UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA E INFORMÁTICA INDUSTRIAL

RONALDO DA SILVA MOURA

DETECÇÃO E CLASSIFICAÇÃO DE CATEGORIAS DE DISFONIAS COM REDES NEURAIS CONVOLUCIONAIS

DISSERTAÇÃO

CURITIBA
2023

RONALDO DA SILVA MOURA

DETECÇÃO E CLASSIFICAÇÃO DE CATEGORIAS DE DISFONIAS COM REDES NEURAIS CONVOLUCIONAIS

Detection And Classification Of Categories of Dysphonia With Convolutional Neural Networks

Dissertação apresentada como requisito para obtenção do título (grau) de Mestre em Engenharia Elétrica e Informática Industrialda Universidade Tecnológica Federal do Paraná (UTFPR).

Orientador: Dr. Joaquim Miguel Maia
Coorientadora: Dra. María Eugenia Dajer

CURITIBA

2023

Ministério da Educação
Universidade Tecnológica Federal do Paraná
Campus Curitiba

RONALDO DA SILVA MOURA

DETECÇÃO E CLASSIFICAÇÃO DE CATEGORIAS DE DISFONIAS COM REDES NEURAIS CONVOLUCIONAIS

Trabalho de pesquisa de mestrado apresentado como requisito para obtenção do título de Mestre Em Ciências da Universidade Tecnológica Federal do Paraná (UTFPR). Área de concentração: Engenharia Biomédica.

Data de aprovação: 28 de Fevereiro de 2023

Dr. Joaquim Miguel Maia, Doutorado - Universidade Tecnológica Federal do Paraná

Dr. Andre Eugenio Lazzaretti, Doutorado - Universidade Tecnológica Federal do Paraná

Dr. Guilherme Nunes Nogueira Neto, Doutorado - Pontifícia Universidade Católica do Paraná (Pucpr)

Documento gerado pelo Sistema Acadêmico da UTFPR a partir dos dados da Ata de Defesa em 28/02/2023.

Dedico este trabalho aos meus pais, por sempre me proporcionarem a oportunidade de continuar estudando.

AGRADECIMENTOS

Agradeço primeiramente a Deus, pela dádiva da vida, e por ter me concedido saúde e forças para produzir este trabalho;

Ao meu orientador, Prof. Dr. Joaquim Miguel Maia, minha coorientadora, Prof. Dra. María Eugenia Dajer, e também ao professor Dr. André Eugênio Lazzaretti, por todos os ensinamentos prestados, e por sempre estarem disponíveis e dispostos a me ajudar durante todo o desenvolvimento deste trabalho;

Aos meus pais, Digenal José de Moura e Maria Aparecida da Silva Moura, por terem me encorajado a entrar na pós-graduação, e por todo apoio emocional e financeiro concedido;

À minha querida namorada, Isabelle Medeiros Guimarães, assim como aos meus familiares e amigos, por todo incentivo e compreensão, sobretudo nos momentos de ausência;

À Universidade Tecnológica Federal do Paraná, pela disponibilização de orçamento para a compra do computador utilizado na elaboração deste estudo;

Agradeço também a todos os que de alguma forma contribuíram para a realização deste trabalho;

Por fim, agradeço a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), por meio do Processo 001, pelo apoio financeiro concedido.

