Universidade Federal de Campina Grande
Centro de Engenharia Elétrica e Informática
Coordenação de Pós-Graduação em Ciência da Computação

Dissertação de Mestrado

Análise Acústica da Fala para Auxílio à Detecção e à Classificação de Distúrbios da Voz

Gabriel Almeida Azevedo

Campina Grande, Paraíba, Brasil
Fevereiro/2025

Universidade Federal de Campina Grande
Centro de Engenharia Elétrica e Informática
Coordenação de Pós-Graduação em Ciência da Computação

Análise Acústica da Fala para Auxílio à Detecção e à Classificação de Distúrbios da Voz

Gabriel Almeida Azevedo

Dissertação submetida à Coordenação do Curso de Pós-Graduação em Ciência da Computação da Universidade Federal de Campina Grande - Campus I como parte dos requisitos necessários para obtenção do grau de Mestre em Ciência da Computação.

Área de Concentração: Ciência da Computação
Linha de Pesquisa: Processamento Digital de Sinais

Profa. Dra. Joseana Macêdo Fechine Régis de Araújo
(Orientadora)

Campina Grande, Paraíba, Brasil
©Gabriel Almeida Azevedo, 03/02/2025

A994a Azevedo, Gabriel Almeida.
Análise acústica da fala para auxílio à detecção e à classificação de distúrbios da voz / Gabriel Almeida Azevedo – Campina Grande, 2025.
87 f. : il. color.

Dissertação (Mestrado em Ciência da Computação) - Universidade Federal de Campina Grande, Centro de Engenharia Elétrica e Informática, 2025.
"Orientação: Profa. Dra. Joseana Macêdo Fechine Régis de Araújo."
Referências.

1. Redes Neurais Profundas. 2. Distúrbios da Voz. 3. Disturbios - Detecção. 4. Disturbios - Classificação. 5. Processamento de Sinais de Voz. 6. Redes de Classificação Binária. I. Araújo, Joseana Macêdo Fechine Régis de. II. Título.

CDU 004.032.26(043)

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECÁRIA ITAPUANA SOARES DIAS GONÇALVES CRB-15/093

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE CAMPINA GRANDE
POS-GRADUACAO EM CIENCIA DA COMPUTACAO
Rua Aprígio Veloso, 882, Edifício Telmo Silva de Araújo, Bloco CG1, - Bairro Universitário, Campina Grande/PB, CEP 58429-900
Telefone: 2101-1122 - (83) 2101-1123 - (83) 2101-1124
Site: http://computacao.ufcg.edu.br - E-mail: secpg@computacao.ufcg.edu.br

FOLHA DE ASSINATURA PARA TESES E DISSERTAÇÕES

GABRIEL ALMEIDA AZEVEDO

ANÁLISE ACÚSTICA DA FALA PARA AUXÍLIO À DETECÇÃO E À CLASSIFICAÇÃO DE DISTÚRBIOS DA VOZ

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação como pré-requisito para obtenção do título de Mestre em Ciência da Computação.

Aprovada em: 03/02/2025

Profa. Dra. JOSEANA MACÊDO FECHINE RÉGIS DE ARAÚJO, Orientadora, UFCG

Prof. Dr. HERMAN MARTINS GOMES, Examinador Interno, UFCG

Prof. Dr. LEONARDO VIDAL BATISTA, Examinador Interno, UFPB

Prof. Dr. EDMAR CANDEIA GURJÃO, Examinador Externo, UFCG

Documento assinado eletronicamente por **JOSEANA MACEDO FECHINE**, **PROFESSOR(A) DO MAGISTERIO SUPERIOR**, em 06/02/2025, às 13:34, conforme horário oficial de Brasília, com fundamento no art. 8º, caput, da Portaria SEI nº 002, de 25 de outubro de 2018.

Documento assinado eletronicamente por **Leonardo Vidal Batista**, **Usuário**

Externo, em 06/02/2025, às 17:53, conforme horário oficial de Brasília, com fundamento no art. 8º, caput, da Portaria SEI nº 002, de 25 de outubro de 2018.

Documento assinado eletronicamente por **HERMAN MARTINS GOMES**, **PROFESSOR 3 GRAU**, em 06/02/2025, às 17:59, conforme horário oficial de Brasília, com fundamento no art. 8º, caput, da Portaria SEI nº 002, de 25 de outubro de 2018.

Documento assinado eletronicamente por **EDMAR CANDEIA GURJAO**, **PROFESSOR 3 GRAU**, em 06/02/2025, às 20:14, conforme horário oficial de Brasília, com fundamento no art. 8º, caput, da Portaria SEI nº 002, de 25 de outubro de 2018.

A autenticidade deste documento pode ser conferida no site https://sei.ufcg.edu.br/autenticidade, informando o código verificador **5200094** e o código CRC **1063A171**.

Referência: Processo nº 23096.003711/2025-86 SEI nº 5200094

Resumo

A voz é um dos meios mais importantes de comunicação do ser humano. Por meio da fala, pode-se transmitir facilmente uma mensagem. Como toda parte do corpo humano, o sistema fonatório pode ser acometido por doenças, que são comumente chamadas de patologias da voz, dentre as quais tem-se os distúrbios do trato vocal (também denominados distúrbios da voz), que incluem disfonia, laringite, pólipo e paralisia, foco da pesquisa. Em grande parte das vezes, o diagnóstico precoce é essencial para conter o agravamento da situação clínica do paciente. Entretanto, a tarefa de detectar e classificar esses distúrbios, por vezes é demorada e requer *expertise* do médico. Além disso, alguns dos exames são invasivos, gerando desconforto ao paciente. Diante do exposto e visando auxiliar o diagnóstico médico, acelerando-o e colaborando no embasamento necessário, a pesquisa ora descrita investiga o uso de redes neurais profundas para a classificação automática de sinais de voz, nas categorias saudável e patológica (ou com distúrbios), com o objetivo de distinguir entre disfonia, laringite, pólipo e paralisia, com adoção de técnicas não invasivas para aquisição de informações. Dados como espectrogramas mel, *zero crossing rate* (ZCR), *root mean square energy* (RMSE) e coeficientes MFCC foram utilizados como fontes de informação para redes pré-treinadas CNN e redes híbridas CNN-RNN LSTM. Técnicas para aumento de dados, como *time stretch, time shifting* e injeção de ruído branco (*white noise*) foram aplicadas nos dados extraídos da base utilizada (*Saarbruecken Voice Database - SVD*) para superar o problema de insuficiência de dados. Cada uma das abordagens propostas foi construída em duas versões, uma para vozes femininas e outra para vozes masculinas. O desempenho foi avaliado a partir das métricas acurácia, perda, precisão, sensibilidade (*recall*) e *F1-score*. As redes de classificação binária alcançaram taxas de acurácia de 99,33% (vozes masculinas) e 99,50% (vozes femininas) e as redes de multi classificação chegaram a apresentar acurácia de 96,40% (vozes femininas) e 89,20% (vozes masculinas), representando avanço importante e contribuição na área de detecção e classificação automática de distúrbios do trato vocal e potencial para uso clínico.

Palavras-chave: Distúrbios da Voz, Detecção, Classificação, Processamento de Sinais de Voz, Redes Neurais Profundas, *Saarbruecken Voice Database* (SVD), Analisador Automático da Condição da Voz.

Abstract

The voice is one of the most important means of human communication. Through speech, a message can be easily transmitted. Like any part of the human body, the phonatory system can be affected by diseases, which are commonly called voice pathologies, among which are vocal tract disorders (also called voice disorders), which include dysphonia, laryngitis, polyps and paralysis, the focus of the research. In most cases, early diagnosis is essential to contain the worsening of the patient's clinical condition. However, the task of detecting and classifying these disorders is sometimes time-consuming and requires expertise from the doctor. In addition, some of the tests are invasive, causing discomfort to the patient. In view of the above and aiming to assist in medical diagnosis, accelerating it and collaborating in the necessary basis, the research described here investigates the use of deep neural networks for the automatic classification of voice signals, in the healthy and pathological (or disordered) categories, with the objective of distinguishing between dysphonia, laryngitis, polyps and paralysis, with the adoption of non-invasive techniques for acquiring information. Data such as mel spectrograms, zero crossing rate (ZCR), root mean square energy (RMSE) and MFCC coefficients were used as sources of information for pre-trained CNN networks and hybrid CNN-RNN LSTM networks. Techniques for data augmentation, such as time stretching, time shifting and white noise injection were applied to the extracted data of the database used (Saarbruecken Voice Database - SVD) to overcome the problem of insufficient data. Each of the proposed approaches was built in two versions, one for female voices and another for male voices, and their performance was evaluated using the metrics accuracy, loss, precision, sensitivity (recall) and F1-score. The performance was evaluated using the metrics accuracy, loss, precision, sensitivity recall and F1-score. The binary classification networks achieved accuracy rates of 99,33% (male voices) and 99,50% (female voices), and the multi-classification networks achieved accuracy rates of 96,40% (female voices) and 89,20% (male voices), representing an important advance and contribution in the area of automatic detection and classification of vocal tract disorders and potential for clinical use.

Keywords: Voice disorder, Detection, Classification, Voice Signal Processing, Deep Neural Networks, Saarbruecken Voice Database (SVD), Automatic Voice Condition Analyzer.

Agradecimentos

Agradeço primeiramente a Deus que me deu saúde e garra durante toda essa jornada do mestrado.

Agradeço também aos meus pais, Profa. Dra. Márcia Rejane de Queiroz Almeida Azevedo e Prof. Dr. Carlos Alberto Vieira de Azevedo, que sempre se fizeram presentes e me incentivaram, desde pequeno, a estudar e reconhecer a importância do professor na formação de um cidadão e da sociedade.

Agradeço a minha orientadora Profa. Dra. Joseana Macêdo Fechine Régis de Araújo pela paciência, compreensão e orientação durante toda essa jornada.

Agradeço a minha irmã, Bianca Almeida Azevedo, aos meus amigos, Thiago Moura, Igor Lucena, Everton Catão, Ionnara Lima, Núbia Morais e tantos outros que, mesmo em momentos difíceis, estiveram ao meu lado, fazendo com que eu perseverasse.

"É justo que muito custe o que muito vale."

Santa Teresa de Jesus

Conteúdo

Lista de Siglas e Abreviaturas

AUC - *Area Under the Curve*

AVCA - *Automatic Voice Condition Analyzer*

AVD - *Arabic Voice pathology Dataset*

AVFAD - *Advanced Voice Function Assessment Databases*

CNN - *Convolutional Neural Network*

CQCC - *Constant Q Cepstral Coefficients*

DF - Disfonia Espasmódica

EGG - *Electroglottography*

eGeMAPS - *Geneva Minimalistic Acoustic Parameter Set*

FFT - *Fast Fourier Transform*

FP - Falsos Positivos

FN - Falsos Negativos

GDPR - *General Data Protection Regulation*

GradCAM - *Gradient-weighted Class Activation Mapping*

GRU - *Gated Recurrent Units*

HF - *Heart Failure*

Hz - *Hertz*

HOS - *Higher-Order Statistics*

HHT - *Hilbert-Huang Transform*

HUPA - *Hospital Príncipe de Asturias*

IS2010 - *INTERSPEECH 2010 Paralinguistic Challenge feature Set*

KNN - *K-Nearest Neighbors*

LGPD - Lei Geral de Proteção de Dados

LIME - *Local Interpretable Model-agnostic Explanations*

LSTM - *Long Short-Term Memory*

LPCC - *Linear Prediction Cepstrum Coefficients*

MFCC - *Mel-Frequency Cepstral Coefficients*

MEEI - *Massachussets Ear and Eye Infirmary*

MLP - *Multilayer Perceptron*

PLPC - *Perceptual Linear Prediction Cepstral Coefficients*

RLNP - *Recurrent Laryngeal Nerve Paralysis*

RNN - *Recurrent Neural Network*

ROC - *Receiver Operating Characteristic*

RMSE - *Root Mean Squared Energy*

RQEM - Raíz Quadrática da Energia Média

SAD - Sistemas Automáticos de Apoio à Decisão

SAHP - *Shapley Additive Explanations*

SVD - *Saarbrücken Voice Database*

STFT - *Short-Time Fourier Transform*

SVM - *Support Vector Machine*

VP - Verdadeiros Positivos

VN - Verdadeiros Negativos

VGG - *Visual Geometry Group*

WST - *Wavelet Scattering Transform*

ZCR - *Zero Crossing Rate*

Lista de Figuras

Lista de Tabelas

Lista de Quadros

Capítulo 1

Introdução

A voz é um dos meios mais importantes de comunicação do ser humano. Por meio da fala, pode-se transmitir facilmente uma mensagem [32]. A emissão da voz é um processo complexo que envolve a coordenação dos sistemas respiratório, fonatório, ressonante, articulatório e nervoso. O sistema fonatório é o mais importante no processo, pois tem a função de gerar o som. Seu principal órgão é a laringe, que possui um mecanismo responsável por regular a passagem de ar vindo dos pulmões (comumente chamado de pregas, cordas ou dobras vocais). Quando as pregas vocais se aproximam, o fluxo de ar as faz vibrar, gerando assim o som [26].

Como toda parte do corpo humano, o sistema fonatório pode ser acometido por doenças que são comumente chamadas de distúrbios (ou patologias) da voz. Em grande parte das vezes, o diagnóstico precoce é essencial para conter o agravamento da situação clínica do paciente. Para detectar e classificar os distúrbios da voz, os médicos comumente elaboram seus diagnósticos baseando-se em exames instrumentais, como eletroglotografia e videolaringoscopia (buscando explorar as estruturas laríngeas), análise perceptiva do discurso do paciente, análise de sintomas e do histórico de saúde [25]. Essa atividade por vezes é demorada e requer *expertise* do médico para propor boas hipóteses e testá-las. Além disso, alguns dos exames são invasivos, gerando desconforto ao paciente [29].

Nesse sentido, o estudo centrado na criação de sistemas para detecção e classificação de distúrbios da voz, por meio da análise do som emitido pelo paciente, se apresenta como uma alternativa viável para o apoio ao diagnóstico médico, por ser um método não invasivo, rápido e barato, atraindo assim a atenção de profissionais da saúde.

Em algumas áreas da medicina, a utilização de sistemas automatizados de apoio à decisão já é uma realidade. Os SAD (Sistemas Automáticos de Apoio à Decisão) têm como objetivos agilizar o atendimento, fornecendo ferramentas que embasem a decisão médica e apresentem evidências úteis para classificação do caso clínico, reduzindo o tempo de diagnóstico [17]. Ferramentas como o EpicCare [1], *software* de prontuário eletrônico desenvolvido pela Epic Systems, que oferece apoio à decisão alertando sobre interações medicamentosas e protocolos pautados em diretrizes, ou ainda, o IBM Watson for Oncology [60], que recomenda opções de terapias específicas baseadas no perfil genético do tumor do paciente, são exemplos desses sistemas automatizados.

No âmbito da otorrinolaringologia existe a necessidade de acelerar o diagnóstico sem aplicar técnicas invasivas. Diante dessa necessidade, a área da computação se dedica a algumas décadas ao estudo e desenvolvimento de sistemas para análise automática da voz, visto que, a partir do sinal digital da voz, pode-se aplicar técnicas de transformação e processamento do dado original, para extrair características e aplicar técnicas de aprendizagem de máquina [49] para classificação desses sinais.

Pesquisas apresentadas nos artigos intitulados *Pathological Voice Detection Based on Phase Reconstitution and Convolutional Neural Network* [20] e *Analysis and Detection of Pathological Voice Using Glottal Source Features* [32] descrevem estudos e propõem diferentes abordagens para detecção de distúrbios a partir de sinais de voz. A pesquisa descrita no artigo intitulado *On the design of automatic voice condition analysis systems. Part I:Review of concepts and an insight to the state of the art* [25] traz uma revisão sobre o estado da arte e elenca as técnicas e características comumente utilizadas para este propósito. Algumas pesquisas apresentadas em [35], [55] e [29] abordam a etapa de classificação de distúrbios.

Diante do exposto, a dissertação ora descrita tem por objetivo construir uma abordagem, utilizando redes neurais, que possa oferecer apoio ao diagnóstico, sendo capaz de classificar um sinal de voz com distúrbio ou saudável e também, distinguir os distúrbios presentes no sinal da voz, dentre quatro categorias: disfonia, laringite, pólipo e paralisia.

Pesquisas nessa área apresentam alguns desafios, que podem ser divididos em desafios técnicos ou ainda em desafios vindos do âmbito da medicina. Desafios técnicos tratam do processo de desenvolvimento do sistema de classificação (alguns exemplos estão elencados

[1] https://epicsolutions.ie/epiccare/

abaixo).

1. Desbalanceamento de dados: Para treinamento de uma rede neural, o desbalanceamento de dados pode gerar efeitos indesejados como *overfitting*, concomitantemente, a quantidade de dados disponíveis de sinais de voz saudáveis é muito maior que a quantidade de sinais de voz com distúrbios. Como exemplo, a base *Saarbruecken Voice Database* (SVD) [28], possui um total de 259 registros de sinais de voz masculinas saudáveis em tom normal, enquanto que para a doença disfonia, tem-se apenas 29 registros masculinos neste mesmo tom.

2. Variações por gênero e idade: o trato vocal de homens e mulheres é, por natureza, diferente. O trato vocal masculino é geralmente maior, resultando em frequências de ressonância mais baixas, o que contribui para uma voz mais grave [57]. As pregas vocais também são mais espessas, vibrando mais lentamente e produzindo frequências fundamentais mais baixas (70-150 Hz em homens) em comparação com mulheres (150-250 Hz). Nas mulheres, as frequências formantes são mais altas, o que contribui para uma percepção de voz mais aguda e brilhante [19]. Mulheres também possuem menor capacidade pulmonar, o que pode afetar a sustentação vocal.

3. Extração de características relevantes: a escolha de métricas relevantes é um desafio e pode impactar no aprendizado do modelo. A depender da forma de representação do dado, também pode ocorrer a perda de informação relevante para a classificação.

4. Interpretação do modelo: modelos de aprendizagem profunda de máquina possuem uma certa restrição de aceitação no campo médico, pois são difíceis de interpretar e explicar de forma clara, o que a rede aprendeu e como aprendeu.

5. Generalização: Pode-se levantar os questionamentos a seguir.

    - Como garantir que o modelo não é bom apenas para os dados utilizados?
    - Como garantir que não houve *overffiting*?
    - Como garantir que modelos treinados utilizando sinais de voz em determinado idioma, adquiram capacidade de generalização?

No âmbito da medicina tem-se dois principais desafios, o primeiro trata da aquisição e categorização de dados. Na área da saúde, os dados são regulamentados por leis como LGPD (Lei Geral de Proteção de Dados) e GDPR (*General Data Protection Regulation*) o que dificulta o acesso. Outro ponto é a escassez de dados categorizados; categorizar os sinais de voz exige trabalho manual de especialistas, o que é demorado e, por vezes, não é priorizado frente a demandas mais urgentes dos profissionais da saúde. Ademais, algumas doenças são raras, o que dificulta ainda mais o acesso a exemplos de dados.

O segundo desafio diz respeito à aceitação médica dos Sistemas de Apoio à Decisão (SAD): sistemas automatizados ainda despertam desconfiança pois têm potencial de classificar erroneamente algum dado, gerando consequências indesejadas [36]. Quando são utilizadas redes neurais profundas, a desconfiança ainda é maior pois a área médica exige explicações claras e concisas sobre como o modelo chegou a determinado resultado, o que ele aprendeu e como aprendeu. Responder essas perguntas não é uma tarefa trivial, entretanto técnicas como GradCAM (*Gradient-weighted Class Activation Mapping*), Lime (*Local Interpretable Model-agnostic Explanations*), Shap (*Shapley Additive Explanations*) [43] têm contribuído para facilitar o entendimento do conhecimento aprendido por redes profundas.

1.1 Objetivos

1.1.1 Objetivo Geral

A pesquisa de mestrado ora descrita tem como objetivo geral desenvolver uma abordagem com métodos de aprendizagem profunda que seja capaz de detectar vozes com distúrbios (capaz de discernir entre uma voz saudável ou uma voz patológica) e que possa classificá-las automaticamente, com certo grau de precisão, em determinados grupos de distúrbios do trato vocal, objetivando apoiar o diagnóstico de um paciente.

1.1.2 Objetivos Específicos

Como objetivos específicos, tem-se:

- Selecionar uma base de dados mais representativa para o processo;

- Avaliar as características mais adequadas para representação dos sinais de voz; e
- Selecionar um classificador mais adequado para a construção da abordagem.

1.2 Questões de Pesquisa

A pesquisa tem como objetivo principal desenvolver um classificador capaz de auxiliar no diagnóstico e detecção de distúrbios da voz, que sirva como apoio à decisão do profissional da saúde em seu diagnóstico. Assim, as questões de pesquisa elencadas estão listadas abaixo.

1. Quais são os distúrbios da voz comumente analisados a partir da voz do paciente?
2. Quais são as características mais adequadas para representação de sinais de voz com distúrbio?
3. No âmbito de aprendizagem profunda para identificar e classificar distúrbios da voz, como lidar com a baixa quantidade de dados disponíveis para a etapa de treinamento? Quais são as alternativas?

1.3 Relevância

A análise automática da condição da voz e a classificação automática de patologias vocais são tarefas que englobam as áreas da ciência da computação e da saúde. Na área da saúde os métodos tradicionais de diagnóstico vocal, dependem fortemente de avaliações manuais realizadas por especialistas, o que pode ser subjetivo, demorado e limitado pelo acesso à profissionais qualificados. Nesse contexto, o uso de redes neurais aplicadas à análise do sinal de voz surge como uma alternativa promissora pois, ao combinar a capacidade de aprendizado profundo, com a análise de características acústicas, é possível construir modelos de análise automática da condição da voz que auxiliem os diagnósticos, tornando-os rápidos, objetivos e acessíveis, contribuindo para uma triagem mais eficiente, e potencialmente reduzindo o tempo e o custo do diagnóstico médico.

Nesse sentido, uma parcela significativa das pesquisas da área objetiva propor diferentes formas para construir os detectores de vozes patológicas e sugerem, como pesquisas futuras,

a construção de classificadores capazes de identificar doenças do trato vocal e/ou o grau de severidade dessas, tais como: [61], [38], [32] e [16].

Outras pesquisas abordam a etapa de classificação da possível doença do paciente, a exemplo de: [30], [20] e [15]. Dessa forma, responder o principal questionamento da pesquisa, que é "Para casos de distúbios da voz, é possível definir qual doença o paciente apresenta, por meio da análise da sua voz, atingindo certo grau de precisão?", poderá contribuir para o estado da arte e incentivar outros pesquisadores a se aprofundar não apenas na detecção, mas também na etapa de classificação de distúrbios. No âmbito da medicina, a relevância consiste em prover uma abordagem com potencial para auxiliar os profissionais da saúde, no diagnóstico dos pacientes.

De forma mais específica, a pesquisa ora descrita propõe uma abordagem para detecção e classificação de distúrbios da voz, a partir de um sinal de voz. Esse estudo permite enriquecer a literatura, com abordagens que combinam características de baixo nível do sinal de voz e que consideram a diferença no trato vocal feminino e masculino, para a geração e o treinamento de classificadores. Através desse desenvolvimento, busca-se também incentivar pesquisas na área de classificação dos distúrbios da voz, expondo características relevantes para que um modelo consiga discernir entre os distúrbios, sendo capaz de lidar com os problemas relacionados à quantidade de dados disponíveis.

1.4 Contribuições

As principais contribuições da pesquisa são listadas a seguir.

- Revisão da literatura atualizada: foi utilizado o período de 2017 a 2024, em motores de busca representativos na área da pesquisa, como *Google Schoolar*, ACM, PubMed, Elsevier e Scopus.

- Validação e comparação com estudos anteriores: o estudo segue uma abordagem similar à utilizada por Ksibi (2023), conseguindo superar os resultados em um cenário específico e contribuindo com o desenvolvimento da etapa de classificação de distúrbios.

- Reprodutibilidade: neste trabalho, são descritas, de forma minuciosa (vide repositório

disponível [2]), as atividades desenvolvidas para a criação do modelo, desde a seleção dos sinais de voz (para os quais tem-se identificadores, possibilitando que novas pesquisas utilizem os mesmo sinais), até as bibliotecas utilizadas na criação da rede neural.

