 – Filtros triangulares Mel.

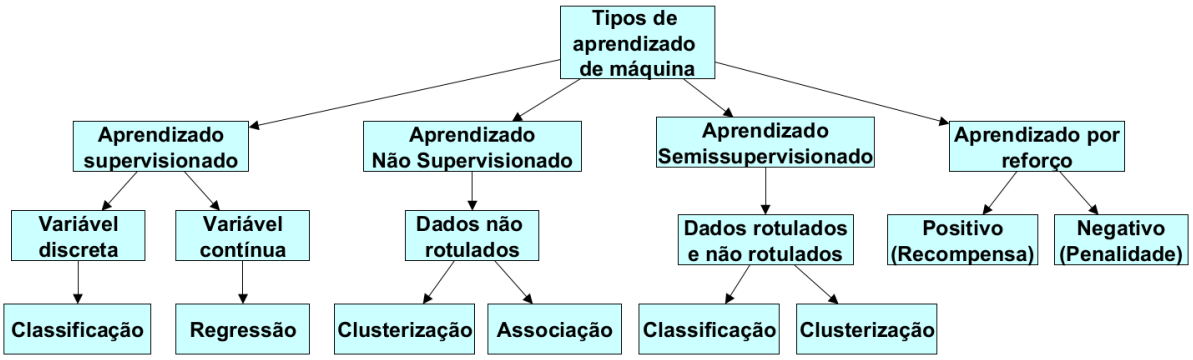


Fonte: Adaptado de (PARLAK; ALTUN, 2022)

2.2 APRENDIZADO DE MÁQUINA E APRENDIZADO PROFUNDO

O aprendizado de máquina é uma subárea da inteligência artificial que consiste na aprendizagem de padrões, sem que o modelo seja programado com os padrões de forma explícita. O aprendizado em geral pode ser dividido nos seguintes paradigmas: supervisionado, não supervisionado, semissupervisionado e por reforço, como apresentado na Figura 9. A seguir, cada um desses tipos é explicado em detalhes (HAN; KAMBER; PEI, 2011; SARKER, 2021).

Figura 9 – Técnicas de aprendizado de máquina.



Fonte: Adaptado de Sarker (2021).

Supervisionado

O aprendizado supervisionado possui como característica o treinamento do modelo com dados que contêm rótulos. Na fase de testes, são inseridas apenas as entradas para realizar o cálculo da saída com base no que o modelo aprendeu. Esse tipo de aprendizado pode ocorrer com variáveis finitas, sendo chamado de classificação, ou com variáveis contínuas, processo este chamado de regressão.

Não supervisionado
O aprendizado não supervisionado trabalha com dados não rotulados, sendo voltado para o agrupamento, ou clusterização, e para a identificação de características em conjuntos de dados, incluindo a detecção de anomalias. Também permite a descoberta de associações entre variáveis, revelando padrões frequentes e relações ocultas nos dados.

Semissupervisionado
O aprendizado semissupervisionado utiliza tanto dados rotulados quanto não rotulados; este método é útil caso os dados que contêm rótulos sejam a minoria e os não rotulados sejam a maioria. Essa abordagem é utilizada, geralmente, em situações em que a rotulagem de dados pode ser um processo custoso e demorado, o que torna as amostras rotuladas raras em bases de dados.

Reforço
O aprendizado por reforço consiste em um modelo aprender por meio de interações com o ambiente e receber recompensas caso faça a decisão correta ou tenha a penalidade se a decisão tomada for incorreta. Esse processo de autoavaliação contínua permite o aprendizado de um conjunto de ações, sendo ideal para automatizar processos na área de robótica.

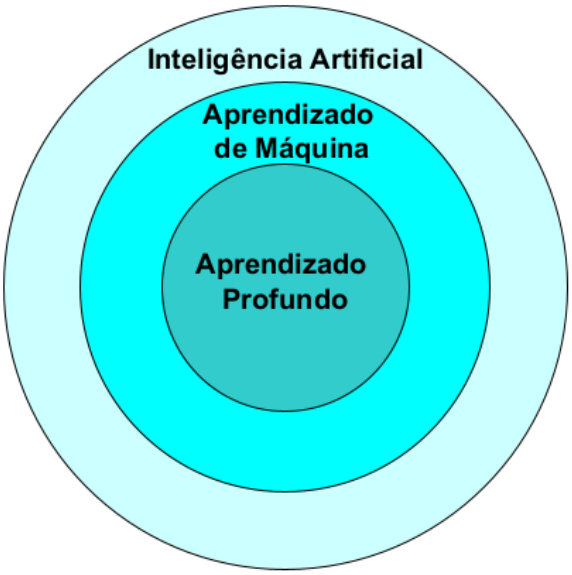
2.2.1 Aprendizado profundo

O aprendizado profundo é uma subárea do aprendizado de máquina, como é possível observar na Figura 10. Chama-se “profundo” devido a diversas camadas ocultas que o modelo possui entre a camada de entrada e a camada de saída. Essas camadas ocultas desempenham o papel de aprendizado das características ao passar a informação em cada camada, sendo que as camadas ocultas iniciais são responsáveis pelo aprendizado de padrões simples e as últimas aprendem padrões complexos (SHINDE; SHAH, 2018).

2.2.2 Arquiteturas de aprendizado profundo

Para que uma rede neural aprenda padrões, foram desenvolvidos neurônios artificiais chamados de perceptrons. O perceptron é um modelo fundamental inspirado

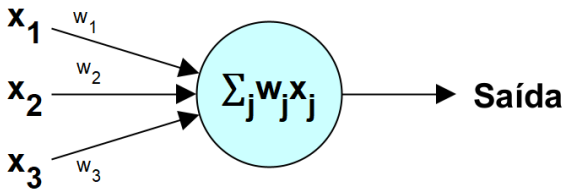
Figura 10 – Campo do aprendizado profundo.



Fonte: Adaptado de Shinde e Shah (2018).

em um neurônio biológico que é representado na Figura 11. Tal modelo recebe as entradas x_1, x_2 e x_3 associadas com seus respectivos pesos w_1, w_2 e w_3. O cálculo realizado pelo perceptron é através da somatória de todas as entradas multiplicadas por seus pesos, descrita pela fórmula $\sum_j w_j x_j$ (NIELSEN, 2015).

Figura 11 – Neurônio artificial perceptron.

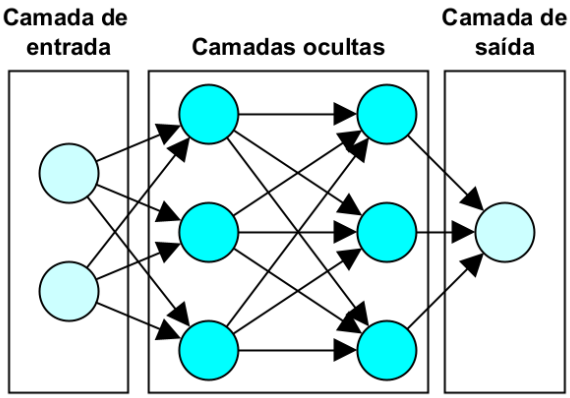


Fonte: Adaptado de Nielsen (2015).

A junção de perceptrons forma uma camada, e as camadas organizadas compõem a rede neural multicamadas. Na Figura 12 é exibida as camadas totalmente conectadas, com dois perceptrons na camada de entrada, duas camadas ocultas com três neurônios cada e uma camada de saída com apenas um perceptron. Como as camadas de entrada e saída não realizam cálculos, este modelo comumente é descrito como um modelo de duas camadas (ZHANG et al., 2023).

A camada oculta atua como uma intermediação entre a camada de entrada e saída, sendo que a quantidade deste tipo de camadas pode variar de acordo com o problema a ser resolvido. De acordo com Uzair e Jamil (2020), a escolha do número

Figura 12 – Representação de camadas ocultas.



Fonte: Adaptado de Zhang et al. (2023).

de camadas ocultas depende do tipo de banco de dados utilizado, de modo que o aumento do número de camadas pode levar a uma possível lentidão no processamento da rede neural, mas contribui para uma melhor precisão.

Em uma classificação, é possível a existência de mais de uma saída; neste caso, os valores devem ser transformados de reais para distribuição de probabilidade, em que a maior probabilidade será escolhida pelo classificador (WANG et al., 2018). A transformação chama-se *softmax* e ocorre de acordo com a Equação 8.

$$f(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{N} e^{x_j}}, \quad \text{para } i = 1, \ldots, N \tag{8}$$

Os termos da Equação (8) são:

- x_i: É o valor real bruto produzido na saída;
- e: Base do logaritmo natural, cujo valor aproximado é 2.71828;
- $\sum_{j=1}^{N} e^{x_j}$: Somatória dos exponenciais de todos os valores de entrada (x_j) para todas as N categorias possíveis.

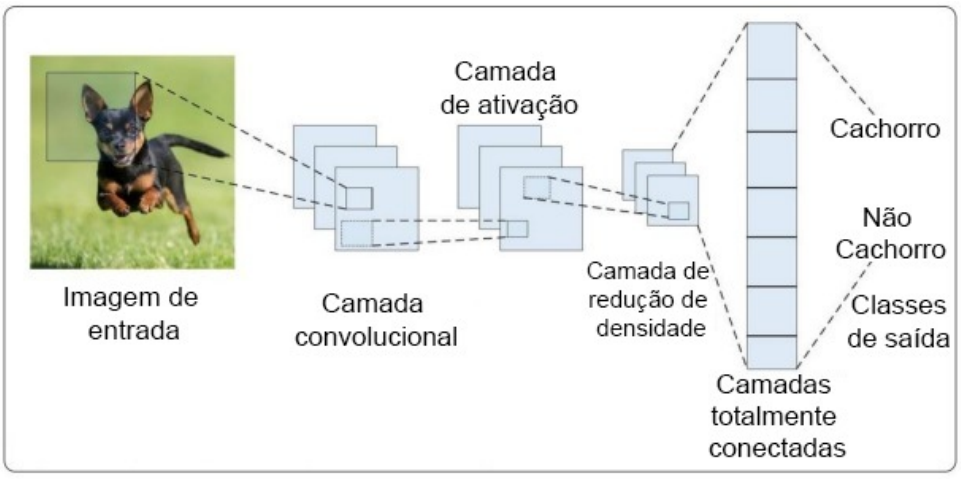
Após o período de pausa por falta de processamento, as técnicas apresentadas só foram possíveis devido ao uso de Unidades de Processamento Gráfico, do inglês *Graphics Processing Unit* (GPU). As técnicas em questão são as redes neurais convolucionais (LECUN; BENGIO; HINTON, 2015).

2.2.3 Redes neurais convolucionais

As redes neurais convolucionais, também chamadas de *ConvNets*, são utilizadas em processamento de múltiplas matrizes, incluindo áudio em 1D, imagens em 2D e vídeos em 3D. O nome convolucional foi designado ao modelo pela varredura de características de acordo com uma convolução baseada em *kernels* que pode variar de tamanho. Os principais usos das *ConvNets* são a detecção de objetos, leitura de caracteres escritos à mão e o reconhecimento facial (LECUN; BENGIO; HINTON, 2015).

De forma geral, a CNN é composta por várias camadas convolucionais, que passam pelo processo de função de ativação não lineares, camadas de redução de dimensionalidade e camadas totalmente conectadas. Como exemplo, é demostrado na Figura 13 o processo para reconhecer um cachorro por meio do processo de CNN. Esse tipo de arquitetura serviu de base para o desenvolvimento de variações mais avançadas, como Redes Residuais (ResNet), que introduzem blocos residuais para facilitar o treinamento de redes profundas, e Redes Neurais de Áudio Pré-treinadas (PANNs), que podem utilizar arquiteturas como a ResNet ou outras redes para tarefas de classificação e detecção de sons.

Figura 13 – Representação da arquitetura CNN.

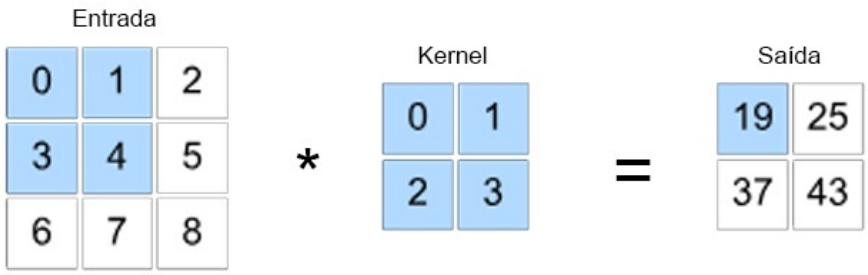


Fonte: Adaptado de Alzubaidi et al. (2021).

O processo consiste em inserir uma entrada no modelo, que será processada em três dimensões: altura, largura e profundidade. A altura e a largura correspondem ao número de pixels verticais e horizontais, respectivamente. A profundidade é discriminada a partir do número de canais da imagem, podendo ter apenas um canal ao ser preta e branca ou três canais no caso de imagens coloridas (RGB). Durante o treinamento, a rede aprende automaticamente filtros úteis em cada camada para extrair características específicas da imagem, como contornos (ALZUBAIDI et al., 2021). Um

exemplo da etapa de convolução para extração de características é apresentado na Figura 14, em que a entrada é segmentada do tamanho do kernel, e são multiplicados os elementos da posição igual das duas matrizes, e por fim, soma-se os resultados.

Figura 14 – Representação do processo de convolução.



Fonte: Adaptado de Zhang et al. (2023).

Após a geração dos mapas de ativação pelas camadas convolucionais, cada camada é especializada em um tipo de característica: camadas iniciais aprendem bordas, camadas intermediárias aprendem texturas e formas e camadas mais profundas aprendem partes complexas da imagem. Em seguida, é aplicada a função de ativação, como por exemplo a *ReLU*, que zera valores negativos e mantém os valores positivos. Este método de não linearidade permite que o modelo aprenda padrões complexos que não são possíveis de ser representados apenas por operações lineares (NAIR; HINTON, 2010; ALZUBAIDI et al., 2021).

Na sequência, é necessário realizar a subamostragem dos mapas gerados por meio das camadas de redução de dimensionalidade *pooling*. Este processo diminui o número de parâmetros da rede e acelera o treinamento, o que torna a identificação da imagem mais precisa, por ignorar pequenas variações que podem acontecer nas imagem. A aplicação da função *pooling* se resume a verificar toda a imagem e determinar o *max pooling* ou *average pooling*, com auxílio de um kernel de altura e largura iguais. Por fim, as camadas totalmente conectadas recebem as características que são passadas para a camada de saída que irá atribuir uma pontuação para realizar a classificação, em que geralmente é utilizada a função *softmax*, em que cada entrada será atrelada a uma saída por meio de probabilidade (ALZUBAIDI et al., 2021). A Figura 15 representa a operação de camadas *pooling*, que utiliza o valor máximo, para redução da dimensionalidade.

2.2.4 Aprendizado de métricas

No campo do aprendizado de máquina, o aprendizado de métricas visa aprender uma métrica de distância para quantificar amostras de dados similares ou não similares. O objetivo desta abordagem é a otimização de uma métrica para que possa

Figura 15 – Representação do processo de pooling.

Fonte: Adaptado de Zhang et al. (2023).

diminuir a distância entre elementos com características semelhantes e aumentar a distância entre elementos com características diferentes. Os estudos da literatura sobre aprendizado de métricas têm relação com a distância, em que o espaço original equivale à distância euclidiana em um espaço linear transformado por uma matriz de projeção (KAYA; BILGE, 2019).

Ao tentar capturar métricas complexas em problemas do mundo real, existe uma limitação ao utilizar métodos lineares euclidianos, pois estes resolvem apenas problemas com características lineares. Para resolver esta situação, foram criados métodos baseados em *kernel*, apropriados para problemas em espaço não linear, mas com possibilidade de *overfitting*. A solução mais robusta criada foi o aprendizado profundo de métricas, que consiste em extrair as características das arquiteturas de aprendizado profundo com o auxílio dos princípios do aprendizado de métricas. Essas redes tendem a ser mais precisas para identificar características semelhantes das amostras, o que gera agrupamentos e classificações mais precisas. Este processo tornou-se eficaz por aprender a métrica de distância dos dados brutos, o que justifica o melhor desempenho do modelo (KAYA; BILGE, 2019).

2.2.5 Redes siamesas

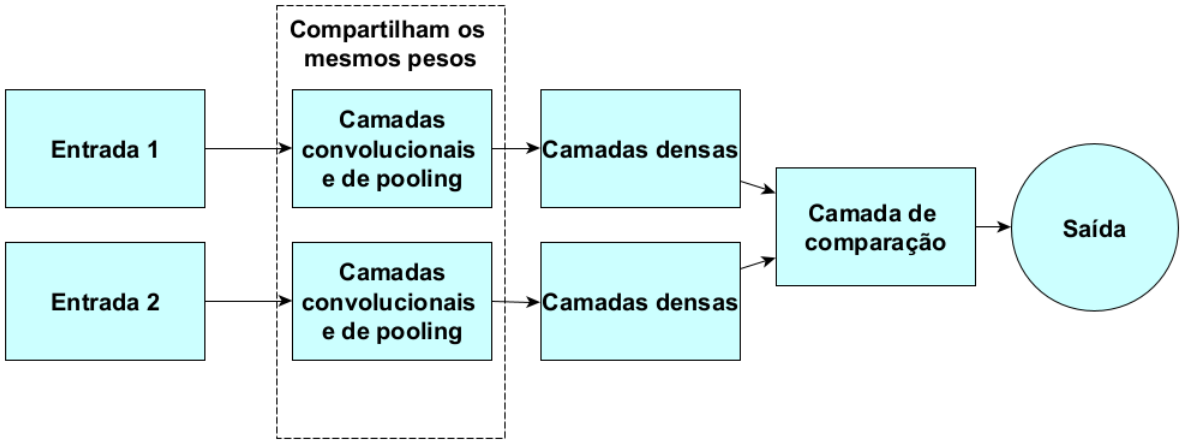
As redes siamesas são redes neurais que contêm duas sub-redes idênticas que compartilham os mesmos pesos. Seu objetivo é avaliar a similaridade entre os dados de entrada processados por cada sub-rede. Cada sub-rede recebe uma entrada diferente a ser processado e, após o processamento com os pesos compartilhados, as saídas são comparadas para verificar se os objetos são semelhantes (SAINI et al., 2021).

A construção da rede siamesa é feita comumente baseada em camadas convolucionais e de *pooling*, mas sem a camada de *softmax*, que é típica das redes de classificação. A ativação é feita geralmente por uma função sigmoide, que produz um

valor entre 0 e 1, função esta que indica o grau de similaridade entre as entradas, sendo 0 entradas diferentes e 1 entradas similares (SAINI et al., 2021).

Na Figura 16 é representada a estrutura de uma rede siamesa, em que recebe duas diferentes entradas, logo após passam pelas camadas convolucionais e de *pooling* que são idênticas e compartilham dos mesmos pesos, depois pelas camadas densas, e por fim, ocorre a comparação e o resultado da saída.

Figura 16 – Representação da rede siamesa



Fonte: Elaborado pelo autor.

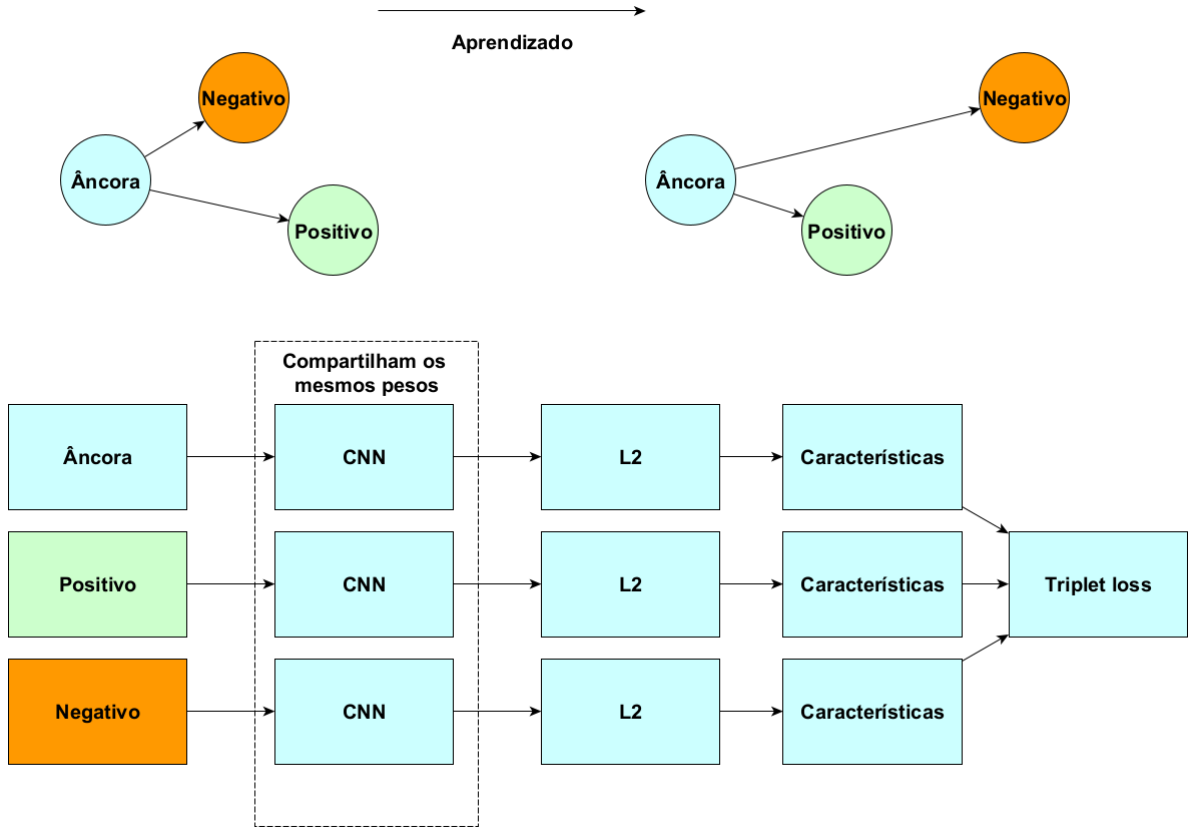
2.2.6 Redes triplas

A proposta das redes triplas é parecida com as redes siamesas, neste caso, é feito o mapeamento das entradas em três instâncias idênticas de uma mesma rede que compartilham pesos, que ao final, a rede aprende a representar amostras de forma que as da mesma classe fiquem próximas. O mapeamento é feito no espaço euclidiano, em que a rede deve calcular a distância entre os vetores de saída e verificar se são próximas, o que indica se as amostras são semelhantes (BASSI; RAMPAZZO; ATTUX, 2019).