RESUMO

Pesquisas conduzidas ao redor do mundo mostram que entre 16,9% e 35,8% da população possui ou afirmam já terem possuído algum grau de distúrbio vocal. Entretanto, indisponibilidades de profissionais treinados ou equipamentos para diagnóstico, dentre outros fatores, podem resultar no não-tratamento de pacientes e consequente piora em suas qualidades de vida. Avanços recentes na ciência computacional possibilitaram a utilização de metodologias de detecção automática de disfonias baseadas em aprendizado de máquina, como forma de complementar a avaliação clínica. No entanto, tais metodologias exploram apenas a distinção binária entre vozes saudáveis e com disfonia, ou realizam uma etapa de classificação limitada às disfonias com maior representatividade nas bases de dados. Por conta disso, o presente trabalho avalia uma nova metodologia de classificação de disfonias, a partir do seu agrupamento em três categorias: Disfonias Funcionais, Disfonias Orgânicas, e Disfonias Organofuncionais. Este agrupamento foi aplicado às gravações presentes em duas bases de dados: a Base de dados de voz de Saarbruecken, do inglês *Saarbruecken Voice Database* (SVD), e o Banco de Dados de Avaliação Avançada da Função de Voz, do inglês *Advanced Voice Function Assessment Database* (AVFAD). Após este agrupamento, foram realizadas etapas de extração de características dos sinais de áudio, com a utilização de espectrogramas, e classificação, com a utilização de redes neurais convolucionais. A partir dos resultados obtidos, pode-se afirmar que o método possui eficácia para a detecção de disfonias orgânicas e organofuncionais, atingindo acurácias de teste de 76,1% e 72,2%, respectivamente, para a SVD, e 82,8% e 77,3% para a AVFAD. Porém, não foi possível distinguir com êxito disfonias funcionais, por estarem pouco representadas nas bases de dados, o que impactou negativamente o desempenho geral do classificador, que foi de 53,2% para os dados da SVD, e 59,8% para os da AVFAD. Contudo, um aperfeiçoamento desta metodologia pode ampliar a capacidade de detecção de disfonias funcionais, aprimorando seu desempenho.

Palavras-chave: Desordens Vocais. Classificação de disfonias. Espectrogramas de voz. Redes Neurais Convolucionais. Saarbruecken Voice Database.

ABSTRACT

Conducted surveys worldwide show that 16.9% to 35.8% of the general population experience or claim to have experienced some degree of vocal disorder. However, the unavailability of trained professionals or diagnostic equipment, among other factors, may result in the non-treatment of patients and the consequent worsening of their quality of life. Recent advances in computational science have enabled the use of methodologies for automatically detecting dysphonia based on machine learning techniques to complement clinical evaluation. However, such methods only explore the binary distinction between healthy voices and dysphonia or perform a limited classification step using the most represented dysphonia types in the databases. Because of this, the present work evaluates a new methodology for classifying dysphonia based on its grouping into three categories: Functional Dysphonia, Organic Dysphonia, and Organofunctional Dysphonia. This grouping was applied to voice recordings of two databases: the Saarbruecken Voice Database (SVD) and the Advanced Voice Function Assessment Database (AVFAD). After this grouping, a feature extraction step was applied to the audio signals using spectrograms, followed by a classification step using convolutional neural networks. From the results obtained, it is valid to state that the method effectively detects organic and organofunctional dysphonia, reaching test accuracies of 76.1% and 72.2%, respectively, for the SVD and 82. 8% and 77.3% for AVFAD. However, it was impossible to successfully distinguish functional dysphonia, since they are underrepresented in the databases, which negatively impacted the overall performance of the classifier, which reached 53.2% for SVD, and 59.8% for AVFAD. However, an improvement of this methodology can increase the capacity to detect functional dysphonia, improving its performance.

Keywords: Vocal Disorders. Dysphonia Classification. Voice Spectrograms. Convolutional Neural Networks. Saarbruecken Voice Database.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS, SIGLAS E ACRÔNIMOS