- Na área da medicina: esta dissertação descreve um modelo que tem potencial para auxiliar diagnósticos clínicos, tornando-os mais rápidos e acessíveis (pode ser utilizado como triagem inicial, antes de exames mais detalhados e como insumo para embasar diagnósticos).

1.5 Estrutura do Documento

Neste capítulo de introdução foram apresentados os princípios que delinearam a realização da pesquisa. Nos demais capítulos, a pesquisa será detalhada. No Capítulo 2, são descritos fundamentos relevantes para a compreensão da pesquisa. No Capítulo 3, são apresentadas as principais pesquisas relacionadas ao tema, no período de 2017 a 2024, com discussão das abordagens utilizadas e resultados obtidos. No Capítulo 4, é apresentada a metodologia utilizada para desenvolvimento da abordagem proposta. O Capítulo 5 trata da apresentação e análise dos resultados obtidos. No Capítulo 6, o estudo é condensado, com apresentação das considerações finais, contribuições, limitações e sugestões para pesquisas futuras. No Apêndice A, encontra-se a lista de identificadores dos sinais de voz utilizados nesta pesquisa.

[2] https://github.com/gabriellmd/automatic-voice-condition-analyser

Capítulo 2

Fundamentação Teórica

Neste capítulo, inicialmente, é apresentada uma descrição do mecanismo de produção da voz e caracterização dos principais distúrbios vocais. Em seguida, são apresentados os conceitos relevantes para a construção de um *Automatic Voice Condition Analyzer* (AVCA), para detecção e classificação de sinais de voz com distúrbios.

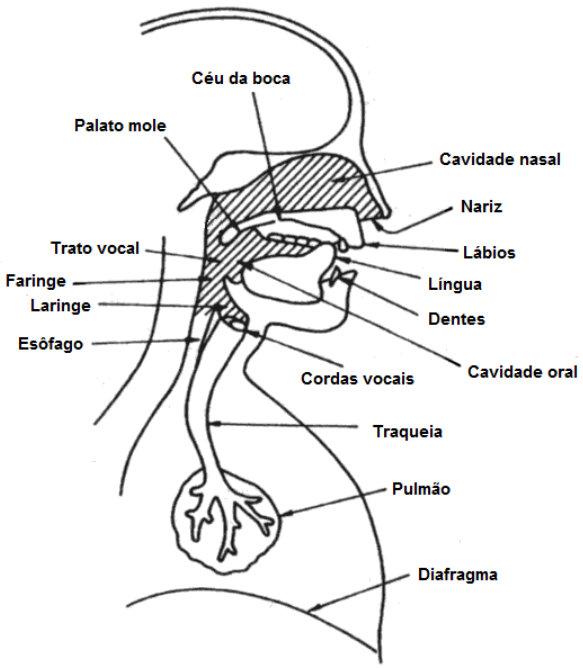
2.1 Produção e Distúrbios da Voz

A emissão da voz é um processo complexo que envolve a coordenação dos sistemas respiratório, fonatório, digestivo, ressonante, articulatório e nervoso. O sistema fonatório é o mais importante no processo, pois tem a função de gerar o som. Seu principal órgão é a laringe, que possui um mecanismo (comumente chamado de pregas, cordas ou dobras vocais) responsável por regular a passagem de ar vindo dos pulmões. Quando as pregas vocais se aproximam, o fluxo de ar as faz vibrar, gerando assim o som [26].

Na Figura 2.1, são especificados os órgãos e estruturas que impactam na produção da voz. Neste processo, o principal músculo envolvido é o diafragma, tendo a função de controlar o fluxo de ar usado na produção vocal; sua movimentação provoca a entrada e saída de ar dos pulmões. A traqueia serve como tubo condutor do ar entre os pulmões e a laringe. A laringe é responsável por controlar a abertura e fechamento das pregas vocais. A faringe, juntamente com as cavidades nasal e oral, atuam como cavidades ressonadoras. Por último, os lábios, língua, dentes e palato modulam o som, formando palavras e sons específicos [22].

Como toda parte do corpo humano, o sistema fonatório pode ser acometido por doenças

Figura 2.1: Sistema Fonatório Humano.



Fonte: Adaptado de [22].

que são comumente chamadas de distúrbios (ou patologias) da voz. Estes distúrbios podem ser classificados como sendo de caráter benigno ou maligno, assim como nas categorias de caráter funcional, orgânico ou neurológico. Como distúrbios neurológicos, tem-se: paralisia das pregas vocais, disfonia espasmódica e doença de Parkinson; como orgânicos, tem-se: pólipos, nódulos, laringite, papiloma, leucoplasia e câncer e, como funcionais, pode-se citar: fonação hiperfuncional, disfonia por tensão muscular e afonia de conversão [53].

A disfonia afeta a qualidade da voz e ocorre quando há impedimento na emissão natural da voz, podendo fazer com que a voz soe rouca, áspera, fraca, estridente [10]. Pode ser causada por abuso vocal, presença de nódulos, pólipos, tumores, laringite, ou por causas emocionais (disfonia psicogênica). Na Figura 2.2, tem-se um caso de disfonia psicogênica, com a laringe normal em todos os aspectos, com exceção do fechamento glótico com fenda paralela [13].

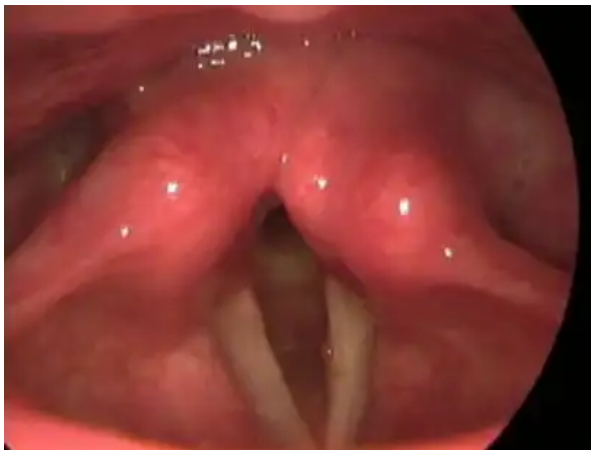
Pólipos são massas benignas encontradas logo abaixo da superfície membrana da corda vocal e afetam a vibração das pregas vocais [30]. São causados por abuso vocal ou exposição a substâncias irritantes, como a fumaça de cigarro. Seus sintomas geralmente são rouquidão, voz grave ou voz sussurrada [18]. Na Figura 2.3, é apresentado um pólipo (ponto vermelho)

na borda da prega vocal direita.

Figura 2.2: Disfonia Psicogênica.

Fonte: Extraído de [13].

Figura 2.3: Pólipo Vocal.



Fonte: Extraído de [13].

Laringite é uma inflamação da laringe que pode ser ocasionada por abuso vocal, alergias, infecção viral, refluxo de ácidos estomacais ou exposição a substâncias irritantes [18]. Seus sintomas são rouquidão ou perda total da voz, dor na garganta, dificuldade de deglutição e tosse. Na Figura 2.4, é apresentado um caso de laringite aguda viral.

A laringe possui os nervos laríngeos recorrentes esquerdo e direito. Estes nervos tem a responsabilidade de controlar as pregas vocais [23]. A paralisia vocal é uma condição em que um ou os dois nervos apresentam mal funcionamento, podendo ser fruto de lesões, doenças neurológicas, infecções ou tumores. A Figura 2.5 apresenta um caso de paralisia do nervo

recorrente direito.

Figura 2.4: Laringite aguda viral.

Fonte: Extraído de [13].

Figura 2.5: Paralisia direita.

Fonte: Extraído de [13].

Para se extrair dados do sinal de voz com os distúrbios citados acima, são comumente utilizadas métricas, tais como, taxa de cruzamento por zero, raiz quadrática da energia média (RQEM) e coeficientes cepstrais de frequência mel, selecionadas em função dos aspectos destacados a seguir.

- Distúrbios como disfonia ou laringite, podem levar a um aumento na instabilidade do sinal vocal, o que pode resultar em uma taxa de cruzamento por zero mais elevada, especialmente em casos em que há ruído ou interrupções na voz e levar a uma dimi-

nuição na intensidade vocal ou a uma variação irregular na intensidade, o que se reflete em uma RQEM mais baixo.

- A paralisia vocal pode resultar em uma voz com menos energia, e a disfonia pode ser caracterizada por uma variação na intensidade e na qualidade da energia vocal;
- Essas patologias podem gerar alterações na energia espectral ou irregularidades na produção de som, que são capturadas pelos coeficientes de Mel.

2.2 Analisador Automático da Condição da Voz

A construção de um AVCA se divide em duas fases, a de treinamento (na qual também é realizada a validação) e a de teste. As etapas básicas para o desenvolvimento do AVCA são: aquisição dos sinais de voz, pré-processamento dos sinais, extração de características, aprendizagem/construção de padrões (apenas na fase de treinamento) e classificação. Na Figura 2.6, são apresentadas as etapas básicas das duas fases.

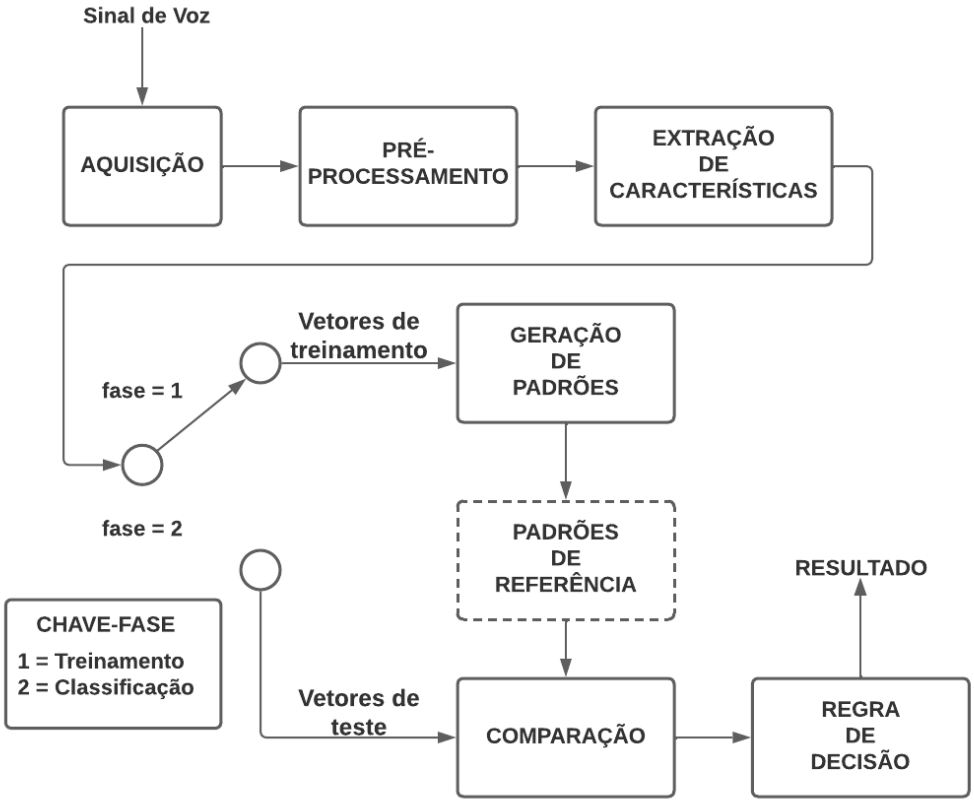
Para a pesquisa ora apresentada, métodos como *noise injection*, *time stretch* e *time shifting* [34] foram utilizados a fim de realizar o aumento de dados. *Zero Crossing Rate* (ZCR), *Root-Mean-Square Energy* (RMSE) [34], *Mel-Frequency Cepstral Coefficients* (MFCC) [3] e Espectrogramas Mel [39] foram utilizados para extração de características. Redes neurais profundas do tipo *Convolutional Neural Network* (CNN) [42] e *Recurrent Neural Network* (RNN) [33] foram empregadas como classificadores.

A seguir, a descrição das etapas necessárias à construção do AVCA adotado na pesquisa.

2.2.1 *Pré-Processamento*

O conceito de segmentação consiste em dividir um sinal contínuo (como uma gravação de voz) em pequenos segmentos temporais. Este passo é feito para facilitar a análise pois o sinal de voz varia estatisticamente com o tempo [14] mas assume características estacionárias entre 16 e 32 ms [50]. O janelamento consiste na multiplicação do segmento por uma função janela no domínio do tempo [40] tendo como objetivo controlar as descontinuidades do sinal e reduzir problemas como o vazamento espectral durante a análise no domínio da frequência. O vazamento espectral ocorre quando um sinal é truncado no domínio do tempo.

Figura 2.6: Diagrama em blocos de um AVCA.



Fonte: Adaptado de [49].

Isso faz com que sua representação no domínio da frequência apresente energia espalhada para frequências vizinhas, em vez de estar concentrada em uma única frequência. Esse efeito acontece porque a DFT assume que o sinal analisado é periódico dentro da janela escolhida. Se a janela não captura um número inteiro de ciclos da onda, a transição brusca nas bordas da janela introduz descontinuidades, que geram componentes de alta frequência na FFT. As janelas de Hamming, Hanning e Blackman [48] são as mais utilizadas devido ao equilíbrio entre suavização de bordas e controle do vazamento espectral.

No contexto do processamento digital de sinais de voz, pode-se aplicar transformações a partir de alterações na duração e no tom do sinal (*time stretching*, *time shifting* e *pitch shifting*), adicionando ruído (branco ou ambiental) ou aplicando filtros no sinal de voz.

- Adição de ruído (*noise injection*): O ruído é tipicamente injetado de forma controlada para simular perturbações que podem ocorrer no mundo real. Seus benefícios são: aumenta o grau de generalização do modelo, age como forma de regularização e pode

ajudar a lidar com mínimos locais na etapa de otimização. Porém, a eficácia da injeção de ruído depende do domínio e da tarefa proposta. O ruído branco é um tipo de sinal que possui potência constante em todas as frequências do espectro, em outras palavras, esse tipo de ruído contempla todas as frequências audíveis com a mesma intensidade.

- *Time Stretch*: Essa técnica altera a duração de um sinal de voz sem modificar sua frequência fundamental. O *time stretch* pode ser realizado prolongando-se a duração do sinal (aumentado o tempo de reprodução, tornando o sinal mais lento) ou comprimindo (reduzindo o tempo de reprodução, tornando o sinal mais rápido). Essa técnica ajuda a aumentar a diversidade do conjunto de dados e a reduz o risco de *overfitting*, pois o modelo estará exposto a dados com variações de duração. Entretanto, caso as mudanças sejam muito extremas, o sinal poderá sofrer perda de naturalidade.

- *Time Shifting*: Envolve o deslocamento de um sinal no eixo do tempo, sem alterar sua duração ou propriedades, como frequência ou tom. Essa técnica é utilizada para simular atrasos/adiantamentos temporais e também ajuda a aumentar a diversidade do conjunto de dados e consequentemente, a robustez do modelo.

2.2.2 Extração de Características

A etapa de extração de características busca reduzir a dimensionalidade dos dados e extrair padrões ou atributos específicos que capturam a essência do problema. No âmbito de processamento de voz, técnicas como taxa de cruzamento por zero e a raiz quadrática média da energia, podem ser empregadas. A extração busca reduzir a dimensionalidade dos dados e destacar padrões relevantes para o problema.

Taxa de Cruzamento por Zero

A taxa de cruzamento por zero (ou *zero crossing rate - ZCR*) é uma métrica que indica a frequência com que o sinal de voz (no domínio do tempo) cruza o eixo zero (eixo x). Esta métrica geralmente é calculada em janelas temporais (*frames*), de curta duração, para capturar variações ao longo do tempo. A ZCR é descrita pela Equação 2.1 [34].

$$Z(i) = \frac{1}{2WL} \sum_{n=1}^{WL} |\text{sgn}(x_i(n)) - \text{sgn}(x_i(n-1))|\,, \tag{2.1}$$

em que:

- $Z(i)$: Taxa de cruzamento por zero no segmento (i);
- WL: Comprimento da janela usada para análise;
- $x_i(n)$: Valor da n-ésima amostra do segmento i;
- sgn: Função de sinal (retorna 1 para valores positivos, -1 para valores negativos, e 0 para zero);
- $\sum_{n=1}^{WL}$: Somatório das diferenças entre sinais consecutivos, iniciando de $n = 1$ até WL.

***Raiz Quadrática da Energia Média* (RQEM)**

A RQEM mede a energia média de um sinal em uma janela de tempo específica, e pode ser útil para caracterizar mudanças de intensidade (amplitude) do sinal de voz (por exemplo, distinguir entre sons fracos e fortes). A RQEM é calculada a partir da Equação 2.2 [34].

$$RQEM = \sqrt{\frac{1}{N} \sum_{n=1}^{N} x(n)^2}, \tag{2.2}$$

em que:

- $x(n)$: Valor da n-ésima amostra do sinal de voz.
- N: Número total de amostras do sinal ou da janela de análise.
- $x(n)^2$: O quadrado do valor da amostra $x(n)$, que representa a energia instantânea do sinal em cada amostras.
- $\frac{1}{N} \sum_{n=1}^{N} x(n)^2$: Média dos valores quadrados das amostras.
- A raiz quadrada ($\sqrt{\cdot}$) é calculada a partir da média dos quadrados para obtenção da energia total do sinal, em termos de unidades de amplitude.

Coeficientes Cepstrais de Frequência Mel (MFCC)

Um espectrograma é uma representação do espectro de frequências de um sinal de voz, ao longo do tempo [39]. O espectrograma mel é um espectrograma mapeado para a escala Mel (escala baseada na percepção auditiva humana [47]).

Estes espectrogramas são muito populares porque capturam características importantes de um sinal de voz que são invariantes a mudanças no domínio do tempo, permitindo a obtenção de resultados melhores em tarefas baseadas em aprendizado de máquina, como reconhecimento ou síntese de voz.

A seguir, as etapas necessárias para geração de um espectrograma mel.

1. Aplicar a segmentação do sinal de voz;

2. Aplicar a transformada de Fourier ao sinal segmentado: o sinal de voz, que geralmente está no domínio do tempo, é convertido para o domínio da frequência usando a transformada de Fourier de curto prazo (STFT - *Short-Time Fourier Transform*);

3. Mapear as frequências: as frequências obtidas a partir da STFT são convertidas para a escala Mel. A escala Mel é definida a partir da Equação 2.3 [12].

$$f_{\text{mel}} = 1127 \ln\left(1 + \frac{f}{700}\right). \tag{2.3}$$

em que:

- f: Frequência em Hertz.

Os MFCC são uma representação compacta do espectro de potência de um sinal de voz, projetada para aproximar a percepção auditiva humana [12]. Eles são obtidos aplicando uma série de transformações que convertem o sinal de voz original em um conjunto de coeficientes que capturam características espectrais. São utilizados em tarefas de reconhecimento de fala, como autenticação de locutor, de classificação de músicas (permitem diferenciar gêneros musicais ou identificar instrumentos) e tarefas de detecção de emoções (auxiliam na análise de entonações).

O cálculo dos MFCC é dividido nas etapas de pré-enfase, divisão em quadros, janelamento, transformada de fourier, aplicação do banco de filtros mel, logaritmo e transformada discreta do cosseno [12] [3].

Inicialmente, é feita a pré-ênfase, que consiste em aplicar um filtro para aumentar a magnitude da energia nas frequências mais altas.

A pré-ênfase pode ser descrita pela Equação 2.4 [4].

$$y(n) = x(n) - \alpha x(n-1), \tag{2.4}$$

em que:

- $x(n)$: sinal de entrada;
- $y(n)$: sinal filtrado;
- α: fator de pré-ênfase, geralmente $\alpha \approx 0,97$.

Em seguida, é feita a divisão do sinal de voz em quadros para capturar características de curta duração. Posteriormente, é aplicada uma janela (como a janela de Hamming) a cada quadro para minimizar descontinuidades nas bordas. A definição formal da janela de Hamming é dada pela Equação 2.5 [40].

$$w(n) = 0,54 - 0,46 \cos\left(\frac{2\pi n}{N-1}\right). \tag{2.5}$$

A Equação 2.6 [12] ilustra como o quadro com a janela aplicada pode ser descrito.

$$x_w(n) = x(n)w(n), \tag{2.6}$$

em que:

- $w(n)$: função janela;
- $x(n)$: amostras do sinal;
- $x_w(n)$: sinal com janela aplicada.

Para cada janela *m*, estima-se o espectro *S(w, m)*, utilizando a FFT (*Fast Fourier Transform*). O espectro então é passado por um banco de filtros. O espectro modificado é representado pela Equação 2.7 [12].

$$P(i) = \sum_{k=0}^{N/2} |S(k,m)|^2 \, H_i \left(\frac{2\pi k}{N} \right), \tag{2.7}$$

em que:

- $P(i)$: energia de saída do filtro i;
- N: número de pontos da transformada discreta de Fourier;
- $|S(k,m)|$: módulo da amplitude na frequência do k-ésimo ponto da m-ésima janela;
- $H_i(w)$: função resposta em frequência do i-ésimo filtro.

Em seguida, define-se o conjunto de pontos *E(k)* dado pela Equação 2.8 [12].

$$E(k) = \begin{cases} \log(P(i)), & \text{se } k = k_i \\ 0, & \text{se } k \in [0, N-1] \end{cases} \tag{2.8}$$

O logaritmo da energia em cada filtro é calculado a fim de melhorar a separação das características de voz. Por fim, é aplicada a transformada discreta do cosseno para, então, obter os coeficientes mel-cepstrais [58] [44]. A Equação 2.9 [12] representa o coeficiente mel-cepstral *n*.

$$C_{\text{mel}}(n) = \sum_{i=1}^{N_f} E(k_i) \cos \left(\frac{2\pi}{N} k_i n \right), \tag{2.9}$$

em que:

- N_f: número de filtros;
- k_i: centro do i-ésimo filtro.

2.2.3 Treinamento e Classificação

Durante o treinamento é possível aplicar uma validação do modelo após cada iteração ou época. Essa validação ajuda a monitorar o desempenho do modelo em dados que não foram utilizados durante o treinamento, permitindo verificar se o modelo está generalizando bem e não está sofrendo *overfitting*.

A técnica de ***k-fold cross-validation*** é uma das técnicas de validação mais populares. Esta técnica produz 'k' conjuntos disjuntos de tamanho N/k, chamados *folds*, com 'N' representando o número de observações. No total, k iterações são realizadas, usando em cada caso um subconjunto diferente para fins de teste e o restante (k - 1) para o treinamento. As medidas de desempenho são, então, avaliadas a partir do valor médio calculado entre as iterações [25].

A técnica ***leave-one-out*** é um caso específico da validação cruzada , a partir da qual 'N' é igual a 'k'. Neste caso, apenas uma observação é utilizada para teste e os registros restantes são empregados para treinamento, repetindo este procedimento N vezes. A validação *leave-one-out* geralmente é preferida à validação cruzada quando os tamanhos do conjunto de dados são pequenos, pois permite maximizar o tamanho da partição de treinamento [25].

A geração de padrões de referência e o teste de comparação são realizadas pelo classificador escolhido para o AVCA. Existem vários tipos de classificadores como por exemplo o SVM (*Support Vector Machine*) [6], KNN (*k-nearest neighbors*) [8], redes neurais.