O mecanismo utilizado para ajustar os parâmetros é a minimização da função de perda triplet loss, com uma entrada sendo a âncora, uma amostra da mesma classe da âncora, chamada de amostra positiva e uma amostra diferente da âncora, chamada de amostra negativa. Após as amostras serem representadas no espaço euclidiano, a rede deve aproximar a distância entre a âncora e a amostra positiva, e afastar destas duas a amostra negativa. Ao realizar o treinamento, as triplas são enviadas por lotes e ocorre a retropropagação do erro para ajustes dos parâmetros de acordo com a função de otimização escolhida (BASSI; RAMPAZZO; ATTUX, 2019).

A Figura 17 representa a aproximação de características positivas e o afastamento de características negativas, de forma que na sequência, está a arquitetura de treinamento que utiliza Triplet Loss. Nesse esquema, três entradas são processadas simultaneamente: a Âncora, uma entrada semelhante a âncora (Positivo) e uma imagem de diferente da âncora (Negativo). As três entradas passam pela mesma rede convolucional com pesos compartilhados, que extrai características normalizadas. Além disso, a rede tripla utiliza uma margem, a qual garante que a distância entre a âncora e o negativo seja maior do que a distância entre a âncora e o positivo, margem esta que evita as instâncias de serem empurradas ao infinito.

Figura 17 – Representação da rede tripla



Fonte: Adaptado de Gao, Tian e Liu (2025).

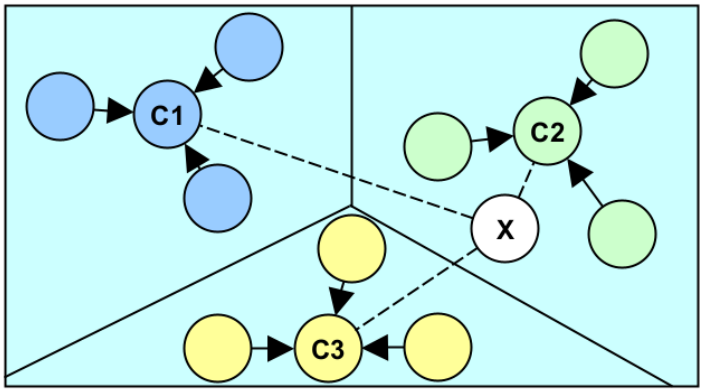
2.2.7 Redes prototípicas

Em cenários com pouca disponibilidade de dados, como na classificação com poucas amostras, a função *softmax* tende a apresentar um desempenho inferior. Como alternativa a essa limitação, foram propostas as redes prototípicas. Essas redes consistem em criar um espaço métrico em que a classificação é realizada através do cálculo da distância entre os pontos de consulta e os protótipos médios de cada classe. Cada protótipo é definido pelo vetor médio dos exemplos de suporte dispos-

tos para cada classe durante o aprendizado. Ao utilizar a distância euclidiana como métrica, é possível separar os conjuntos de forma eficaz e tende a obter melhores resultados em comparação com métodos binários, como redes siamesas, em ambientes com poucos dados (SNELL; SWERSKY; ZEMEL, 2017).

Na Figura 18, é representada uma rede prototípica com poucas amostras. Os pontos $C1$, $C2$ e $C3$ são protótipos de suas respectivas classes, calculados de acordo com a média dos exemplos de suporte. O ponto X representa um vetor de características a ser classificado, em que a classificação é realizada com referência à proximidade dos protótipos de cada classe, de acordo com a distância euclidiana. Devido à maior proximidade de X com o protótipo $C2$, o vetor X é classificado como pertencente à classe $C2$.

Figura 18 – Rede prototípica de poucas amostras.



Fonte: Adaptado de Snell, Swersky e Zemel (2017).

2.2.8 Aam-softmax

A AAM-Softmax (*Additive Angular Margin Softmax*) é uma função de perda que utiliza uma margem angular adicional na classificação, que busca aumentar a separação entre as classes no espaço de características. Diferente da Softmax convencional, a AAM-Softmax modifica a similaridade angular da classe correta por meio da adição de uma margem m, o que torna a tarefa mais discriminativa. Essa abordagem é utilizada em reconhecimento e verificação, nas quais a obtenção de representações com maior separação entre as classes é relevante (DENG et al., 2019).

2.2.9 Métricas de avaliação

As métricas de avaliação desempenham um papel fundamental na análise do desempenho do sistema desenvolvido. No contexto da verificação de locutor, tem-se como destaque a Taxa de Aceitação Falsa, do inglês *False Acceptance Rate* (FAR), e a Taxa de Rejeição Falsa, em inglês *False Rejection Rate* (FRR), as quais consistem, respectivamente, em aceitar erroneamente que um par de áudios pertence ao mesmo locutor, entretanto, pertencem a locutores diferentes, e rejeitar de maneira incorreta um par de áudios que realmente pertence ao mesmo locutor (ZAIDI et al., 2021).

O cálculo do FAR é representado pela Equação (9), em que Nar é o número de áudios incorretos aceitos erroneamente e $Ntar$ é o número total de áudios incorretos.

$$\text{FAR} = \frac{Nar}{Ntar} \tag{9}$$

Em relação ao FRR, tal taxa pode ser definida por meio da Equação (10), em que Nac é o número de áudios corretos que foram rejeitados e $Ntac$ é o número total de áudios corretos.

$$\text{FRR} = \frac{Nac}{Ntac} \tag{10}$$

A FAR está associada à segurança, enquanto FRR está ligada à usabilidade do ambiente desenvolvido, sendo que quanto menores os valores, melhores são as métricas. Dessa forma, valores baixos de FAR indicam uma segurança mais rigorosa, mas, ao mesmo tempo, podem resultar numa diminuição da usabilidade, pois aumenta o FRR. De forma contrária, com uma FRR baixa, tem-se uma melhor usabilidade, porém com a possibilidade de aumentar a taxa de aceitação de áudios ilegítimos e, como consequência, mais falhas de segurança. Por conta disso, faz-se necessário buscar um ponto de equilíbrio, no qual FAR e FRR se igualam (ZAIDI et al., 2021).

A igualdade entre as duas taxas de FAR e FRR, quando obtida, é denominada Taxa de Erro Igual, do inglês *Equal Error Rate* (EER).

2.3 RECONHECIMENTO E VERIFICAÇÃO DE LOCUTOR

O entendimento da verificação de voz depende da compreensão das diferenças entre identificação e verificação do falante. A identificação consiste no modelo receber diversos áudios de locutores e extrair as principais características de cada falante. Após o treino, as características do áudio a ser analisado são comparadas com as características salvas e o modelo indica com qual dos locutores salvos o locutor do

áudio de entrada se assemelha, ou seja, checa-se se o áudio em análise pertence a algum locutor previamente conhecido pelo modelo. Na verificação, o objetivo é afirmar ou negar se os áudios apresentados pertencem ao mesmo locutor. Diferentemente da identificação de locutor, em que o modelo calcula a probabilidade de qual locutor é mais semelhante ao áudio apresentado, a verificação compara as características entre um par de áudios, com o objetivo de verificar se foram produzidos pela mesma pessoa, sendo aprovados mediante a compatibilidade suficiente das características em comum (HANIFA; ISA; MOHAMAD, 2021).

Com a definição de verificação, também deve-se definir as duas categorias para verificar um falante, sendo elas a dependência e a independência de texto. A dependência refere-se a pessoa falar exatamente o mesmo texto treinado no modelo, enquanto a independência não impõe ao locutor recitar a mesma frase dita no treinamento, sendo a segunda situação mais desafiadora e próxima da realidade (BAI; ZHANG, 2021).

2.3.1 Representação da verificação de locutor

Na Figura 19 é apresentada uma representação da verificação de locutor, que é iniciada na fase de treinamento com a inserção do discurso conhecido no modelo, que irá extrair as principais características do falante, como a geração de espectrogramas. Após gerar as características, estas seguem para uma rede neural que armazena o locutor alvo. Na fase de avaliação, acontece o mesmo processo de extração de características, mas o modelo irá medir a semelhança e tomar a decisão baseada em um limite de igualdade que irá aceitar ou rejeitar o falante desconhecido (BÄCKSTRÖM et al., 2022).

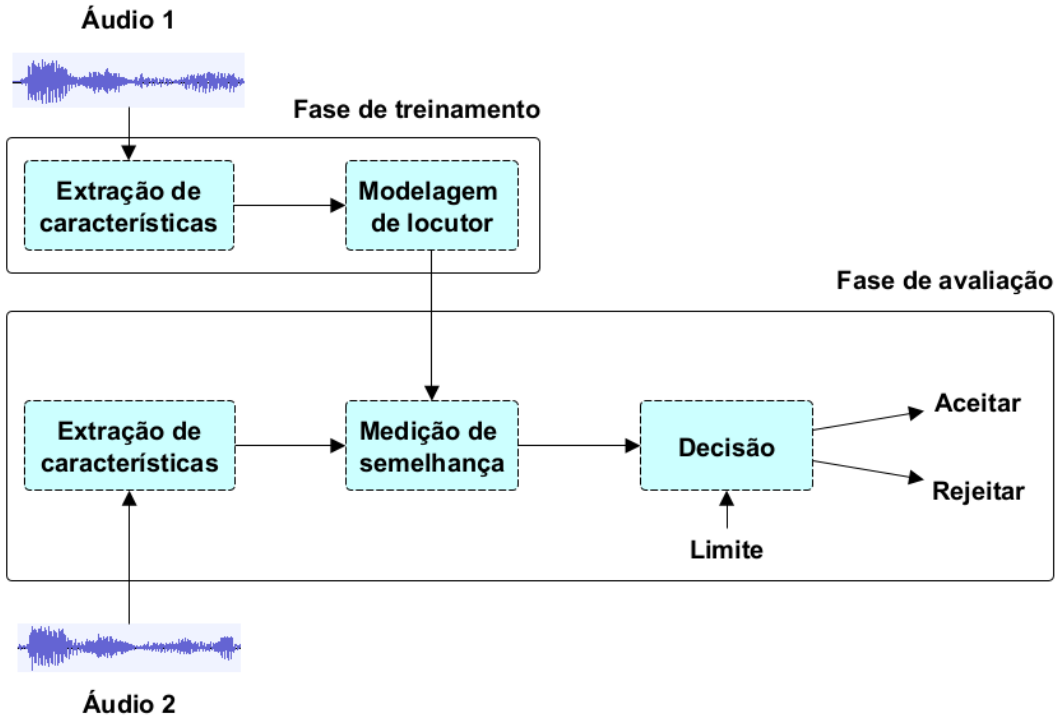
2.3.2 Desafios

Dentre os desafios enfrentados para a extração do sinal de fala, destaca-se a possibilidade de o áudio ser afetado por ruídos e reverberações durante o percurso da saída da voz do indivíduo até a captação da fala pelo microfone (AL-KARAWI; AHMED, 2021).

O ruído, segundo a atual definição da *International Commission on Biological Effects of Noise* (2025), é considerado um som indesejável e/ou prejudicial. O trecho “prejudicial” refere-se a ruídos que, dependendo da intensidade, podem afetar a saúde humana em fatores básicos como perturbação do sono e da concentração, até o risco de perda auditiva (FINK, 2020).

A reverberação é resultado da degradação do áudio de acordo com o ambiente em que o som foi reproduzido, ambiente este que possui objetos que refletem as

Figura 19 – Representação em blocos de um sistema simples de verificação de locutor.



Fonte: Adaptado de Bäckström et al. (2022).

ondas de áudio com atraso para o microfone (KOUMURA; FURUKAWA, 2017).

Além dos fatores relacionados à coleta da amostra com interferências do ambiente, o falante também pode apresentar variações de fala, por exemplo, a mudança de estado emocional, velocidade de fala, resfriados, entre outros. O modelo que irá analisar os fatores deve minimizar o aprendizado dessas alterações indesejadas e maximizar os fatores fundamentais da voz do locutor (SAI et al., 2023).

2.3.3 Métricas específicas para verificação de locutor

Além das métricas comentadas, outras medidas tradicionais da área de aprendizado de máquina podem ser usadas na validação do sistema, listadas a seguir (ZAIDI et al., 2021; LIU; LANG, 2019):

Acurácia: Ao relacionar o contexto deste projeto, é utilizada como base a verificação do locutor entre um par de áudios. Considera-se o número de amostras corretas em relação ao número total de amostras. As amostras corretas consistem na soma dos verdadeiros positivos (VP), quando ambos os áudios pertencem ao mesmo locutor e o sistema identifica corretamente, e verdadeiros negativos (VN), ou seja, valores identificados corretamente que são de pessoas diferentes. O número total

de amostras é composto pelos verdadeiros positivos, verdadeiros negativos, falsos positivos (FP), que são locutores distintos identificados como a mesma pessoa, e falsos negativos (FN), em que o sistema não reconhece áudios que são do mesmo locutor. Desta forma, tal métrica é definida pela Equação (11).

$$\text{Acurácia} = \frac{VP + VN}{VP + VN + FP + FN} \tag{11}$$

Precisão: consiste em calcular as amostras de dois áudios classificados de maneira correta como positivas pelo número total de pares classificados pelo sistema como positivos, na qual engloba tanto os verdadeiros positivos quanto os falsos positivos. A fórmula matemática é apresentada na Equação (12).

$$\text{Precisão} = \frac{VP}{VP + FP} \tag{12}$$

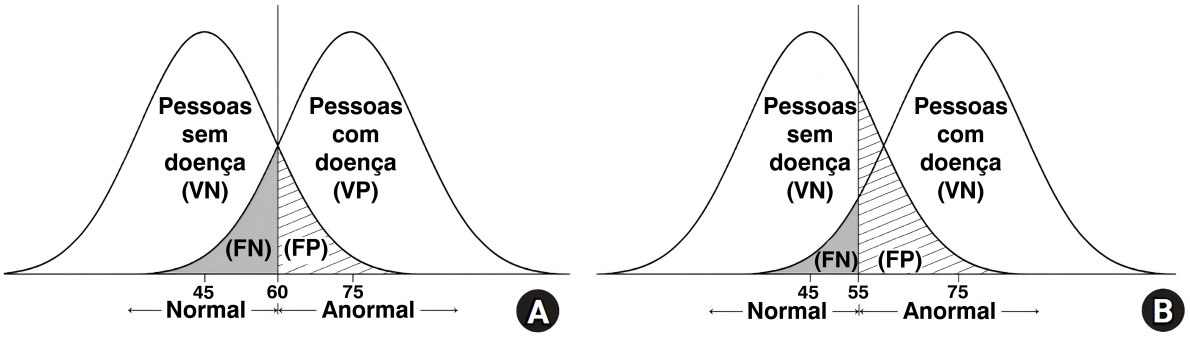
Recall: realiza o cálculo da proporção de vozes verdadeiras positivas pelo total de vozes classificadas como positivas. Esse total consiste na soma dos VP e dos FN, sendo FN os dois trechos do mesmo locutor classificados de maneira errada como locutores diferentes. No âmbito da verificação do locutor, o *Recall* é equivalente à Taxa de Aceitação Verdadeira, do inglês *True Acceptance Rate* (TAR), definida pela quantidade de áudios classificados corretamente como verdadeiros dentre todos os áudios legítimos. Dessa forma, a fórmula é apresentada na Equação (13).

$$\textit{Recall} = \frac{VP}{VP + FN} \tag{13}$$

Curva *Receiver Operating Characteristics* (ROC): um gráfico no qual são apresentadas nos eixos as métricas TAR e FAR. O objetivo é analisar o desempenho em relação a essas métricas para diferentes limiares. Além disso, a área sob a Curva Característica de Operação do Receptor, do inglês Receiver Operating Characteristic (ROC), denominada Área Sob a Curva, do inglês Area Under the Curve (AUC), determina o desempenho de um sistema em um único valor ao realizar o cálculo da área entre a curva ROC e o eixo *x* do gráfico.

Para ilustrar a curva ROC, primeiramente é feita a simulação de duas distribuições: uma para pacientes com a doença e outra para pacientes sem a doença na Figura 20A, em que a média de doentes e não doentes é 75 e 45, respectivamente. Se a nota de corte for a média entre estes valores, ou seja, 60, pessoas com a doença abaixo deste valor serão classificadas como falsos negativos, mas se a nota de corte for reduzida para 55, com o objetivo de aumentar a sensibilidade, o número de pessoas que têm a doença corretamente identificadas aumenta, juntamente com a quantidade de falsos positivos, como representado em FP na Figura 20B.

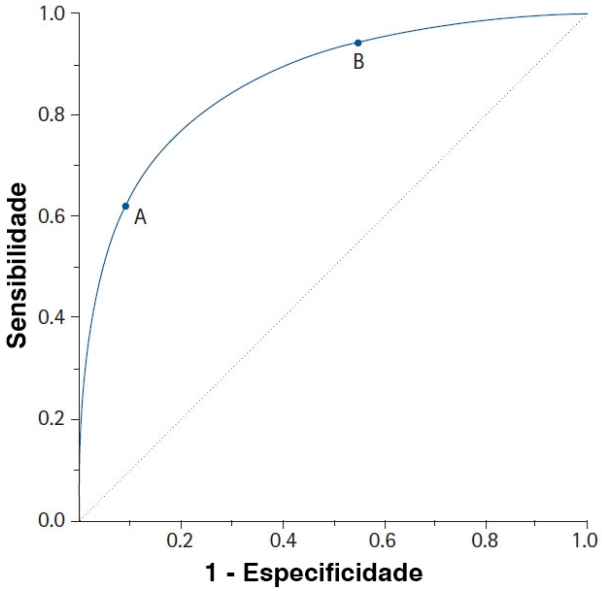
Figura 20 – Distribuição de pacientes hipotéticos com doenças.



Fonte: Adaptado de Nahm (2022).

Com base na distribuição, é possível desenhar a curva ROC representada na Figura 21. Nessa curva, o eixo *y* representa a sensibilidade, ou *recall*, que corresponde à proporção dos indivíduos positivos que foram corretamente identificados pelo modelo, ou seja, à taxa de verdadeiros positivos (VP). O eixo *x* representa 1 - Especificidade. A especificidade corresponde à proporção de indivíduos negativos corretamente classificados, ou seja, à taxa de verdadeiros negativos (VN). Assim, 1 - Especificidade representa a taxa de falsos positivos (FP). Os pontos A e B apresentam os efeitos da mudança nos valores da nota de corte indicados na Figura 20, que, ao reduzir a nota de corte de 60 para 55, a sensibilidade para detectar doentes aumenta, mas também eleva a indicação de alarmes falsos, o que resulta em mais ocorrências com resultado de falso positivo.

Figura 21 – Curva ROC de pacientes hipotéticos com doenças.



Fonte: Adaptado de Nahm (2022).

2.3.4 Bases de dados

Os conjuntos de dados para o processo de verificação de locutor são usados tanto para o treinamento quanto para a avaliação do modelo em si. A maioria dos trabalhos pertencentes à literatura utiliza como base o *VoxCeleb*, atualmente na segunda versão (BEVEREN; SERGIDOU; ZIABARI, 2024). Tal conjunto foi construído por meio de técnicas de visão computacional que extraíram de forma automática milhares de vídeos de celebridades no YouTube (BAI; ZHANG, 2021).

As principais diferenças entre a versão 1 e 2 são diversidade e quantidade de amostras disponíveis. Enquanto a primeira versão contém cerca de 150.000 declarações de 1.251 celebridades, a segunda versão expande para mais de 1.000.000 de declarações espalhadas entre 6.000 celebridades, aproximadamente. Os vídeos incluem variações de ambientes, desde salas destinadas a gravações até estádios lotados. Além disso, os registros incluem falantes de diferentes etnias, sotaques e idades (NAGRANI; CHUNG; ZISSERMAN, 2017; CHUNG; NAGRANI; ZISSERMAN, 2018).