SIGLAS

AVPD	*Arabic Voice Pathology Database*
CDBN	Rede Convolucional de Crença Profunda, do inglês *Convolutional Deep Belief Network*
CNN	Rede Neural Convolucional, do inglês *Convolutional Neural Network*
DRGE	Doença do Refluxo Gastroesofágico
EQM	Erro Quadrático Médio
FFT	Transformada Rápida de Fourier, do inglês *Fast Fourier Transform*
GMM	Modelo de Misturas Gaussianas, do inglês *Gaussian Mixture Model*
IDP	Padrões Derivados Entrelaçados, do inglês *Interlaced Derivative Patterns*
MFCC	Coeficientes Cepstrais de Frequência de Mel, do inglês *Mel-Frequency Cepstral Coefficients*
MFSC	Logaritmo dos Coeficientes Espectrais de Frequência de Mel, do inglês *Log Mel-Frequency Spectral Coefficients*
NNE	Taxa de Ruído Normalizada, do inglês *Normalized Noise Ratio*
RNA	Redes Neurais Artificiais
SGD	Gradiente-Descendente Estocástico, do inglês *Stochastic Gradient Descent*
SVD	Base de dados de voz de Saarbruecken, do inglês *Saarbruecken Voice Database*
SVM	Máquina de Vetores de Suporte, do inglês *Support Vector Machine*

ACRÔNIMOS

AVFAD	Banco de Dados de Avaliação Avançada da Função de Voz, do inglês *Advanced Voice Function Assessment Database*
CAPE-V	Avaliação Perceptivo-auditiva Consensual da Voz, do inglês *Consensus Auditory-Perceptual Evaluation of Voice*
CAPES	Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
MEEI	Enfermaria de Olhos e Ouvidos de Massachusetts, do inglês *Massachusetts Eye and Ear Infirmary*

SUMÁRIO

1 INTRODUÇÃO

Pesquisas conduzidas em diferentes localidades concluem que uma parte considerável da população pode apresentar disfonias em algum momento de suas vidas. De acordo com Roy *et al*. (2005) em uma pesquisa realizada nos EUA, 29,9% da população adulta afirmou possuírem ou já terem possuído algum distúrbio vocal. Já no estudo de Åhlander *et al*. (2019), conduzido na Suécia com mais de 74 mil indivíduos acima de 18 anos de idade, 16,9% alegaram estar na mesma condição. Além disso, em uma pesquisa realizada no Brasil, Behlau *et al*. (2012) afirmaram que a profissão de professor, que está associada a um uso mais intenso da voz, causa maior prevalência de disfonias, uma vez que 63,3% dos professores entrevistados relataram a terem contraído, versus 35,8% dos não-professores entrevistados.

Uma disfonia pode ser categorizada tanto pelo grau de intensidade do distúrbio vocal como pela sua causa. Para o primeiro caso, é necessário que o paciente reproduza frases ou vogais específicas em um ambiente não-ruidoso, e um profissional treinado o avalia de acordo com uma escala perceptivo-auditiva, como a Avaliação Perceptivo-auditiva Consensual da Voz, do inglês *Consensus Auditory-Perceptual Evaluation of Voice* (CAPE-V) (ZRAICK *et al.*, 2011). Já para o segundo caso, é necessária a realização de um exame por imagem, como a endoscopia, ou a laringoscopia (SULICA, 2013), para que possam ser identificadas alterações nos órgãos envolvidos na produção vocal, como inchaços, lesões, ou anomalias, como cistos ou pólipos. Assim, nota-se que o resultado da avaliação é condicionado ao nível de preparo do profissional avaliador, e quando utilizado um exame por imagem, na disponibilidade do equipamento.

Estudos como os de Mehta *et al*. (1994) e Haux (2010) demonstram que a informática passou a ser utilizada na área médica desde a década de 1950, e passou a facilitar processos de coleta e análise de dados, diagnósticos, realização de exames e cirurgias, dentre outras tarefas. De acordo com Ravì *et al*. (2017), técnicas de aprendizado de máquina também passaram a ser implementadas na obtenção e análise de sinais biológicos, como o ECG, bem como de imagens médicas.

Com isso, diversos estudos, como os de Muhammad *et al*. (2017) e Belenko *et al*. (2020), foram realizados com o intuito de desenvolver métodos computadorizados para detectar e classificar disfonias a partir da aquisição da própria fala dos pacientes, para que possam ser usados como complementação não invasiva à avaliação médica. Contudo, apesar de conseguirem realizar suas tarefas, tais métodos ainda são limitados, pois realizam apenas uma classificação

binária entre vozes saudáveis e com disfonia, ou uma distinção entre um número limitado de disfonias com maior representatividade nas bases de dados (WU *et al.*, 2018), e também não possuem uma taxa de assertividade suficientemente elevada para que sejam utilizados como procedimento clínico.