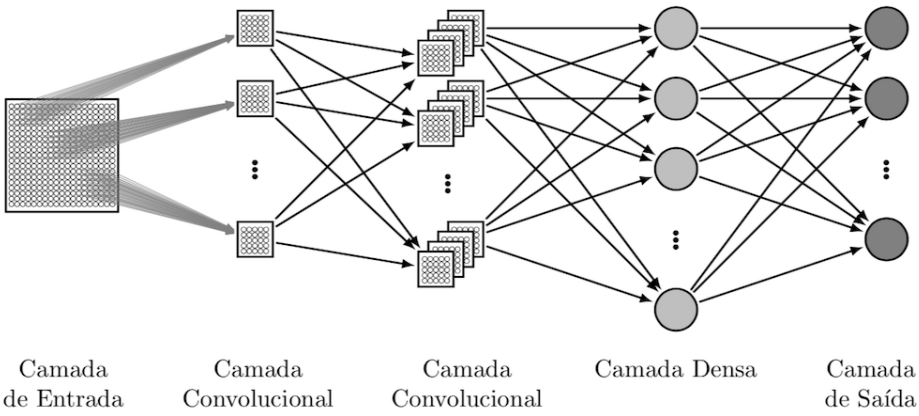
Redes Neurais Convolucionais (CNN)

A CNN (*Convolutional Neural Network*) é uma rede neural profunda *feed-forward* (cada camada se conecta à próxima camada, porém não há caminho de volta), capaz de processar dados estruturados em grades, como imagens. Esta rede utiliza operações matemáticas de convolução ao invés de multiplicação de matrizes [42] e sua estrutura base é composta por camadas de convolução, camadas de *pooling* e camadas de conexão completa que otimizam uma função de perda em tarefas como reconhecimento de imagens e sinais.

O classificador CNN também pode ser utilizado com dados gerados a partir de um sinal de voz pois, este sinal, em sua essência, é uma série temporal. Imagens como espectrogramas mel podem alimentar este tipo de rede. Na Figura 2.7, é apresentada a topologia básica de

uma CNN.

Figura 2.7: Topologia básica de uma CNN.



Fonte: Extraído de [52].

As camadas convolucionais capturam características locais, permitindo que a rede identifique padrões mais simples antes de combiná-los em representações mais complexas. As operações de *pooling* reduzem a resolução espacial e ajudam a tornar o modelo mais robusto a pequenas variações dos dados. Por meio das operações de convolução e *pooling*, as redes CNN são capazes de identificar padrões em uma entrada, independentemente de sua localização (como um objeto em diferentes posições dentro de uma imagem). A essência das redes CNN está em aprender automaticamente representações de alto nível a partir dos dados, construindo redes com múltiplas camadas ocultas e utilizando operações de convolução para melhorar a classificação ou a precisão das previsões [20].

Redes Neurais Recorrentes (RNN)

Redes neurais recorrentes (RNN - *Recurrent Neural Network*), são redes neurais construídas para trabalhar com dados sequenciais, mantendo um estado interno de memória [33]. Esta memória captura informações sobre elementos anteriores da sequência, permitindo que decisões atuais considerem contextos passados. A RNN é ideal para tarefas de processamento de linguagem natural, reconhecimento de fala e previsão de séries temporais. Entretanto, este modelo pode sofrer com os problemas de desvanecimento e explosão de gradientes. Para lidar com isto, foram desenvolvidas variações de RNN chamadas *Long Short-Term Memory*

(LSTM) e *Gated Recurrent Units* (GRU) [9]. As LSTM RNN e GRU RNN conseguem capturar dependências temporais de curto e longo prazo que são importantes para identificar padrões associados à distúrbios da voz.

- LSTM: apresenta uma arquitetura que introduz mecanismos de portas para controlar o fluxo de informações, mitigando problemas de desvanecimento de gradientes e permitindo a captura de dependências de longo prazo.
- GRU: é uma variante mais simples que também utiliza portas para gerenciar informações, oferecendo desempenho comparável ao LSTM com menor complexidade computacional.

Modelos Pré-Treinados

Redes neurais pré-treinadas nada mais são que modelos de redes neurais que já passaram por um processo de treinamento inicial em um grande conjunto de dados amplo e genérico, o que permite que essas aprendam representações robustas e generalizáveis. Esses modelos podem ser reutilizados ou ajustados para resolver tarefas específicas em outros problemas, muitas vezes relacionados através do *transfer learning* [37]. O *transfer learning* pode ser baseado em extração de características ou ajuste fino (*fine-tunning*) [2]:

1. Extração de características: Ocorre o aumento da arquitetura original, adicionando camadas ao final da rede pré-treinada. Essas novas camadas são responsáveis por extrair características específicas do problema em questão. Apenas essas camadas adicionais necessitam ser treinadas [46].
2. Ajuste Fino: Algumas ou todas as camadas do modelo pré-treinado são "descongeladas"para que seus pesos possam ser atualizados e a camada de saída é substituída para adaptar-se à tarefa desejada.

As vantagens de se utilizar redes pré-treinadas são: economia de tempo (reduz o tempo necessário para desenvolver e treinar um modelo), melhor desempenho (aproveita o aprendizado de características avançadas que seriam difíceis de aprender em um conjunto de dados limitado), melhoria da eficiência dos dados (requer menos dados de treinamento para alcançar bons resultados em um novo problema).

Por outro lado, as desvantagens são: dependência da base de dados original (a qualidade e relevância do aprendizado depende dos dados usados no treinamento inicial), risco de *overfitting*, tamanho do modelo (a maioria dos modelos exigem muitos recursos de *hardware* para serem ajustados e utilizados).

Um exemplo de modelo de rede pré-treinada é o modelo VGG16, que foi proposto por Simonyan et al. (2014) da Universidade de Oxford, no artigo *Very Deep Convolutional Networks for Large-Scale Image Recognition* [54]. Esta é uma rede neural convolucional que possui uma profundidade de 16 camadas treináveis e foi projetada para tarefas de classificação de imagens. Seu treino ocorreu no conjunto de dados ImageNet, que contém mais de 14 milhões de imagens distribuídas em 1000 (mil) categorias. Sua popularidade deve-se à disponibilidade de seus pesos pré-treinados, facilitando a adaptação para novas tarefas por meio de aprendizado por transferência.

2.3 Métricas e Resultados

- Acurácia: mede a proporção de previsões corretas feitas por um modelo em relação ao total de previsões realizadas. Entretanto, é sensivel ao desbalanceamento de classes (cenário que não ocorre nesta pesquisa).

- Precisão: avalia a proporção de previsões positivas feitas pelo modelo que estão corretas. É importante quando as consequências de prever algo incorretamente como positivo são críticas.

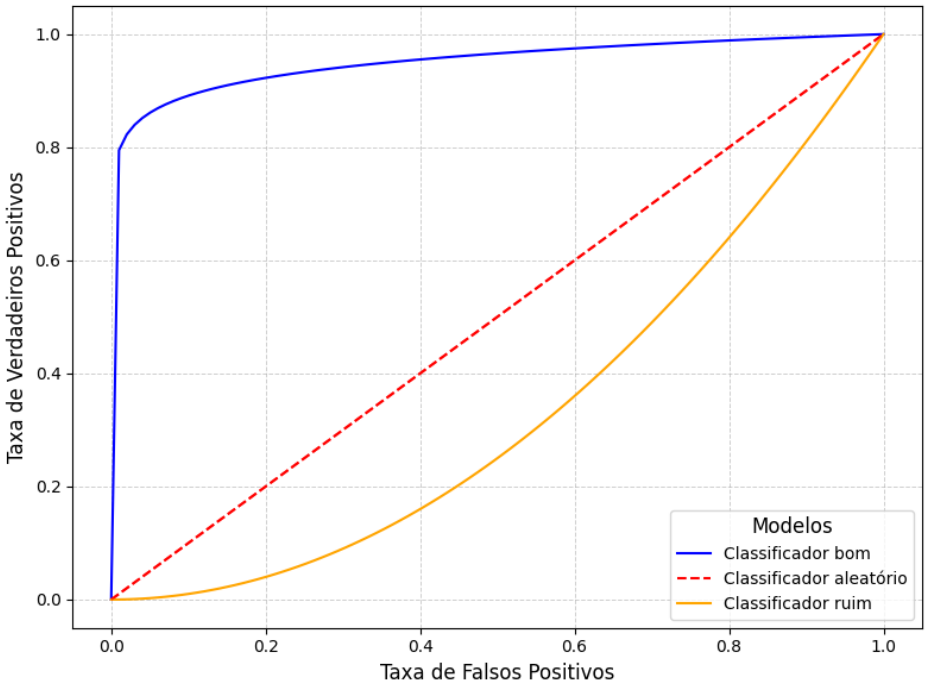
- *Recall* (sensibilidade): avalia a exatidão das previsões positivas feitas pelo modelo. É importante quando perder exemplos positivos é crítico.

- *F1-score*: combina as métricas de precisão e de *recall* em um único valor, representando a média harmônica entre essas. É valiosa quando se faz necessário equilibrar a importância de minimizar falsos positivos e falsos negativos.

Algumas representações visuais como as curvas ROC ajudam a interpretar os resultados. A Figura 2.8 ilustra a curva ROC. Para interpretar esta curva, é necessário entender que quanto mais próximo do canto superior esquerdo, melhor é o desempenho do modelo

(se a área abaixo da curva (AUC - *area under the curve*) se aproximar de 1,0, indica um desempenho excelente enquanto que próximo de 0,5, indica um modelo fraco que poderia ser substituído por um classificador aleatório).

Figura 2.8: Curva ROC.



Fonte: Autoria própria.

Outra representação muito útil é a matriz de confusão, que busca apresentar quantitativamente os acertos e erros nas previsões, ou seja, descreve o desempenho de um modelo de classificação, comparando suas previsões com os valores reais. Na Tabela 2.1, tem-se uma breve descrição de uma matriz de confusão, com a informação do que cada uma de suas células representa.

Tabela 2.1: Matriz de Confusão.

	Predição Positiva	**Predição Negativa**
Real Positivo	Verdadeiros Positivos (VP)	Falsos Negativos (FN)
Real Negativo	Falsos Positivos (FP)	Verdadeiros Negativos (VN)

Fonte: Autoria Própria.

A seguir, a descrição dos elementos da Matriz de Confusão.

- Verdadeiros Positivos (VP): o número de instâncias corretamente classificadas como positivas (o modelo previu positivo e o valor real era positivo). Para a análise em questão, ocorre quando o modelo classifica uma voz como saudável e a voz é realmente saudável.

- Falsos Positivos (FP): o número de instâncias incorretamente classificadas como positivas (o modelo previu positivo, mas o valor real era negativo).

- Falsos Negativos (FN): o número de instâncias incorretamente classificadas como negativas (o modelo previu negativo, mas o valor real era positivo).

- Verdadeiros Negativos (VN): o número de instâncias corretamente classificadas como negativas (o modelo previu negativo e o valor real era negativo).

- Diagonal principal (VP e VN): indica as previsões corretas, ou seja, os casos em que o modelo acertou a classe. Quanto maior é a soma dos valores dessa diagonal, melhor é o desempenho do modelo.

- Diagonal secundária (FP e FN): indica os erros cometidos pelo modelo.

2.4 Considerações Finais

Neste capítulo, foram apresentados os principais conceitos e métodos relacionados a um analisador automático da condição da voz, evidenciando suas etapas. Métodos para aumento de dados, extração de características, e tipo de classificadores foram evidenciados. No próximo capítulo é apresentada uma revisão bibliográfica das pesquisas correlatas.

Capítulo 3

Pesquisas Correlatas

No presente capítulo é apresentada a revisão bibliográfica sistemática realizada sobre processamento automático de sinais de voz para detecção e classificação de distúrbios da fala, focando em pesquisas correlatas que utilizaram a base de dados SVD.

3.1 Pesquisas Selecionadas

A revisão bibliográfica foi realizada a partir do método de revisão narrativa. Esta revisão foi centralizada em artigos e *surveys* disponíveis nos motores de busca do Google Acadêmico, e nas bibliotecas digitais do IEEE (IEEExplore), da ACM e Elsevier. As *strings* de busca foram elaboradas a partir de termos: "*Speech analysis*", "*Pathological voice*", "*Pathology detection*", "*Pathology classification*", "*Voice Disorders*", "*Automatic voice condition analysis*", "*AVCA system*", "deteccção de distúrbios da voz", "detecção de patologias da voz", "classificação automática de patologias da voz". Trabalhos com mais de sete anos de publicação são descartados (ano de publicação mínimo 2017). Trabalhos que utilizaram a base *Saarbruecken Voice Database* (SVD) e/ou expõem uma revisão sistemática do estado da arte, para detecção e/ou classificação de distúrbios da voz, foram priorizados.

O primeiro passo da revisão foi a filtragem dos artigos a partir dos critérios de exclusão e inclusão. O segundo foi a realização de leituras exploratórias dos artigos selecionados. O terceiro passo foi a revisão dos conceitos básicos e essenciais dessa área, a exemplo de aspectos vocais (produção da voz, trato vocal, tom, vibração), pré-processamento, redução de dimensionalidade e redes neurais. O quarto passo foi o estudo das técnicas aplicadas nos

modelos descritos nos artigos.

No Quadro 1, apresentado no final desta seção, são sumarizadas as principais pesquisas selecionadas a partir da revisão da literatura, evidenciando a base utilizada, o classificador, características extraídas e os melhores resultados alcançados. Uma parcela significativa desses trabalhos foca em propor diferentes formas para construir os detectores e sugerem, como trabalhos futuros, a construção de classificadores capazes de identificar as doenças e/ou o grau de severidade, tais como: [38], [32] e [16]. Alguns outros já abordam a etapa de classificação da possível doença do paciente, a exemplo de: [34], [30], [20] e [15].

Os artigos mais relevantes para esta pesquisa foram [34] e [20]. O artigo *On the design of automatic voice condition analysis systems. Part I: Review of concepts and an insight to the state of the art* [25] não está no Quadro 1 mas é de grande importância, pois está voltado para uma revisão sistemática do estado da arte na área de sistemas automáticos de análise da condição da voz. O objetivo desse artigo é fornecer uma revisão dos sistemas de análise automática de condição de voz (AVCA), detalhando cada um de seus blocos constituintes, introduzir os principais conceitos relacionados aos distúrbios vocais e discutir sobre alguns fatores que afetam os AVCA. Ao longo do artigo é apresentado um estudo dos aspectos fonatórios da fala, abordando a disfonia. Conceitos relacionados à voz como entonação, volume, ressonância e sopro são descritos e comentados. Etapas básicas de um sistema AVCA (pré-processamento, caracterização, redução de dimensionalidade, aprendizado de máquina, teste/análise e decisão) são explorados. Também são citadas as bases de dados mais utilizadas.

A respeito da etapa de validação do classificador, alguns artigos reportam ter feito apenas a separação de parte da base de dados para treino e outra parte para teste. Esta abordagem não se mostra adequada pois não é possível admitir a generalização dos dados em cada amostra e, eventualmente, uma amostra pode não ser representativa, afetando a capacidade do classificador. A maioria dos artigos reporta o uso da técnica *K-fold Cross-Validation* (validação cruzada de k conjuntos). A técnica de *Leave One Out* também é mencionada. Curvas ROC e AUC foram métodos empregados por alguns autores como métricas de avaliação e são considerados adequados.

O artigo *Voice pathology detection using a two-level classifier based on combined cnn-rnn architecture* [34] se mostrou relevante ao contexto da pesquisa e foi utilizado como

base para o desenvolvimento da arquitetura de rede binária descrita nessa dissertação. No artigo é apresentado um modelo de classificador em cascata de dois níveis, sendo o primeiro, responsável por distinguir se o sinal é de uma voz masculina ou feminina e o segundo, para categorizar o sinal de voz entre saudável ou patológico. O classificador proposto consiste em uma rede neural híbrida CNN-RNN alimentada por características ZCR, RMSE e MFCC extraídas dos sinais de voz originados da base SVD.

No modelo proposto, a CNN é responsável por capturar características espaciais dos dados de entrada, enquanto a RNN captura dependências temporais dos sinais. Essa combinação permite que o modelo processe e classifique os sinais para detecção de distúrbios. Os resultados alcançados por este artigo foram uma acurácia de 88,83% e um *F1-score* de 87,39%. Como trabalhos futuros, sugere-se expandir a quantidade de dados (utilizando vogais além da vogal sustentada "a") para dispor de mais variações dos sinais, tornando o modelo mais robusto e também analisar a possibilidade de se explorar a integração de recursos adicionais, como informações prosódicas e linguísticas, para melhorar ainda mais a precisão da classificação. O estudo dessa dissertação faz uso da mesma arquitetura proposta nesse artigo, aumentando a quantidade de dados utilizada e evoluindo o modelo para realizar também a etapa de classificação do distúrbio.

Os autores do artigo *Classification of Voice Disorders Using a One-Dimensional Convolutional Neural Network* [21] propõem uma nova abordagem que integra a reconstrução do espaço de fase com CNNs para melhorar a classificação de vozes normais e com distúrbios. O modelo proposto é similar ao VGG e aproveita a reconstrução do espaço de fase para converter sinais de voz em representações de alta dimensão, que podem ser efetivamente processadas pela CNN. Os conceitos de *back-propagation*, *cross-entropy* e gradiente estocástico descendente são apresentados. Este artigo inova ao utilizar a combinação da reconstrução do espaço de fase com um modelo baseado em CNN, aprensentando resultados de acurácia de 88,83%.

O artigo *Experimental Evaluation of Deep Learning Methods for an Intelligent Pathological Voice Detection System Using the Saarbruecken Voice Database* [38] propõe o uso de múltiplas técnicas de aprendizagem profunda em combinação com características extraídas de amostras vocais para distinguir entre vozes normais e patológicas. O artigo foca na utilização de coeficientes cepstrais de frequência mel (MFCCs) e parâmetros estatísticos de

ordem superior (HOS) aplicados a redes RNN e CNN e realiza os experimentos separados para os gêneros masculino e feminino. O artigo traz contribuição ao utilizar parâmetros de alta ordem como forma de extração de características.

No artigo *Voice disorder identification by using Hilbert-Huang transform (HHT) and Knearest neighbor (KNN)*[8] foi realizada a extração de características utilizando LPCC e HHT, gerando no total 21 coeficientes. O estudo apresenta comparação entre os classificadores KNN, *random forest* e *extra trees* atingindo precisão de 93% e *F1-score* de 94%, ao utilizar o KNN. Sua contribuição consiste em apresentar um método utilizando abordagens clássicas que atinge o patamar de 90% de acurácia utilizando um conjunto de dados de apenas 208 registros.

O artigo *A comparison of cepstral features in the detection of pathological voices by varying the input and filterbank of the cepstrum computation* [51] aborda o desafio de detectar distúrbios vocais, como disfonia, doença de Parkinson, laringite e insuficiência cardíaca, utilizando características cepstrais. Seu intuito é comparar o desempenho de variantes do MFCC por intermédio da alteração de dois fatores: os dados de insumo (sinal de voz e, sinal glotal e do trato vocal extraídos através do método de filtragem inversa glótica na fase quase fechada) e o banco de filtros (filtros de frequência mel e de frequência linear) no cálculo do cepstrum. A análise revelou que, embora os coeficientes cepstrais baseados em sinais de voz produzissem resultados fortes para um banco de dados específico, as características derivadas da fonte glótica e do trato vocal frequentemente os superavam. A contribuição do artigo se dá pela descoberta/indício de uma vantagem dos bancos de filtros lineares sobre os bancos de filtros mel, particularmente ao extrair características de sinais do trato vocal apresentando taxas de acurácia da ordem de 95% com a utilização de um classificador CNN.

O artigo *Automatic Classification of Neurological Voice Disorders using Wavelet Scattering Features* [59] destaca a importância da detecção precisa de distúrbios de voz para diagnóstico e tratamento eficazes. A classificação de distúrbios vocais, como disfonia espasmódica (DS) e paralisia do nervo laríngeo recorrente (RLNP), apresenta desafios devido às características diferenciadas das variações da fala causadas por deficiências neurológicas. O estudo tem como objetivo demonstrar a eficácia do WST na distinção entre vozes saudáveis e desordenadas (classificação binária) e entre diferentes tipos de distúrbios vocais (classificação multiclasse) e conclui que o WST não apenas melhora a capacidade de cap-

turar características relevantes de distúrbios de voz, mas também melhora a generalização dos modelos de classificação. Em essência, o artigo contribui para a área ao apresentar um novo método de extração de características (WST) que supera as técnicas tradicionais no contexto da classificação de distúrbios de voz, abordando assim lacunas na literatura e nas metodologias existentes.

O artigo *Voice pathology detection using convolutional neural networks with electroglottographic (EGG) and speech signals* [30] aborda as limitações dos métodos convencionais de detecção de patologias de voz, enfatizando que informações críticas podem ser perdidas nas etapas de pré-processamento e filtragem. Além disso, as técnicas tradicionais dependem frequentemente de características acústicas selecionadas, levando potencialmente à perda de dados relevantes. Diante desse problema os autores propõem o uso de sinais eletroglotográficos (EGG) e sinais de fala como entrada para uma rede CNN a fim de aumentar a precisão da detecção. Este artigo destaca o poder do aprendizado profundo ao tentar minimizar a etapa de pré-processamento. Entretanto, embora a CNN possa extrair características automaticamente, ela pode não capturar características patológicas únicas, que seriam reveladas por métodos específicos de extração de características.

Por fim, vale ainda citar o artigo *Pathological Voice Detection Based on Phase Reconstitution and Convolutional Neural Network* [20], o qual descreve um novo método para distinguir entre vozes normais e patológicas, aproveitando a reconstrução do espaço de fase aliada às redes neurais convolucionais. A rede descrita é alimentada com imagens da trajetória do sinal de voz no espaço de fase bidimensional. Neste artigo, são realizados experimentos com dados de três bases distintas (MEEI, SVD e uma base de autoria própria), com os seguintes resultados: 99,42% de acurácia na base MEEI, 97,30% na base SVD e 95,88% na base desenvolvida. Como trabalhos futuros, menciona-se aumento dos dados para e o uso de discurso contínuo, visando aumentar o poder de generalização do modelo.

Diante do exposto, conclui-se que há possibilidades para melhorias das bases de dados, pois ainda é necessário registrar conjuntos de dados maiores, mais equilibrados em termos de distúrbios, idade ou sexo. A revisão da literatura também revelou que a maioria dos trabalhos emprega fonação sustentada devido à sua simplicidade, apesar do potencial que a fala contínua apresenta. Verificou-se, também, que os efeitos da extralinguística ou paralinguística (como idade, sexo, sotaque), na grande maioria dos sistemas relatados, raramente são

considerados.

Nesta dissertação, é apresentada uma abordagem baseada em redes pré-treinadas cujos dados de entrada são espectrogramas mel derivados do sinal de voz, objetivando a detecção de vozes com distúrbios e, outra abordagem, baseada em redes neurais híbridas (CNN-RNN LSTM), que combina técnicas para aumento de dados, para a classificação de vozes com distúrbios, considerando separadamente, vozes masculinas e femininas, o que proporciona uma análise mais específica e precisa.

Diferentemente das pesquisas correlatas apresentadas, o estudo ora descrito trata da classificação entre distúrbios. Dessa forma, ao responder o principal questionamento da pesquisa, busca-se contribuir para o estado da arte e, incentivar outros pesquisadores a se aprofundar na classificação dos distúrbios.

Quadro 1: Análise comparativa das pesquisas correlatas.