Outro conjunto de dados do contexto da verificação de locutor é o *RS2015*, criado com o objetivo de ser um banco de dados para a subtarefa de dependência de texto, ou seja, é uma autenticação de locutor em que é solicitado ao falante que pronuncie frases específicas que são previamente conhecidas pelo sistema, diferente de independente de texto, que analisa a voz independentemente das frases que são pronunciadas. O conjunto consiste em cerca de 300 falantes, gravados com diferentes dispositivos móveis, totalizando aproximadamente 150 horas. Além disso, tal base é dividida em 3 subconjuntos: gravações para verificação de senhas curtas, registros de comandos específicos e checagem de sequência de dígitos aleatoriamente (LARCHER et al., 2014).

Para a mesma subtarefa de dependência de texto, tem-se o conjunto de dados *Hi-Mia*. Em comparação ao *RS2015*, no qual foi elaborado em um ambiente controlado, com microfones posicionados próximos dos locutores, o *Hi-Mia* utiliza cenários de casas inteligentes, com frases de ativação e gravações registradas por múltiplos microfones, tanto próximos quanto distantes dos falantes. Essa base possui duas divisões: uma com gravações de 250 locutores pronunciando as palavras de ativação, totalizando cerca de 1500 horas, e a outra com 86 falantes, contendo enunciados registrados com e sem ruídos (QIN; BU; LI, 2020).

Por último, num escopo maior abrangendo diversos idiomas, tem-se o conjunto *Common Voice*. Tal banco de dados inclui cerca de 2500 horas de gravações realizadas por 50.000 falantes em 38 idiomas, sendo o maior corpus de áudio para reconhecimento de fala. Os registros são sentenças pré-definidas, lidas pelos próprios

usuários por meio de site ou aplicativo do sistema (ARDILA et al., 2020).

2.4 CONSIDERAÇÕES FINAIS

Neste capítulo, foram apresentados os tópicos pertinentes para o entendimento do projeto, que trata desde os conceitos básicos relacionados ao som e à voz, até as técnicas modernas de processamento do áudio para o meio digital, análise de características com auxílio de inteligência artificial e as principais métricas utilizadas para verificação do desempenho do modelo.

3 TRABALHOS CORRELATOS

Existem diversos trabalhos relacionados ao processamento de fala. O foco deste capítulo está na comparação com artigos desenvolvidos na área de processamento de fala, mais especificamente em pesquisas relacionadas às subáreas de reconhecimento de fala e de verificação de locutor. Distinguem-se, no caso do reconhecimento de fala, a necessidade do conhecimento de uma base de dados composta por locutores previamente cadastrados, a fim de que o modelo possa classificar com qual locutor o áudio a ser verificado se assemelha. Por outro lado, o objetivo da verificação é reconhecer, com base em um par de áudios, se ambos pertencem ao mesmo locutor.

A seguir, são descritas as técnicas, métodos e resultados que os principais autores que compõem os estudos relacionados utilizaram para o avanço do conhecimento da área.

3.1 ANÁLISE DE ESPECTROGRAMA DE LARGURA DE BANDA DUPLA PARA VERIFICAÇÃO DE FALANTE

No estudo elaborado por Virgilli et al. (2024), que é a base utilizada para realizar este trabalho, é proposto o treinamento da arquitetura de rede neural *Fast ResNet-34* com a utilização de espectrogramas de banda dupla, representados pela janela de banda curta para observar as características únicas do falante com o tamanho de 5 ms e 480 amostras e a janela de banda larga que fornece informações ao longo do tempo com o tamanho de 30 ms e 80 amostras. Em relação às métricas de desempenho, foram utilizadas a *Equal Error Rate* (EER) e o *minimum of the Detection Cost Function* (minDCF).

Para a execução do trabalho, a rede *Fast ResNet-34* foi implementada com a utilização do código original, mas com alterações na camada inicial que foi adaptada para receber os dois espectrogramas em vez de um único. Para realizar a convolução, os autores inseriram o *Self-Attentive Pooling* (SAP) para o modelo determinar as partes importantes do áudio analisado. A função de perda *Prototypical Loss Function* foi utilizada para separar as classes com base no ângulo entre um protótipo aprendido, após a passagem do trecho das vozes pela rede neural. Os dados utilizados para o treino e teste foram retirados da base de dados *VoxCeleb2*; este conjunto contém diversos trechos de áudios retirados de mais de um milhão de vídeos do YouTube relacionados à fala de celebridades (VASWANI et al., 2017; SNELL; SWERSKY; ZEMEL,

2017).

Como resultado, após o treinamento e teste com o uso de entradas *Mel-Spectrogram*, o modelo Dual-Bandwidth alcançou uma taxa EER mínima de $1.64\,\%$ $\pm\ 0.13\,\%$. Em relação ao modelo original, houve uma redução de EER de $2.22\,\%$ para $1.64\,\%$, demonstrando uma melhora significativa em relação ao trabalho anterior. Além disso, o espectrograma Mel de banda curta obteve um desempenho melhor comparado ao espectrograma Mel de banda larga, o que indica que a banda estreita carrega características mais valiosas ao modelo comparadas à banda larga.

3.2 CACRN-NET: UM ESPECTROGRAMA DE LOG MEL 3D BASEADO EM REDE NEURAL RECORRENTE CONVOLUCIONAL DE ATENÇÃO DE CANAL PARA IDENTIFICAÇÃO DE LOCUTORES DE POUCOS EXEMPLOS

O artigo de Saritha et al. (2024a) aborda os desafios na identificação de locutores em cenários do mundo real, especialmente em contextos com quantidade reduzida de áudios disponíveis. Esta pesquisa torna-se relevante em situações de segurança ou investigação, onde as amostras coletadas são escassas, e entre os principais obstáculos estão a presença de ruído, variação entre as gravações e problemas de compatibilidade entre o conjunto de treino e teste.

Para resolver os problemas citados, os autores desenvolveram a *CACRN-Net*, uma rede neural que propõe utilizar técnicas de aprendizado em poucos disparos para identificação do locutor. O funcionamento do sistema utiliza a extração de características da fala em um formato de espectro de log mel tridimensional, com um mecanismo de atenção para melhor identificação dos locutores desconhecidos a partir de poucas amostras, além de usar a rede prototípica atenta para o aprendizado por atenção.

O desenvolvimento do modelo envolveu a extração do trecho da gravação a ser analisado, logo em seguida, a amostra é separada em três espectrogramas log mel, sendo o log-mel, a primeira e a segunda derivada do log-mel. Após a separação da entrada, são extraídas as características na rede Rede Neural Recorrente Convolucional, do inglês Convolutional Recurrent Neural Network (CRNN), com o canal de atenção e, por fim, é feita a classificação por meio da perda prototípica e o resultado passado para o *softmax* para a identificação do locutor.

Os dados utilizados no modelo foram obtidos das seguintes bases de áudio:

VoxCeleb1, contendo as gravações de entrevistas em ambiente real; *VCTK*, com gravações livres de ruídos e padrões diversos; e o *IITG-MV*, que é um banco de dados indiano que serviu para identificações de falantes com idioma diferente do inglês. As métricas para análise do modelo foram a precisão, *recall*, *F1-score* e a acurácia; e o otimizador Estimativa Adaptativa de Momento, do inglês Adaptive Moment Estimation (ADAM), foi utilizado para ajustes de peso da rede neural (KINGMA; BA, 2015).

Ao analisar o desempenho, o *CACRN-Net* teve uma acurácia média de $2\,\%$ a mais que o modelo base, utilizando a estratégia de poucos dados com 5 classes e 5 exemplos rotulados de cada. Em comparação com a rede prototípica, ao colocar o trecho de atenção, os resultados utilizando as três bases de dados foram superiores comparados ao modelo sem atenção. O espectrograma Mel 3D contribuiu para o aumento de $0.75\,\%$ na precisão, comparado com as outras entradas, pelo maior detalhamento das características fundamentais derivadas.

3.3 REPRESENTAÇÃO DE ÁUDIO DISCRETA COMO UMA ALTERNATIVA AOS ESPECTROGRAMAS MEL PARA RECONHECIMENTO DE FALA E DE LOCUTOR

No projeto desenvolvido por Puvvada et al. (2024), é demonstrada a discretização de áudio por meio de *tokens* baseados em compressão neural, processo este que realiza a compressão do áudio, como no modelo denominado *EnCodec*, então transforma-se trechos do som em *tokens* semânticos para representar o conteúdo do áudio de forma compactada. As contribuições citadas consistem em estudar os *tokens* baseados em compressões; apresentar os resultados competitivos entre os *tokens* e o espectrograma Mel; demonstrar a solidez do *EnCodec* nas tarefas de reconhecimento de falantes; mostrar as possibilidades de compressão de até 20 vezes sem perder as características fundamentais do áudio e concluir que o modelo de *tokenizador* expõe uma frequência passa-baixa que pode ser útil em trabalhos futuros.

Dentre as características do modelo utilizado para comprimir os dados sonoros, pode-se citar a transformação da taxa de amostragem de 24 kHz para 75 Hz, com uma representação compacta comparada ao original. Além deste fato, os cálculos são feitos com auxílio de 32 *codebooks*, onde armazena-se 32×(d×75) *tokens*, sendo d a quantidade de segundos da amostra. Para comparação, foi utilizado o espectrograma Mel de 80 dimensões, 512 amostras usadas na transformada rápida de Fourier e o janelamento de Hann com 25 ms do tamanho da janela e 10 ms de deslocamento.

O modelo de extração de características e os conjuntos de dados utilizados foram o *TitaNet* com os dados de *VoxCeleb1&2* e *NIST-SRE 2004-2008*; o *En Fast Conformer* com 960 horas de áudio do conjunto *LibriSpeech* e o *En-Es Fast Conformer* com a base *LibriSpeech* e 776 horas de amostras do *Multilingual LibriSpeech*. As tarefas avaliadas foram a verificação dos falantes, que consiste em verificar se a gravação pertence a um falante específico; diarização, que segmenta um arquivo de áudio em partes correspondentes a diferentes falantes; e o reconhecimento de fala com a conversão de áudio em texto.

Ao final dos testes, conclui-se que na verificação de falantes o modelo treinado com *tokens* foi inferior ao espectrograma Mel, com um resultado superior apenas ao treinar com as bases *VoxCeleb 1&2* e testar no modelo *NIST-SRE 18*. Na diarização, apesar de estimar com mais precisão o número de falantes em alguns casos, em geral, o espectrograma Mel obteve melhores estimativas. No reconhecimento de fala, as duas entradas obtiveram o mesmo desempenho com os dados sem ruído, mas com a adição de ruído, os conjuntos de validação e teste obtiveram uma diferença de 0.47 e 0.35, respectivamente, com vantagem para o espectrograma Mel.

3.4 IDENTIFICAÇÃO DE LOCUTOR DE PONTA A PONTA BASEADA EM APRENDIZADO PROFUNDO USANDO REPRESENTAÇÃO DE TEMPO-FREQUÊNCIA DO SINAL DE FALA

O trabalho publicado por Saritha et al. (2024b) consiste na entrada dos dados por meio de espectrogramas em vez de áudios unidimensionais e na construção de um modelo de rede neural com uma menor quantidade de parâmetros. Dentre as contribuições, foram citadas uma melhor representação do áudio ao utilizar tempo e frequência em vez de dados brutos; a redução do tempo de treinamento com a utilização de espectrogramas comparados ao processamento dos dados brutos; utilização da rede *SpectroNet* com a arquitetura *MobileNetV2*, que possui menos parâmetros comparada a outras arquiteturas pelo seu uso eficiente em dispositivos móveis; e a comparação e análise entre vários modelos de redes neurais convolucionais profundas.

Os dados utilizados para o treinamento e teste foram obtidos da primeira parte da base *RSR2015* e do *VoxCeleb1*. O intuito dos autores ao escolher essas bases é o fato de o *RSR2015* ser um conjunto que contém apenas os dados mínimos para reconhecer o falante e a *VoxCeleb1* representar condições ruidosas, por sua natureza de ser trechos de áudio obtidos do YouTube com ruídos capturados em ambiente real.

Após a coleta dos dados, os sinais de fala são pré-processados a uma taxa de amostragem de 16 kHz e segmentados em um período de duração de 20 ms. Dentre os modelos de janela possíveis, utilizou-se a janela de Hamming para minimizar distorções espectrais. Para a análise do modelo, foram usados três tipos de espectrogramas, sendo eles: espectrograma, espectrograma log Mel e o cocleagrama. De acordo com os três espectrogramas gerados, cada um terá uma entrada na rede que segue o formato tridimensional 224x224x3, obtido por meio da convolução 1x1 aplicada a cada espectrograma 224x224x1.

Por fim, o resultado do experimento é validado pelas seguintes medidas de desempenho: Precisão, *Recall*, *F-score* e Acurácia. Entre os espectrogramas, os melhores resultados na base de dados RSR2015 foram obtidos ao utilizar o cocleagrama, com a acurácia de $98.23\,\%$, precisão de $98.19\,\%$, *recall* de $98.58\,\%$ e *F-score* de $98.38\,\%$. Na comparação entre os trabalhos correlatos dos autores, esta pesquisa alcançou o melhor desempenho ao usar a base *VoxCeleb1*, com a acurácia de $99.26\,\%$. Demonstra-se com este projeto que o *SpectroNet* obteve maior êxito em comparação com as demais redes, e que a maior precisão foi obtida com o uso do cocleagrama.

3.5 UM MODELO DE REDE NEURAL PROFUNDA PARA IDENTIFICAÇÃO DE LOCUTOR

No artigo de Ye e Yang (2021), é criado um modelo de rede neural profunda que combina as características de extração de informação da rede neural convolucional e a capacidade de aprendizado de séries temporais de uma unidade recorrente com portões. Dentre as razões para as escolhas, é citado que as Redes Neurais Convolucionais, do inglês Convolutional Neural Network (CNN) são reconhecidas por identificar padrões em imagens, sendo que o espectrograma é uma representação visual do som; desta forma, a CNN é utilizada para a extração das características relacionadas à imagem do som. E a Unidade Recorrente com Portões, do inglês Gated Recurrent Unit (GRU), foi selecionada por ter um desempenho superior em relação às outras Redes Neurais Recorrentes, do inglês Recurrent Neural Networks (RNNs) tradicionais e, diante deste fato, é empregada para a análise detalhada dos trechos de áudio.

Após a análise da junção dos modelos propostos, é realizado o pré-processamento dos dados para criar os espectrogramas; logo após, a imagem gerada passa pela CNN, onde são extraídas as peculiaridades do falante. Em seguida, transfere-se para as camadas GRU empilhadas, onde o aprendizado temporal é realizado. Por fim, as nuances de cada voz são obtidas e a camada de *softmax* pontua valores para identificar o locutor.

Na fase de experimentos, foi utilizado o conjunto de dados *Aishell*, que consiste em uma base de dados com falas em mandarim gravadas por 400 falantes com diferentes sotaques chinês. Os equipamentos de gravação foram um microfone, que grava áudios em taxa de amostragem de 44,1 kHz, um celular Android e o outro iOS com gravações a 16 kHz e quantização nos três dispositivos de 16 bits; contudo, a taxa dos sons gravados pelo microfone foi reduzida para 16 kHz para normalizar e produzir o conjunto de dados. A divisão dos dados seguiu-se na razão de $90\,\%$ para treinamento, $5\,\%$ validação e $5\,\%$ teste devido ao grande volume de dados utilizados de aproximadamente 142 mil trechos de áudio. Ao final, a geração do espectrograma Mel foi realizada com 512 amostras por janela, no tamanho de 32 milissegundos e sobreposição a cada 16 milissegundos, com o tipo de janela Hamming.

No treinamento realizado, utilizou-se uma taxa de aprendizado decrescente linear de 10^{-4} em 80 épocas e 32 minilotes. Nos testes, a precisão do modelo estudado chegou a $98.96\,\%$, e em comparação com os modelos tradicionais de Memória de Longo Prazo e Curto Prazo, do inglês Long Short-Term Memory (LSTM) e CNN, todos foram superados, inclusive com dados ruidosos. Ao final, as conclusões destacam a eficiência do modelo, com a combinação bem-sucedida das redes CNN e GRU. Como recomendação de trabalhos futuros, é sugerido realizar novos estudos em ambientes com ruídos reais.

3.6 ESPECTROGRAMAS MULTICANAL PARA APLICAÇÕES DE PROCESSAMENTO DE FALA USANDO MÉTODOS DE APRENDIZAGEM PROFUNDA

A pesquisa elaborada por Arias-Vergara et al. (2021) apresenta o uso de três diferentes representações de espectrogramas na entrada do modelo de rede neural, para detecção de fala de usuários com implante coclear; dentre eles, o espectrograma Mel, o cocleagrama e a transformada wavelet contínua.

Para a obtenção das duas primeiras entradas citadas, deve-se converter o som por meio da transformada de Fourier de curta duração, que divide a amostra em janelas de 40 ms e as obtém a cada 10 ms, para garantir a sobreposição e a melhor análise temporal. A última transformada não é processada pelo tamanho da janela, mas sim ao longo do sinal, onde a escala da wavelet é ajustada para capturar as variações específicas das frequências vocais ao longo do tempo.

Após a definição das entradas, a execução da pesquisa segue em dois modelos para o processamento da informação, o primeiro é uma CNN inspirada na arqui-

tetura da LeNet-5, baseada em extrair os detalhes fundamentais da voz; a otimização dos pesos desta rede foi feita pelo ADAM, com a taxa de aprendizado de 10^{-4}. O segundo é uma arquitetura aprimorada do GRU, que é chamada de Unidade Recorrente com Portões e Camada Convolucional, do inglês Convolutional Gated Recurrent Unit (CGRU), responsável por identificar os fonemas.

Os áudios utilizados para validação do sistema foram gravados em uma clínica com usuários de implante coclear e também extraídos da base de dados *PhonDat1*. Neste contexto, as amostras passaram por pré-processamentos para redução de ruído e compressão de áudio. Alguns dos atributos escolhidos para verificação são os sons nasais, as vogais, as trinadas, fricativas, entre outros.

Nos experimentos, a comparação entre os três espectrogramas revela que a precisão, *recall* e *F1-score* são superiores ao utilizar a entrada 3D comparada às entradas únicas de 1D. O modelo de CNN para reconhecimento da fala desordenada de usuários com implantes cocleares obteve o desempenho F1 de 0,84. No reconhecimento de fonemas com o modelo CGRU, a combinação das três representações do espectrograma não foi um fator decisivo para melhorar o desempenho da rede.

Conclui-se que os três canais foram relevantes na detecção da fala desordenada, mas a adição de múltiplos canais de entrada não demonstrou um melhor desempenho comparado aos canais únicos, que abrem possibilidades para os trabalhos futuros explorarem outras representações de tempo-frequência.

3.7 ARQUITETURAS DE EXTRAÇÃO DE CARACTERÍSTICAS: PANN, ECAPA-TDNN E ECAPA2

Para extrair as características nos sistemas de verificação de locutor, diferentes arquiteturas tem sido propostas na literatura, com diversos vieses indutivos. A rede PANN (*Pre-trained Audio Neural Network*), especificamente a variante CNN14, proposta por Kong et al. (2020), é baseada em convoluções bidimensionais (2D) e tem se destacado por processar espectrogramas de áudio parecido com o processamento de imagens, com a captura eficiente de padrões de tempo-frequência.

Além das arquiteturas convolucionais clássicas, o estado da arte foi impulsionado pelas Redes Neurais de Atraso de Tempo (TDNN). A arquitetura ECAPA-TDNN, introduzida por Desplanques, Thienpondt e Demuynck (2020), trouxe melhorias no processamento das características temporais 1D, com a utilização de blocos residuais multiescala e canais de atenção (*Squeeze-and-Excitation*) para capturar dependências de longo alcance no sinal de fala.

Mais recentemente, a evolução dessa arquitetura gerou o ECAPA2 desen-

volvido por Thienpondt e Demuynck (2023). Embora a origem utilize o processamento temporal do TDNN, o ECAPA2 incorporou mecanismos de atenção avançados e representações que aproximam o modelo do processamento 2D. Essa capacidade de capturar a estrutura de tempo-frequência tornou o ECAPA2 uma das abordagens mais robustas atualmente, o que justifica sua escola, juntamente com a rede PANN (CNN14) e o ECAPA-TDNN original, como *backbones* nos experimentos desenvolvidos neste trabalho.

3.8 CONSIDERAÇÕES FINAIS

As abordagens utilizadas remetem ao uso de mais de um espectrograma na entrada, com a justificativa de verificar diferentes posições da amostra, muitas vezes com análises em janelas que se sobrepõem.

O maior problema enfrentado na área é a frequente possibilidade da amostra de áudio conter ruído, que atrapalha o processo de análise do modelo. Para suavização dos gradientes e proporcionar estabilidade ao modelo, utilizou-se do otimizador ADAM e suas variações, como AdamW e NAdam, para ajustar os pesos dos modelos e controlar a atualização de descida eficiente de gradientes.

Em relação às bases de dados, fica clara a vantagem da *VoxCeleb2* em relação às demais, por apresentar a maior quantidade de amostras e locutores de diferentes idiomas em que a captação do som acontece no ambiente real. Por fim, conclui-se que os trabalhos futuros recomendados são direcionados a melhorar as representações de entrada, quantidades de espectrogramas, e exploração de condições ruidosas.

Por fim, foram apresentadas as arquiteturas padrão de extração de características utilizadas neste trabalho, que serviram como base para a análise e comparação dos modelos avaliados.