1.1 OBJETIVOS

1.1.1 Objetivo geral

O objetivo geral do presente trabalho foi desenvolver uma nova metodologia para efetuar a classificação de categorias de disfonias a partir de sinais de áudio, separando-as de acordo com sua etiologia, de modo que não seja necessário desconsiderar as que possuem pouca representatividade entre os dados.

1.1.2 Objetivos específicos

- Encontrar bases de dados relevantes para a realização deste trabalho;
- Separar as disfonias presentes nas bases de dados em categorias, de acordo com sua etiologia;
- Desenvolver um algoritmo de classificação, utilizando uma Rede Neural Convolucional, do inglês *Convolutional Neural Network* (CNN);
- Utilizar este algoritmo para testar a metodologia desenvolvida.

1.2 ESTRUTURA DO TRABALHO

A presente dissertação está dividida em 5 capítulos. São eles: *Introdução*, *Referencial Teórico*, *Material e Métodos*, *Resultados e Discussão*, e *Conclusões e Perspectivas*.

- O primeiro capítulo contém uma introdução ao tema deste trabalho, assim como suas motivações e seus objetivos;
- Já o segundo capítulo apresenta os conceitos teóricos necessários à compreensão da metodologia desenvolvida, bem como um resumo dos trabalhos que compõe o estado da arte sobre este tema, destacando seus resultados;

- O terceiro capítulo detalha esta metodologia, que também apresenta as métricas empregadas na sua avaliação, e os materiais utilizados no seu desenvolvimento;
- No quarto capítulo, são apresentados os resultados, bem como sua discussão;
- Por fim, o quinto capítulo apresenta as conclusões deste trabalho, bem como assuntos que podem ser explorados em pesquisas futuras.

2 REFERENCIAL TEÓRICO

Este capítulo contém os conceitos teóricos utilizados no desenvolvimento deste trabalho, assim como um estado da arte sobre o tema abordado.

2.1 CARACTERIZAÇÃO DAS DISFONIAS

De acordo com Omori (2011), o processo de vocalização é iniciado com um impulso nervoso proveniente do córtex cerebral para contrair, ao mesmo tempo, os músculos do sistema respiratório e o nervo laríngeo-recorrente, resultando no fechamento da glote e no deslocamento do ar contido nos pulmões à traqueia, o que, por consequência, provoca um aumento de pressão na região sub-glotal da laringe. Ao ultrapassar um determinado limiar de pressão, o ar começa a escapar pela glote, causando assim o movimento de vibração das pregas vocais, que segundo Rosen *et al*. (2020), produz um sinal sonoro complexo, composto por uma frequência fundamental, chamada de "tom", e diversos harmônicos.