Artigo	**Ano**	**Base**	**Dados**	**Classificador**	**Características Extraídas**	**Resultados (acurácia)**
Automatic classification of neurological voice disorders using wavelet scattering features [59]	2024	SVD	1800	SVM e RNN	MFCC, eGeMAPS, IS2010 features, WST features	80,83% (SVM saudável x SD); 82,58% (FNN saudável x SD).
Voice Pathology Detection Using a Two-Level Classifier Based on Combined CNN–RNN Architecture[34]	2023	SVD	654 originais	CNN-RNN	ZCR, RMSE, MFCC	88,83% (binário); 80,70% (multiclasse)
Pathological Voice Detection Based on Phase Reconstitution and Convolutional Neural Network [20]	2022	MEEI, SVD e base de autoria própria	8939	VGG (CNN)	Imagem da trajetória do sinal de fases em três dimensões	MEEI: 96,04%; SVD: 92,27%

Artigo	**Ano**	**Base**	**Dados**	**Classificador**	**Características Extraídas**	**Resultados (acurácia)**
Voice pathology detection using convolutional neural networks with electroglottographic (EGG) and speech signals [30]	2022	SVD	215	CNN	Sinail de voz e sinal EGG	80,30% (voz); 88,67% (EGG)
Classification of Voice Disorders Using a One-Dimensional Convolutional Neural Network [21]	2022	Base de autoria própria	1377	CNN	Sinal de voz particionado	88,83%
Experimental Evaluation of Deep Learning Methods for an Intelligent Pathological Voice Detection System Using the Saarbruecken Voice Database [38]	2021	SVD	518	RNN e CNN	MFCC, LPCC e HOS	80,30% (sinal de voz); 82,69% (CNN com LPCC)

Artigo	**Ano**	**Base**	**Dados**	**Classificador**	**Características Extraídas**	**Resultados (acurácia)**
Voice disorder identification by using Hilbert-Huang transform (HHT) and K nearest neighbor (KNN) [8]	2021	Voice ICar federico II	114	KNN	LPCC, 12 características HHT	93,30%
A comparison of cepstral features in the detection of pathological voices by varying the input and filterbank of the cepstrum computation [51]	2021	HUPA, Neurovoz, PC-GITA, SVD-HkD, SVD-Laryngitis, HF	-	SVM e CNN	MFCC, LPCC, QCCC, PLPCC	95,44% (CNN)

Artigo	**Ano**	**Base**	**Dados**	**Classificador**	**Características Extraídas**	**Resultados (acurácia)**
Detecção de patologias laríngeas com base na análise dinâmica de sinais de voz utilizando redes neurais profundas [16]	2020	SVD	643	RNN	Sinal particionado	85,55% ± 4,39%
Discriminação entre sinais de vozes saudáveis e patológicos por meio da análise da imagem do espaço de fase reconstruído [15]	2017	MEEI	149	MLP	Contagem de caixas ponderadas e método da similaridade.	99,00%

Fonte: Autoria própria.

3.2 Considerações Finais

Neste capítulo, foi apresentada a revisão bibliográfica adotada, com destaque para os avanços, desafios e limitações das abordagens selecionadas. Em seguida, foram analisadas as possibilidades de melhorias e a contribuição que o presente trabalho traz. No próximo capítulo, será especificada a metodologia adotada para desenvolver a pesquisa em questão, evidenciando a descrição das etapas necessárias à construção das etapas para os processos de detecção e classificação dos sinais, com ênfase às decisões para definição da quantidade de dados, camadas das redes neurais, parâmetros e funções utilizadas.

Capítulo 4

Metodologia

Neste capítulo, são apresentados os estudos realizados na pesquisa, evidenciando-se as etapas de pré-processamento, extração de características, arquitetura da rede, separação de dados e validação dos treinamentos. Para cada etapa, evidencia-se a definição de métodos e funções utilizadas a fim de tornar a pesquisa reprodutível.

4.1 Abordagens Desenvolvidas

A pesquisa teve como intuito o desenvolvimento de abordagens capazes de auxiliar os profissionais da saúde no diagnóstico de distúrbios da voz. Para tanto, foram desenvolvidos duas abordagens. A primeira trata da utilização de redes neurais convolucionais pré-treinadas, alimentando-as com espectrogramas gerados a partir do sinal de voz. A segunda abordagem, trata da utilização de uma rede neural híbrida (CNN e RNN) alimentada por matrizes de dados geradas a partir da extração das seguintes características: *Zero Crossing Rate*, *Root-Mean-Square Energy* e *Mel-Frequency Cepstral Coefficients*. Para ambas, foi realizada a separação entre dados masculinos e femininos para que as redes neurais pudessem ser mais precisas, visto que, como comentado anteriormente, sinais de voz do sexo masculino apresentam frequências fundamentais mais baixas (70-150 Hz) em comparação com mulheres (150-250 Hz). A primeira abordagem objetiva detecção de sinais de voz com distúrbios enquanto que a segunda objetiva tanto a detecção quanto a classificação.

As duas abordagens foram implementadas na linguagem *Python*, com bibliotecas para aprendizagem de máquina, como *scikit-learn*[45] e *TensorFlow* [1].

4.1.1 Abordagem com Rede Pré-Treinada

Na primeira abordagem, foram tratados os problemas relacionados ao volume de dados, necessidade de poder computacional elevado e *overfitting*. Como a quantidade de dados presente na base SVD é escassa, uma rede pré-treinada poderia superar estes desafios, pois com a realização do treinamento prévio, aprenderia características gerais do domínio e suas camadas mais profundas já estariam regularizadas com conhecimento generalizado. Assim, mesmo possuindo poucos dados, ao utilizar a rede neural pré-treinada, aplicando a técnica de *transfer learning* [27] (congelando os pesos e treinamento de camadas mais profundas, adicionando novas camadas específicas para o problema e ajustando-as), é possível que a rede neural aprenda padrões específicos da tarefa/classificação desejada, utilizando-os em conjunto com os padrões genéricos.

Desse modo, essa abordagem permitiu tratar questionamentos como:

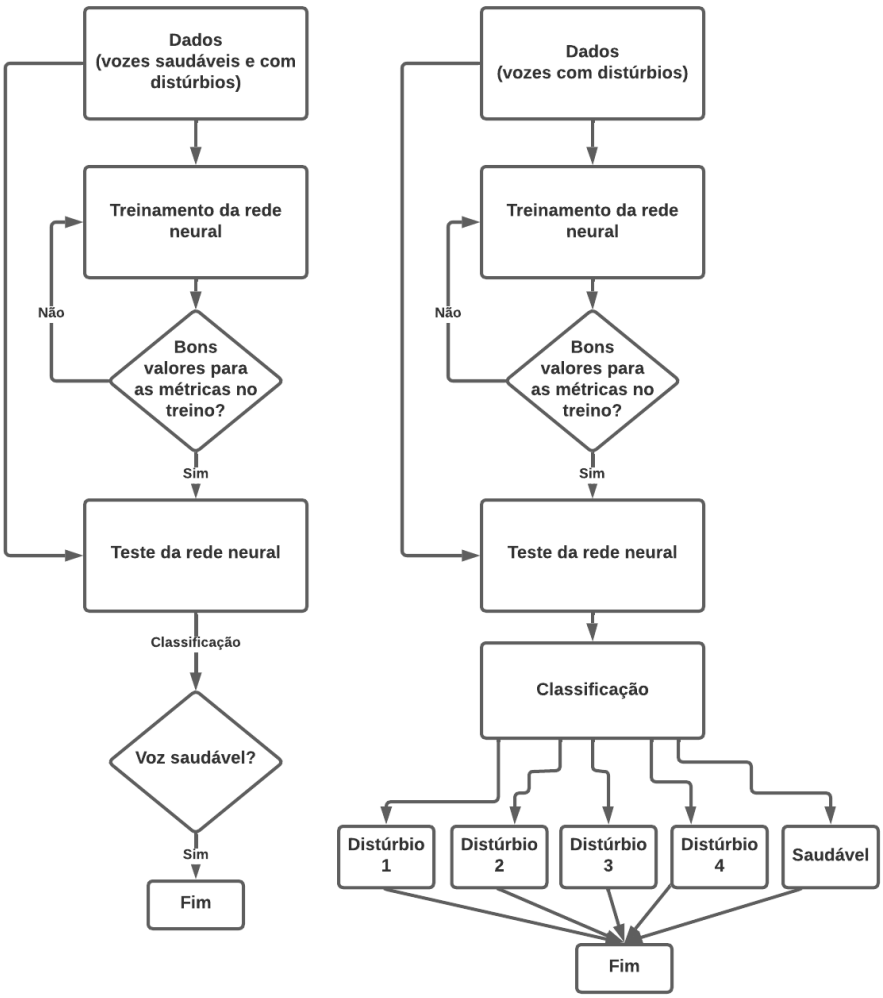
- Espectrogramas Mel são suficientes para que uma rede neural consiga distinguir entre sinais de voz com distúrbios e vozes saudáveis?
- Redes pré-treinadas conseguem se ajustar bem a este problema de classificação binária?

4.1.2 Abordagem Rede Híbrida CNN-RNN

A respeito do segundo estudo, a rede híbrida buscou aliar características da CNN, tais como, capacidade de processar dados estruturados em grades (matrizes) e capacidade de identificar padrões independentemente de onde estão localizados na entrada [7], com a característica de memória da rede neural recorrente, RNN.

A abordagem adotada consistiu em criar dois classificadores isolados (um binário e um Multiclasse). O binário é responsável por distinguir se um sinal de voz é classificado como patológico ou saudável, o multiclasse é responsável por classificar o sinal de voz entre as seguintes categorias: disfonia, laringite, paralisia, pólipo ou saudável. Esse segundo estudo, buscou utilizar a mesma proposta de arquitetura utilizada por ksibi (2023) a fim de utilizá-lo como base para comparação. Na Figura 4.1, é apresentado um modelo simplificado da abordagem desenvolvida.

Figura 4.1: Modelo dos classificadores binário e Multiclasse.



Fonte: Adaptado de [30].

4.2 Coleta e Análise dos Dados

Segundo Garcia (2019) [25], que apresenta uma revisão sistemática da literatura, são apresentadas as principais bases utilizadas em pesquisas na área, sendo essas: *Massachussets Ear and Eye Infirmary* (MEEI), *Saarbrücken Voice Database* (SVD), Hospital Príncipe de Asturias (HUPA), *Arabic Voice pathology Dataset* (AVD), e a base de dados obtida nos hospitais de Marseilles e Aix-en- Provence (França).

Outra base que vem ganhando visibilidade é a AVFAD (*Advanced Voice Function As-*

sessment Databases) [31] desenvolvida em Portugal, em 2017, com dados de 709 pacientes, dos quais 346 foram diagnosticados com distúrbios da voz e 363 com vozes saudáveis. Esta base apresenta registros vocais de fonação sustentada das vogais /a/, /i/ e /u/ além da leitura de frases, texto e fala espontânea de cada indivíduo, totalizando 8648 arquivos de áudio, registros da extração de 19 parâmetros através da função *Voice Report* do Praat e registros de 16 parâmetros clínicos do paciente.

Sabe-se que a construção de uma base de dados de qualidade demanda uma série de fatores, que podem dificultar esse processo, tais como: ambiente restrito, controle de ruído, bons equipamentos para captura do sinal, tempo disponível do falante para repetição de gravação com entonações diferentes, cuidado para não haver estresse/cansaço da voz.

Dessa forma, buscando limitar dificuldades relacionadas à aquisição de uma base de dados, bem como facilitar a análise comparativa com outras pesquisas, o trabalho ora descrito fez uso da base de dados SVD. Esta base foi a escolhida pois é a que oferece uma melhor separação entre entonações (normal, grave, aguda, grave-alta-baixa), fonemas (/a/, /i/, /u/), sexo do falante e possui uma quantidade representativa de dados para alguns tipos de distúrbios, tais como: disfonia, laringite e paralisia. Outra vantagem dessa base reside no fato de que, além de oferecer o sinal de voz, também oferece o sinal de eletroglotografia (EGG). Os artigos Albadr (2024), Kisibi (2023), Tirronen (2024) e Yagnavajjula (2024) também fizeram uso da SVD.

Na Tabela 4.1, tem-se a descrição da quantidade de sinais de voz que foi extraída do repositório da base de dados SVD. Na pesquisa foram utilizados apenas sinais de voz saudáveis ou que foram diagnosticados com apenas um único distúrbio da voz. Embora a base SVD disponibilize áudios com a elocução das vogais /a/, /i/ e /u/, cada um em três tons (baixo, neutro e alto), por falante, foram extraídos apenas áudios referentes a elocução vocal /a/. No estudo de Cordeiro (2016) [11], é mencionado que a vogal /a/ apresenta, no domínio do tempo, picos de maior amplitude e menor abertura, facilitando a análise acústica. Também é destacado que a simplicidade na produção dessa vogal, torna sua pronúncia uma tarefa acessível para pacientes, independentemente de suas habilidades físicas ou cognitivas. Ademais, o uso dessa vogal sustentada é comum em diversos trabalhos sobre reconhecimento de vozes com distúrbios, servindo como padrão para comparação entre estudos.

Vale salientar ainda, que na tabela 4.1, nota-se o valor de 126 (destacado em negrito),

expressando a quantidade de sinais para a classe de paralisia, em tom baixo, do sexo feminino. Na base SVD, idealmente, todos os sinais de voz deveriam ter seus representantes em tom baixo, neutro e alto, entretanto o sinal de voz de identificador 1631 não é disponibilizado para o tom baixo. No Apêndice A, é apresentada uma lista com os nomes dos sinais de voz utilizados para cada categoria, evidenciando os seus identificadores, a fim de facilitar a reprodutibilidade do estudo.

Tabela 4.1: Quantidade de sinais de voz extraída da base SVD para cada classificação.

Diagnósticos dos Distúrbios	**Quantidade de sinais**					
	Feminino			**Masculino**		
	Baixo	**Neutro**	**Alto**	**Baixo**	**Neutro**	**Alto**
Disfonia	41	41	41	29	29	29
Laringite	32	32	32	50	50	50
Pólipo vocal	10	10	10	17	17	17
Paralisia	**126**	127	127	70	70	70
Subtotal	209	210	210	166	166	166
Sinais de voz saudáveis	428	428	428	259	259	259
Total	637	638	638	425	425	425

Fonte: Autoria própria.

4.3 Pré-Processamento dos Sinais

No âmbito da aprendizagem de máquina, o pré-processamento dos dados é uma etapa crucial para garantir que os modelos consigam aprender de forma eficiente e precisa. Ao lidar com sinais de voz, técnicas como remoção de silêncio e detecção de interrupção e de ruídos podem ser aplicadas para garantir que apenas os fragmentos gerados durante a vibração das pregas vocais sejam utilizados. Uma das justificativas para a escolha da base SVD é que esta base já apresenta tratamento para os sinais disponibilizados, mitigando assim a presença de intervalos de silêncio no sinal de voz.

Diante disso, para o primeiro estudo (abordagem utilizando redes pré-treinadas), a etapa

de pré-processamento limitou-se à seleção dos sinais de voz em tom neutro pois este tom reflete a frequência fundamental ($F0$) na qual as pregas vocais vibram mais confortavelmente. Foram utilizados todos os sinais de voz com distúrbios existentes para este tom. Para o sexo masculino, o total de sinais com distúrbios foi de 166 (cento e sessenta e seis) e, para o feminino, foi de 210 (duzentos e dez). Foi selecionada a mesma quantidade de sinais de voz saudáveis, de forma aleatória. Nenhum processo para aumento de dados foi realizado, pois desejava-se analisar se a rede pré-treinada iria proporcionar bons resultados mesmo com a pequena quantidade de dados.

Para o segundo estudo (abordagem com classificador CNN-RNN), foi definido que seriam utilizados 1.000 (mil) sinais de voz com distúrbios e 1.000 (mil) sinais de voz saudáveis para cada sexo, resultando em uma base de dados com 4.000 (quatro mil) sinais. Como demonstrado previamente na Tabela 4.1, os sinais com distúrbios, estão limitados ao máximo de 629 registros femininos e 498 registros masculinos (considerando os três tons). Portanto, fez-se necessário aplicar técnicas para aumento de dados.

As técnicas aplicadas foram: injeção de ruído branco (*white noise*), *time strech* e *time shifting*. Para aplicá-las foram utilizadas funções da biblioteca librosa [41] [1].

Na Tabela 4.2, são apresentadas as quantidades de dados existentes e criados, por categoria. A respeito dos sinais de vozes femininas, a quantidade que representa vozes saudáveis e com paralisia (evidenciados em negrito) ultrapassa o limite definido (mil sinais de áudio por categoria), portanto, para essas categorias não foi necessário realizar o aumento de dados, sendo selecionados 250 (duzentos e cinquenta) sinais de voz com paralisia e 1.000 (mil) sinais de vozes saudáveis, de forma aleatória.

4.4 Extração de Características

A extração de características é uma etapa por vezes importante, pois permite transformar dados brutos, muitas vezes complexos e de grande dimensionalidade, em informações mais relevantes e de fácil interpretação para o modelo. Ao extrair características significativas, como padrões, tendências ou informações discriminativas, é possível reduzir a variabilidade e o ruído nos dados, facilitando o processo de aprendizagem da rede neural [24]. Para o

[1] Librosa disponível em: https://librosa.org

Tabela 4.2: Quantidade de sinais de voz existentes e criados.

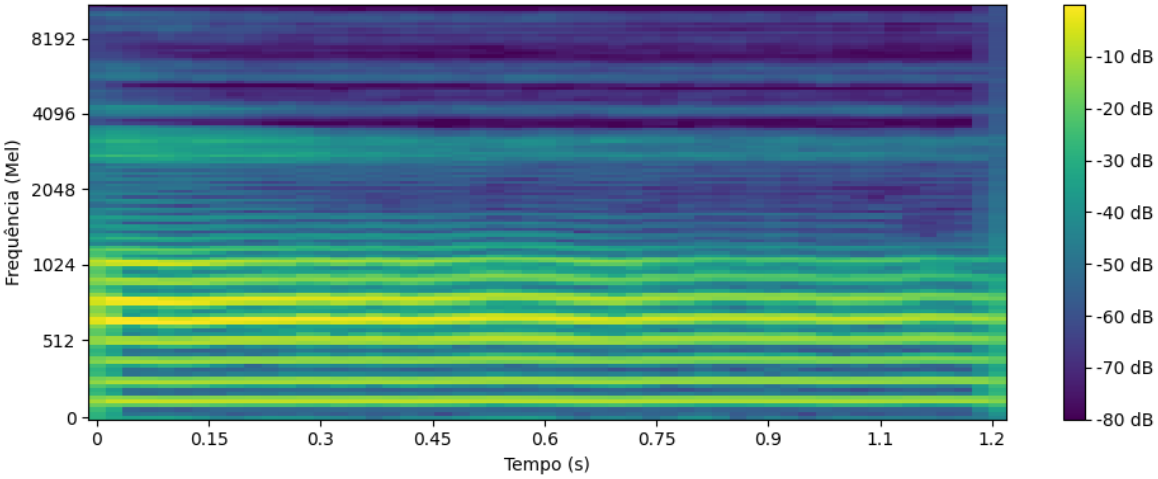
Diagnósticos dos Distúrbios	**Quantidade de sinais**			
	Feminino		**Masculino**	
	Existente	**Criado**	**Existente**	**Criado**
Disfonia	123	127	87	163
Laringite	96	154	150	100
Pólipo vocal	30	220	51	199
Paralisia	**380**	-	210	40
Sinais de voz saudáveis	**1284**	-	777	223

Fonte: Autoria própria.

estudo com redes pré-treinadas, a extração de características se deu por meio da geração dos espectrogramas mel. Conforme seção de Fundamentação Teórica (Capítulo 2), estes são uma representação visual do sinal de voz no domínio da frequência, utilizando a escala Mel, uma escala logarítmica projetada para refletir a percepção de frequências acústicas por humanos. Para extraí-los, foram utilizadas as seguintes funções da biblioteca librosa: *feature melspectrogram* e *power to db*.

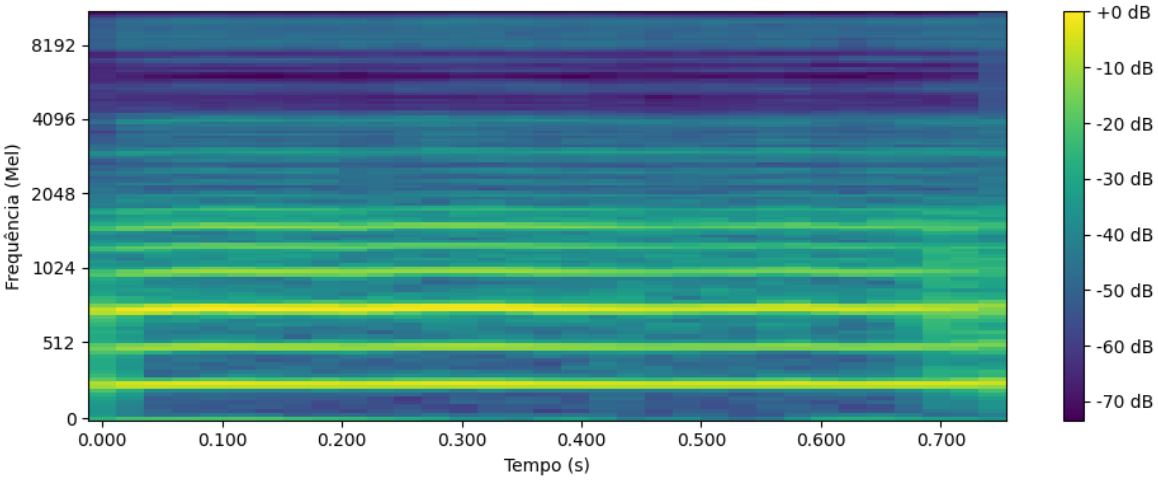
Nas Figuras 4.2 e 4.3 são apresentados espectrogramas mel gerados a partir de sinais de vozes masculinas e femininas, com laringite, pronunciados em tom neutro. Nas Figuras 4.4 e 4.5 são exibidos espectrogramas mel gerados a partir de sinais de vozes saudáveis.

Figura 4.2: Espectrograma Mel de voz masculina com laringite.



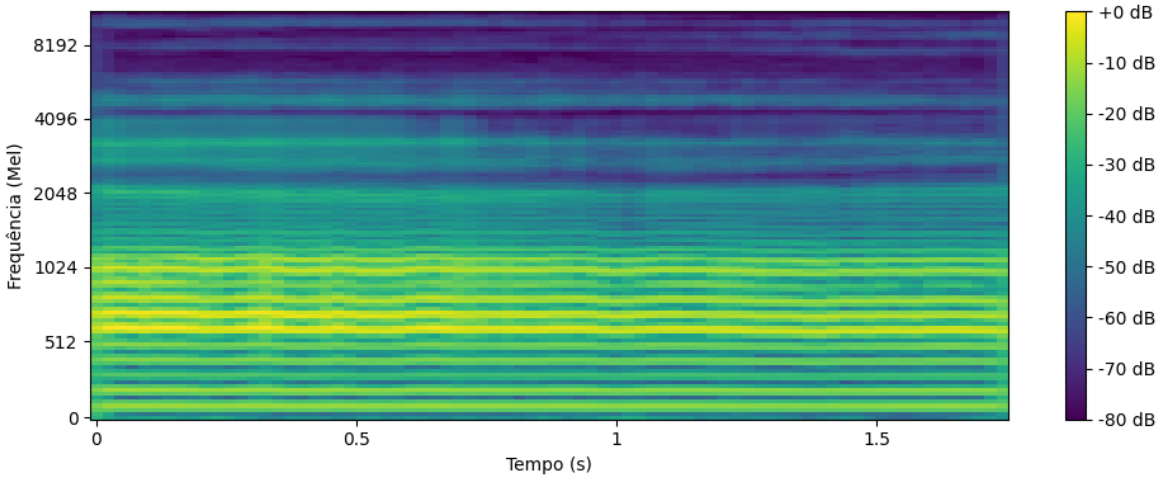
Fonte: Autoria Própria.

Figura 4.3: Espectrograma Mel de voz feminina com laringite.



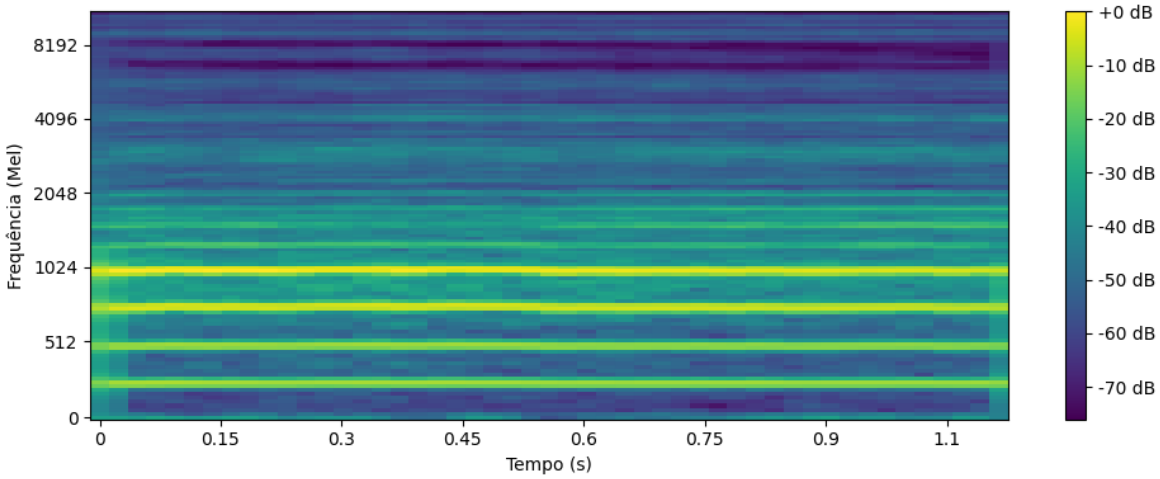
Fonte: Autoria Própria.