4 MATERIAIS E MÉTODOS

Neste capítulo, são apresentadas a proposta e as configurações iniciais utilizadas na realização dos experimentos, bem como a descrição dos experimentos realizados e a análise dos resultados obtidos, a fim de obter a validação da hipótese de que o uso de espectrogramas Mel multirresolução, especificamente com três resoluções de frequência, otimiza o desempenho de verificação de locutor baseado em aprendizado profundo.

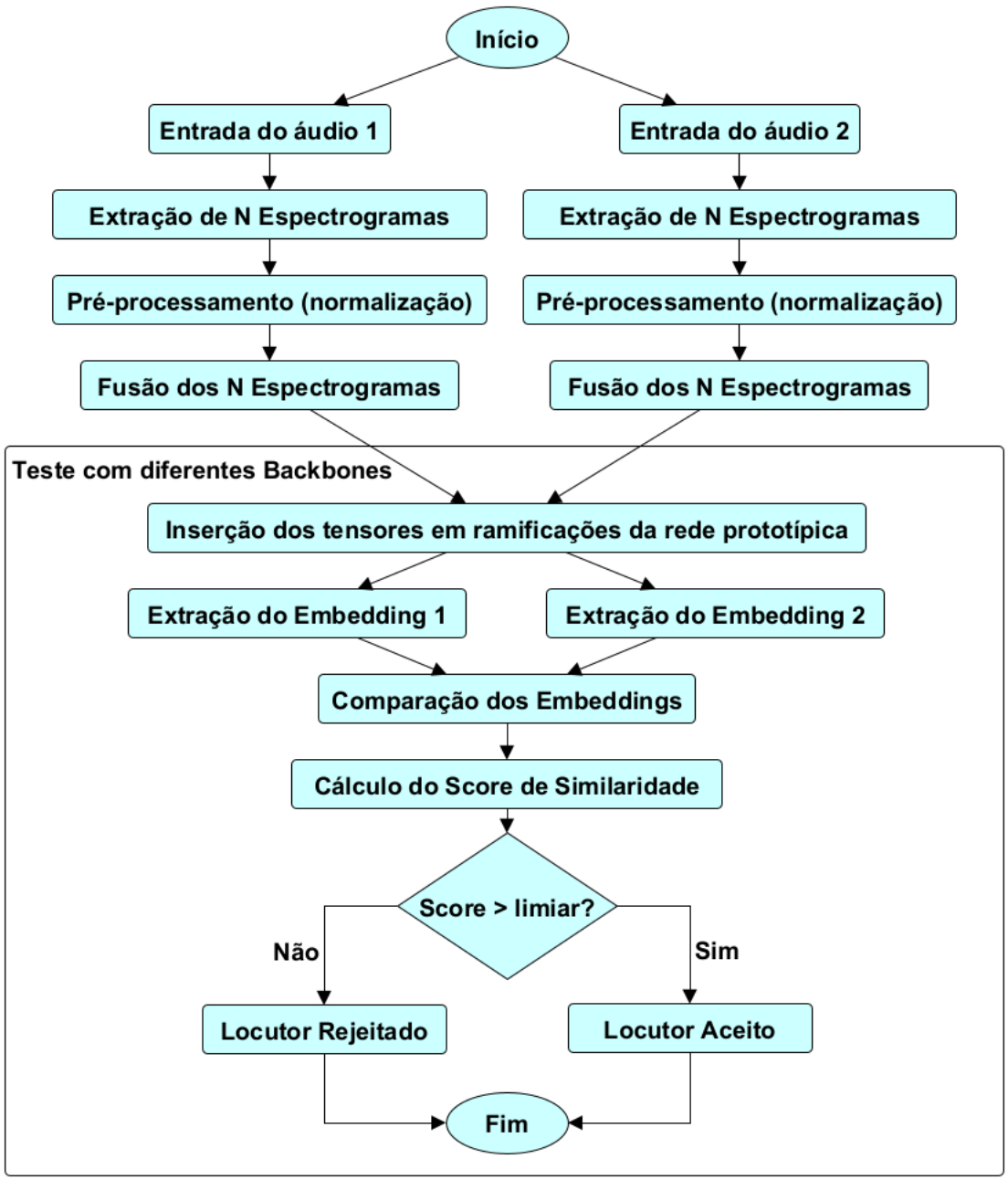
4.1 PROPOSTA INICIAL

Este projeto visa uma versão aprimorada do estudo de Virgilli et al. (2024), que inclui testes com a utilização de diferentes *backbones*, além de permitir variações nas larguras de resolução dos dados de entrada.

O diagrama exibido na Figura 22 ilustra o processo em que os áudios irão percorrer até a aprovação ou reprovação da similaridade do locutor. O processo inicial descreve a inserção de ambos os áudios, por exemplo, o primeiro áudio sendo do locutor autorizado e o segundo do indivíduo que está tentando se autenticar. Após a entrada dos áudios, foram extraídos os N espectrogramas de cada amostra, com a utilização de larguras de resoluções distintas para analisar diferentes representações do mesmo sinal. O pré-processamento é responsável pela padronização da escala, do tamanho e do formato das amostras. A etapa de fusão é a junção dos tensores dos espectrogramas, que, ao final, gera um único tensor de cada áudio.

A etapa seguinte do diagrama é caracterizada pela utilização de diferentes *backbones*. No trabalho base de Virgilli et al. (2024), o *backbone* utilizado foi o ResNet-34, em que dois tensores gerados pelos áudios de entrada transitaram em redes neurais prototípicas. Cada ramificação gera um vetor de características ao extrair os *embeddings* de cada amostra. Com a extração das características, é realizada a comparação com a utilização de uma função de similaridade, como a distância euclidiana, para o cálculo do *score* de similaridade. Esse *score* define se os áudios pertencem ao mesmo locutor ou não, pois, caso o *score* obtido seja maior que o limiar predefinido, considera-se que os áudios pertencem ao mesmo locutor; caso o *score* esteja abaixo do limiar, a autenticação é rejeitada.

Figura 22 – Diagrama do processo para verificação de locutor proposto.



Fonte: Elaborado pelo autor.

4.2 REPRESENTAÇÃO ESPECTRAL MULTIRRESOLUÇÃO

O sinal de entrada, amostrado a 16 kHz, é decomposto em múltiplos espectrogramas log Mel que utilizam diferentes tamanhos de janelas temporais (8 ms, 16 ms, 32 ms) (JOSHI; PAREEK; AMBATKAR, 2023; AYVAZ et al., 2022; BOULAL et al., 2024). As janelas curtas servem para reproduzir espectrogramas de banda larga, que

exibem uma melhor representação do domínio do tempo, o que permite identificar com precisão o momento em que um som ocorreu. No entanto, esta configuração dificulta distinguir quais frequências estavam presentes naquele instante. Em relação à janela longa, que gera o espectrograma de banda estreita, é proporcionada uma visualização mais nítida das frequências, mas com uma menor resolução temporal, o que prejudica a identificação de quando uma frequência inicia e termina.

Assim, cada tamanho de janela influencia na construção de informações complementares do falante, em que o deslocamento dos quadros é proporcional ao tamanho da janela, o que permite uma correspondência preliminar em relação ao tempo nas representações. Para garantir um alinhamento temporal exato e sincronizado antes do processamento convolucional, os espectrogramas foram truncados para a dimensão temporal mínima comum encontrada entre as múltiplas escalas.

4.2.1 Aprendizagem consistente por características em diferentes resoluções

Cada representação de espectro é processada por um caminho independente, sem redução temporal, o que preserva a estrutura de cada escala e evita a dominância de uma resolução sobre as demais. Após o processo inicial, as representações foram alinhadas e fundidas, o que representa a fusão das características. Este processo força o modelo a aprender padrões em diferentes resoluções, em vez de depender da resolução temporal.

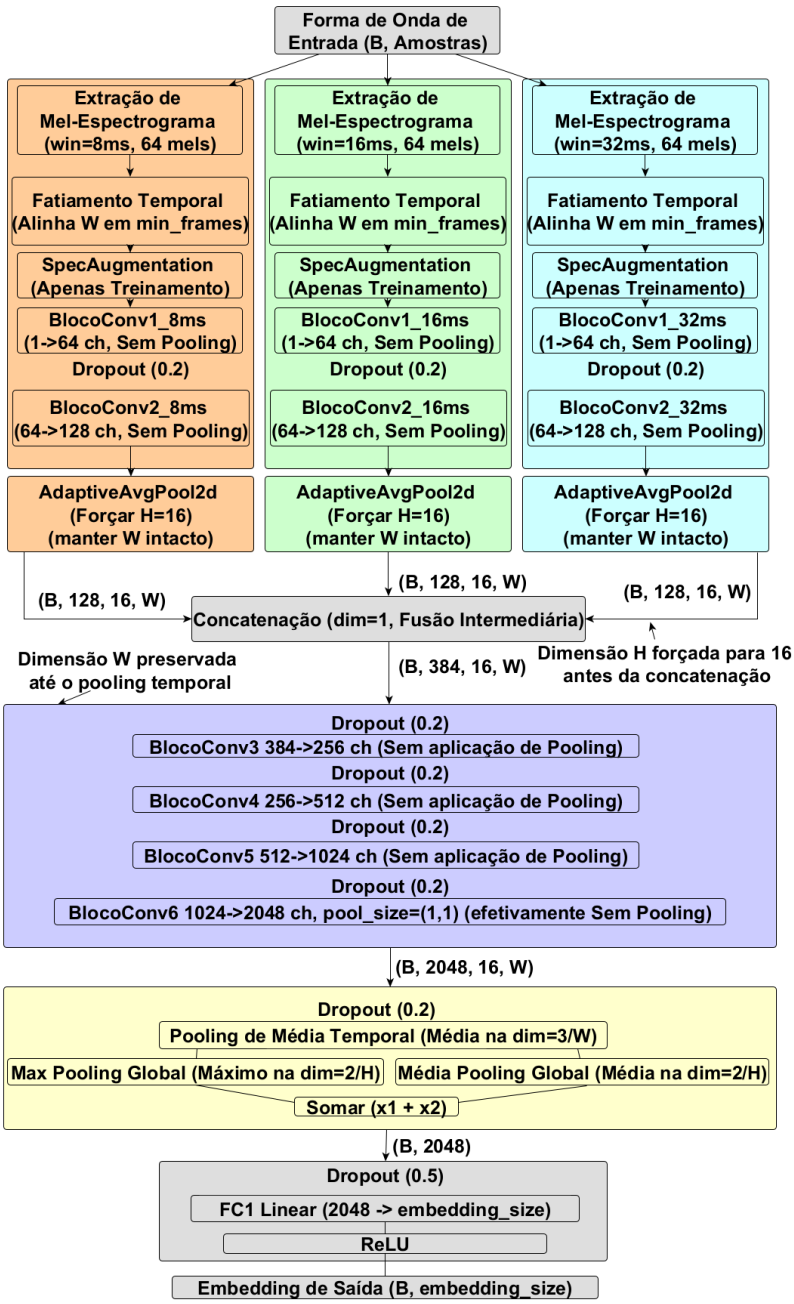
Antes da fusão, as representações foram alinhadas na dimensão de frequência por meio de pooling médio adaptativo, mantendo a dimensão temporal sem alterações. Em seguida, realiza-se a fusão das representações por meio da concatenação das representações ao longo da dimensão de canais.

Após a fusão, a entrada unificada é processada por blocos convolucionais adicionais, responsáveis por modelar características de alto nível associadas à identidade do locutor.

Por fim, é aplicada uma etapa de Temporal Double Aggregation Pooling (TDAP), em que realiza-se um average pooling global ao longo do eixo temporal para condensar a dinâmica do sinal. Em seguida, aplica-se a combinação de max pooling e average pooling ao longo do eixo de frequência remanescente, somando-se os resultados para capturar simultaneamente as ressonâncias dominantes e as estatísticas globais do espectro. A representação resultante é projetada em um espaço vetorial de dimensão fixa por meio de uma camada totalmente conectada, produzindo o embedding final utilizado na verificação do locutor (KAYA; BILGE, 2019; LU; HU; ZHOU, 2017; WANG; LIU, 2021).

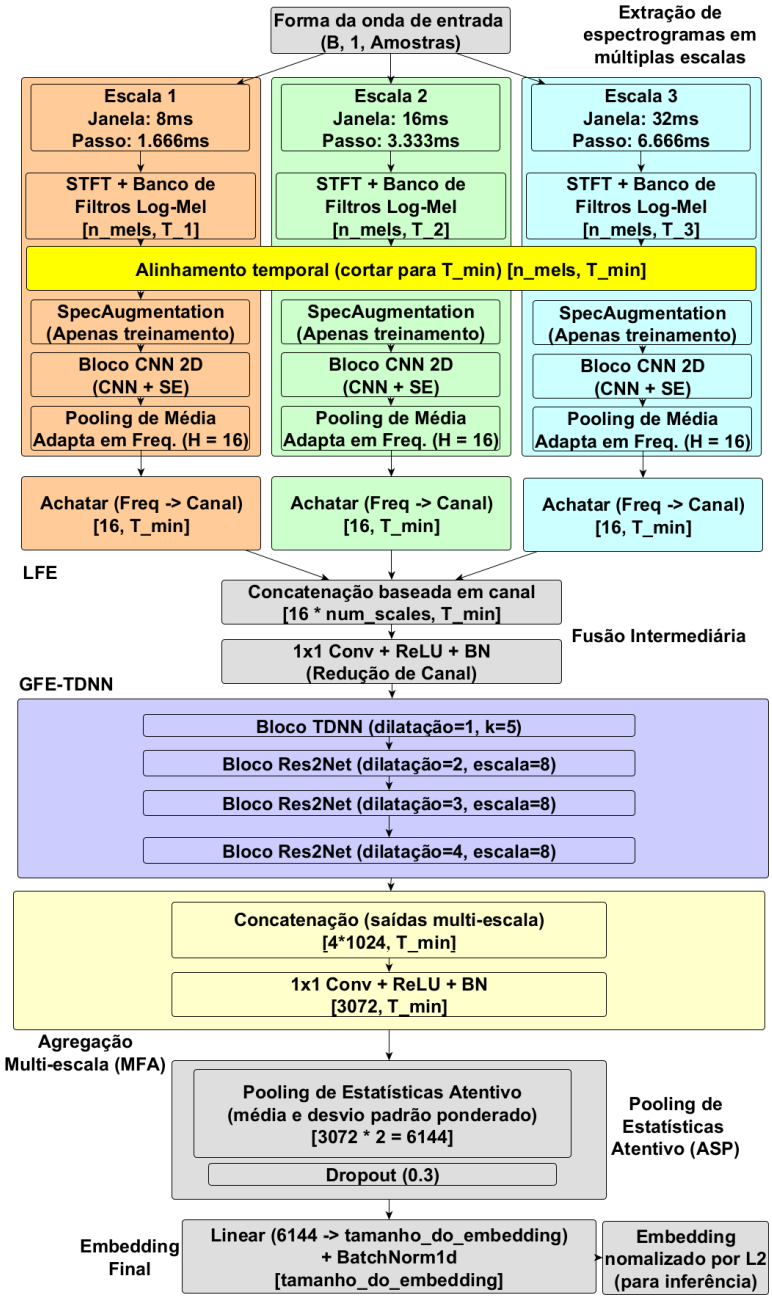
Na Figura 23 é possível observar a arquitetura proposta da rede PANN. A seguir, apresenta-se a arquitetura do ECAPA2, que também utiliza três entradas, conforme ilustrado na Figura 24. Embora ambas utilizem a mesma estratégia de fusão para as entradas multibanda, elas se diferenciam no modelo principal utilizado para a extração de características. Enquanto a PANN (CNN14) é baseada em redes neurais convolucionais (CNNs) projetadas para o reconhecimento geral de padrões de áudio, o ECAPA2 utiliza redes neurais de atraso de tempo (TDNNs) projetadas e otimizadas especificamente para a tarefa de verificação de locutor.

Figura 23 – Arquitetura proposta da rede PANN (CNN14).



Fonte: Elaborado pelo autor.

Figura 24 – Arquitetura proposta da rede ECAPA2.



Fonte: Elaborado pelo autor.

4.2.2 Materiais e métodos

Dentre os materiais, a base de dados utilizada no estudo é a VoxCeleb1 por ser a mais versátil base de dados disponível atualmente e por conter áudios gravados em diferentes condições de ruído e a VoxCeleb2. O desenvolvimento do código-fonte, treinamento e testes do modelo foi realizado no ambiente do Centro de Excelência em Inteligência Artificial da UFG , um conjunto de hardware que oferece GPUs, como a NVIDIA H100, para aceleração do processamento dos dados. Para o desenvolvi-

mento, foi escolhida a linguagem Python[1], na versão 3.11, com as bibliotecas essenciais: Librosa[2], na versão 0.11.0, para carregar e extrair o espectrograma Mel dos áudios; PyTorch[3], na versão 2.6.0, para implementação e treinamento das redes neurais; e Scikit-Learn[4], na versão 1.6.1, para avaliação do modelo por meio de métricas de validação.

O método que guiou este projeto consiste em utilizar a base de dados VoxCeleb1 e VoxCeleb2, pré-processar os áudios para extração do espectrograma mel por meio da biblioteca librosa, que serve de entrada para o modelo de rede neural implementado com o PyTorch, e utilizar uma rede prototípica com testes de diversos *backbones* para processar as características fundamentais dos áudios de entrada.

Enquanto o trabalho base utilizava apenas dois canais, sendo eles de banda estreita e banda larga, esta dissertação propôs a análise de múltiplos canais, baseados em uma, duas e três resoluções, com o objetivo de analisar o impacto destes canais na verificação de locutor. O treinamento foi realizado com ajustes dos parâmetros, como o número de épocas, com o objetivo de otimizar o desempenho do modelo. Após o treino, foi feita a avaliação de desempenho por meio das métricas específicas e tradicionais para verificação de locutor, que foram utilizadas com o suporte da biblioteca scikit-learn.

A validação dos resultados obtidos pelo modelo proposto é a taxa de falsos aceites, juntamente com a taxa de falsos rejeitos, cuja métrica principal é a taxa de erro igual, que representa o ponto em que as duas taxas se igualam. Além desta métrica, foram utilizadas as métricas específicas para a verificação de locutor, que incluem a acurácia, precisão, *recall* e curva ROC.

Em relação aos testes no modelo, foram utilizados diferentes valores nos parâmetros de entrada, como a variação na quantidade de canais utilizados, além da comparação entre diferentes *backbones* que compõem o estado da arte. No quesito de treinamento, os parâmetros principais ajustados são o número de épocas, o tamanho do lote e a taxa de aprendizado. O ajuste ideal desses parâmetros foi definido com base nos resultados obtidos por meio das métricas mencionadas anteriormente.

4.3 CONFIGURAÇÃO

Os experimentos foram realizados no supercomputador para Inteligência Artificial da Universidade Federal de Goiás (UFG), em uma máquina NVIDIA DGX, com a utilização do nó compartilhado dgx-H100-02, também utilizado em outros projetos de

[1] https://www.python.org/
[2] https://librosa.org/
[3] https://pytorch.org/
[4] https://scikit-learn.org/

pesquisa. Este sistema possui as seguintes configurações: dois processadores Intel Xeon Platinum 8480C, com o total de 112 núcleos físicos e 224 threads, oito GPUs NVIDIA H100, cada uma equipada com 80 GB de memória e 2 TiB de memória RAM disponível. O ambiente de execução dos experimentos foi configurado da seguinte forma conforme apresentado na Tabela 1:

Tabela 1 – Configuração do ambiente computacional utilizado nos experimentos

Especificação	**Valor**
Núcleos físicos do processador	1 núcleo Intel Xeon Platinum 8480C
Núcleos virtuais do processador	16 núcleos virtuais
GPU	1 GPU NVIDIA H100
Memória RAM	64 GiB de memória RAM
Execução do contêiner	Apptainer

Em relação ao tempo de treinamento, em média, os experimentos demoraram, em média, entre duas a três horas para a execução de 200 épocas com o VoxCeleb1, no ambiente descrito na Tabela 1.

4.3.1 Conjuntos de dados

Para realização dos treinos e testes, foi escolhida inicialmente a base de dados de áudio VoxCeleb1, que, devido ao seu tamanho reduzido de 153.516 trechos de áudio de 1.251 locutores, mostra-se ideal para validar a hipótese deste projeto. Para alcançar um desempenho competitivo em relação ao trabalho correlato de Virgilli et al. (2024), o conjunto de dados VoxCeleb2 revela-se a melhor opção, por conter 1.128.246 trechos de áudio de 6.112 locutores.

A separação do conjunto VoxCeleb1 e VoxCeleb2 para treino e teste segue os padrões citados por Nagrani, Chung e Zisserman (2017) e Chung, Nagrani e Zisserman (2018), respectivamente. A divisão é apresentada na Tabela 2:

Tabela 2 – Divisão dos dados de treino e teste

	VoxCeleb1	VoxCeleb2
Locutores		
Treino	1.211	5.994
Teste	40	118
Total	1.251	6.112
Trechos de fala		
Treino	148.642	1.092.009
Teste	4.874	36.237
Total	153.516	1.128.246

Além das bases de áudio citadas, foram utilizados conjuntos adicionais para aumento de dados (data augmentation), incluindo trechos de música, fala e ruído (MUSAN - Music, Speech, and Noise) e o banco de respostas ao impulso (RIRs - Room Impulse Responses) para salas pequenas, médias e grandes, disponibilizados pela Open Speech and Language Resources (OpenSLR).