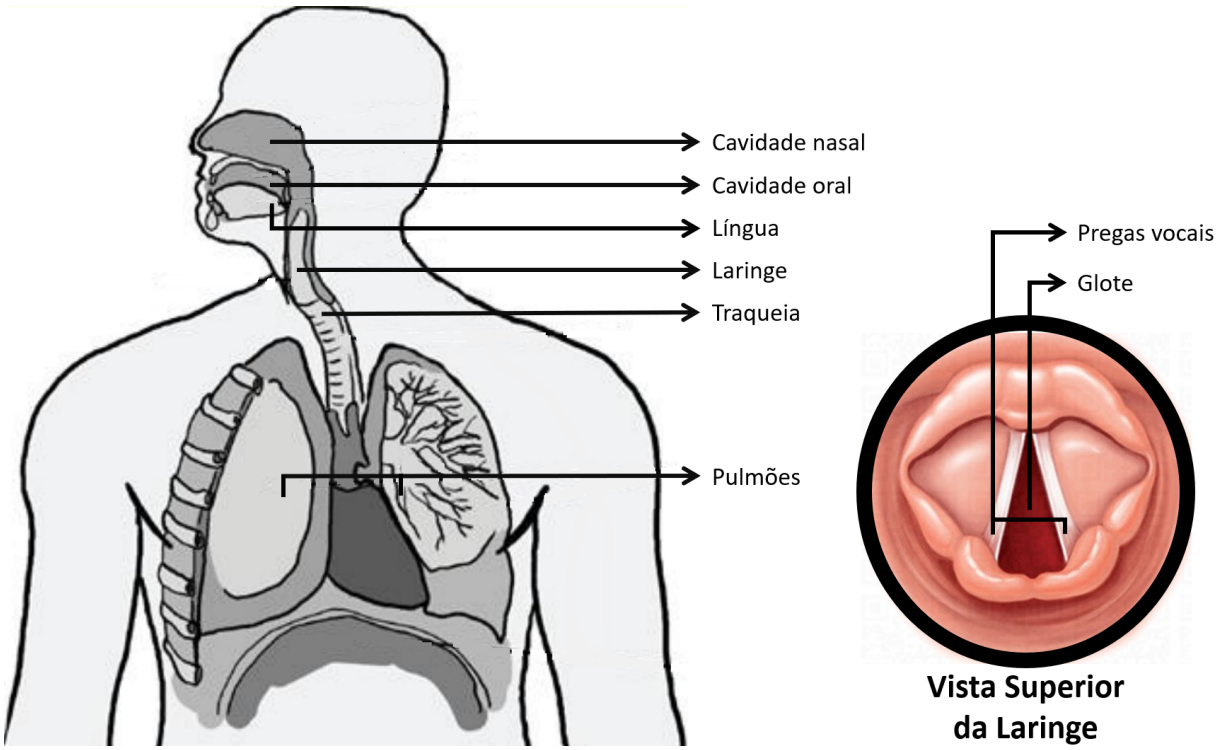
Em seguida, segundo Ghazanfar e Rendall (2008), este som passa ao trato vocal, composto pela língua, faringe, e pelas cavidades oral e nasal, que atuam como ressonadores interconectados, e fazem com que seus harmônicos possam ser atenuados ou amplificados para a produção dos sons vocais. A Figura 1 apresenta uma visão detalhada do sistema de produção vocal, destacando suas principais estruturas.

Assim, pode-se afirmar que qualquer desordem ou alteração nos órgãos e estruturas envolvidos processo de produção vocal pode afetar o padrão vibratório das pregas vocais, causando assim alterações vocais. Ademais, de acordo com Simpson e Fleming (2000), disfonias também podem ter causas não-diretamente relacionadas ao sistema de produção vocal, como no caso de doenças neurológicas, autoimunes, e até mesmo distúrbios comportamentais. Dada a grande variedade de tipos e nomenclaturas de disfonias, Behlau (2001) as separam em três categorias: disfonias funcionais, disfonias orgânicas, e disfonias organofuncionais.

2.1.1 Disfonias Funcionais

São alterações vocais caracterizadas pela ausência de mudanças estruturais significativas ou condições neurológicas associadas à laringe, sendo normalmente associadas ao

Figura 1 – Sistema de produção vocal.



Fonte: Adaptado de Cristofolini (2013).

comportamento, tendo como principal causa o uso incorreto da voz. Também podem decorrer de inadaptações vocais, como incoordenação pneumofônica, alterações respiratórias, ou fatores psicogênicos (ALTMAN *et al.*, 2005; BEHLAU, 2001).

Um de seus tipos mais comuns é a Disfonia Hiperfuncional, também denominada Disfonia de Tensão Muscular. De acordo com Pereira *et al*. (2018), esta disfonia é caracterizada pelo aparecimento de rouquidão e dores na garganta, entre outros sintomas, provocados por uma hiper-adução das pregas vocais. Segundo Altman *et al*. (2005), Sua principal etiologia é o uso excessivo da voz, normalmente devido à profissão do paciente, e pode ser tratada com a utilização de fonoterapia.