Figura 4.4: Espectrograma Mel de voz masculina saudável.



Fonte: Autoria Própria.

Figura 4.5: Espectrograma Mel de voz feminina saudável.



Fonte: Autoria Própria.

Para o estudo com a rede-híbrida, três tipos de características foram extraídas dos sinais com a finalidade de alimentar a rede neural, sendo essas: Taxa de Cruzamento por Zeros (*Zero Crossing Rate* - ZCR), Raíz Quadrática Média da Energia (*Root-Mean-Square Energy* - RMSE) e Coeficientes Cepestrais de Frequência Mel (*Mel-Frequency Cepstral Coefficients - MFCC*). Para extrair o ZCR e o RMSE foram utilizadas as funções *feature zero crossing rate* e *feature rmse* da biblioteca librosa. Para os MFCC, foi utilizada a função *mfcc* da biblioteca *python speech features*, com os seguintes parâmetros (escolhidos baseado em estudo prévio

[5] destinado à criação de um verificador de locutor utilizando coeficientes MFCC):

- *Number of cepstral coefficients*: 20;
- *Number of frames to consider*: 158;
- *Pre-emphasis*: 0,97;
- *Sample Rate*: Original da base SVD: 50 kHz;
- *FFT Size*: 2048;
- *Lifter*: 0,0
- *Window Length*: 0,025;
- *Window Function*: Hamming;
- *Append Energy*: *True*;
- *Window Step*:

$$\text{winstep} = \frac{audiolen - winlen}{num.\ of frames\ to\ consider - 1}. \tag{4.1}$$

Na Tabela 4.3, são apresentadas as pesquisas relacionadas que fizeram uso dos MFCC.

Tabela 4.3: Pesquisas correlatas que utilizaram MFCC.

Artigo	**Características Extraídas**
Voice Pathology Detection Using a Two-Level Classifier Based on Combined CNN–RNN Architecture [34]	MFCC, ZCR, RMSE
Experimental Evaluation of Deep Learning Methods for an Intelligent Pathological Voice Detection System Using the Saarbruecken Voice Database [38]	MFCC, LPCC, HOS (estatísticas de alta ordem)

Continua na próxima página.

Artigo	**Características Extraídas**
Hierarchical Multi-Class Classification of Voice Disorders Using Self-Supervised Models and Glottal Features [55]	MFCC, delta (diferença de 1ª ordem) MFCC e delta-delta (diferença de 2ª ordem) MFCC
The Effect of the MFCC Frame Length in Automatic Voice Pathology Detection [56]	MFCC
Automatic classification of neurological voice disorders using wavelet scattering features [59]	MFCC, eGeMAPS, transformada de espalhamento de onda e conjunto de características do desafio paralinguístico INTERSPEECH 2010

Fonte: Autoria própria.

4.5 Arquiteturas das Redes Neurais

A descrição das arquiteturas das redes utilizadas na pesquisa é apresentada a seguir.

4.5.1 Rede Pré-Treinada

Para o estudo com redes pré-treinadas convolucionais, foi selecionado o modelo de rede VGG16 disponibilizado na biblioteca Keras por ter sido o mesmo modelo de rede pré-treinada, utilizado por Fu et al. (2022), para distinguir entre vozes saudáveis e com distúrbio, através da reconstituição do espaço de fases do sinal de voz, atingindo acurácia de 92,27%. Na Tabela 4.4, é apresentada uma comparação entre os modelos disponíveis nessa biblioteca (as informações foram extraídas da documentação oficial da biblioteca). As acurácias referem-se ao desempenho do modelo no conjunto de dados de validação ImageNet e o tempo por etapa de inferência é a média de 30 *batches* e 10 repetições.

A VGG16 é uma rede neural convolucional desenvolvida pelo *Visual Geometry Group da Universidade de Oxford* e caracteriza-se por sua profundidade de 16 (dezesseis) camadas treináveis. Esta rede foi projetada para tarefas de classificação de imagens e foi treinada com o conjunto de dados *ImageNet*, que contém mais de 14 milhões de imagens distribuídas em

Tabela 4.4: Comparação entre os modelos de redes neurais pré-treinadas da biblioteca Keras.

Model	***Size (MB)***	***Top-1 Accuracy***	***Top-5 Accuracy***	***Parameters***	***Depth***	***Time (ms) per inference step (CPU/GPU)***
Xception	88	79.0%	94.5%	22.9M	81	109.4 / 8.1
VGG16	528	71.3%	90.1%	138.4M	16	69.5 / 4.2
VGG19	549	71.3%	90.0%	143.7M	19	84.8 / 4.4
ResNet50	98	74.9%	92.1%	25.6M	107	58.2 / 4.6
ResNet50V2	98	76.0%	93.0%	25.6M	103	45.6 / 4.6
ResNet101	171	76.4%	92.8%	44.7M	209	89.6 / 5.2
ResNet101V2	171	77.2%	93.8%	44.7M	205	72.7 / 5.4
ResNet152	232	76.6%	93.1%	60.4M	311	127.4 / 6.5
ResNet152V2	232	78.0%	94.2%	60.4M	307	107.5 / 6.6
InceptionV3	92	77.9%	93.7%	23.9M	189	42.2 / 6.9
InceptionResNetV2	215	80.3%	95.3%	55.9M	449	130.2 / 10.0
MobileNet	16	70.4%	89.5%	4.3M	55	22.6 / 3.4
MobileNetV2	14	71.3%	90.1%	3.5M	105	25.9 / 3.8
DenseNet121	33	75.0%	92.3%	8.1M	242	77.1 / 5.4
DenseNet169	57	76.2%	93.2%	14.3M	338	96.4 / 6.3
DenseNet201	80	77.3%	93.6%	20.2M	402	127.2 / 6.7
NASNetMobile	23	74.4%	91.9%	5.3M	389	27.0 / 6.2
NASNetLarge	343	82.5%	96.0%	88.9M	533	344.5 / 20.0

Fonte: Adaptado da documentação do Keras.

1.000 (mil) categorias. A VGG16 tem sido aplicada em diversos estudos acadêmicos, dentre os quais:

- *Using VGG16 Algorithms for classification of lung cancer in CT scans Image* [62], no qual a rede foi utilizada para classificar imagens de tomografias computadorizadas de pulmão, alcançando 91% de acurácia.
- *Evaluation of Convolutional Neural Networks for COVID-19 Classification on Chest X-Rays* [63], no qual a VGG16 foi empregada para identificar pneumonia causada por COVID-19 em radiografias de tórax, obtendo uma acurácia de 85,11%.

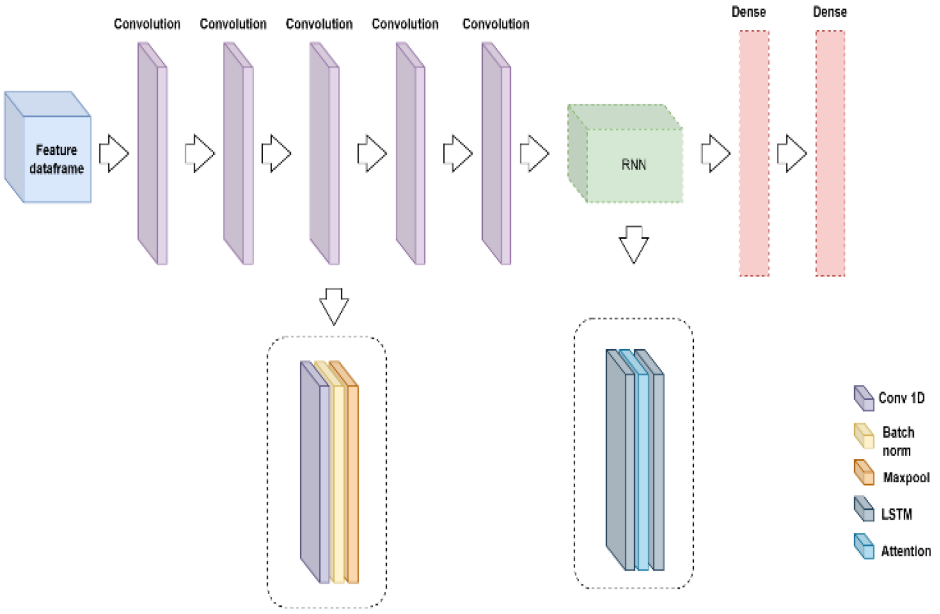
Para adaptar a rede VGG16 à pesquisa ora descrita foi gerado, em *Python*, o código disponível no *gitHub* [2]. A tabela B.1, presente no Apêndice B, apresenta as camadas (topologia) e parâmetros da rede VGG adaptada.

[2]https://github.com/gabriellmd/automatic-voice-condition-analyser

4.5.2 Rede CNN-RNN Binária

A respeito do estudo utilizando a rede neural híbrida CNN-RNN, a arquitetura foi montada baseando-se no artigo [34]. Na Figura 4.6, são ilustrados os blocos da rede híbrida.

Figura 4.6: Blocos da rede CNN-RNN.



Fonte: Extraído de [34].

Esta arquitetura tem início no modelo CNN. Assim, os primeiros cinco blocos são responsáveis pelas convoluções. Cada bloco de convolução possui uma camada de convolução com 64 (sessenta e quatro) neurônios, uma de *batch normalization*, *max pooling* e outra de *dropout*, descritas a seguir.

- A camada de *batch normalization* é utilizada para normalizar cada saída da camada anterior, ou seja, a média e o desvio padrão são calculados para cada *feature*. Os dados normalizados são ajustados para obtenção de uma média de 0 e um desvio padrão de 1. Ao evitar que as ativações se tornem muito grandes ou muito pequenas, o treinamento é estabilizado. A normalização também ajuda a rede a convergir mais rapidamente.

- O *max pooling* é utilizado para diminuir a complexidade do modelo e ajudar a controlar o *overfitting*. Para isso, essa camada reduz o tamanho das ativações, mantendo apenas as características mais relevantes.

- O *dropout* também serve para reduzir o risco de *overfitting* tornando a rede mais generalizável. Sua função é desativar aleatoriamente um percentual de neurônios em cada iteração, ou seja, fazer com que algumas conexões da rede não sejam usadas durante o treinamento, forçando o modelo a aprender representações mais robustas e menos dependentes de neurônios específicos.

Após os cinco blocos, os dados passam para uma arquitetura de rede recorrente, que é composta por uma camada LSTM, uma camada de atenção e termina com outra camada LSTM. A camada LSTM é característica de redes RNN e é responsável por adicionar o conceito de 'memória' à rede. Sua função é capturar a dependência de uma informação, em relação a uma informação prévia.

A camada de atenção permite que o modelo se concentre em partes específicas da entrada, ao invés de processar toda a sequência de maneira igualitária. Para fazer isso, esta camada define pesos distintos para diferentes partes da entrada. Essa técnica é particularmente útil quando os dados contêm informações relevantes em diferentes locais.

Após o bloco da RNN, ocorre a redução dos neurônios, de 128 (cento e vinte e oito) para 64 (sessenta e quatro), aplicando uma nova camada densa com taxa de *dropout* de 0,5 (50% dos neurônios são desativados). A camada final do modelo é adicionada, possuindo apenas um neurônio e ativação *sigmoid*, visto que a classificação final deve ser binária.

O modelo é compilado com o otimizador Adam, utilizando as métricas de acurácia binária, precisão e *recall*. O código fonte está disponível no *gitHub* [3] e a tabela B.2, presente no Apêndice B, apresenta as camadas e os parâmetros da rede CNN-RNN binária desenvolvida.

4.5.3 Rede CNN-RNN Multiclasse

Esta rede foi utilizada para a classificação dos sinais de voz entre os seguintes distúrbios: disfonia, laringite, pólipo, paralisia. Adicionalmente, a categoria saudável também foi considerada. Essa rede consiste numa variação da rede binária. Para tal, foram realizadas as modificações descritas a seguir.

1. A CNN possui seis blocos convolucionais. Cada camada de convolução possui 256 (duzentos e cinquenta e seis) neurônios. Foi adicionada mais uma camada e aumentado

[3] https://github.com/gabriellmd/automatic-voice-condition-analyser

o número de neurônios para que a rede ficasse mais robusta, aprofundando mais seu aprendizado e criando mais conexões entre neurônios.

2. Na última camada densa, antes da saída, o número de neurônios decresceu de 256 (duzentos e cinquenta e seis) para 128 (cento e vinte e oito), a fim de reduzir a dimensionalidade.

3. A camada final possui cinco neurônios pois precisa classificar o dado entre cinco categorias.

4. O modelo foi compilado com a função de perda *categorical crossentropy*, pois é o modelo indicado para tarefas de multiclassificação. Foi utilizado uma *learning rate* de 0,0005.

O código fonte está disponível no *gitHub* e a tabela B.3, presente no Apêndice B, apresenta as camadas e os parâmetros da rede CNN-RNN multiclasse desenvolvida.

4.6 Treinamentos das Redes

Os treinamentos das redes foram realizados numa máquina com sistema operacional Windows 11, placa mãe B550M AORUS ELITE (com placa de vídeo integrada), processador AMD Ryzen 7 5700G, com Radeon Graphics 3,80 GHz e 16 GB de memória RAM. Todos os modelos foram desenvolvidos utilizando a linguagem *Python* (possui várias bibliotecas para suporte ao aprendizado de máquina).

4.6.1 Divisão dos dados (Treinamento, Validação e Teste)

No três modelos, os dados foram divididos em treinamento, validação e teste e foi adotada uma validação cruzada de ordem dez. Após a validação cruzada, o modelo foi re-treinado com todos os dados de treinamento e de validação, para aproveitar ao máximo os dados disponíveis. Os dados de teste foram utilizados apenas no modelo final, o que forneceu uma estimativa imparcial do desempenho do modelo em dados completamente novos.

Para as redes VGG16, como não foi utilizado o processo de *data augmentation*, a divisão foi feita em 72% para treinamento, 8% para validação e, 20% para teste. No modelo

de classificação Multiclasse repetiu-se essa porcentagem de divisão, pois cada classe tem apenas 250 registros. Para os modelos de classificação binária, a divisão foi de 63% para treinamento, 7% para validação e, 30% para teste. Na Tabela 4.5, é apresenta a quantidade absoluta de sinais de voz utilizados em cada separação.

Tabela 4.5: Separação de dados para os modelos.

Modelo	Percentual de Separação	Qtd. de Sinais de Voz			
		Treinamento	Validação	Teste	Total
VVG16 Masc.	~72%:~8%:~20%	239	26	67	332
VGG16 Fem.	~72%:~8%:20%	303	33	84	420
C. Bin. Masc.	63%:7%:30%	1260	140	600	2000
C. Bin. Fem.	63%:7%:30%	1260	140	600	2000
C. Multi. Masc.	72%:8%:20%	900	100	250	1250
C. Multi. Fem.	72%:8%:20%	900	100	250	1250

Fonte: Autoria própria.

4.6.2 Entradas, Hiperparâmetros e *Callbacks*

Os modelos da rede VGG16 foram alimentados com os espectrogramas mel. Para os modelos binário e multiclasse CNN-RNN, cada áudio foi convertido para uma matriz (*frames versus features*), e em seguida, foi feito a padronização do tamanho das sequências, realizando operação de truncamento ou *padding*, resultando em um tensor (estrutura de dados usada para representar *arrays* multidimensionais) da ordem (número de áudios, número máximo de *frames* por sequência, número de *features*). Os tensors são os dados com os quais esses modelos de arquitetura CNN-RNN foram alimentados.

Para realizar a normalização das imagens (espectrogramas) para o intervalo [0, 1], todos os valores da matriz foram divididos por 255. Para normalizar os sinais de voz, fez-se necessário definir um número fixo de *frames*. Para tanto, utilizou-se a média da quantidade de *frames* por áudio. Caso um sinal de voz não possuísse a quantidade mínima de *frames*, esse seria preenchido com a operação de *padding* de zeros ao seu final; caso o sinal possuísse uma quantidade superior ao máximo, esse seria cortado.

A respeito das redes multiclasse, para selecionar os melhores hiperparâmetros, foi utilizada a técnica de *grid search*, com validação cruzada de três *folds*. Os parâmetros avaliados foram: *batch size* (32 e 64), *dropout* (0,2 e 0,3), *learning rate* (0,001 e 0,0005), *LSTM units* (128 e 256) e a quantidade de épocas (10, 20 e 30).

Para a rede com sinais de vozes femininas, a melhor combinação foi: *batch size* de 32, *dropout* de 0,2, *learning rate* de 0,001, 256 unidades LSTM e 20 épocas. Para a rede com sinais de vozes masculinas foi: *batch size* de 32, *dropout* de 0,2, *learning rate* de 0,005, 256 unidades LSTM e 20 épocas.

Os modelos foram monitorados a partir do callback de *early stopping*, que foi configurado para monitorar a perda no treinamento e interrompê-lo caso não houvesse melhoria na perda por dez épocas consecutivas, restaurando os melhores pesos encontrados anteriormente. Essa parada ajuda a evitar *overfitting*.

As redes multiclasse também fizeram uso do *callback reduce on plateau*, que ajusta dinamicamente a taxa de aprendizado durante o treinamento do modelo, reduzindo-a quando a métrica monitorada (no caso específico, a perda) não melhora por um número definido de épocas. O número de épocas consecutivas foi definido para cinco, o fator para 0,2 e o *learning rate* mínimo para 1e-6, todos definidos arbitrariamente. Seu objetivo é evitar que o modelo "fique preso"em um *plateau* de otimização. Ajustar o *learning rate* pode ajudar o modelo a escapar de mínimos locais e melhorar seu desempenho.

4.6.3 Métricas no Treinamento

A descrição das métricas utilizadas na fase de treinamento é apresentada a seguir.

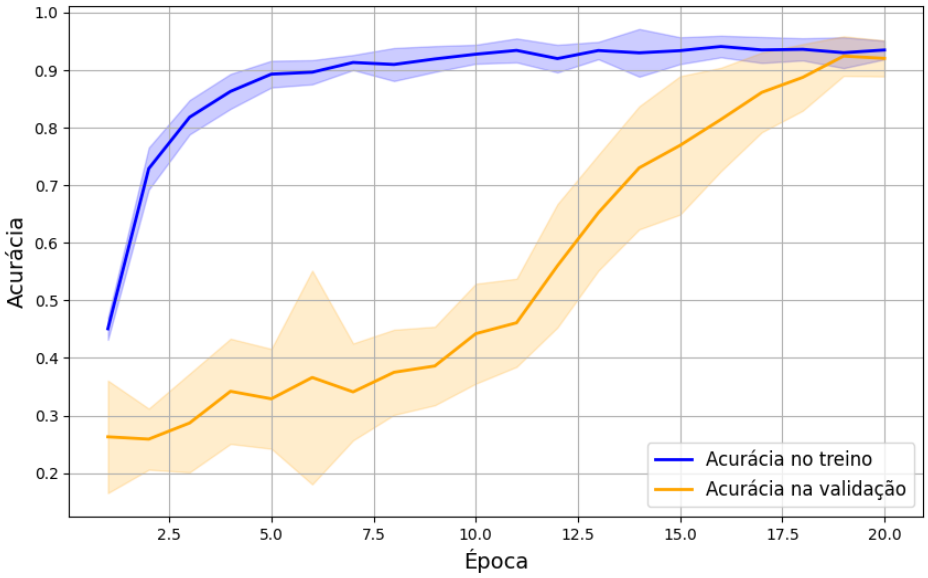
Validação Cruzada

A etapa de validação cruzada serve para avaliar o desempenho do modelo de uma maneira mais robusta. Essa etapa minimiza a possibilidade de obter uma avaliação enviesada do desempenho do modelo, já que alterna os *folds* utilizados para validação (garantindo que cada dado do conjunto seja usado para validação exatamente uma vez) e ajuda a identificar se o modelo está se ajustando bem aos dados em geral (generalização) ou se está apenas memorizando o conjunto de treinamento (sobreajuste).

Na Figura 4.7, é apresentada a acurácia média, por época, para os 10 *folds* do modelo Multiclasse para vozes femininas. A curva de validação média indica o desempenho do modelo em dados não utilizados no treinamento e a faixa sombreada indica a variação entre os diferentes *folds*, o que ajuda a avaliar a consistência do modelo.

O padrão comumente observado é o crescimento de curva de validação seguindo a mesma tendência da curva de treino, entretanto, o padrão apresentado nesta figura é um crescimento rápido no treino e sua estagnação com poucas épocas enquanto a curva de validação continua crescendo até se igualar na época 20. Esse comportamento pode, talvez, ser explicado caso tenha ocorrido um ou mais dos seguintes pontos: uma arquitetura complexa com excesso de neurônios ou camadas, pouca regularização, conjunto de treino pequeno, normalização em escala diferente, *learning rate* alto, insuficiência no aumento de dados ou *overfitting*.

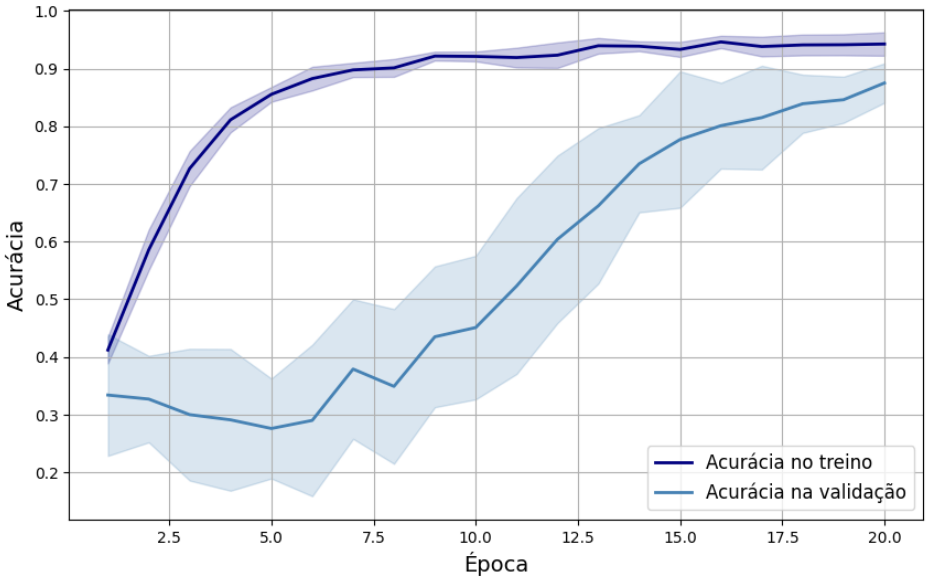
Figura 4.7: Acurácia Média no Treinamento *10-Fold* Fem. Multiclasse



Fonte: Autoria Própria.

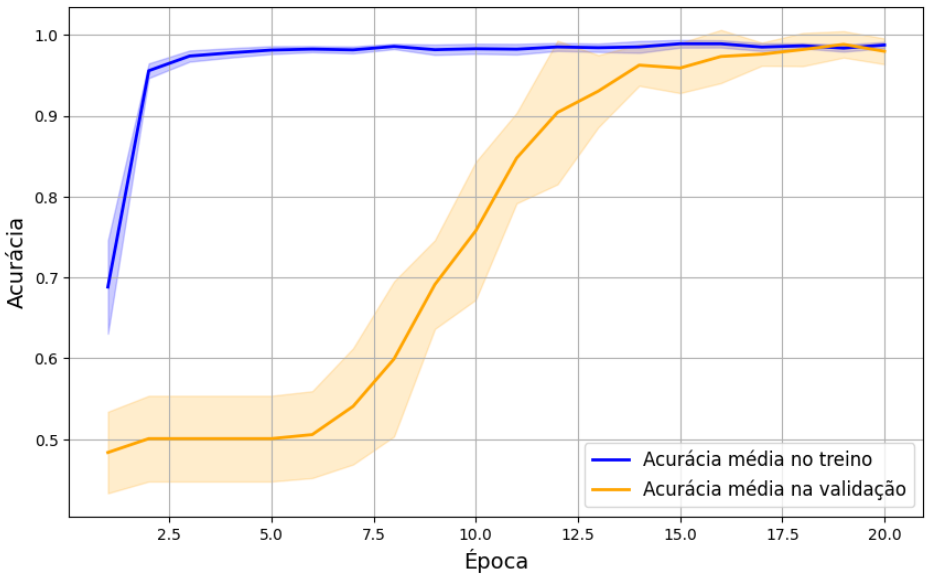
Nas Figuras 4.8, 4.9, 4.10, 4.11 e 4.12, são apresentadas os resultados das demais validações cruzadas para cada modelo, especificando o sexo do falante.