4.3.2 Pré-processamento e geração dos espectrogramas

O pré-processamento deste projeto foi focado em ajustar a entrada do projeto para a geração de três tipos de espectrogramas com janelamento fixo de tamanho 6,25 ms, configurados da seguinte forma:

- Broadband: Janela Hamming de 5 ms;
- Narrowband: Janela Hamming de 30 ms;
- Narrowerband: Janela Hamming de 50 ms.

Sendo que no experimento principal, que obteve o melhor resultado, foi definido o tamanho das janelas e saltos no seguinte formato, com base na escala logarítmica base 2, e com o respaldo de que ao aumentar o tamanho das janelas acima de 30ms não houve ganho significativo:

- Broadband: Janela Hamming de 8 ms com salto aproximado de 1,66 ms;
- Narrowband: Janela Hamming de 16 ms com salto aproximado de 3,33 ms;
- Narrowerband: Janela Hamming de 32 ms com salto aproximado de 6,66 ms.

O Broadband corresponde a uma janela curta, que demonstra uma boa resolução temporal, mas com a resolução em frequência menor; o uso desta janela serve para capturar mudanças rápidas no sinal do áudio.

O Narrowband é a janela média que serve de base para um bom equilíbrio entre a resolução temporal e de frequência; o uso deve-se ao fato de analisar os padrões e frequências estáveis da voz.

O Narrowerband é a janela longa, com foco em demonstrar uma melhor resolução de frequência, mas perde em resolução temporal; útil para verificar os harmônicos e componentes estacionários.

A escolha inicial para utilizar janelas de espectro único de 30 ms e duplos de 5 ms e 30 ms foi fundamentada com base no trabalho de Virgilli et al. (2024), a janela posterior de 25ms foi utilizada com base no trabalho de Thienpondt e Demuynck

(2023). O objetivo principal foi comparar resoluções complementares de janela curta para análise de transições rápidas e janela longa para análise das frequências. As variações milimétricas do tamanho da janela longa tendem a representar um impacto mínimo nos resultados, com a priorização da captura das representações do sinal.

Por fim, o passo dos janelamentos foi especificado fixo para os experimentos complementares e variável para o experimento principal em filtros Mel. Diante da junção de três janelas especificadas, é possível aproveitar simultaneamente todas as resoluções temporais e de frequência, com foco em um melhor desempenho na análise da voz do locutor em relação ao uso de apenas duas janelas.

4.3.3 Arquitetura e estratégia de treinamento

Em relação à arquitetura, os testes iniciais foram realizados com a utilização de redes pré-treinadas PANN na variante CNN-14, uma rede convolucional profunda composta por 14 camadas. O modelo original utiliza apenas um canal de entrada; portanto, foram realizadas adaptações para que o modelo pudesse receber dois e três canais para a realização dos experimentos.

Dentre as estratégias utilizadas, foram realizados treinamentos do zero e a abordagem de *fine-tuning* na rede pré-treinada Cnn14_16k_mAP=0.438, modelo este que foi treinado com amostras de áudio de 16 kHz do conjunto da Google denominado AudioSet. Este conjunto inclui sons humanos e de animais, além de sons instrumentais e do ambiente cotidiano coletados do YouTube. A rede utilizada obteve uma Precisão Média (mAP) de acertos de 0,438 em cada classe.

Ao relacionar as configurações dos hiperparâmetros considerados de maior relevância para este projeto, destacam-se na Tabela 3 os seguintes:

Por fim, as fases de treinamento e teste foram avaliadas em relação à Taxa de Erro Igual - EER e Custo Mínimo de Detecção - MinDCF. A EER é importante no fator de verificação de locutor para definição da taxa de erro do sistema, em que é utilizada sua versão para Verificação da Taxa de Erro Igual - VEER dos testes realizados, enquanto o MinDCF realiza a ponderação dos erros.

Tabela 3 – Configuração dos hiperparâmetros utilizados no projeto

Parâmetro	**Valor**	**Descrição**
Função de Perda	Softmax Prototypical Loss AAM Softmax	Função utilizada para cálculo do erro durante o treinamento.
Otimizador	Adam	Algoritmo de otimização para atualização dos pesos da rede.
Taxa de Aprendizagem Inicial	0.001 (1e-3) 0.0001 (1e-4)	Determina o tamanho do passo de atualização dos pesos na primeira época.
Scheduler de LR	StepLR	Ajusta a taxa de aprendizagem com decaimento de 0.95 a cada 10 épocas.
Tamanho do Batch	200	Número de amostras processadas antes de atualizar os pesos da rede.
Épocas de Treinamento	100/200	Número total de passagens completas pelo conjunto de dados durante o treinamento.

5 RESULTADOS E DISCUSSÃO

Este capítulo é dividido entre a apresentação de experimentos preliminares para explorar as possibilidades de configurações, e experimentos principais que demonstram a hipótese multirresolução em modelos recentes adaptados.

5.1 EXPERIMENTOS PRELIMINARES

O objetivo dos experimentos preliminares é comparar e selecionar as configurações com melhor desempenho, a fim de analisar e extrair hipóteses e evidências que fundamentem o desenvolvimento do presente trabalho. Para isso, os experimentos foram divididos em cinco etapas principais, todas utilizando a base de dados VoxCeleb1 e sua divisão oficial entre os conjuntos de treinamento e teste.

Na primeira etapa, mostra-se a comparação do aprendizado de características aprendidas do zero em relação ao aprendizado previamente realizado a partir de uma base de áudio consolidada, com auxílio do processo de *fine-tuning*. O segundo experimento compara a base de dados original com a versão após a utilização de técnicas de *data augmentation*. O terceiro experimento busca descobrir o melhor desempenho em comparação à janela Narrowerband de 50 ms e 60 ms. O quarto experimento é uma continuação do terceiro experimento, em que são testado tamanhos diferentes de janelas e de deslocamentos. E o quinto experimento é verificar o quão os resultados do VoxCeleb2 são superiores ao VoxCeleb1 em comparação à quantidade de dados entre as bases de dados.

5.1.1 Experimento 1: Treinamento a Partir do Zero vs. *fine-tuning*

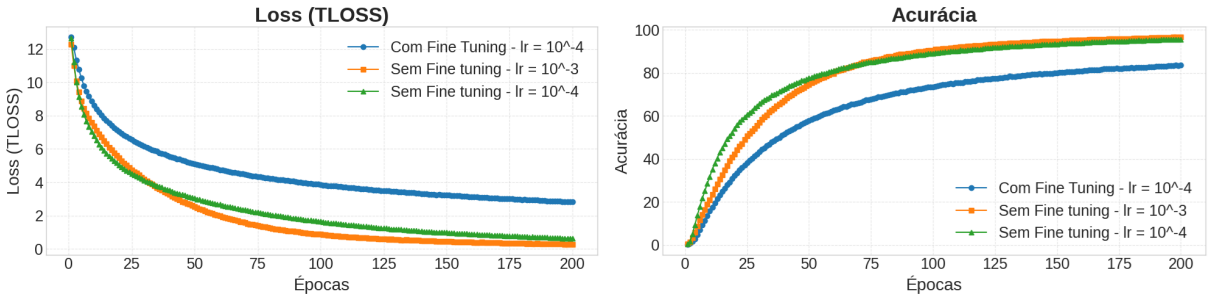
Ao iniciar os experimentos, a primeira decisão para a realização do treinamento, após adquirir o conjunto de dados, consiste em decidir se o treino irá ocorrer somente com a base de dados fornecida, ou se será utilizado um modelo com pesos pré-treinados.

Em relação ao primeiro experimento, foi realizada uma comparação entre o treinamento do modelo a partir do zero (*Sem fine-tuning*) e o treinamento com ajuste fino (*Com fine-tuning*). Foram avaliados três modelos, o primeiro com a utilização de pesos pré-treinados do modelo Cnn14 (16 kHz, mAP = 0,438) e learning rate de 10^{-4}, o segundo treinado do zero sem *fine-tuning* com 10^{-3} e o terceiro com aplicação de *fine-tuning* e learning rate de 10^{-4}. Os treinos e testes dos casos citados foram realizados

no mesmo conjunto de dados VoxCeleb1, sem a utilização de *data augmentation* e com três resoluções (5 ms, 30 ms, 50 ms).

A Figura 25 compara as curvas de Perda (Loss - TLOSS) e Acurácia obtidas durante o treinamento ao longo de 200 épocas.

Figura 25 – Treino *fine-tuning*

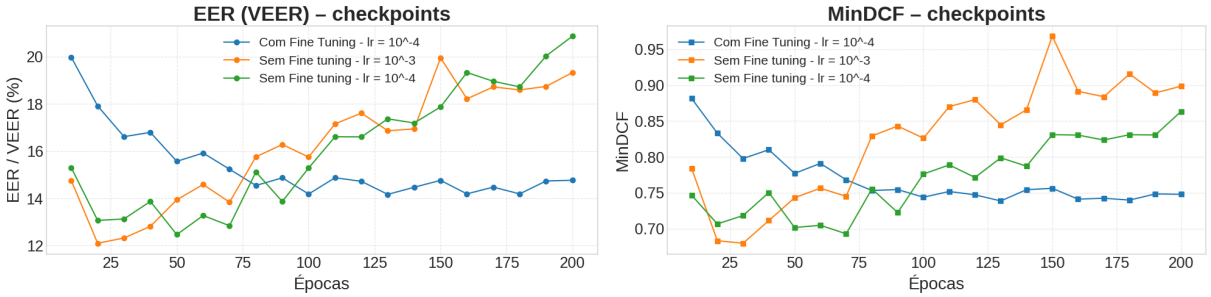


Fonte: Elaborado pelo autor.

Perda (TLOSS): As abordagens Sem *fine-tuning* (linhas verde e laranja) apresentam a convergência mais rápida e alcançam os valores de perda mais baixos. O resultado é compreensível, pois o modelo é treinado no VoxCeleb1, um modelo relativamente pequeno. A abordagem Com *fine-tuning* com $lr = 10^{-4}$ (linha Azul) também converge, mas de forma mais lenta do que as versões Sem *fine-tuning*.

Acurácia: Todas as três curvas de acurácia com $lr = 10^{-4}$ (Com *fine-tuning* 10^{-4}, Sem *fine-tuning* 10^{-4}) demonstram uma boa acurácia de treinamento, todas convergindo para valores acima de $80\,\%$ no final das 200 épocas. A abordagem Com *fine-tuning* (linha azul) apresenta a acurácia mais baixa durante o treinamento, reforçando que é possível extrair mais desempenho ao treinar por mais épocas.

Figura 26 – Teste *fine-tuning*



Fonte: Elaborado pelo autor.

A Figura 26 mostra o desempenho dos modelos com as principais métricas de teste (EER e MinDCF) em *checkpoints*.

EER (VEER): Nesta métrica, o desempenho superior é claramente obtido pelo modelo Com *fine-tuning*, com $lr = 10^{-4}$ (linha azul). Este modelo atinge e mantém o

menor EER, em torno de $15\,\%$, indicando a melhor generalização e menor taxa de erro na verificação de locutor. O modelo Sem *fine-tuning* com $lr = 10^{-4}$ e $lr = 10^{-3}$ (linha verde e laranja) demonstra um desempenho mais volátil e o pior EER, acima de $18\,\%$ após 200 épocas, o que sugere modelos instáveis no teste.

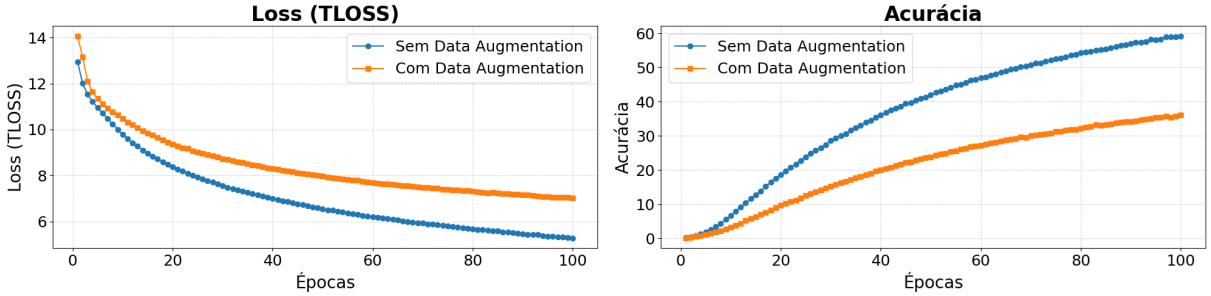
MinDCF: O padrão se repete na métrica MinDCF, que é crucial para sistemas de verificação. O modelo Com *fine-tuning* com $lr = 10^{-4}$ (linha azul) apresenta os valores de MinDCF consistentemente melhores, com a estabilização em torno de 0.75. Os outros dois modelos, Sem *fine-tuning* com LR de 10^{-4} e 10^{-3}, resultam em MinDCF mais altos e menos estáveis após 200 épocas.

5.1.2 Experimento 2: Treinamento com *data augmentation*

O segundo experimento foi realizado com o intuito de verificar o efeito da aplicação de técnicas de *data augmentation* (Aumento de Dados) para aumentar a robustez do modelo a ruídos e reverberações, simulando condições ambientais mais realistas. O aumento de dados utilizou ruídos das bases MUSAN e RIRS, em que cada tipo de perturbação teve $25\,\%$ de chance de ser adicionado aos áudios treinados, sendo: reverberações, músicas, falas e ruídos. As configurações utilizadas na experimentação foram com 1 resolução de 30 ms e learning rate de 10^{-4} em cada caso, com a diferença de aproximadamente 1 dia a mais de treinamento no modelo em que os dados passaram pelo processo de *Augmentation*.

A Figura 27 compara as curvas de Perda (Loss - TLOSS) e Acurácia do treinamento com e sem *data augmentation* ao longo de 100 épocas.

Figura 27 – *Treino Augmentation*



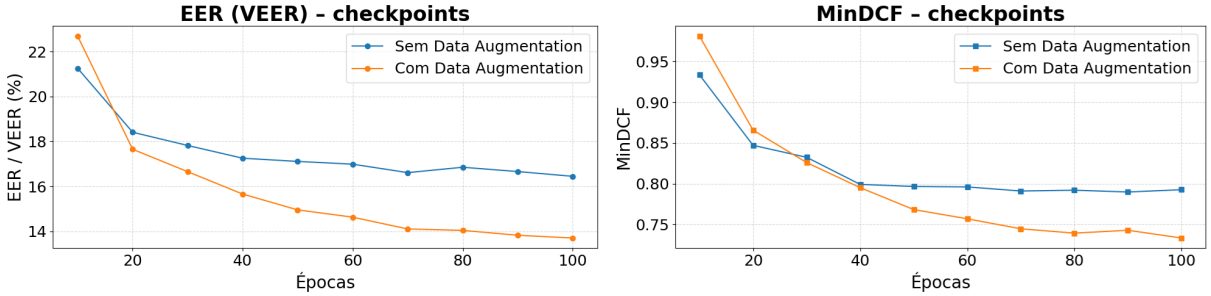
Fonte: Elaborado pelo autor.

Perda (TLOSS): A curva de perda do modelo treinado Com *data augmentation* (laranja) apresenta valores mais altos e uma convergência mais lenta do que o modelo treinado Sem *data augmentation* (azul). Este comportamento é esperado, pois ao incluir ruídos e perturbações, o conjunto de treinamento torna-se mais diverso, o que dificulta o ajuste da rede aos dados.

Acurácia: De forma igual à perda, a acurácia do treinamento Com *data augmentation* (laranja) também foi inferior. Enquanto o modelo Sem *data augmentation* obteve uma acurácia próxima de $60\,\%$ nas 100 épocas, o modelo aumentado estabiliza-se em um valor significativamente menor, próximo de $35\,\%$. Esta diferença mostra que o modelo está sendo forçado a aprender características mais robustas e menos específicas, com consequência de redução do desempenho na tarefa de treinamento em favor da generalização.

A Figura 28 mostra as principais métricas de desempenho, especificamente o EER (VEER) e o MinDCF, avaliados em intervalos de 10 épocas.

Figura 28 – Teste *Augmentation*



Fonte: Elaborado pelo autor.

EER (VEER): O modelo treinado Com *data augmentation* (laranja) exibe uma melhoria no desempenho. O EER da versão aumentada converge para valores consistentemente mais baixos para em torno de $14\,\%$, enquanto a versão Sem *data augmentation* (azul) se mantém em torno de $16\,\%$. O menor EER indica que o modelo aumentado é mais robusto e mais eficaz na avaliação entre locutores no conjunto de teste, que geralmente contém dados mais variáveis e ruidosos do que o conjunto de treinamento original.

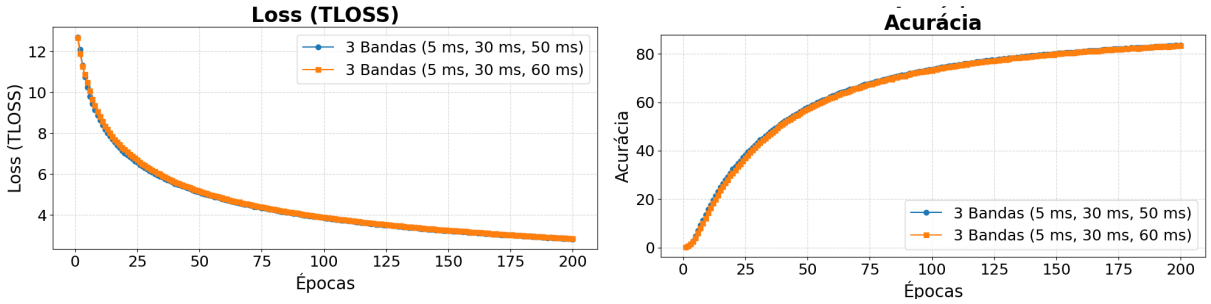
MinDCF: O melhor desempenho da versão aumentada no EER é replicado na métrica MinDCF. O modelo Com *data augmentation* (laranja) atinge valores de MinDCF menores, abaixo de 0.75, comparado ao modelo Sem *data augmentation* (azul), que se mantém próximo aos 0.80.

5.1.3 Experimento 3: Comparação da janela Narrowerband de 50 ms e 60 ms

O terceiro experimento teve como objetivo comparar o impacto de diferentes tamanhos de janela Narrowband (50 ms vs 60 ms) no desempenho do modelo de 3 resoluções (5 ms, 30 ms, X ms), em que X assume os valores de 50 ms e 60 ms. A ideia é verificar qual tamanho de janela oferece o melhor equilíbrio entre captura de informação espectral e estabilidade na classificação.

A Figura 29 apresenta as curvas de Perda (Loss - TLOSS) e Acurácia obtidas durante o treinamento ao longo de 200 épocas.

Figura 29 – Treino Comparação de Narrowerband de 50 ms e 60 ms



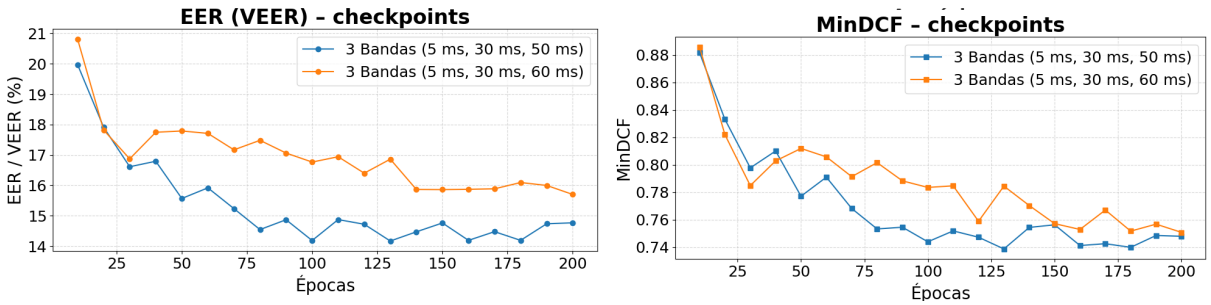
Fonte: Elaborado pelo autor.

Perda (TLOSS): Ambas as configurações (50 ms e 60 ms) apresentaram um comportamento convergente e muito semelhante. A perda diminuiu de forma constante, o que indica que o modelo aprendeu de maneira estável nas duas configurações. Isso sugere que a diferença no tamanho da janela Narrowband (de 50 ms para 60 ms) não teve um impacto significativo na taxa ou estabilidade da convergência de treinamento.

Acurácia: As curvas de acurácia também apresentaram bastante semelhança, ao atingir um valor final muito similar próximo de $80\,\%$. Este resultado obtido durante o treinamento revela que a diferença de 10 ms na janela Narrowband não produziu variações ao classificar corretamente os dados de treino.

A Figura 30 mostra as principais métricas de desempenho, especificamente o EER (VEER) e o MinDCF, avaliados em intervalos de 10 épocas.

Figura 30 – Teste Comparação de Narrowerband de 50 ms e 60 ms



Fonte: Elaborado pelo autor.