Figura 4.8: Acurácia Média no Treinamento *10-Fold* Masc. Multiclasse



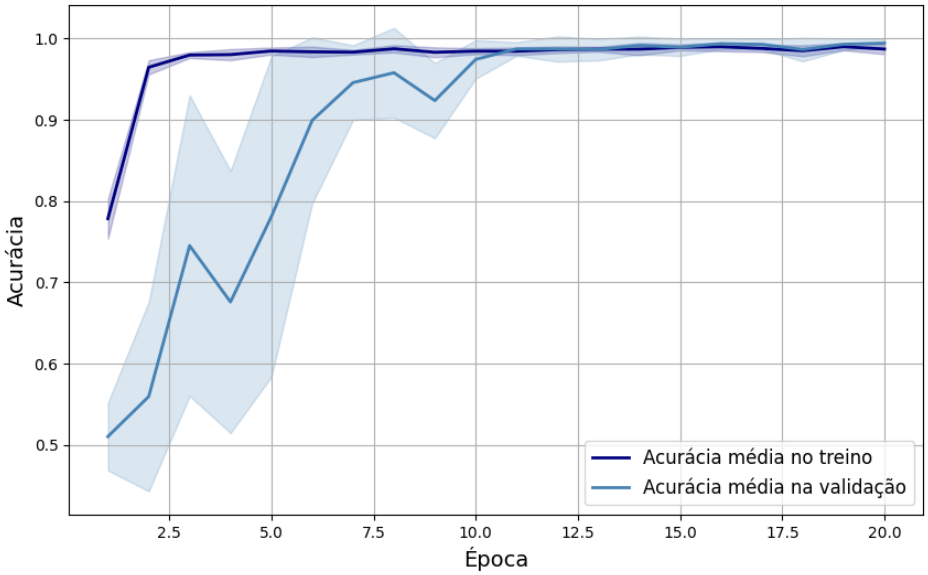
Fonte: Autoria Própria.

Figura 4.9: Acurácia Média no Treinamento *10-Fold* Fem. Binário



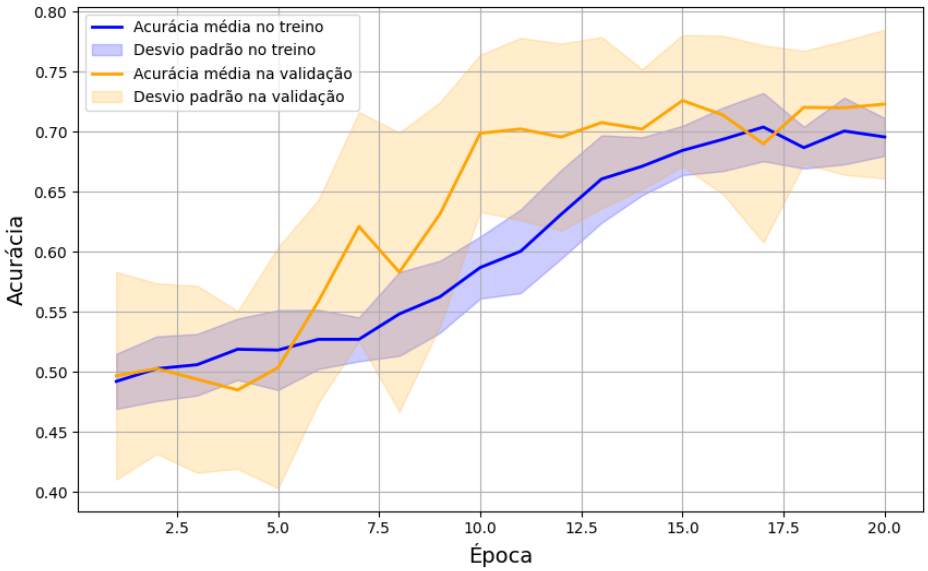
Fonte: Autoria Própria.

Figura 4.10: Acurácia Média no Treinamento *10-Fold* Masc. Binário



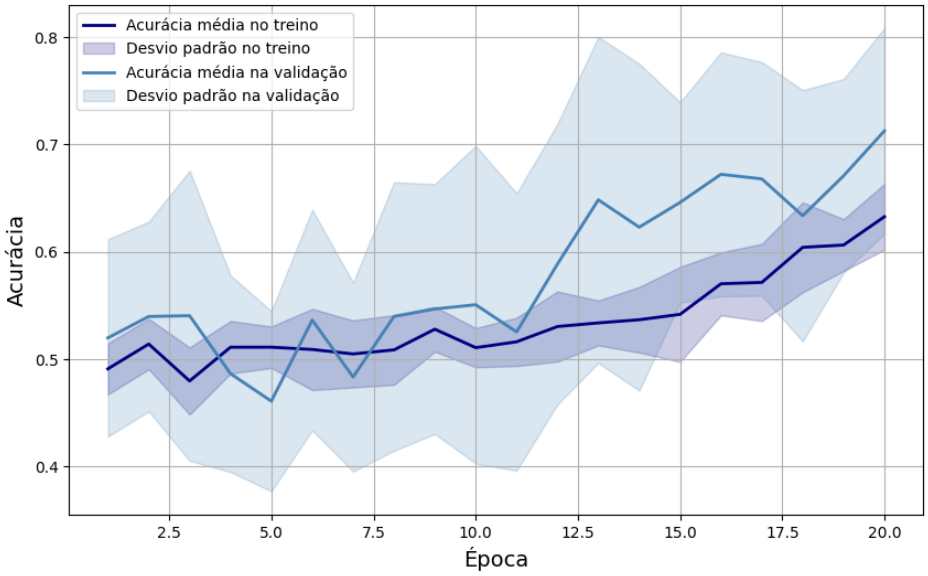
Fonte: Autoria Própria.

Figura 4.11: Acurácia Média no Treinamento *10-Fold* Fem. VGG16



Fonte: Autoria Própria.

Figura 4.12: Acurácia Média no Treinamento *10-Fold* Masc. VGG16



Fonte: Autoria Própria.

4.7 Considerações Finais

Neste capítulo, foram evidenciadas as etapas para construção dos modelos propostos (classificadores binários VGG16 e CNN-RNN e, classificador Multiclasse CNN-RNN). Além disso, especificou-se a etapa de treinamento apresentando a curva de acurácia média de cada modelo durante a validação cruzada de ordem dez. No próximo capítulo, são apresentados os resultados dos modelos na etapa de teste utilizando as métricas de acurácia, precisão, *recall*, *f1-score*, curvas ROC e matrizes de confusão.

Capítulo 5

Resultados e Discussões

Neste capítulo, são apresentados os resultados da pesquisa para cada uma das abordagens (classificador VGG16, classificador CNN-RNN binário e classificador CCNN-RNN multiclasse), evidenciando o sexo do falante. Curvas ROC e matrizes de confusão são apresentadas a fim de facilitar a compreensão e discussão. Ao final, também são apresentadas as ameaças à validade da pesquisa.

5.1 Métricas Utilizadas

Conforme Seção 1.1.1, o objetivo geral dessa pesquisa consiste em desenvolver uma abordagem com métodos de aprendizagem profunda que seja capaz de detectar vozes com distúrbios (capaz de discernir entre uma voz saudável ou uma voz patológica) e que possa classificá-las automaticamente, com certo grau de precisão. Com a finalidade validar os modelos desenvolvidos, estes foram avaliados sob as métricas de acurácia, precisão, *recall* e *F1-Score*. Também são apresentadas as curvas ROC, juntamente com as matrizes de confusão, evidenciando a relação entre classificação original *versus* previsão dos modelos.

Na Tabela 5.1, estão sumarizadas as métricas dos modelos no conjunto de teste. Para fins de validação, os modelos também foram testados unindo os dados de vozes femininas e masculinas a fim de comprovar que as diferenças nos tratos vocais podem causar impacto no modelo. Na tabela esses modelos estão referenciados como C. Mix. Binário e C. Mix. Multiclasse.

Para os modelos VGG16, o desempenho para o sexo feminino foi superior ao masculino.

Tabela 5.1: Desempenho dos modelos no conjunto de teste.

Modelo	**Acurácia %**	**Precisão %**	***Recall %***	***F1-score %***
VGG16 Fem.	70,24	84,00	50,00	62,69
VGG16 Masc.	65,67	67,74	61,76	64,62
C. Fem. Binário	99,50	99,34	99,67	99,50
C. Masc. Binário	99,33	98,74	100,00	99,37
C. Mix. Binário	95,83	95,72	96,04	95,88
C. Fem. Multiclasse	96,40	97,54	95,20	96,35
C. Masc. Multiclasse	89,20	90,83	87,20	88,97
C. Mix. Multiclasse	85,20	87,14	84,00	85,54

Fonte: Autoria própria.

A acurácia e *F1-score* atingiram, respectivamente, 70,24% e 62,69%. Entretanto, 70,24% ainda é um valor relativamente baixo para previsões (não gera credibilidade para o modelo). Este fato levanta a hipótese de que espectrogramas, sozinhos, não possuem características suficientes para que o modelo aprenda a diferenciar vozes patológicas de vozes saudáveis. É importante observar, também, que foi utilizada uma rede pré-treinada CNN, a fim de lidar com a baixa quantidade de dados. Porém, é possível que, mesmo assim, treinar as camadas finais com apenas 303 (trezentos e três) espectrogramas seja um fator limitante, até para redes pré-treinadas.

A respeito dos modelos com classificador binário, as métricas se aproximam de 100% indicando grande capacidade do modelo em lidar com a tarefa proposta. Este modelo em específico é o que possui maior quantidade de dados (vide Tabela 4.5) e, portanto, a parcela de dados para teste foi definida em 30%.

Conforme Seção de Contribuições (1.4), a criação do classificador binário se baseou no estudo do artigo Kisibi (2023). Os resultados apresentados nesse trabalho foram: 88,83% (acurácia), 86,95% (precisão), 87,91% (*recall*) e 87,39% (*F1-score*). Embora o autor afirme que usou uma arquitetura em dois estágios, em que primeiro o sinal de voz passa por um classificador do sexo do falante (masculino ou feminino) e depois o sinal é enviado para o classificador específico de voz saudável *versus* patológica correspondente, os resultados não

são expostos separadamente (para o feminino e para o masculino). Outro ponto reside no fato de que o artigo não fez um estudo aprofundado para multiclassificação, apenas alterou a quantidade de neurônios da camada final e testou sua predição.

Vale salientar ainda que o artigo base usou sempre as partições de 70% para treino, 10% para validação e 20% para teste, 50 épocas, otimizador rmsprops, e um *callback* para diminuição dinâmica do *learning rate*, caso não houvesse mudança na acurácia de validação por três épocas. O limite para o *learning rate* dinâmico foram: inicia-se em 0,5 e pode atingir o mínimo de 0,00001. A respeito dos distúrbios utilizados, o artigo base utilizou disodia, gagueira, disfonia, laringite e voz senil e esta dissertação utilizou laringite, disfonia, pólipo e paralisia.

Na Tabela 5.2, são apresentados os resultados obtidos nesse trabalho.

Tabela 5.2: Resultados obtidos no artigo base.

Modelo	**Acurácia %**	**Precisão %**	***Recall %***	***F1-Score %***
C. Binário	88,83	86,95	87,91	87,39
C. multiclasse	87,00	85,96	84,24	84,99

Fonte: Extraído de [34].

Na atividade de multiclassificação, os resultados obtidos foram muito relevantes (acurácias de 96,40% e 89,20%). Um ponto a se evidenciar é que o classificador que opera com vozes femininas proporcionou resultados superiores ao classificador com vozes masculinas.

Ao comparar os resultados apresentados nas Tabelas 5.1 e 5.2, observa-se que ambos os classificadores, binário e multiclasse, desenvolvidos nesta dissertação, superaram os do artigo base (tanto para vozes masculinas quanto femininas) apresentando melhoria. Vale salientar que o artigo base usou sempre as partições de 70% para treino, 10% para validação e 20% para teste, 50 épocas, otimizador rmsprops, e um *callback* para diminuição dinâmica do *learning rate* caso não houvesse mudança na acurácia de validação por três épocas. O *learning rate* inicia-se em 0,5 e pode atingir o mínimo de 0,00001. Outro ponto importante a se evidenciar é que os resultados apresentados na tabela 5.1 foram obtidos separando os dados de teste após a realização do aumento de dados. Ao separar os dados de teste antes do aumento de dados, os valores de 76,89% (acurácia), 75,23% (precisão), 80% (*recall*) e

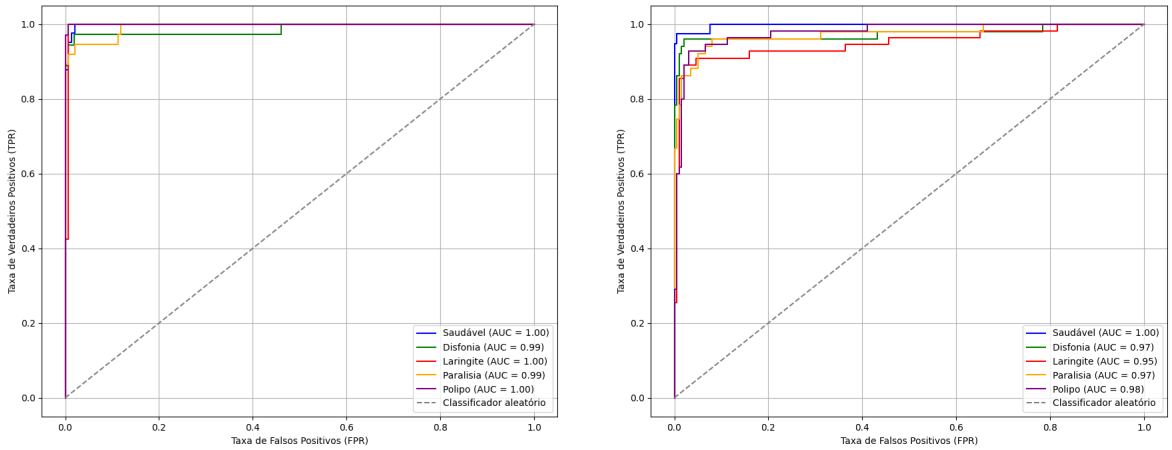
77,5% (*f1-score*) foram obtidos para classificação binária em vozes femininas e os valores 72,96% (acurácia), 72,67% (precisão), 73,33% (*recall*) e 73% (*f1-score*) para classificação binária em vozes masculinas.

5.2 Curvas ROC e Matrizes de Confusão

A seguir, são apresentadas as curvas ROC para cada modelo.

Nas Figuras 5.1a e 5.1b, são exibidas as curvas ROC para os modelos multiclasse. É possível comprovar a qualidade do modelo ao observar que as áreas de todas as categorias se aproximam de 1. Para os sinais de vozes femininas o melhor desempenho foi obtido ao se classificar vozes saudáveis, com laringite ou com pólipo Para os sinais de vozes masculinas, a maior área foi obtida para a categoria saudável seguida por pólipo.

Figura 5.1: Curvas ROC para classificadores multiclasses (Fem. e Masc.).



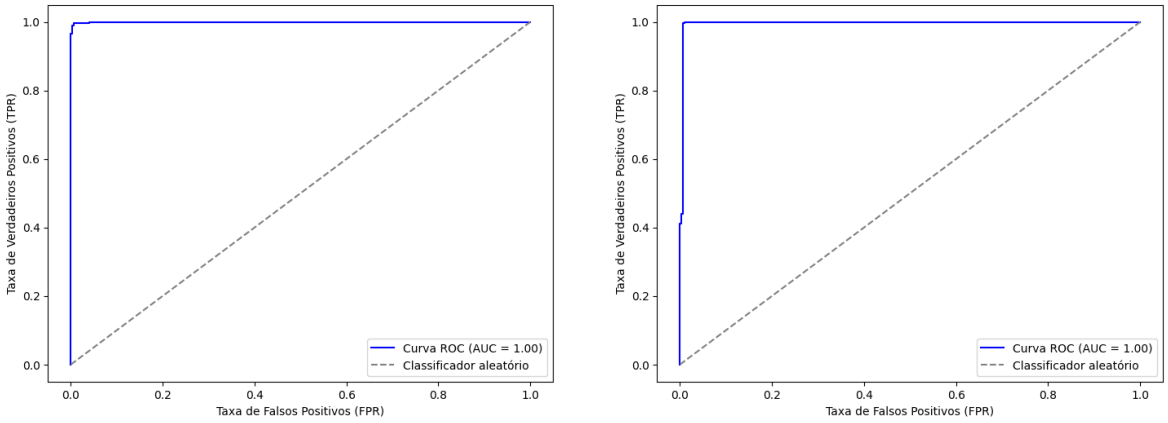
(a) Classificador Fem. multiclasse (b) Classificador Masc. multiclasse

Fonte: Autoria Própria.

Nas Figuras 5.2a e 5.2b, tem-se o desempenho do classificador binário. Observa-se que a área abaixo da curva é 1, evidenciando a capacidade do modelo em aprender a diferenciar vozes saudáveis de vozes com distúrbios.

A partir das Figuras 5.3a e 5.3b, observa-se o desempenho para os modelos que utilizam a rede VGG16. Observa-se que a área abaixo da curva é 0,7 (modelo feminino) e 0,66 (modelo masculino). Estes valores estão relativamente próximos da linha do classificador aleatório, indicando que os modelos não se adaptaram bem à tarefa de classificação.

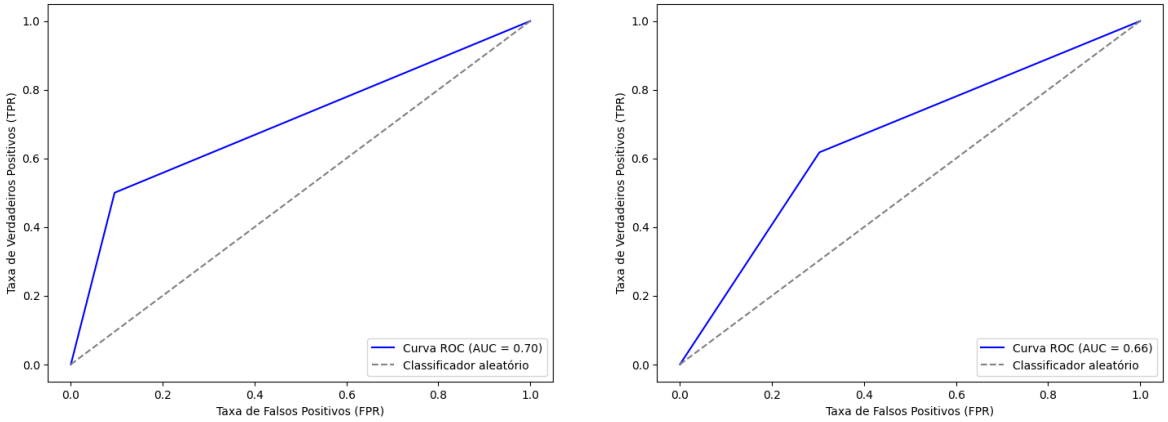
Figura 5.2: Curvas ROC para classificadores binários (Fem. e Masc.).



(a) Classificador Fem. Binário (b) Classificador Masc. Binário

Fonte: Autoria Própria.

Figura 5.3: Curvas ROC para classificadores Fem. e Masc. VGG16.

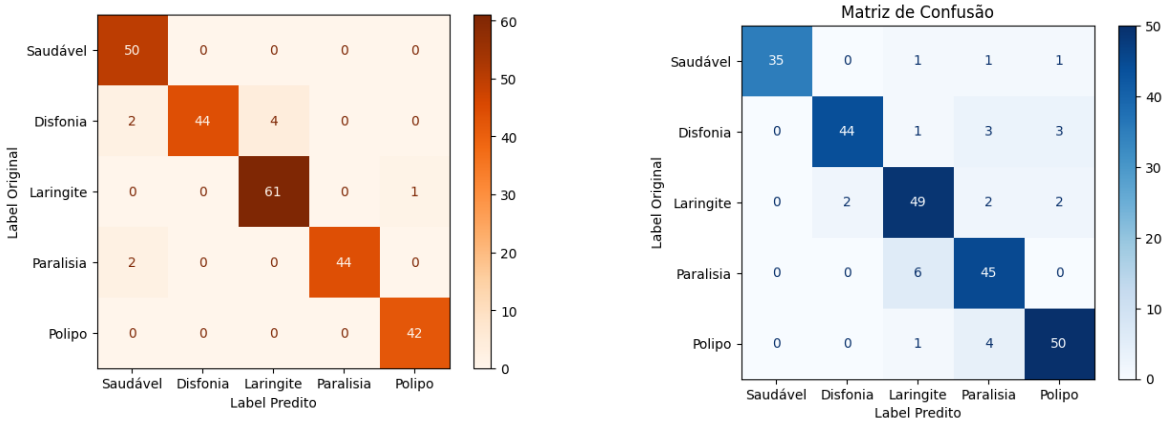


(a) Classificador Fem. VGG16 (b) Classificador Masc. VGG16

Fonte: Autoria Própria.

Como o intuito dos modelos é apoiar a decisão médica, é de suma importância avaliar a quantidade de falsos negativos (número de instâncias classificadas incorretamente como negativas) pois, caso isto ocorra, o modelo estará julgando erroneamente que a voz do paciente é saudável, desencorajando os médicos a iniciar o tratamento. A matriz de confusão é ideal para representar e possibilitar essa análise. Na Figura 5.4 (a e b), são exibidas as matrizes para o classificador multiclasse. A partir dessas, é possível observar que, para vozes femininas, as classes saudável e pólipo não apresentam falsos negativos. O erro mais presente foi classificar a disfonia como laringite. Para vozes masculinas, nenhuma classe apresentou acerto de 100% e a taxa de falsos negativos foi maior que a de falsos positivos.

Figura 5.4: Matrizes de Confusão para os classificadores multiclasse (Fem. e Masc.).



(a) Classificador Fem. Multiclasse

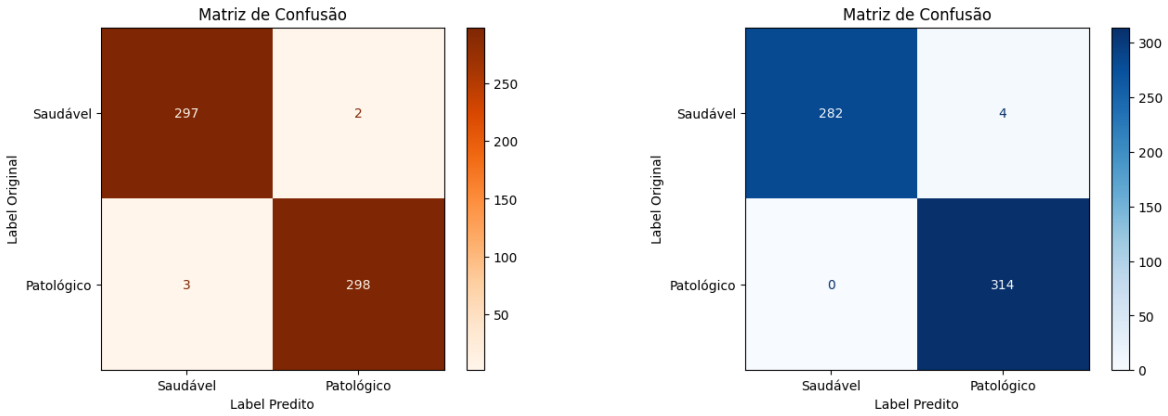
(b) Classificador Masc. Multiclasse

Fonte: Autoria Própria.