EER (VEER): Nesta métrica, o modelo com janela Narrowband de 50 ms (linha azul) demonstrou um desempenho superior com valores mais baixos. A curva de 50 ms estabilizou abaixo de $15\,\%$, enquanto a de 60 ms (linha laranja) permaneceu mais alta, com variação em torno dos $16\,\%$. Isso indica que a janela de 50 ms é mais

eficaz na minimização da taxa de erro, alcançando um melhor equilíbrio entre falsas aceitações e falsas rejeições.

MinDCF: Semelhantemente ao EER, a configuração de 50 ms (linha azul) também apresentou valores de MinDCF menores ao longo de quase todo o treinamento de teste. O desempenho do modelo com janela de 60 ms (linha laranja) se manteve pior na fase de teste, o que sugere que a janela de 50 ms captura características mais robustas para a tarefa de verificação de locutor.

5.1.4 Experimento 4: Comparação Entre Tamanhos de Janela e Deslocamento

O quarto experimento teve como objetivo analisar o desempenho de diferentes combinações de tamanhos de janela e deslocamentos (hop) no modelo PANN. Foram comparadas três configurações:

a) Janelas de 5 ms, 30 ms e 50 ms com deslocamento fixo de 6,25 ms;

b) Janelas de 8 ms, 16 ms e 32 ms com deslocamento fixo de 6,25 ms;

c) Janelas de 8 ms, 16 ms e 32 ms com deslocamentos proporcionais aos respectivos tamanhos das janelas.

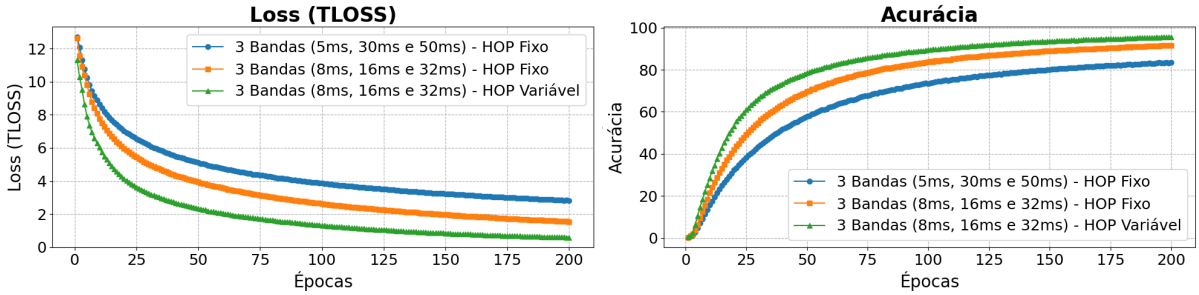
O objetivo foi avaliar se a utilização de um deslocamento proporcional às janelas resulta em uma representação temporal mais adequada e em melhorias no desempenho da verificação de locutor.

Os deslocamentos foram obtidos mantendo a mesma razão entre janela e deslocamento utilizada no trabalho de Virgilli et al. (2024), que utiliza uma janela de 30 ms e deslocamento de 6,25 ms. Logo, foi preservada a razão de 4,8:1 entre janela e deslocamento, da qual foram calculados os deslocamentos de 1,6666 ms, 3,3333 ms e 6,6666 ms para as janelas de 8 ms, 16 ms e 32 ms, respectivamente.

A Figura 31 apresenta as curvas de Perda (Loss - TLOSS) e Acurácia obtidas durante o treinamento ao longo de 200 épocas.

Perda (TLOSS): As três configurações apresentaram um comportamento com diferentes velocidades de convergência. A configuração que utiliza janelas de 8 ms, 16 ms e 32 ms com deslocamentos proporcionais apresentou a redução mais rápida da perda e atingiu os menores valores finais de TLOSS. A configuração com as mesmas janelas e deslocamento fixo de 6,25 ms também superou a configuração composta pelas janelas de 5 ms, 30 ms e 50 ms, o que indica que a combinação de janelas menores, principalmente quando associada a deslocamentos proporcionais, favorece a convergência do modelo.

Figura 31 – Treino Comparação Entre Tamanho de Janela e Deslocamento

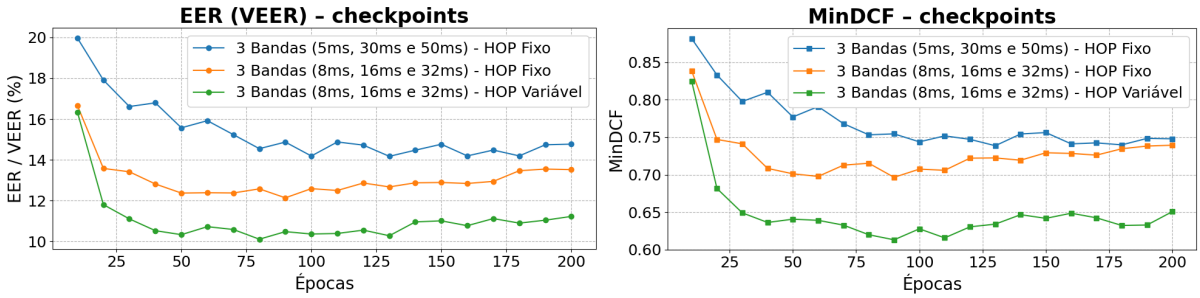


Fonte: Elaborado pelo autor.

Acurácia: As curvas de acurácia apresentaram diferenças entre as três configurações, sendo que a configuração com janelas de 8 ms, 16 ms e 32 ms e deslocamentos proporcionais atingiu o maior valor final, próximo de $95\,\%$. A configuração com as mesmas janelas e deslocamento fixo também apresentou desempenho superior à configuração de 5 ms, 30 ms e 50 ms. Esse resultado indica que, a utilização de janelas menores com deslocamentos proporcionais melhora a capacidade do modelo de classificar os dados de treinamento.

A Figura 32 mostra as principais métricas de desempenho, especificamente o EER (VEER) e o MinDCF, avaliados em intervalos de 10 épocas.

Figura 32 – Teste Comparação Entre Tamanho de Janela e Deslocamento



Fonte: Elaborado pelo autor.

EER (VEER): Nesta métrica, a configuração com janelas de 8 ms, 16 ms e 32 ms e o deslocamentos proporcionais(linha verde) apresentaram os menores valores de EER ao longo do teste, com valores mínimos próximo de $10\,\%$. A configuração com as mesmas janelas e deslocamento fixo(linha laranja) apresentou desempenho mediano, enquanto a configuração de 5 ms, 30 ms e 50 ms(linha azul) obteve os maiores valores de EER. Esse resultado indica que a utilização dos deslocamentos proporcionais contibuí para um melhor equilíbrio entre falsas aceitações e falsas rejeições.

MinDCF: Semelhante ao EER, a configuração com janelas de 8 ms, 16 ms e 32 ms e deslocamentos proporcionais apresentou os menores valores de MinDCF

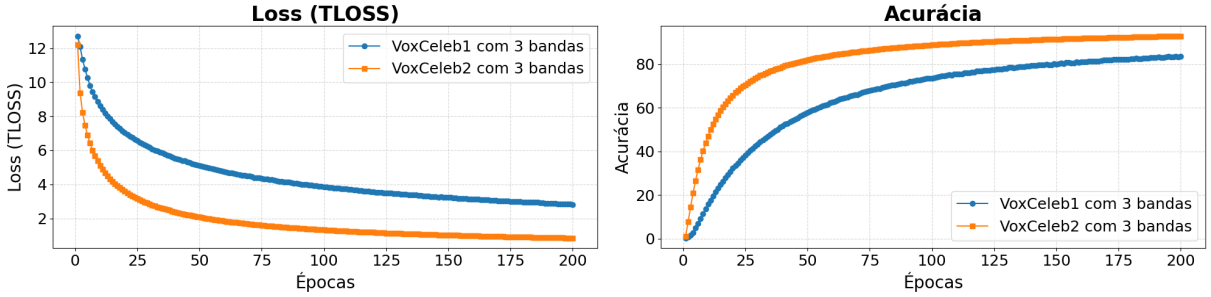
durante praticamente todo o treinamento de teste. A configuração com deslocamento fixo apresentou desempenho intermediário, e a configuração de 5 ms, 30 ms e 50 ms registrou os maiores valores ao longo das épocas. Esses resultados reforçam que a utilização de janelas menores em conjunto com deslocamentos proporcionais demonstra uma representação mais robusta para a tarefa de verificação de locutor.

5.1.5 Experimento 5: VoxCeleb1 vs VoxCeleb2

O quinto experimento avaliou o desempenho no uso de diferentes conjuntos de dados de treinamento, comparando o VoxCeleb1 e o VoxCeleb2, em que ambos utilizaram a arquitetura otimizada de 3 resoluções (5 ms, 30 ms, 50 ms). O objetivo é verificar se o volume de dados maior e a diversidade do VoxCeleb2 geram um modelo com maior capacidade de generalização para a tarefa de verificação de locutor.

A Figura 33 apresenta as curvas de Perda (Loss - TLOSS) e Acurácia obtidas durante o treinamento ao longo de 200 épocas.

Figura 33 – Treino VoxCeleb1 vs VoxCeleb2



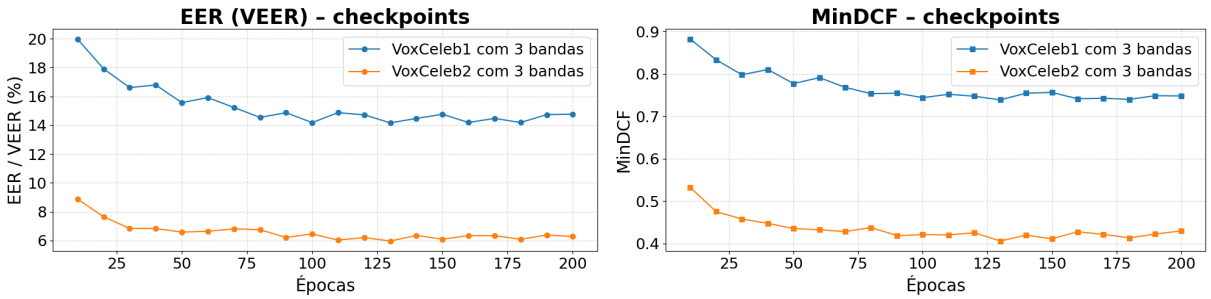
Fonte: Elaborado pelo autor.

Perda (TLOSS): A curva de perda do modelo treinado no VoxCeleb2 (laranja) converge para um valor final menor e de forma mais rápida do que a curva do modelo treinado no VoxCeleb1 (azul). Os gráficos indicam que o modelo treinado com o VoxCeleb2 consegue modelar os dados de treinamento com maior precisão e confiança.

Acurácia: A curva de acurácia do modelo treinado no VoxCeleb2 (laranja) também demonstra ser superior ao VoxCeleb1 (azul), alcançando valores acima de $90\,\%$, enquanto o modelo VoxCeleb1 estabiliza próximo aos $80\,\%$. A acurácia elevada sugere que o conjunto de dados VoxCeleb2, devido ao seu maior volume de dados, permite ao modelo aprender características mais robustas e específicas para a classificação durante a fase de treinamento.

A Figura 34 mostra as principais métricas de desempenho, especificamente o EER (VEER) e o MinDCF, avaliados em intervalos de 10 épocas.

Figura 34 – Teste VoxCeleb1 vs VoxCeleb2



Fonte: Elaborado pelo autor.

EER (VEER): O modelo treinado no VoxCeleb2 (laranja) apresenta um desempenho expressivamente superior, em que o EER se estabiliza em torno de $6\,\%$. Por outro lado, o modelo treinado no VoxCeleb1 (azul) tem um EER que se mantém em patamares muito mais altos, próximo de $15\,\%$. Esta diferença evidencia a grande vantagem do VoxCeleb2 em produzir um modelo com menor taxa de erro na fase de teste, indicando uma generalização muito mais eficaz.

MinDCF: A superioridade é replicada na métrica MinDCF. O modelo treinado no VoxCeleb2 (laranja) alcança um MinDCF baixo, em torno de 0.4 a 0.5, enquanto o modelo VoxCeleb1 (azul) permanece com valores mais altos, entre 0.7 e 0.8. O resultado do MinDCF favorece claramente o treinamento com o VoxCeleb2, o que evidencia melhores resultados na captura de características mais robustas.

5.2 EXPERIMENTOS PRINCIPAIS

Com base nos resultados obtidos na seção anterior, adotou-se a configuração da base de dados VoxCeleb1, devido ao tempo de treino e testes elevados do VoxCeleb2 e alto consumo computacional, com três resoluções espectrais com janelas de 8 ms, 16 ms e 32 ms e a utilização de técnicas de *data augmentation*, por apresentar o melhor desempenho entre as configurações avaliadas. A partir dessa configuração, os experimentos desta seção têm como objetivo avaliar o impacto das multirresoluções em arquiteturas consolidadas e verificar o desempenho do modelo construído para verificação de locutor.

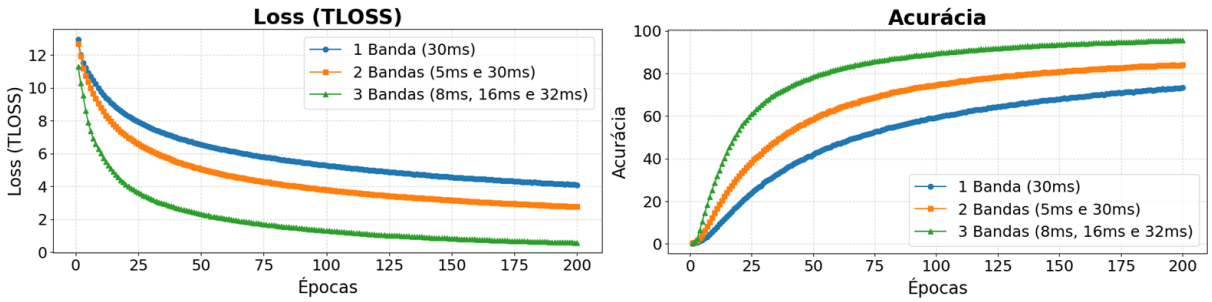
5.2.1 Experimento 1: Estudo da Entrada Multirresolução na Rede PANN (CNN14)

O primeiro experimento principal teve como foco investigar a influência do número e da configuração das resoluções de entrada multirresolução no desempenho do modelo, mantendo a taxa de aprendizado em 10^{-4}. Foram testadas três

configurações, que utilizam como referência os resultados de Virgilli et al. (2024), que demonstraram que a configuração de resolução única com janela de 30 ms apresentou o melhor desempenho dentre todas as configurações de resolução única avaliadas: 1 resolução (30 ms), 2 resoluções (5 ms e 30 ms) e 3 resoluções (8 ms, 16 ms e 32 ms), em relação ao salto de janelamento, a configuração de 3 resoluções foi configurada com aproximadamente 1,66 ms, 3,33 ms e 6,66 ms, respectivamente. O objetivo é determinar se as três resoluções interferem positivamente para adquirir informações do contexto com janela curta, média e longa; por fim, determinar se houve uma melhora no desempenho do sistema.

A Figura 35 compara as curvas de Perda (Loss - TLOSS) e Acurácia obtidas no treinamento ao longo de 200 épocas para as três configurações de resolução.

Figura 35 – Treino Três Bandas



Fonte: Elaborado pelo autor.

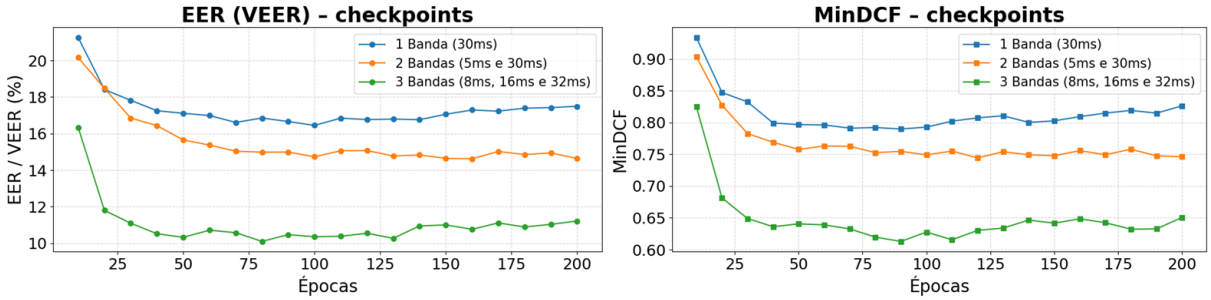
Perda (TLOSS): A configuração de 3 resoluções (verde) apresentou o maior destaque em relação a perda comparado com as outras resoluções, com a convergência de perda mais rápida e alcançou o valores de perda final mais baixo, seguido pela configuração de 2 resoluções (laranja), que obteve o desempenho intermediário no teste. A configuração de 1 resolução (azul) foi mais lenta e atingiu o valor de perda mais alto, o que sugere que uma única resolução temporal não é suficiente para otimizar eficientemente os parâmetros do modelo.

Acurácia: A curva de 3 resoluções (verde) alcançou a maior acurácia de treinamento, com valor superior a $95\,\%$, em seguida, 2 resoluções (laranja) ficaram próximo a $80\,\%$. A acurácia do modelo de 1 resolução (azul) foi a mais baixa e a mais lenta para convergir, o que reforça a ideia de que a utilização de multirresolução facilita o aprendizado durante o treinamento.

A Figura 36 mostra os resultados no conjunto de teste para as métricas EER e MinDCF para o experimento multirresolução.

EER (VEER): O modelo 3 resoluções (verde) conseguiu alcançar os menores valores de EER, em torno de $10\,\%$ nos menores patamares, sendo consistentemente inferior ao modelo de 1 resolução e de 2 resoluções. O modelo de 2 resoluções

Figura 36 – Teste Três Bandas



Fonte: Elaborado pelo autor.

(laranja) também apresentou uma melhoria significativa em relação à 1 resolução, estabilizando-se ligeiramente acima do modelo de 3 resoluções na média. O modelo de 1 resolução (azul) teve o pior EER, acima de $17\,\%$, o que indica o melhor desempenho com multirresoluções para verificar características de forma mais robusta.

MinDCF: O modelo 3 resoluções (verde) também obteve o melhor MinDCF, convergindo para o valor mais baixo e estável, abaixo de 0.65. A configuração de 2 resoluções (laranja) fica em segundo lugar e o pior desempenho, que segue a tendência do EER, foi do modelo que utiliza 1 resolução apenas.

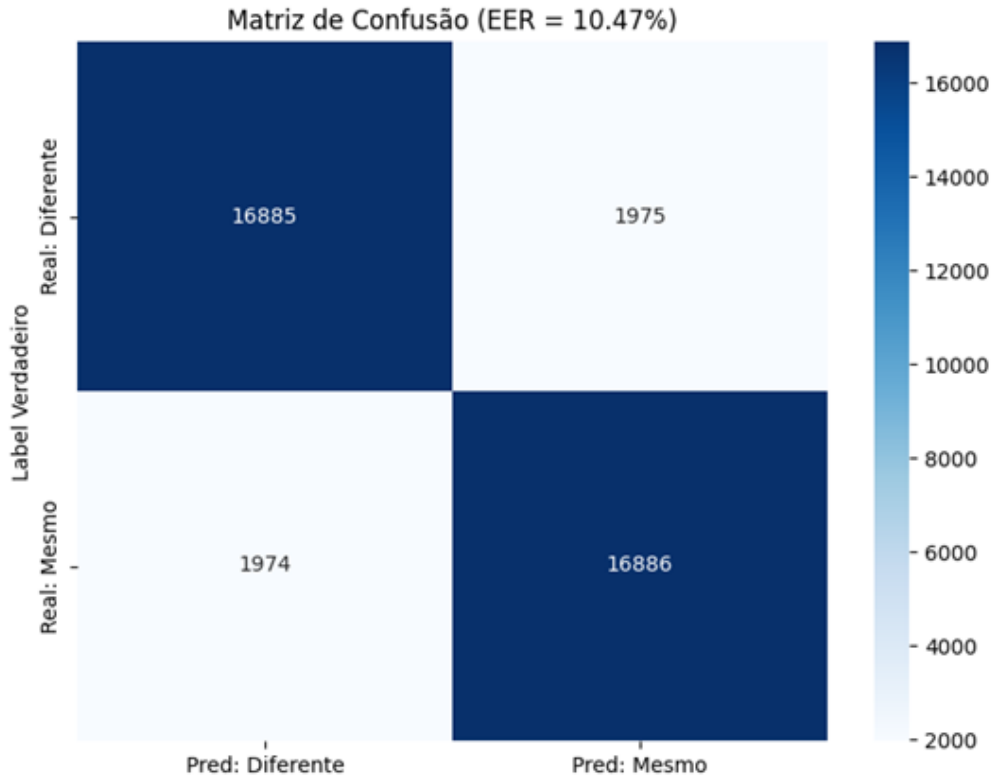
Além dos gráficos de treino e teste, é possível observar a matriz de confusão do modelo de três resoluções em um dos melhores valores de ERR, apresentado na Figura 37.