Na Figura 5.5 (a e b), são apresentadas as matrizes para classificação binária. A partir dessas, verifica-se o quão bom foram os resultados, pois quase nenhum dado de teste foi erroneamente classificado, indicando uma boa capacidade de generalização do modelo.

Na Figura 5.6 (a e b), são apresentadas as matrizes de confusão para os modelos VGG16. Para o classificador de vozes femininas, é possível perceber que a quantidade de predições corretas para o *label* com distúrbios vocais (pólipo, laringite, disfonia e paralisia) é igual à quantidade de predições incorretas, ou seja, o modelo não aprendeu a extrair características importantes para classificar uma voz como patológica. O modelo para vozes masculinas apresenta erros menos discrepantes, porém também não é satisfatório.

Figura 5.5: Matrizes de Confusão para os classificadores binários (Fem. e Masc.)

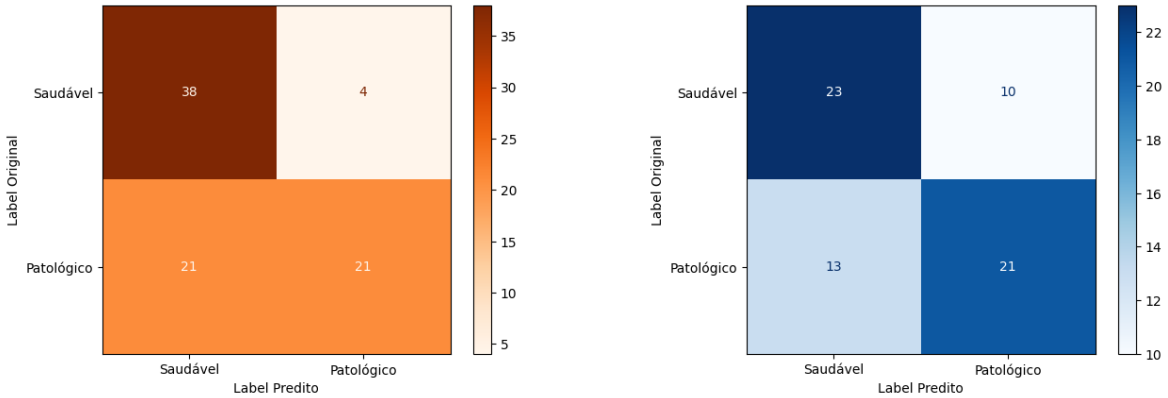


(a) Classificador Fem. Binário

(b) Classificador Masc. Binário

Fonte: Autoria Própria.

Figura 5.6: Matrizes de Confusão para os classificadores VGG16 (Fem. e Masc.)



(a) Classificador VGG16 Fem.

(b) Classificador VGG16 Masc.

Fonte: Autoria Própria.

5.3 Ameaças à Validade

Nesta seção, são tratados os fatores que podem comprometer os resultados e a interpretação desse estudo, a partir das ameças à validade.

5.3.1 Validade Interna

A validade interna é crucial porque garante que as conclusões sobre a relação causa-efeito sejam confiáveis. Assim, para que haja alta validade interna, é necessário controlar fatores como qualidade e enviesamento dos dados.

- Qualidade dos Dados: A base SVD foi utilizada pois é uma base que adquiriu os sinais em ambiente controlado. Para o desenvolvimento dessa dissertação os dados foram baixados diretamente do site oficial.

- Balanceamento dos dados: Quando os dados não estão balanceados o modelo tenderá a aprender características apenas da classe majoritária, ou ainda, classificará os dados sempre como a etiqueta dos dados majoritários, não aprendendo as reais diferenças. Nos três modelos, as classes foram balanceadas (houve necessidade de realizar técnicas de *data augmentation*) a fim de evitar que os modelos ficassem enviesados para a classe majoritária.

- Rotulação dos Dados: A rotulação tem potencial para ameaçar a validade interna pois a classificação do sinais de voz (como saudável ou com determinado distúrbio) depende da precisão do diagnóstico médico e, as vezes, pode ser subjetiva. Neste estudo, assumiu-se que as classificações postas pela base SVD estão corretas.

- *Overfitting (sobreajuste)*: Quando ocorre o *overfitting*, o modelo se ajusta fortemente aos dados de treinamento, passando a memorizar ruídos e padrões espúrios presentes nos dados que não representam relações verdadeiras. Neste estudo as técnicas de *data augmentation*, como *time stretch* ou ainda o *callback* de *early stopping*, foram aplicados direta ou indiretamente, para evitar o *overfitting*. O bom desempenho nos dados de teste, para os modelos de classificação binária e multiclasse criados, corroboram como indícios contrários de sobreajuste.

- Separação dos dados: Nos experimentos executados, os dados para teste foram separados após ter sido feito o aumento de dados. Essa conduta pode ameaçar a validade interna da pesquisa. Em novo teste, fazendo a separação antes do aumento e, garantindo que todos os dados de um mesmo falante (pronuncia da vogal /a/ nos três tons, o que gera três áudios distintos) estejam exclusivamente no conjunto de treino ou no de teste, os resultados para classificação binária apresentaram acurácias variando de 73% a 76%.

5.3.2 Validade Externa

A validade externa relaciona-se à capacidade de generalizar os resultados para outros contextos ou populações. A seguir, estão pontos importantes a se considerar.

- Generalização do Modelo: sabe-se que o modelo pode não ser generalizável para populações diferentes daquelas usadas no treinamento (vozes de diferentes faixas etárias, gêneros ou sotaques). Entretanto, como nesta dissertação, são apresentados para os três modelos, classificadores para vozes masculinas e femininas, o problema de gênero deve ter sido superado. A respeito de sotaques, o uso da vogal /a/ sustentada em detrimento de frases, ajuda a mitigar sua influência.

- Variedade de Distúrbios: se o conjunto de dados contiver um número limitado de tipos de distúrbios, o modelo pode ter dificuldade em generalizar para outras condições médicas. Neste ponto, é possível afirmar que os modelos binários e multiclasse propostos são eficientes ao lidar com os distúrbios estudados (laringite, disfonia, pólipo e paralisia). Nada se pode afirmar para outros distúrbios.

- Contextos Ambientais: o desempenho do modelo pode ser inferior ao esperado se os dados forem capturados em ambientes diferentes, como clínicas ou situações informais, visto que os equipamentos podem não ser tão precisos e não é possível garantir o controle para mitigar a presença de ruído. Nas abordagens propostas, não foram aplicadas técnicas como filtros passa-baixa, passa-alta ou *denoising algorithms* (transformada *wavelet*) para mitigar o efeito de ruídos.

5.3.3 Validade de Construção

A validade de construção refere-se a quão bem o estudo utiliza os conceitos teóricos. A seguir, os principais pontos identificados na pesquisa.

- Escolha das características: as características extraídas das vozes podem não representar adequadamente as diferenças entre vozes saudáveis e com distúrbios. Entretanto, a escolha foi baseada nas características utilizadas nas pesquisas correlatas.

- Arquitetura da Rede Neural: a escolha de uma arquitetura inadequada ou mal ajustada (a exemplo da definição do número de camadas ou parâmetros) pode comprometer os resultados. Neste quesito, a arquitetura final da rede neural VGG16 talvez pudesse ter sido mais bem avaliada. Seus hiperparâmetros também poderiam ter sido avaliados pelo *grid search*, a fim de trazer mais certezas sobre a potencialidade de espectrogramas em carregar características suficientemente importantes para classificação de vozes e também, a respeito da quantidade de dados necessária para uma rede pré-treinada.

- Métricas de Avaliação: nesta dissertação foram apresentadas várias métricas que se complementam, como precisão, *recall*, curva ROC e AUC e matrizes de confusão, mitigando interpretações tendenciosas.

5.3.4 Outras Limitações

Ameaças à reprodutibilidade e ameaças éticas também podem ser mencionadas.

- Descrição Incompleta: um das preocupações da pesquisa foi tornar possível sua reprodução, assim, o capítulo que trata da metodologia adotada foi cuidadosamente escrito, expondo técnicas, parâmetros e até códigos utilizados no desenvolvimento dos modelos.

- Dependência de Dados Privados: este não é um problema para este trabalho, pois os dados da SVD estão disponíveis para uso público, na internet.

- Consentimento e Privacidade: todos os falantes que cederam sua voz para a base SVD tiveram sua identidade preservada e são identificados apenas por um número.

5.4 Considerações Finais

Neste capítulo, foram apresentados e discutidos os resultados da pesquisa. Inicialmente, foi exposto o conceito de cada métrica utilizada para análise. Em seguida, os resultados foram sumarizados na Tabela 5.1 e comparados com o artigo base dessa pesquisa. Diagramas visuais como matrizes de confusão foram utilizadas para tornar a análise dinâmica e simples. Por fim foram apresentadas as ameaças à validade da pesquisa.

No próximo capítulo, são apresentadas as considerações finais sobre a pesquisa e oportunidades para trabalhos futuros.

Capítulo 6

Considerações Finais

Neste capítulo, são apresentadas as considerações finais sobre a pesquisa, sintetizando os resultados alcançados e relacionando-os aos objetivos definidos inicialmente.

Com base nos métodos e análises desenvolvidos ao longo da pesquisa, buscou-se avaliar a eficácia de modelos de redes neurais profundas na classificação de distúrbios da voz. Além disso, são discutidas as limitações do estudo, as contribuições para a área e possíveis direções para trabalhos futuros.

Inicialmente, são retomados os principais objetivos traçados no início do trabalho, analisando em que medida foram atingidos. Em seguida, são destacados os achados mais relevantes, incluindo os avanços teóricos e práticos obtidos. Por fim, são apresentadas recomendações e perspectivas para o aprofundamento do tema, com ênfase no aumento da diversidade da base de dados e comparação do modelo de rede pré-treinada VGG16 com outras redes pré-treinadas (ResNet50, InceptionV3).

6.1 Sumário da Pesquisa

A partir dos resultados apresentados no Capítulo 5, para os modelos de classificação multiclasse, pode-se afirmar que, pelo menos para os distúrbios: laringite, disfonia, pólipo e paralisia (aos quais se restringe o estudo dessa dissertação), é possível obter resultados relevantes utilizando redes neurais na construção de classificadores de distúrbios da voz. Fazendo alusão à Tabela 5.1, a acurácia e *F1-score* obtidos por esse tipo de modelo foram, respectivamente, 96,40% e 96,35% para sinais de vozes femininas e 89,20% e 88,97% para

sinais de vozes masculinas. Essas taxas comprovam que o modelo cumpre com sua finalidade.

Este estudo também tem como questão de pesquisa descobrir quais características são mais adequadas para a representação do sinal de voz com distúrbios do trato vocal. À luz dos resultados, é possível evidenciar que as características ZCR, RMSE e MFCC carregam informações importantes e suficientes para a tarefa de classificação de distúrbios. Ao analisar restritamente o desempenho dos modelos gerados com a rede VGG16, conclui-se que o espectrograma pode não ser uma fonte de dados tão útil para a rede, se utilizado sozinho.

Outra questão de pesquisa que motivou esta dissertação foi: "No âmbito de aprendizagem profunda para identificar e classificar distúrbios da voz, como lidar com a baixa quantidade de dados disponíveis para a etapa de treinamento? Quais são as alternativas?". Sabe-se que, ao lidar com redes neurais profundas, a quantidade de dados disponíveis é fator fundamental para o treinamento da rede e que a base SVD não fornece quantidade suficiente de sinais de voz, com distúrbios do trato vocal, para esse fim. Assim, esta dissertação apresentou formas eficientes para realizar o aumento dos dados, aplicando as técnicas de *time stretch, time shifting* e injeção de ruído branco, para superar tal desafio.

A respeito da etapa de detecção (a qual se restringem muitas das pesquisas presentes), pode-se concluir, a partir dos resultados alcançados, que o problema de detecção entre vozes com distúrbios e saudáveis foi superado (atingindo 99,34% de precisão e 99,50% de acurácia para vozes femininas e, 98,74% de precisão e 99,33% de acurácia para vozes masculinas) e que, dois dos três modelos propostos, superam abordagens presentes na literatura como as descritas em: [55] (75,65% +- 5,81% - acurácia), [34] (86,95% - precisão e 88,83% - acurácia), [20] (92,27% - acurácia) e [29] (88,67% - acurácia).

6.2 Contribuições da Pesquisa

A pesquisa ora descrita contribuiu para a área de classificação de distúrbios da voz ao validar, especificar e implementar uma abordagem de classificação multiclasse de distúrbios da voz, de forma automática, que utiliza uma adaptação da arquitetura proposta por Ksibi (2023). Além disso, os resultados alcançados com o modelo (precisão na faixa de 90%) proporcionam desempenho superior em comparação com abordagens previamente estudadas,

contribuindo para o avanço das aplicações clínicas na identificação de distúrbios da voz. A forma adotada para a descrição da metodologia também contribui para a reprodutibilidade dos experimentos realizados.

Metodologicamente, ao responder o principal questionamento desta pesquisa, busca-se contribuir para o estado da arte e, incentivar outros pesquisadores a se aprofundar na classificação dos distúrbios do trato vocal.

Em suma, como contribuições, pode-se citar:

- Adaptação de uma metodologia presente na literatura que buscou classificar os distúrbios da voz;
- Superação de modelos existentes;
- Criação de algoritmos que podem ser usados por outros pesquisadores ou profissionais;
- Modelagem com potencial para agilizar a etapa de diagnóstico de distúrbios do trato vocal, tornando-o mais eficiente e rápido, especialmente em ambientes clínicos;
- Reprodutibilidade da pesquisa (metodologia e descrição dos sinais de voz utilizados);
- Incentivo para estudos voltados à classificação multiclasse.

6.3 Limitações da Pesquisa

Apesar das contribuições alcançadas, a pesquisa apresenta limitações. Primeiramente, a base de dados utilizada, embora amplamente reconhecida na literatura, apresenta representatividade limitada, no que se refere a distúrbios do trato vocal que são menos comuns.

Além disso, o modelo multiclasse desenvolvido, embora eficiente, requer mais validação em cenários clínicos reais, para confirmar sua aplicabilidade prática. Outra limitação está relacionada à necessidade de maior robustez frente a sinais de voz com ruído ou gravações de baixa qualidade, que são comuns em contextos fora de um ambiente controlado. Esses fatores abrem caminhos para estudos futuros que possam abordar essas lacunas.

Por fim, pode-se ainda evidenciar a baixa variedade de distúrbios abordados, visto que o estudo se manteve restrito à somente quatro tipos.

6.4 Sugestões para Pesquisas Futuras

Com base nos resultados alcançados, diversas direções podem ser sugeridas para pesquisas futuras. Uma possibilidade é a ampliação do modelo multiclasse para lidar com mais doenças presentes na base SVD. Outra possibilidade consiste em realizar um estudo mais voltado à abordagem com redes pré-treinadas, realizando a validação dos hiperparâmetros e o uso de redes ResNet50 e InceptionV3.

Pode-se também analisar, com apoio de um especialista, os casos de erro apontados nas matrizes de confusão, buscando identificar eventuais erros de rotulagem ou a possibilidade de uso de outras características para facilitar a discriminação.

A validação prática dos modelos, em parceria com profissionais de saúde, avaliando sua aplicabilidade em diagnósticos clínicos e monitoramento de pacientes, também se apresenta como evolução dessa pesquisa. Por fim, o desenvolvimento de ferramentas em tempo real, integradas a dispositivos portáteis ou aplicativos, poderia ampliar o impacto prático das soluções propostas, especialmente em regiões com acesso limitado a especialistas.

Outros pontos também podem ser estudados, tais como:

- Melhoria na robustez: evoluir os modelos gerados aplicando técnicas como filtragem passa-baixa, passa-alta ou *denoising algorithms* (transformada *wavelet*) para mitigar o efeito de ruídos e a utilização de regularizadores na função de perda do modelo;

- Uso de dados multimodais: investigar a combinação de diferentes tipos de dados, como sinais de voz e imagens da laringe (videolaringoscopia) ou sinais eletroglotográficos (EGG), a fim de identificar características complementares;

- Extração de dados em ambientes reais: propor a coleta de gravações em ambientes clínicos reais, que incluem ruídos e condições menos controladas;

- Análise interpretável: Investigar e aplicar métodos que tornem os modelos de redes neurais mais interpretáveis tais como GradCAM, Lime e Shap, ajudando profissionais da saúde a compreender as decisões dos modelos;

- Análise de progressão: investigar como os modelos podem ser usados para monitorar a progressão de um distúrbio ao longo do tempo.

Referências Bibliográficas

[1] Martín Abadi, Ashish Agarwal, Paul Barham, et al. Tensorflow: Large-scale machine learning on heterogeneous systems. https://www.tensorflow.org/?hl=pt, 2015. Software available from tensorflow.org.

[2] Mohammad Aljanabi Saad Abbas Abed Ahmed Hussein Ali, Mohanad G. Yaseen. Transfer learning: A new promising technique. *Mesopotamian Journal of Big Data*, 2023:31–32, 2023.

[3] Shalbbya Ali, Safdar Tanweer, Syed Sibtain Khalid, and Naseem Rao. Mel frequency cepstral coefficient: a review. *ICIDSSD*, 2020.

[4] V Arpitha, K Samvrudhi, G Manjula, J Sowmya, and GB Thanushree. Diagnosis of disordered speech using automatic speech recognition. *International Journal of Engineering Research and Technology*, pages 126–132, 2020.

[5] Gabriel Almeida Azevedo. Controle de acesso a ambiente restrito, a partir da identidade vocal, utilizando coeficientes mfcc e classificador k-means. 2021.

[6] Kishor Barasu Bhangale and K Mohanaprasad. A review on speech processing using machine learning paradigm. *International Journal of Speech Technology*, 24:367–388, 2021.

[7] Alberto Bietti and Julien Mairal. Group invariance, stability to deformations, and complexity of deep convolutional representations. *Journal of Machine Learning Research*, 20(25):1–49, 2019.

[8] Lili Chen, Chaoyu Wang, Junjiang Chen, Zejun Xiang, and Xue Hu. Voice disor-

der identification by using hilbert-huang transform (hht) and k nearest neighbor (knn). *Journal of Voice*, 35(6):932–e1, 2021.

[9] Nuo Chen. Exploring the development and application of lstm variants. *Applied and Computational Engineering*, 53(1):103–107, 2024.

[10] CMEB, C.O. Doenças da voz: Laringite, nódulos da pregas vocais e disfonias. https://cmeb.med.br/tratamento/doencas-da-voz-laringite-nodulos-da-prega-vocal-e-disfonias/, 2024. Disponível em: 26/01/2025.

[11] Hugo Tito Cordeiro. *Reconhecimento de patologias da voz usando técnicas de processamento da fala*. PhD thesis, Universidade NOVA de Lisboa (Portugal), 2016.

[12] Carlos Daniel Riquelme Cuadros. Reconhecimento de voz e de locutor em ambientes ruidosos: comparação das técnicas mfcc e zcpa. Master's thesis, Universidade Federal Fluminense, 2008.

[13] CÓSER, C.O. Laringoscopia. Disponível em: https://www.clinicacoser.com/veja-fotos-de/laringoscopia/, 2024. Acessado em: 26/01/2025.

[14] Vladmir Frabegas Surigue De Alencar. *Atributos e domínios de interpolação eficientes em reconhecimento de voz distribuído*. PhD thesis, Pontifícia Universidade Católica do Rio de Janeiro - PUC Rio, 2005.

[15] Pablo HU de Pinho, Maria FKB Couras, Silvana LNC Costa, and Suzete EN Correia. Discriminação entre sinais de vozes saudáveis e patológicos por meio da análise da imagem do espaço de fase reconstruído. *XXXV Simpósio Brasileiro de Telecomunicações e Processamento de Sinais*, 2017.

[16] Lucas C Dias, Luana R Barros, Suzete EN Correia, and Silvana L do NC Costa. Detecção de patologias laríngeas com base na análise dinâmica de sinais de voz utilizando redes neurais profundas. *XXXV Simpósio Brasileiro de Telecomunicações e Processamento de Sinais*, 2020.

[17] Kenneth D Doig, Andrew Fellowes, Anthony H Bell, Andrei Seleznev, David Ma, Jason Ellul, Jason Li, Maria A Doyle, Ella R Thompson, Amit Kumar, et al. Pathos: a

decision support system for reporting high throughput sequencing of cancers in clinical diagnostic laboratories. *Genome medicine*, 9:1–16, 2017.

[18] Drugs.com. Vocal cord disorders. https://www.drugs.com/health-guide/vocal-cord-disorders.html, 2024. Disponível em: 26/01/2025.

[19] Gunnar Fant. Acoustic theory of speech production. *The Hague*, 1960.

[20] Deli Fu, Xuehui Zhang, Dandan Chen, and Weiping Hu. Pathological voice detection based on phase reconstitution and convolutional neural network. *Journal of Voice*, 2022.

[21] Shintaro Fujimura, Tsuyoshi Kojima, Yusuke Okanoue, Kazuhiko Shoji, Masato Inoue, Koichi Omori, and Ryusuke Hori. Classification of voice disorders using a one-dimensional convolutional neural network. *Journal of Voice*, 36(1):15–20, 2022.

[22] S. Furui. *Digital Speech Processing: Synthesis, and Recognition, Second Edition,*. Signal Processing and Communications. Taylor & Francis, 2000.

[23] Marcelo de Mattos Garcia, Fabiana Pizanni Magalhães, Gabriela Bijos Dadalto, and Marina Vimieiro Timponi de Moura. Avaliação por imagem da paralisia de pregas vocais. *Radiologia Brasileira*, 42:321–326, 2009.

[24] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.

[25] J.A. Gómez-García, L. Moro-Velázquez, and J.I. Godino-Llorente. On the design of automatic voice condition analysis systems. part i: Review of concepts and an insight to the state of the art. *Biomedical Signal Processing and Control*, 51:181–199, 2019.

[26] Jorge Andrés Gómez García. *Contributions to the design of automatic voice quality analysis systems using speech technologies*. PhD thesis, Telecomunicacion, 2018.

[27] Asmaul Hosna, Ethel Merry, Jigmey Gyalmo, Zulfikar Alom, Zeyar Aung, and Mohammad Abdul Azim. Transfer learning: a friendly introduction. *Journal of Big Data*, 9(1):102, 2022.

[28] Saarland University Institute of Phonetics. Saarbruecken voice database. https://www.stimmdatenbank.coli.uni-saarland.de/, 2007. Accessado em: 24/12/2024.

[29] Rumana Islam, Esam Abdel-Raheem, and Mohammed Tarique. Voice pathology detection using convolutional neural networks with electroglottographic (egg) and speech signals. *Computer Methods and Programs in Biomedicine Update*, 2:100074, 2022.

[30] Rumana Islam, Esam Abdel Raheem, and Mohammed Tarique. Voice pathology detection using convolutional neural networks with electroglottographic (egg) and speech signals. *Computer Methods and Programs in Biomedicine Update*, 2:100074, 2022.

[31] Luis MT Jesus, Inês Belo, Jessica Machado, and Andreia Hall. The advanced voice function assessment databases (avfad): Tools for voice clinicians and speech research. In *Advances in speech-language pathology*. IntechOpen, 2017.

[32] Sudarsana Reddy Kadiri and Paavo Alku. Analysis and detection of pathological voice using glottal source features. *IEEE Journal of Selected Topics in Signal Processing*, 14:367–379, 2020.

[33] Franz Kernic. Recurrent neural networks. In *Reinforcement Learning for Finance: Solve Problems in Finance with CNN and RNN Using the TensorFlow Library*, chapter 3. Springer, 2022.

[34] Amel Ksibi, Nada Ali Hakami, Nazik Alturki, Mashael M Asiri, Mohammed Zakariah, and Manel Ayadi. Voice pathology detection using a two-level classifier based on combined cnn–rnn architecture. *Sustainability*, 15(4):3204, 2023.

[35] S Pravin Kumar, Nanthini Narayanan, Janaki Ramachandran, and Bhavadharani Thangavel. Convolutional neural network for voice disorders classification using kymograms. *Biomedical Signal Processing and Control*, 86:105159, 2023.