5.2.2 Experimento 2: Estudo da Entrada Multirresolução no ECAPA-TDNN

O segundo experimento principal não apresenta ganhos significativos na aplicação de técnicas multirresolução temporal na extração de características para o modelo ECAPA-TDNN. Devido à codificação da arquitetura, o processamento convolucional do modelo é realizado de forma 1D ao longo do eixo temporal, que não explora as vantagens de janelas com diferentes espaços de tempo. Para avaliar a hipótese, comparou-se a entrada com uma única janela padrão de 25 ms com a entrada de três janelas de 8 ms, 16 ms e 32 ms. Os resultados evidenciam a falta de ganhos de desempenho ao aumentar a quantidade de janelas na entrada do algoritmo ECAPA-TDNN, que é um dos precursores do atual estado da arte em relação à verificação de locutor. Este comportamento pode estar relacionado à natureza 1D do processamento do ECAPA-TDNN que difere da abordagem de 2D baseada em espectrogramas.

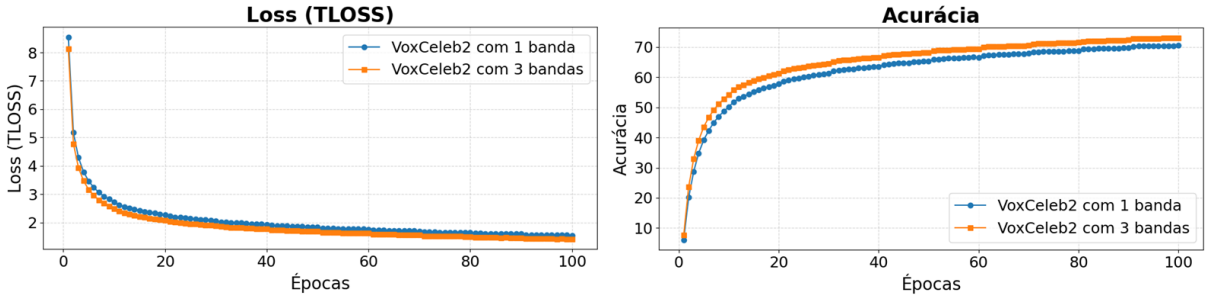
A Figura 38 compara as curvas de Perda (Loss - TLOSS) e Acurácia obtidas no treinamento ao longo de 100 épocas para uma e três entradas de resolução.

Figura 37 – Matriz de confusão do modelo com três resoluções



Fonte: Elaborado pelo autor.

Figura 38 – Treino ECAPA-TDNN 1 e 3 resoluções



Fonte: Elaborado pelo autor.

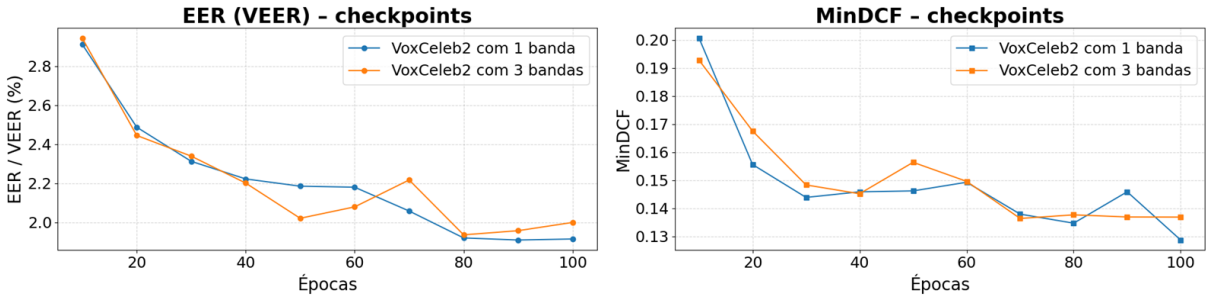
Perda (TLOSS): A configuração de 3 resoluções (laranja) se comportou de forma semelhante a configuração de 1 resolução (azul) com uma leve desvantagem para a variante de uma resolução que atingiu o valor de perda maior, o que sugere uma leve vantagem para a configuração com 3 resoluções.

Acurácia: Segue o mesmo padrão entre os dois modelos observados, em que a acurácia maior foi obtida pela versão com 3 resoluções (laranja) e a acurácia menor

pela versão de 1 resolução (azul).

A Figura 39 mostra o resultado obtido com as métricas EER e MinDCF para a experimentação multirresolução na rede ECAPA-TDNN.

Figura 39 – Teste ECAPA-TDNN 1 e 3 resoluções



Fonte: Elaborado pelo autor.

EER (VEER): O modelo 3 resoluções (laranja) alcançou patamares semelhantes ao modelo de 1 resolução (azul) em EER, abaixo de $2\,\%$. O modelo de 1 resolução (azul) teve o menor EER registrado, acima de $17\,\%$, o que indica o desempenho similar com resolução única e multirresoluções para verificar características.

MinDCF: O modelo 3 resoluções (laranja) também obteve o desempenho semelhante do modelo de 1 resolução (azul) MinDCF, ambos convergiram para o valores abaixo de 0.14. O que representa a falta de ganho ao introduzir os métodos multirresolução.

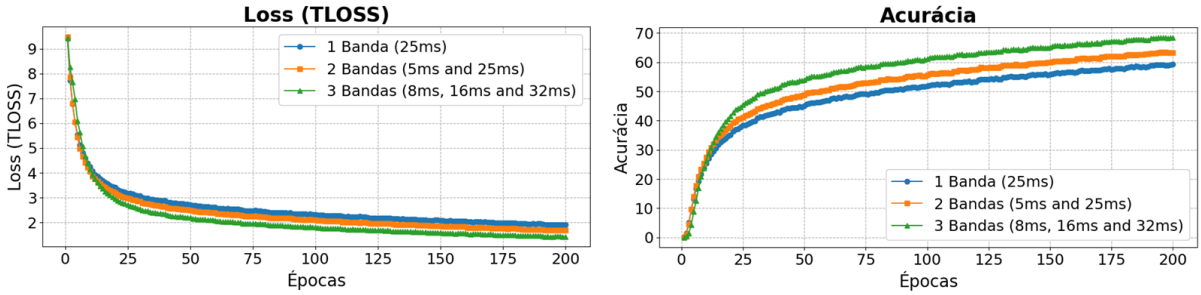
5.2.3 Experimento 3: Estudo da Entrada Multirresolução no ECAPA2

O terceiro experimento principal segue o mesmo raciocínio do primeiro experimento principal, mas com a diferença de utilizar o modelo ECAPA2 ao invés da PANN e o único experimento a utilizar a função de perda *AAM Softmax*, com margem angular de 0,3 e fator de escala de 15. A taxa de aprendizado manteve-se em 10^{-4}, e foram utilizadas as três seguintes configurações: 1 resolução (25 ms), 2 resoluções (5 ms e 25 ms) e 3 resoluções (8 ms, 16 ms e 32 ms), no salto de janelamento, a configuração de 3 resoluções manteve-se em 1,66 ms, 3,33 ms e 6,66 ms, respectivamente. O objetivo é verificar se as três resoluções serão mais eficientes, assim como no primeiro experimento; assim, determinar se houve melhora independente do modelo utilizado.

A Figura 40 compara as curvas de Perda (Loss - TLOSS) e Acurácia obtidas no treinamento ao longo de 200 épocas para as três configurações de janelas no modelo ECAPA2.

Perda (TLOSS): A configuração de 3 resoluções (verde) apresentou o maior destaque em relação a perda comparado com as outras resoluções, com a con-

Figura 40 – Treino ECAPA2 com 1, 2 e 3 resoluções



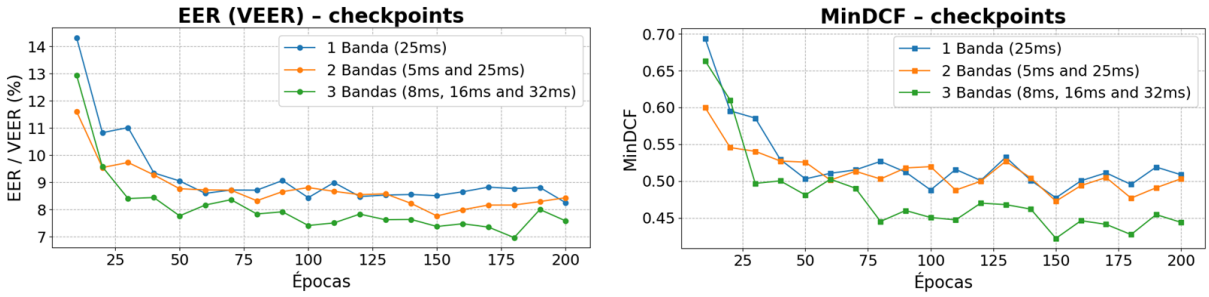
Fonte: Elaborado pelo autor.

vergência de perda mais rápida e alcançou o valores de perda final mais baixo, seguido pela configuração de 2 resoluções (laranja), que obteve o desempenho intermediário no teste. A configuração de 1 resolução (azul) foi mais lenta e atingiu o valor de perda mais alto, o que sugere que uma única resolução temporal não é suficiente para otimizar eficientemente os parâmetros do modelo.

Acurácia: A curva de 3 resoluções (verde) alcançou a maior acurácia de treinamento, com valor superior a $95\,\%$, em seguida, 2 resoluções (laranja) ficaram próximo a $80\,\%$. A acurácia do modelo de 1 resolução (azul) foi a mais baixa e a mais lenta para convergir, o que reforça a ideia de que a utilização de multirresolução facilita o aprendizado durante o treinamento.

A Figura 41 mostra os resultados no conjunto de teste para as métricas EER e MinDCF para o experimento multirresolução.

Figura 41 – Teste ECAPA2 com 1, 2 e 3 resoluções



Fonte: Elaborado pelo autor.

EER (VEER): O modelo 3 resoluções (verde) conseguiu alcançar os menores valores de EER, em torno de $10\,\%$ nos menores patamares, sendo consistentemente inferior ao modelo de 1 resolução e de 2 resoluções. O modelo de 2 resoluções (laranja) também apresentou uma melhoria significativa em relação à 1 resolução, estabilizando-se ligeiramente acima do modelo de 3 resoluções na média. O modelo de 1 resolução (azul) teve o pior EER, acima de $17\,\%$, o que indica o melhor desem-

penho com multirresoluções para verificar características de forma mais robusta.

MinDCF: O modelo 3 resoluções (verde) também obteve o melhor MinDCF, convergindo para o valor mais baixo e estável, abaixo de 0.65. A configuração de 2 resoluções (laranja) fica em segundo lugar e o pior desempenho, que segue a tendência do EER, foi do modelo que utiliza 1 resolução apenas.

5.3 RESULTADOS E ANÁLISE

No primeiro experimento, foram demonstradas que técnicas de *fine-tuning* são superiores ao treinamento do zero, pois o aprendizado se manteve constante, e após 100 épocas, o modelo estabilizou, diferentemente dos modelos sem *fine-tuning*, que se tornaram instáveis com desempenho bom apenas nas primeiras épocas, mas logo após começaram a sofrer com *overfitting*. Logo, na rede PANN (CNN14), as vantagens de ser uma rede pré-treinada prevalecem ao realizar a comparação com o treinamento do zero. Esta estratégia é singular ao uso da PANN, então não foi utilizada nos experimentos subsequentes, pois o *VoxCeleb2* permite um treinamento eficiente sem a necessidade de *fine-tuning*, devido à sua grande quantidade de dados.

No segundo experimento, a verificação de locutor com a utilização de *data augmentation* se mostra superior. Embora as perturbações no áudio tenham elevado a perda de treinamento e diminuído a acurácia de treinamento, ela resultou em um modelo com uma capacidade de generalização relativamente melhor. A redução consistente e significativa tanto no EER quanto no MinDCF no conjunto de teste prova que a exposição a dados ruidosos durante o treinamento torna o modelo mais robusto para variações do mundo real. Entretanto, a rede demora cerca de 12 vezes mais tempo para realizar o treinamento até 100 épocas. Esta técnica não foi utilizada por Virgilli et al. (2024), mas foi utilizada nos experimentos futuros deste trabalho.

No terceiro experimento, embora as curvas de treinamento tenham sido muito parecidas para as duas configurações, os resultados no conjunto de teste revelaram uma vantagem da janela Narrowband de 50 ms em relação à de 60 ms. A configuração com 50 ms atingiu taxas de EER e MinDCF inferiores, o que indica uma melhor capacidade de generalização e um desempenho superior no cenário de verificação de locutor. Desta forma, o modelo de 3 resoluções (8 ms, 16 ms, 32 ms) do terceiro experimento foi a configuração preferencial para experimentos subsequentes.

No quarto experimento, os resultados demonstraram que a utilização de janelas menores, combinadas com deslocamentos proporcionais ao tamanho das janelas, obteve o melhor desempenho em comparação com as configurações testadas. A configuração de 3 resoluções (8 ms, 16 ms e 32 ms) com hop variável apresentou uma menor perda durante o treinamento, maior acurácia e os menores valores de EER e

MinDCF no conjunto de teste, o que indica uma melhor capacidade de generalização e maior desempenho em realizar a verificação do locutor. Portanto, a configuração de 3 resoluções (8 ms, 16 ms e 32 ms) com deslocamentos proporcionais foi escolhida como configuração dos experimentos principais.

A quinta experimentação demonstrou que o conjunto de dados utilizado para o treinamento tem um grande impacto no desempenho final do modelo. O treinamento com o VoxCeleb2 resultou em um modelo com convergência de perda superior, acurácia de treinamento muito mais alta e um desempenho de teste radicalmente melhor em termos de EER, com redução de $15\,\%$ para $6\,\%$. Essa melhoria é atribuída ao maior volume de dados e à diversidade de locutores e ambientes do VoxCeleb2, que fornecem ao modelo uma capacidade de generalização muito mais robusta em comparação com o VoxCeleb1. Com este fato, o conjunto de dados VoxCeleb2 foi utilizado como base para as avaliações de desempenho.

Ao realizar a análise dos experimentos principais, apresentam-se os seguintes resultados complementares do estudo:

O primeiro experimento principal confirma o benefício da utilização de entradas multirresolução. A adição da janela curta (8 ms) e, principalmente, da janela intermediária (16 ms) com a janela longa de (32 ms) resultou em uma convergência de treinamento superior e uma redução significativa do EER e do MinDCF no conjunto de teste. O modelo com 3 resoluções (8 ms, 16 ms e 32 ms) se mostrou mais eficaz em comparação com 2 resoluções (5 ms, 30 ms). Portanto, a configuração de 3 resoluções foi a arquitetura preferencial para os próximos experimentos com diferentes backbones. Também foi utilizado o conceito de processar cada espectrograma separadamente com seus respectivos saltos na proporção de 4,8, que em 8 ms foi utilizado 1,66 ms; 16 ms utilizou 3,33 ms e 32 ms com 6,66 ms. e por fim, juntá-los na terceira camada da rede.

O segundo experimento principal evidencia que a entrada multirresolução não é relevante em modelos de uma dimensão. Ao transformar o Ecapa-TDNN para receber até 3 entradas, a melhoria é ignorada e não é apresentada uma melhora significativa entre uma única janela padrão de 25 ms comparado a entrada de três janelas de 8 ms, 16 ms e 32 ms.

No terceiro experimento principal, a entrada multirresolução obteve um desempenho superior as entradas de 1 resolução e 2 resoluções. Assim como representado no primeiro experimento principal, a sequência de desempenho aumenta perante a adição de novas resoluções. O pior desempenho foi registrado com 1 resolução (25 ms), o desempenho intermediário foi registrado ao utilizar 2 resoluções (5 ms e 25 ms) e o melhor desempenho foi alcançado com 3 resoluções (8 ms, 16 ms e 32 ms). A hipótese apresentada de melhor desempenho ao utilizar multirresolução é validada na

arquitetura ECAPA2, com saltos na proporção de 4,8, que em 8 ms foi utilizado 1,66 ms; 16 ms utilizou 3,33 ms e 32 ms com 6,66 ms de tempo de salto.

Um fator determinante para os resultados obtidos está relacionado ao viés indutivo das arquiteturas avaliadas. A rede PANN (CNN14), por ser baseada em convoluções bidimensionais (2D), processa o espectrograma de forma parecida com uma imagem, o que favorece o aproveitamento das múltiplas resoluções tempo-frequência fornecidas pela entrada multirresolução. Entretanto, a arquitetura ECAPA-TDNN possui um viés indutivo orientado ao processamento sequencial unidimensional (1D). Ao fundir as características, a informação pode não ser igualmente explorada pela arquitetura, o que pode explicar a ausência de ganhos significativos na comparação entre três resoluções. Por fim, apesar da herança TDNN, o ECAPA2 incorpora mecanismos de atenção que permitem uma exploração mais abrangente das representações. Esse comportamento pode ter contribuído para que o modelo obtivesse ganhos significativos na métrica EER em comparação com o ECAPA-TDNN.

5.4 CONSIDERAÇÕES FINAIS

Neste capítulo, pode-se observar os resultados preliminares dos experimentos e concluir que, apesar da base VoxCeleb1 conter um número razoável de trechos de áudio e locutores, os resultados demonstram que é necessária uma base maior para obter resultados competitivos. Por este motivo, os próximos testes serão realizados na base VoxCeleb2.

Em comparação à utilização de três resoluções de entrada, houve um desempenho superior ao comparar com a utilização de uma e duas resoluções. Os testes iniciais foram limitados à rede PANN (CNN14) para um maior controle de quais fatores seriam cruciais para melhores resultados, ao passo que os testes finais utilizaram o backbone no estado da arte (ECAPA2) para medir e comprovar a efetividade da adição das janelas, que demonstrou de forma semelhante à rede PANN, que 3 resoluções de entrada são superiores a 2 resoluções e 1 resolução.

6 CONCLUSÃO

Este trabalho apresentou a introdução ao tema abordado, a fundamentação teórica necessária para a compreensão do projeto, a revisão dos trabalhos relacionados ao trabalho proposto e a descrição da proposta com explicação em detalhes que serão seguidos na confecção da dissertação e os resultados preliminares.

Devido ao avanço da área de segurança, em especial nas subáreas voltadas ao reconhecimento biométrico por meio de características únicas do indivíduo, torna-se necessário o aprimoramento dessas técnicas, com o objetivo de alcançar a maior precisão e, consequentemente, tornar os sistemas mais seguros e resistentes a tentativas de invasão. Apesar dos avanços da biometria por voz, as soluções atuais ainda carecem de estudos em relação às definições de quais ferramentas, configurações e representações oferecem maior desempenho na tarefa de verificação do locutor.

Conclui-se que o presente trabalho atingiu os objetivos propostos ao investigar a extração das características intrínsecas da voz e a verificação do locutor por meio de testes em diferentes redes neurais e múltiplas janelas de tempo para análise do sinal de áudio. A utilização de três resoluções espectrais nas arquiteturas PANN e ECAPA2 demonstrou melhorias em comparação às configurações com uma e duas resoluções devido à estrutura de convolução 2D, o que evidencia o potencial da abordagem multirresolução para a representação de características discriminativas da voz. Na arquitetura ECAPA-TDNN não houve benefícios ao comparar três resoluções espectrais com uma resolução, possivelmente devido ao processamento convolucional 1D.

Os objetivos relacionados à avaliação de diferentes arquiteturas e à análise do impacto de diferentes resoluções espectrais foram atendidos, permitindo identificar configurações mais eficientes para a tarefa de verificação de locutor. Como contribuição científica, este trabalho apresenta uma análise da utilização de múltiplas resoluções espectrais em modelos de aprendizado profundo, o que contribuí para o desenvolvimento de abordagens mais robustas e eficientes.

6.1 LIMITAÇÕES DO ESTUDO

Apesar dos resultados obtidos, é importante reconhecer as limitações práticas e metodológicas desta pesquisa, que abrem caminhos para trabalhos futuros:

- **Desenho Experimental e Ablação:** O experimento que compara o desempenho

de duas bandas (5 ms e 25 ms) com três bandas (8 ms, 16 ms e 32 ms), houve dois passos de ablação avaliados conjuntamente: a alteração da quantidade de bandas e do tamanho das janelas temporais. Pode ser necessária uma ablação perfeitamente aninhada (ex: 32 ms; 8 ms e 32 ms; 8 ms, 16 ms e 32 ms) para validar a superioridade de forma definitiva.

- **Função de Perda:** Nos modelos CNN14 e ECAPA1, utilizou-se a *Softmax Prototypical Loss*, visando manter a consistência com o trabalho de Virgilli et al. (2024). Para o ECAPA2, foi utilizada a *AAM Softmax*. Embora a AAM Softmax seja uma função de perda utilizada no estado da arte, sua utilização apenas no ECAPA2 constitui uma limitação para a comparação direta entre as arquiteturas. Dessa forma, permanece a dúvida sobre em que medida os ganhos observados com a abordagem multirresolução são exclusivamente do número de bandas ou são influenciados também pela escolha da função de perda.

- **Custo e *Trade-off* Computacional:** Aumentar o número de bandas resulta em um *trade-off* direto com o custo computacional. A extração e fusão de múltiplos espectrogramas eleva o número de parâmetros nas camadas iniciais e aumenta o tempo de inferência para processar os trechos de áudio em cerca de três vezes ao utilizar três espectros. Embora exista ganho de desempenho em precisão, esse custo adicional deve ser considerado em aplicações no mundo real.

6.2 TRABALHOS FUTUROS

Embora os resultados obtidos tenham atendido ao que foi proposto nesta pesquisa, diversas possibilidades podem ser exploradas em trabalhos futuros com o intuito de ampliar a robustez da solução proposta.

Uma das possibilidades consiste na aplicação do método apresentado em novas arquiteturas de aprendizado profundo. Modelos mais recentes poderão ser comparados à abordagem proposta, com o objetivo de verificar e melhorar o desempenho, a precisão e o consumo computacional.