[36] Gabrielle dos Santos Leandro and Claudia Maria Cabral Moro. Métodos para avaliação de sistemas de apoio à decisão. *Revista Eletrônica de Sistemas de Informação (RESI)*, 18(1):1–21, 2019.

[37] Angie Lee. O que é um modelo de ai pré-treinado? https://blog.nvidia.com.br/blog/o-que-e-um-modelo-de-ai-pre-treinado/, 2023. Disponível em: 25/01/2025.

[38] Ji-Yeoun Lee. Experimental evaluation of deep learning methods for an intelligent pathological voice detection system using the saarbruecken voice database. *Applied Sciences*, 2021.

[39] Danilo Rangel Arruda Leite. *Desenvolvimento de um modelo de classificação da tipologia dos sinais vocais com base no Deep Learning*. PhD thesis, 2022.

[40] João Vilian de Moraes Lima MARINUS. *Classificação de sinais de voz afetada por patologia nas pregas vocais utilizando reconstrução do espaço de fases.* PhD thesis, Universidade Federal de Campina Grande, 2019.

[41] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th Python in Science Conference*, pages 18–25, 2015.

[42] Arslan Munir, Joonho Kong, and Mahmood Azhar Qureshi. Overview of convolutional neural networks. 2024.

[43] Chandana Panati, Simon Wagner, and Stefan Brüggenwirth. Feature relevance evaluation using grad-cam, lime and shap for deep learning sar data classification. In *2022 23rd International Radar Symposium (IRS)*, pages 457–462, 2022.

[44] Ibrahim Patel and Y. Srinivasa Rao. Speech recognition using hidden markov model with mfcc-subband technique. In *2010 International Conference on Recent Trends in Information, Telecommunication and Computing*, pages 168–172, 2010.

[45] F. Pedregosa, G. Varoquaux, A. Gramfort, et al. Scikit-learn: Machine learning in python. https://scikit-learn.org/stable/, 2011. Journal of Machine Learning Research, 12, pp. 2825–2830.

[46] Matthew E Peters, Sebastian Ruder, and Noah A Smith. To tune or not to tune? adapting pretrained representations to diverse tasks. *arXiv preprint arXiv:1903.05987*, 2019.

[47] Joseph W Picone. Signal modeling techniques in speech recognition. *Proceedings of the IEEE*, 81(9):1215–1247, 1993.

[48] Prajoy Podder, Tanvir Zaman Khan, Mamdudul Haque Khan, and M Muktadir Rahman. Comparative performance analysis of hamming, hanning and blackman window. *International Journal of Computer Applications*, 96(18):1–7, 2014.

[49] L. Rabiner and R. Schafer. *Theory and Applications of Digital Speech Processing*. Pearson Education, 2011.

[50] Lawrence Rabiner. Fundamentals of speech recognition. *Prentice Hall google schola*, 2:447–453, 1993.

[51] Mittapalle Kiran Reddy and Paavo Alku. A comparison of cepstral features in the detection of pathological voices by varying the input and filterbank of the cepstrum computation. *IEEE Access*, 9:135953–135963, 2021.

[52] Wesley Sakurai. Cnn: Um paralelo com o mapreduce. https://www.sakurai.dev.br/cnn-mapreduce/, 2024. Disponível em: 29/12/2024.

[53] C. Sapienza and B. Hoffman. *Voice Disorders, Fourth Edition*. Plural Publishing, Incorporated, 2020.

[54] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[55] Saska Tirronen, Sudarsana Reddy Kadiri, and Paavo Alku. Hierarchical multi-class classification of voice disorders using self-supervised models and glottal features. *IEEE Open Journal of Signal Processing*, 4:80–88, 2023.

[56] Saska Tirronen, Sudarsana Reddy Kadiri, and Paavo Alku. The effect of the mfcc frame length in automatic voice pathology detection. *Journal of Voice*, 38(5):975–982, 2024.

[57] Ingo R. Titze. Physiologic and acoustic differences between male and female voices. *Journal of the Acoustical Society of America*, 95(5):2595–2604, 1994.

[58] Hesdras Oliveira Viana. Descritor de voz invariante ao ruído. Master's thesis, Universidade Federal de Pernambuco, 2013.

[59] Madhu Keerthana Yagnavajjula, Kiran Reddy Mittapalle, Paavo Alku, Pabitra Mitra, et al. Automatic classification of neurological voice disorders using wavelet scattering features. *Speech Communication*, 157:103040, 2024.

[60] Xiao Yang, Ying Zhang, Xiaoying Zeng, Chunyu Zhao, Jie Gao, Xia Wan, Jin Wang, Hong Hu, Yu Wang, Yun Zhang, et al. Concordance study between ibm watson for oncology and clinical practice for patients with cancer in china. *Oncologist*, 24(6):812–819, 2019.

[61] Wujian Ye, Zixing Jiang, Qi Li, Yijun Liu, and Zhiwei Mou. A hybrid model for pathological voice recognition of post-stroke dysarthria by using 1dcnn and double-lstm networks. *Applied Acoustics*, 197:108934, 2022.

[62] Hasan Hejbari Zargar, Saha Hejbari Zargar, Raziye Mehri, and Farzane Tajidini. Using vgg16 algorithms for classification of lung cancer in ct scans image. *arXiv preprint arXiv:2305.18367*, 2023.

[63] Felipe André Zeiser, Cristiano André da Costa, Gabriel de Oliveira Ramos, Henrique Bohn, Ismael Santos, and Rodrigo da Rosa Righi. Evaluation of convolutional neural networks for covid-19 classification on chest x-rays. In *Brazilian conference on intelligent systems*, pages 121–132. Springer, 2021.

Apêndice A

Identificadores dos Sinais de Voz Extraídos da Base SVD

Na Tabela A.1 estão listados os identificadores dos sinais de vozes femininas da base de dados SVD utilizados na pesquisa. A base de dados não disponibiliza o sinal de voz 1631 para paralisia em tom baixo.

Tabela A.1: Identificadores dos sinais de vozes femininas.

Categoria	**Identificadores**
Disfonia	368, 445, 449, 451, 494, 561, 674, 1057, 1187, 1276, 1312, 1485, 1502, 1633, 1833, 2019, 2087, 2106, 2117, 2119, 2127, 2147, 2217, 2238, 2266, 2311, 2347, 2350, 2366, 2369, 2388, 2389, 2391, 2392, 2405, 2422, 2428, 2537, 2544, 2583, 2607
Pólipo	562, 932, 1052, 1164, 1779, 1957, 2005, 2056, 2300, 2587
Laringite	498, 568, 844, 871, 919, 931, 1119, 1192, 1228, 1233, 1257, 1260, 1265, 1295, 1300, 1311, 1388, 1404, 1440, 1806, 1809, 1965, 2021, 2128, 2276, 2315, 2510, 2514, 2516, 2578, 2602, 2610

Continua na próxima página.

Categoria	Identificadores
Paralisia	105, 108, 120, 130, 138, 142, 150, 152, 356, 362, 448, 450, 565, 633, 671, 712, 716, 725, 728, 825, 855, 864, 869, 874, 879, 880, 887, 893, 924, 926, 928, 929, 1040, 1049, 1054, 1166, 1249, 1256, 1262, 1263, 1272, 1292, 1293, 1330, 1390, 1397, 1402, 1407, 1437, 1444, 1470, 1484, 1489, 1547, 1553, 1570, 1604, 1620, 1624, 1626, 1631, 1644, 1655, 1658, 1664, 1676, 1691, 1713, 1743, 1760, 1776, 1784, 1785, 1788, 1803, 1820, 1825, 1826, 1829, 1835, 1836, 1863, 1867, 1892, 1901, 1939, 1955, 1970, 2002, 2014, 2024, 2033, 2073, 2083, 2086, 2109, 2115, 2150, 2153, 2155, 2160, 2190, 2216, 2218, 2237, 2247, 2274, 2330, 2394, 2414, 2430, 2460, 2464, 2465, 2469, 2476, 2493, 2496, 2501, 2507, 2512, 2536, 2538, 2562, 2563, 2565, 2581

Continua na próxima página.

Categoria	Identificadores
Saudável	1, 2, 3, 6, 7, 8, 10, 12, 13, 14, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 30, 31, 33, 34, 35, 36, 37, 38, 39, 42, 44, 45, 46, 47, 48, 49, 50, 52, 54, 55, 56, 57, 58, 62, 64, 65, 70, 71, 72, 77, 78, 79, 80, 82, 83, 86, 88, 89, 90, 91, 93, 94, 95, 97, 99, 102, 104, 112, 113, 114, 115, 116, 117, 121, 122, 123, 124, 125, 126, 133, 135, 136, 137, 145, 153, 157, 158, 676, 677, 678, 682, 683, 684, 685, 686, 687, 688, 689, 691, 695, 696, 697, 699, 700, 702, 703, 705, 707, 709, 710, 711, 732, 733, 734, 735, 736, 737, 738, 743, 745, 746, 747, 831, 832, 833, 834, 835, 836, 837, 838, 839, 840, 841, 842, 843, 852, 865, 867, 944, 962, 990, 1006, 1022, 1024, 1029, 1059, 1060, 1065, 1066, 1067, 1068, 1094, 1095, 1096, 1097, 1098, 1099, 1100, 1101, 1102, 1103, 1104, 1105, 1106, 1107, 1109, 1110, 1112, 1121, 1122, 1123, 1124, 1127, 1128, 1129, 1130, 1131, 1132, 1133, 1134, 1135, 1137, 1139, 1144, 1145, 1146, 1147, 1148, 1149, 1150, 1151, 1167, 1168, 1169, 1170, 1171, 1172, 1174, 1175, 1176, 1177, 1178, 1179, 1180, 1182, 1184, 1185, 1207, 1208, 1209, 1210, 1211, 1212, 1214, 1215, 1216, 1278, 1320, 1342, 1343, 1346, 1347, 1349, 1350, 1351, 1352, 1353, 1355, 1356, 1357, 1359, 1361, 1362, 1363, 1364, 1365, 1366, 1367, 1368, 1369, 1370, 1371, 1372, 1373, 1374, 1375, 1504, 1505, 1506, 1508, 1509, 1510, 1511, 1513, 1514, 1515, 1516, 1518, 1519, 1523, 1527, 1528, 1529, 1530, 1531, 1533, 1534, 1535, 1536, 1538, 1540, 1541, 1542, 1543, 1544, 1579, 1583, 1584, 1598, 1605, 1612, 1623, 1695, 1696, 1699, 1703, 1707, 1708, 1710, 1711, 1712, 1725, 1726, 1727, 1728, 1729, 1731, 1732, 1733, 1734, 1737, 1739, 1740, 1815, 1837, 1838, 1839, 1840, 1843, 1844, 1845, 1846, 1848, 1850, 1852, 1853, 1855, 1856, 1857, 1858, 1860, 1865, 1870, 1872, 1873, 1874, 1875, 1878, 1879, 1880, 1881, 1883, 1884, 1885, 1886, 1915, 1916, 1920, 1922, 1924, 1925, 1926, 1947, 1969, 1995, 1996, 2018, 2036, 2037, 2038, 2039, 2040, 2041, 2042, 2043, 2044, 2046, 2047, 2049, 2050, 2051, 2052, 2053, 2054, 2084, 2108, 2162, 2163, 2164, 2165, 2166, 2167, 2168, 2169, 2170, 2171, 2172, 2173, 2174, 2175, 2177, 2178, 2179, 2180, 2196, 2198, 2199, 2200, 2201, 2203, 2204, 2205, 2206, 2207,

Continua na próxima página.

Categoria	**Identificadores**
Saudável	2208, 2209, 2210, 2248, 2250, 2252, 2253, 2254, 2255, 2256, 2257, 2258, 2259, 2260, 2261, 2263, 2264, 2278, 2279, 2280, 2281, 2282, 2283, 2284, 2285, 2286, 2407, 2462, 2495, 2532, 2561

Na Tabela A.2, estão listados os identificadores dos sinais de vozes masculinas da base de dados SVD utilizados na pesquisa.

Tabela A.2: Identificadores dos sinais de vozes masculinas.

Categoria	**Identificadores**
Disfonia	916, 925, 1086, 1194, 1219, 1250, 1455, 1457, 1597, 1657, 1718, 1751, 1768, 1984, 2000, 2016, 2066, 2125, 2323, 2325, 2348, 2356, 2387, 2436, 2461, 2488, 2530, 2559, 2608
Pólipo	109, 501, 1157, 1317, 1334, 1389, 1433, 1576, 1621, 1622, 1639, 1719, 1764, 2099, 2157, 2557, 2592
Laringite	107, 139, 141, 493, 563, 824, 904, 918, 1161, 1235, 1237, 1246, 1259, 1264, 1269, 1283, 1301, 1307, 1315, 1414, 1426, 1456, 1463, 1554, 1567, 1578, 1614, 1795, 1796, 1930, 1953, 1973, 2156, 2191, 2296, 2301, 2321, 2328, 2404, 2424, 2435, 2511, 2541, 2542, 2567, 2574, 2582, 2585, 2600, 2605
Paralisia	128, 155, 358, 364, 365, 492, 629, 670, 726, 849, 883, 896, 897, 908, 912, 914, 1120, 1223, 1270, 1271, 1277, 1289, 1304, 1314, 1325, 1399, 1424, 1466, 1548, 1549, 1555, 1564, 1601, 1630, 1638, 1641, 1662, 1671, 1683, 1748, 1755, 1789, 1818, 1830, 1832, 2007, 2009, 2058, 2072, 2078, 2088, 2100, 2148, 2182, 2192, 2235, 2243, 2341, 2342, 2346, 2393, 2395, 2408, 2418, 2450, 2455, 2508, 2527, 2539, 2570

Continua na próxima página.

Categoria	**Identificadores**
Saudável	4, 5, 9, 11, 15, 29, 32, 40, 41, 43, 53, 59, 60, 61, 63, 66, 67, 68, 69, 74, 81, 84, 85, 87, 92, 96, 98, 100, 103, 132, 134, 156, 675, 679, 680, 681, 690, 694, 698, 701, 704, 706, 708, 739, 740, 744, 782, 803, 804, 805, 806, 807, 808, 809, 810, 811, 812, 813, 860, 861, 862, 941, 942, 943, 945, 946, 947, 948, 949, 950, 951, 952, 953, 954, 955, 956, 957, 958, 959, 960, 961, 963, 964, 965, 967, 968, 969, 970, 971, 972, 973, 974, 975, 976, 977, 978, 979, 980, 981, 982, 983, 984, 985, 986, 987, 988, 989, 991, 992, 993, 994, 995, 996, 997, 998, 999, 1000, 1002, 1003, 1004, 1005, 1007, 1008, 1009, 1010, 1011, 1012, 1013, 1014, 1015, 1016, 1017, 1018, 1019, 1020, 1021, 1023, 1025, 1026, 1027, 1028, 1030, 1031, 1032, 1033, 1034, 1035, 1036, 1061, 1062, 1063, 1064, 1069, 1070, 1071, 1072, 1073, 1074, 1075, 1076, 1077, 1078, 1079, 1080, 1081, 1082, 1091, 1092, 1093, 1108, 1111, 1125, 1126, 1136, 1138, 1140, 1141, 1142, 1143, 1153, 1155, 1173, 1181, 1183, 1308, 1336, 1344, 1345, 1348, 1354, 1358, 1360, 1377, 1427, 1460, 1507, 1512, 1517, 1520, 1521, 1522, 1524, 1526, 1532, 1537, 1539, 1573, 1577, 1582, 1585, 1586, 1587, 1697, 1698, 1700, 1701, 1702, 1704, 1705, 1706, 1709, 1724, 1730, 1735, 1736, 1738, 1775, 1841, 1842, 1847, 1849, 1851, 1854, 1859, 1871, 1876, 1877, 1882, 1917, 1918, 1919, 1921, 1923, 1956, 1974, 2045, 2048, 2142, 2176, 2181, 2195, 2197, 2202, 2244, 2249, 2251, 2262, 2478, 2576

Apêndice B

Camadas e Parâmetros das Redes Neurais utilizadas

As tabelas B.1, B.2 e B.3 expõem as camadas das redes construídas para as abordagens de classificação binária utilizando VGG16, classificação binária utilizando rede CNN-RNN e classificação multiclasse utilizando rede CNN-RNN.

Tabela B.1: Camadas da Rede VGG16.

Layer (type)	**Output Shape**	**Param #**
input_layer (InputLayer)	(None, 370, 969, 3)	0
block1_conv1 (Conv2D)	(None, 370, 969, 64)	1,792
block1_conv2 (Conv2D)	(None, 370, 969, 64)	36,928
block1_pool (MaxPooling2D)	(None, 185, 484, 64)	0
block2_conv1 (Conv2D)	(None, 185, 484, 128)	73,856
block2_conv2 (Conv2D)	(None, 185, 484, 128)	147,584
block2_pool (MaxPooling2D)	(None, 92, 242, 128)	0
block3_conv1 (Conv2D)	(None, 92, 242, 256)	295,168
block3_conv2 (Conv2D)	(None, 92, 242, 256)	590,080
block3_conv3 (Conv2D)	(None, 92, 242, 256)	590,080
block3_pool (MaxPooling2D)	(None, 46, 121, 256)	0
block4_conv1 (Conv2D)	(None, 46, 121, 512)	1,180,160
block4_conv2 (Conv2D)	(None, 46, 121, 512)	2,359,808
block4_conv3 (Conv2D)	(None, 46, 121, 512)	2,359,808
block4_pool (MaxPooling2D)	(None, 23, 60, 512)	0
block5_conv1 (Conv2D)	(None, 23, 60, 512)	2,359,808
block5_conv2 (Conv2D)	(None, 23, 60, 512)	2,359,808
block5_conv3 (Conv2D)	(None, 23, 60, 512)	2,359,808
block5_pool (MaxPooling2D)	(None, 11, 30, 512)	0
global_average_pooling2d_1 (GlobalAveragePooling2D)	(None, 512)	0
dense (Dense)	(None, 1024)	525,312
dropout (Dropout)	(None, 1024)	0
dense_1 (Dense)	(None, 512)	524,800
dropout_1 (Dropout)	(None, 512)	0
dense_2 (Dense)	(None, 256)	131,328
dropout_2 (Dropout)	(None, 256)	0
dense_3 (Dense)	(None, 128)	32,896
dropout_3 (Dense)	(None, 128)	0
dense_4 (Dense)	(None, 2)	258

Fonte: Autoria Própria.

Tabela B.2: Camadas da Rede CNN-RNN binária.

Layer (type)	**Output Shape**	**Param #**	**Connected to**
input_layer	(None, None, 22)	0	-
conv1d (Conv1D)	(None, None, 64)	4,288	input_layer[0][0]
batch_normalization	(None, None, 64)	256	conv1d[0][0]
max_pooling1d (MaxPooling1D)	(None, None, 64)	0	batch_normalization[0]
dropout (Dropout)	(None, None, 64)	0	max_pooling1d[0]
conv1d_1 (Conv1D)	(None, None, 64)	12,352	dropout[0][0]
batch_normalization	(None, None, 64)	256	conv1d_1[0][0]
max_pooling1d_1 (MaxPooling1D)	(None, None, 64)	0	batch_normalization[0]
dropout_1 (Dropout)	(None, None, 64)	0	max_pooling1d_1[0]
conv1d_2 (Conv1D)	(None, None, 64)	12,352	dropout_1[0][0]
batch_normalization	(None, None, 64)	256	conv1d_2[0][0]
max_pooling1d_2 (MaxPooling1D)	(None, None, 64)	0	batch_normalization[0]
dropout_2 (Dropout)	(None, None, 64)	0	max_pooling1d_2[0]
conv1d_3 (Conv1D)	(None, None, 64)	12,352	dropout_2[0][0]
batch_normalization	(None, None, 64)	256	conv1d_3[0][0]
max_pooling1d_3 (MaxPooling1D)	(None, None, 64)	0	batch_normalization[0]
dropout_3 (Dropout)	(None, None, 64)	0	max_pooling1d_3[0]
conv1d_4 (Conv1D)	(None, None, 64)	12,352	dropout_3[0][0]
batch_normalization	(None, None, 64)	256	conv1d_4[0][0]
max_pooling1d_4 (MaxPooling1D)	(None, None, 64)	0	batch_normalization[0]
dropout_4 (Dropout)	(None, None, 64)	0	max_pooling1d_4[0]
lstm (LSTM)	(None, None, 128)	98,816	dropout_4[0][0]
dense (Dense)	(None, None, 1)	129	lstm[0][0]
flatten (Flatten)	(None, None)	0	dense[0][0]
activation (Activation)	(None, None)	0	flatten[0][0]
repeat_vector (RepeatVector)	(None, 128, None)	0	activation[0][0]
permute (Permute)	(None, None, 128)	0	repeat_vector[0][0]
multiply (Multiply)	(None, None, 128)	98,816	lstm[0][0], permute[0][0]
lstm_1 (LSTM)	(None, 128)	131,584	multiply[0][0]
dense_1 (Dense)	(None, 64)	8,256	lstm_1[0][0]
dropout_5 (Dropout)	(None, 64)	0	dense_1[0][0]
dense_2 (Dense)	(None, 1)	65	dropout_5[0][0]

Fonte: Autoria Própria.

Tabela B.3: Camadas da Rede CNN-RNN Multiclasse.

Layer (type)	**Output Shape**	**Param #**	**Connected to**
input_layer	(None, None, 22)	0	-
conv1d (Conv1D)	(None, None, 128)	8,576	input_layer[0][0]
batch_normalization	(None, None, 128)	512	conv1d[0][0]
max_pooling1d	(None, None, 128)	0	batch_normalization
dropout (Dropout)	(None, None, 128)	0	max_pooling1d[0]
conv1d_1 (Conv1D)	(None, None, 128)	49,280	dropout[0][0]
batch_normalization	(None, None, 128)	512	conv1d_1[0][0]
max_pooling1d_1	(None, None, 128)	0	batch_normalization
dropout_1 (Dropout)	(None, None, 128)	0	max_pooling1d_1[0]
conv1d_2 (Conv1D)	(None, None, 128)	49,280	dropout_1[0][0]
batch_normalization	(None, None, 128)	512	conv1d_2[0][0]
max_pooling1d_2	(None, None, 128)	0	batch_normalization
dropout_2 (Dropout)	(None, None, 128)	0	max_pooling1d_2[0]
conv1d_3 (Conv1D)	(None, None, 128)	49,280	dropout_2[0][0]
batch_normalization	(None, None, 128)	512	conv1d_3[0][0]
max_pooling1d_3	(None, None, 128)	0	batch_normalization
dropout_3 (Dropout)	(None, None, 128)	0	max_pooling1d_3[0]
conv1d_4 (Conv1D)	(None, None, 128)	49,280	dropout_3[0][0]
batch_normalization	(None, None, 128)	512	conv1d_4[0][0]
max_pooling1d_4	(None, None, 128)	0	batch_normalization
dropout_4 (Dropout)	(None, None, 128)	0	max_pooling1d_4[0]
conv1d_5 (Conv1D)	(None, None, 128)	49,280	dropout_4[0][0]
batch_normalization	(None, None, 128)	512	conv1d_5[0][0]
max_pooling1d_5	(None, None, 128)	0	batch_normalization
dropout_5 (Dropout)	(None, None, 128)	0	max_pooling1d_5
lstm (LSTM)	(None, None, 256)	394,240	dropout_5[0][0]
dense (Dense)	(None, None, 256)	65,792	lstm[0][0]
dense_1 (Dense)	(None, None, 256)	65,792	lstm[0][0]
dot (Dot)	(None, None, None)	0	dense[0][0], dense_1[0][0]
softmax (Softmax)	(None, None, None)	0	dot[0][0]
dot_1 (Dot)	(None, None, 256)	0	softmax[0][0], lstm[0][0]
add (Add)	(None, None, 256)	0	lstm[0][0], dot_1[0][0]
lstm_1 (LSTM)	(None, 256)	525,312	add[0][0]
dense_2 (Dense)	(None, 128)	32,896	lstm_1[0][0]
dropout_6 (Dropout)	(None, 128)	0	dense_2[0][0]
dense_3 (Dense)	(None, 5)	645	dropout_6[0][0]

Fonte: Autoria Própria.