Outra linha a ser estudada consiste em comparar diferentes técnicas de pré-processamento e aumento de dados, como adição de ruídos e alteração na velocidade da fala. Essas técnicas podem aumentar a generalização do modelo em diferentes condições acústicas, o que aproxima a utilização em ambientes reais.

É sugerida a investigação de leis de escala aplicadas a modelos de verificação de locutor baseados em espectrogramas multiescala, por meio de uma análise que relacione o ganho de desempenho do sistema com a quantidade de dados do conjunto

de treinamento e o número de resoluções espectrais. Essa abordagem permitiria extrapolar os resultados, além de prever o comportamento e os custos computacionais.

Também é indicado como trabalho futuro uma investigação do uso de diferentes resoluções espectrais, em particular 8 ms, 16 ms e 32 ms, a fim de avaliar seus impactos no desempenho e no custo computacional do modelo. Além disso, recomenda-se a realização de testes estatísticos e o cálculo de intervalos de confiança para as métricas obtidas, o que permite verificar as diferenças observadas entre as configurações avaliadas e fornecer maior robustez às conclusões apresentadas.

Por fim, é recomendada a investigação da metodologia em aplicações reais de autenticação por voz, com a análise da escalabilidade, tempo de resposta, segurança e integração com outros sistemas. Essas investigações poderão contribuir para tornar a solução mais adequada para aplicações práticas.

6.3 DECLARAÇÃO DE USO DE IA GENERATIVA

Ferramentas de IA generativa foram utilizadas para auxiliar na formatação da dissertação e em pequenos ajustes linguísticos, com o objetivo de melhorar a legibilidade.

Nenhuma IA generativa foi utilizada para produzir parte significativa do conteúdo central ou dos resultados da dissertação. O autor assume total responsabilidade pelo trabalho apresentado e declara que todas as correções realizadas por IA foram revisadas por humanos.

Referências

AL-KARAWI, K. A.; AHMED, S. T. Model selection toward robustness speaker verification in reverberant conditions. *Multimedia Tools and Applications*, v. 80, n. 30, p. 36549–36566, Dec 2021. ISSN 1573-7721. Disponível em: https://doi.org/10.1007/s11042-021-11356-3.

ALZUBAIDI, L. et al. Review of deep learning: concepts, cnn architectures, challenges, applications, future directions. *Journal of Big Data*, v. 8, n. 1, p. 53, Mar 2021. ISSN 2196-1115. Disponível em: https://doi.org/10.1186/s40537-021-00444-8.

AMBAYE, G. A. Time and frequency domain analysis of signals: A review. *International Journal of Engineering Research & Technology (IJERT)*, v. 9, n. 12, December 2020. Disponível em: https://www.ijert.org/time-and-frequency-domain-analysis-of-signals-a-review.

ARDILA, R. et al. Common voice: A massively-multilingual speech corpus. In: CALZOLARI, N. et al. (Ed.). *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, 2020. p. 4218–4222. ISBN 979-10-95546-34-4. Disponível em: https://aclanthology.org/2020.lrec-1.520/.

ARIAS-VERGARA, T. et al. Multi-channel spectrograms for speech processing applications using deep learning methods. *Pattern Analysis and Applications*, v. 24, n. 2, p. 423–431, May 2021. ISSN 1433-755X. Disponível em: https://doi.org/10.1007/s10044-020-00921-5.

AUSTERLITZ, H. Chapter 4 - analog/digital conversions. In: AUSTERLITZ, H. (Ed.). *Data Acquisition Techniques Using PCs (Second Edition)*. Second edition. San Diego: Academic Press, 2003. p. 51–77. ISBN 978-0-12-068377-2. Disponível em: https://www.sciencedirect.com/science/article/pii/B9780120683772500048.

AYVAZ, U. et al. Automatic speaker recognition using mel-frequency cepstral coefficients through machine learning. *Computers, Materials & Continua*, v. 71, n. 3, 2022.

BÄCKSTRÖM, T. et al. *Introduction to Speech Processing*. 2. ed. Aalto University, 2022. Disponível em: https://speechprocessingbook.aalto.fi.

BAI, Z.; ZHANG, X.-L. Speaker recognition based on deep learning: An overview. *Neural Networks*, v. 140, p. 65–99, 2021. ISSN 0893-6080. Disponível em: https://www.sciencedirect.com/science/article/pii/S0893608021000848.

BASSI, P.; RAMPAZZO, W.; ATTUX, R. Redes neurais profundas triplet aplicadas à classificação de sinais em interfaces cérebro-computador. In: . [S.l.: s.n.], 2019.

BEVEREN, I.; SERGIDOU, E.; ZIABARI, S. S. M. Evaluating deep learning-based speaker verification systems: A comparative study across open-source and forensic datasets. In: *Proceedings of the 17th Multi-Disciplinary International Conference on Artificial Intelligence (MIWAI 2024)*. Thailand: [s.n.], 2024.

BOULAL, H. et al. Amazigh cnn speech recognition system based on mel spectrogram feature extraction method. *International Journal of Speech Technology*, v. 27, p. 287–296, 2024. Disponível em: https://doi.org/10.1007/s10772-024-10100-0.

BRACEWELL, R. N. *The Fourier Transform and Its Applications*. 3. ed. New York: McGraw-Hill, 2000.

CHEN, G. et al. Who is real bob? adversarial attacks on speaker recognition systems. In: *2021 IEEE Symposium on Security and Privacy (SP)*. [S.l.: s.n.], 2021. p. 694–711.

CHUNG, J. S.; NAGRANI, A.; ZISSERMAN, A. Voxceleb2: Deep speaker recognition. In: *Interspeech 2018*. [S.l.: s.n.], 2018. p. 1086–1090. ISSN 2958-1796.

CONSTANTINESCU, C.; BRAD, R. An overview on sound features in time and frequency domain. *International Journal of Advanced Statistics and IT&C for Economics and Life Sciences*, v. 13, n. 1, p. 45–58, 2023. Disponível em: https://doi.org/10.2478/ijasitels-2023-0006.

DENG, J. et al. Arcface: Additive angular margin loss for deep face recognition. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2019. p. 4685–4694.

DESPLANQUES, B.; THIENPONDT, J.; DEMUYNCK, K. ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification. In: *INTERSPEECH 2020*. [S.l.: s.n.], 2020. p. 3830–3834.

DIAS, F. F.; PONTI, M. A.; MINGHIM, R. Enhancing sound-based classification of birds and anurans with spectrogram representations and acoustic indices in neural network architectures. *Ecological Informatics*, v. 90, p. 103232, 2025. ISSN 1574-9541. Disponível em: https://www.sciencedirect.com/science/article/pii/S1574954125002419.

FINK, D. A new definition of noise: noise is unwanted and/or harmful sound. noise is the new 'secondhand smoke'. *Proceedings of Meetings on Acoustics*, v. 39, n. 1, p. 050002, 02 2020. ISSN 1939-800X. Disponível em: https://doi.org/10.1121/2.0001186.

FOURIER, J. *Théorie analytique de la chaleur*. Guillaume Koebner, 1883. Disponível em: https://books.google.com.br/books?id=JXtJAAAAYAAJ.

GAO, L.; TIAN, H.; LIU, K. A method for extracting black box models based on interpretable attention. *Expert Systems*, Wiley, v. 42, n. 7, 2025. Disponível em: https://onlinelibrary.wiley.com/doi/10.1111/exsy.70084.

GRAMI, A. Chapter 3 - signals, systems, and spectral analysis. In: GRAMI, A. (Ed.). *Introduction to Digital Communications*. Boston: Academic Press, 2016. p. 41–150. ISBN 978-0-12-407682-2. Disponível em: https://www.sciencedirect.com/science/article/pii/B978012407682200003X.

HAN, J.; KAMBER, M.; PEI, J. *Data Mining: Concepts and Techniques*. 3rd. ed. Waltham, MA: Morgan Kaufmann, 2011. (The Morgan Kaufmann Series in Data Management Systems). Published in 2012, Copyright ©2011 Elsevier Inc. All rights reserved. ISBN 978-0-12-381479-1.

HANIFA, R. M.; ISA, K.; MOHAMAD, S. A review on speaker recognition: Technology and challenges. *Computers & Electrical Engineering*, v. 90, p. 107005, 2021. ISSN 0045-7906. Disponível em: https://www.sciencedirect.com/science/article/pii/S0045790621000318.

HANSEN, J. H.; HASAN, T. Speaker recognition by machines and humans: A tutorial review. *IEEE Signal Processing Magazine*, v. 32, n. 6, p. 74–99, 2015.

International Commission on Biological Effects of Noise. *About ICBEN*. 2025. https://www.icben.org/About.html. Accessed: 2025-04-18.

JAKUBEC, M. et al. Deep speaker embeddings for speaker verification: Review and experimental comparison. *Engineering Applications of Artificial Intelligence*, v. 127, p. 107232, 2024. ISSN 0952-1976. Disponível em: https://www.sciencedirect.com/science/article/pii/S0952197623014161.

JOSHI, D.; PAREEK, J.; AMBATKAR, P. Comparative study of mfcc and mel spectrogram for raga classification using cnn. *Indian Journal of Science and Technology*, v. 16, n. 11, p. 816–822, 2023. Disponível em: https://doi.org/10.17485/IJST/v16i11.1809.

JUNEJA, K. Two-level noise robust and block featured pnn model for speaker recognition in real environment. *Wireless Personal Communications*, v. 125, n. 4, p. 3741–3771, Aug 2022. ISSN 1572-834X. Disponível em: https://doi.org/10.1007/s11277-022-09734-7.

KABIR, M. M. et al. A survey of speaker recognition: Fundamental theories, recognition methods and opportunities. *IEEE Access*, v. 9, p. 79236–79263, 2021.

KAYA, M.; BILGE, H. S. Deep metric learning: A survey. *Symmetry*, v. 11, n. 9, 2019. ISSN 2073-8994. Disponível em: https://www.mdpi.com/2073-8994/11/9/1066.

KINGMA, D. P.; BA, J. Adam: A method for stochastic optimization. In: *International Conference on Learning Representations (ICLR)*. [s.n.], 2015. Disponível em: https://arxiv.org/abs/1412.6980.

KONG, Q. et al. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, v. 28, p. 2880–2894, 2020.

KOUMURA, T.; FURUKAWA, S. Context-dependent effect of reverberation on material perception from impact sound. *Scientific Reports*, v. 7, n. 1, p. 16455, Nov 2017. ISSN 2045-2322. Disponível em: https://doi.org/10.1038/s41598-017-16651-4.

LACERDA, T. et al. Deep learning and mel-spectrograms for physica violence detection in audio. In: *Anais do XVIII Encontro Nacional de Inteligência Artificial e Computacional*. Porto Alegre, RS, Brasil: SBC, 2021. p. 268–279. ISSN 2763-9061. Disponível em: https://sol.sbc.org.br/index.php/eniac/article/view/18259.

LAMBAMO, W.; SRINIVASAGAN, R.; JIFARA, W. Analyzing noise robustness of cochleogram and mel spectrogram features in deep learning based speaker recognition. *Applied Sciences*, v. 13, n. 1, 2023. ISSN 2076-3417. Disponível em: https://www.mdpi.com/2076-3417/13/1/569.

LARCHER, A. et al. Text-dependent speaker verification: Classifiers, databases and rsr2015. *Speech Communication*, v. 60, p. 56–77, 2014. ISSN 0167-6393. Disponível em: https://www.sciencedirect.com/science/article/pii/S0167639314000156.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. *Nature*, v. 521, n. 7553, p. 436–444, May 2015. ISSN 1476-4687. Disponível em: https://doi.org/10.1038/nature14539.

LIU, H.; LANG, B. Machine learning and deep learning methods for intrusion detection systems: A survey. *Applied Sciences*, v. 9, n. 20, 2019. ISSN 2076-3417. Disponível em: https://www.mdpi.com/2076-3417/9/20/4396.

LU, J.; HU, J.; ZHOU, J. Deep metric learning for visual understanding: An overview of recent advances. *IEEE Signal Processing Magazine*, v. 34, n. 6, p. 76–84, 2017.

MÜLLER, M. *Fundamentals of Music Processing: Audio, Analysis, Algorithms, Applications*. 1. ed. Springer Cham, 2015. XXIX, 487 p. Springer International Publishing Switzerland, eBook, Computer Science Package. ISBN 978-3-319-21945-5. Disponível em: https://doi.org/10.1007/978-3-319-21945-5.

NAGRANI, A.; CHUNG, J. S.; ZISSERMAN, A. Voxceleb: A large-scale speaker identification dataset. In: *Interspeech 2017*. [S.l.: s.n.], 2017. p. 2616–2620. ISSN 2958-1796.

NAHM, F. S. Receiver operating characteristic curve: overview and practical use for clinicians. *Korean Journal of Anesthesiology*, Korean J Anesthesiol, Korea (South), v. 75, n. 1, p. 25–36, February 2022. ISSN 2005-7563. Epub 2022 Jan 18. Review. No potential conflict of interest relevant to this article was reported. Disponível em: https://doi.org/10.4097/kja.21209.

NAIR, V.; HINTON, G. E. Rectified linear units improve restricted boltzmann machines. In: *Proceedings of the 27th International Conference on International Conference on Machine Learning*. Madison, WI, USA: Omnipress, 2010. (ICML'10), p. 807–814. ISBN 9781605589077.

NIELSEN, M. A. *Neural Networks and Deep Learning*. [S.l.]: Determination Press, 2015. Last update: Thu Dec 26 15:26:33 2019. Licensed under Creative Commons Attribution-NonCommercial 3.0 Unported License.

NYQUIST, H. Certain topics in telegraph transmission theory. *Transactions of the American Institute of Electrical Engineers*, v. 47, n. 2, p. 617–644, 1928.

OPPENHEIM, A.; SCHAFER, R.; BUCK, J. *Discrete-time Signal Processing*. Prentice Hall, 1999. (Prentice Hall International Editions Series). ISBN 9780130834430. Disponível em: https://books.google.com.br/books?id=cR3CQgAACAAJ.

PARLAK, C.; ALTUN, Y. A novel filter bank design for speech emotion recognition. *Research Square*, Research Square, nov 2022. PREPRINT (Version 3). Disponível em: https://doi.org/10.21203/rs.3.rs-1993444/v3.

PIERCE, A. D. *Acoustics: An Introduction to Its Physical Principles and Applications*. 3. ed. Springer Cham, 2019. XLI, 768 p. Physics and Astronomy eBook Package. © The Author(s), under exclusive license to Springer Nature Switzerland AG 2019. ISBN 978-3-030-11214-1. Disponível em: https://doi.org/10.1007/978-3-030-11214-1.

PUVVADA, K. C. et al. Discrete audio representation as an alternative to mel-spectrograms for speaker and speech recognition. In: *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. [s.n.], 2024. p. 12111–12115. Disponível em: https://doi.org/10.1109/ICASSP48485.2024.10446998.

QIN, X.; BU, H.; LI, M. Hi-mia: A far-field text-dependent speaker verification database and the baselines. In: *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. [S.l.: s.n.], 2020. p. 7609–7613.

SAI, N. L. et al. Speaker verification in real-world applications: Advances and emerging trends. *International Journal of Engineering Research & Technology (IJERT)*, IJERT, v. 12, n. 06, June 2023. Disponível em: https://www.ijert.org/speaker-verification-in-real-world-applications-advances-and-emerging-trends.

SAINI, D. et al. A comparative analysis of automatic classification and grading methods for knee osteoarthritis focussing on x-ray images. *Biocybernetics and Biomedical Engineering*, v. 41, n. 2, p. 419–444, 2021. ISSN 0208-5216. Disponível em: https://www.sciencedirect.com/science/article/pii/S0208521621000206.

SARITHA, B. et al. Cacrn-net: A 3d log mel spectrogram based channel attention convolutional recurrent neural network for few-shot speaker identification. *Computers and Electrical Engineering*, v. 115, p. 109100, 2024. ISSN 0045-7906. Disponível em: https://www.sciencedirect.com/science/article/pii/S0045790624000284.

SARITHA, B. et al. Deep learning-based end-to-end speaker identification using time–frequency representation of speech signal. *Circuits, Systems, and Signal Processing*, v. 43, n. 3, p. 1839–1861, Mar 2024. ISSN 1531-5878. Disponível em: https://doi.org/10.1007/s00034-023-02542-9.

SARKER, I. H. Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science*, v. 2, n. 3, p. 160, Mar 2021. ISSN 2661-8907. Disponível em: https://doi.org/10.1007/s42979-021-00592-x.

SAXENA, N.; VARSHNEY, D. Smart home security solutions using facial authentication and speaker recognition through artificial neural networks. *International Journal of Cognitive Computing in Engineering*, v. 2, p. 154–164, 2021. ISSN 2666-3074. Disponível em: https://www.sciencedirect.com/science/article/pii/S266630742100019X.

SERWAY, R. A.; JEWETT, J. W. *Physics for Scientists and Engineers with Modern Physics*. 7th. ed. [S.l.]: Thomson Brooks/Cole, 2008. ISBN 978-0-495-11259-1.

SHINDE, P. P.; SHAH, S. A review of machine learning and deep learning applications. In: *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*. [S.l.: s.n.], 2018. p. 1–6.

SIDHU, M. S.; LATIB, N. A. A.; SIDHU, K. K. Mfcc in audio signal processing for voice disorder: a review. *Multimedia Tools and Applications*, v. 84, n. 10, p. 8015–8035, Mar 2025. ISSN 1573-7721. Disponível em: https://doi.org/10.1007/s11042-024-19253-1.

SNELL, J.; SWERSKY, K.; ZEMEL, R. S. Prototypical networks for few-shot learning. In: *Advances in Neural Information Processing Systems (NeurIPS)*. [s.n.], 2017. v. 30, p. 4077–4087. Disponível em: https://papers.nips.cc/paper_files/paper/2017/file/cb8da6767461f2812ae4290eac7cbc42-Paper.pdf.

STEVENS, S. S. A scale for the measurement of the psychological magnitude of pitch. *The Journal of the Acoustical Society of America*, Acoustical Society of America, v. 8, n. 1, p. 1–11, 1937.

THIENPONDT, J.; DEMUYNCK, K. Ecapa2: A hybrid neural network architecture and training strategy for robust speaker embeddings. In: *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. [S.l.: s.n.], 2023. p. 1–8.

UZAIR, M.; JAMIL, N. Effects of hidden layers on the efficiency of neural networks. In: *2020 IEEE 23rd International Multitopic Conference (INMIC)*. [S.l.: s.n.], 2020. p. 1–6.

VASWANI, A. et al. Attention is all you need. In: GUYON, I. et al. (Ed.). *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. v. 30. Disponível em: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

VIRGILLI, R. et al. Dual-bandwidth spectrogram analysis for speaker verification. In: PAES, A.; VERRI, F. A. N. (Ed.). *Intelligent Systems*. Cham: Springer Nature Switzerland, 2024. p. 340–351. ISBN 978-3-031-79029-4. Disponível em: https://doi.org/10.1007/978-3-031-79029-4_24.

WANG, F.; LIU, H. Understanding the behaviour of contrastive loss. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2021. p. 2495–2504.

WANG, M. et al. A high-speed and low-complexity architecture for softmax function in deep learning. In: *2018 IEEE Asia Pacific Conference on Circuits and Systems (APCCAS)*. [S.l.: s.n.], 2018. p. 223–226.

YAROSLAVSKY, L. P. Discretization and quantization of signals. In: ____. *Digital Picture Processing: An Introduction*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1985. p. 40–74. ISBN 978-3-642-81929-2. Disponível em: https://doi.org/10.1007/978-3-642-81929-2_3.

YE, F.; YANG, J. A deep neural network model for speaker identification. *Applied Sciences*, v. 11, n. 8, 2021. ISSN 2076-3417. Disponível em: https://doi.org/10.3390/app11083603.

ZAIDI, A. Z. et al. Touch-based continuous mobile device authentication: State-of-the-art, challenges and opportunities. *Journal of Network and Computer Applications*, v. 191, p. 103162, 2021. ISSN 1084-8045. Disponível em: https://www.sciencedirect.com/science/article/pii/S1084804521001740.

ZHANG, A. et al. *Dive into Deep Learning*. Cambridge University Press, 2023. Disponível em: https://D2L.ai.

DADOS CURRICULARES

IDENTIFICAÇÃO	
	Gustavo Pereira Pazzini Data de nascimento: 20/06/1999
Nacionalidade	Brasileiro
Nome em citações bibliográficas	PAZZINI, GUSTAVO PEREIRA PAZZINI, G. P.
Currículo Lattes	http://lattes.cnpq.br/7712897962582944
ORCID	https://orcid.org/0009-0007-5417-7820
FORMAÇÃO ACADÊMICA	
2018/2023	Bacharel em Ciência da Computação Universidade Estadual Paulista “Júlio de Mesquita Filho”
2024/2026	Mestre em Ciência da Computação Universidade Estadual Paulista “Júlio de Mesquita Filho”
PRODUÇÃO BIBLIOGRÁFICA	
PARTICIPAÇÃO EM BANCAS E ORIENTAÇÕES	
Bancas de trabalhos de conclusão	
Orientações	
PARTICIPAÇÃO EM EVENTOS CIENTÍFICOS	