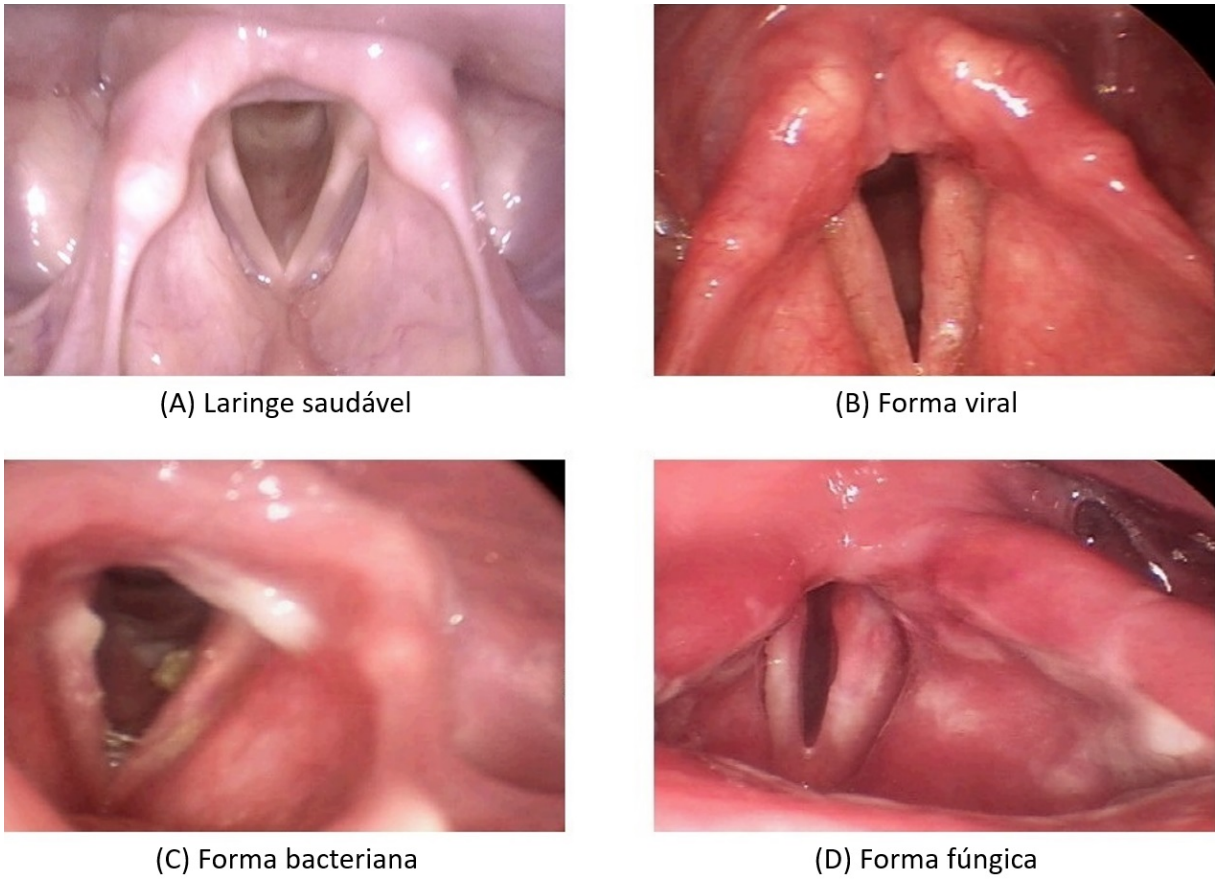
Conforme explica Kosztyła-Hojna *et al*. (2018), outro tipo comum é a Disfonia Psicogênica. Normalmente, esta disfonia é caracterizada por uma hiperfunção das pregas vocais, da mesma forma como na disfonia hiperfuncional, embora alguns casos possam ser caracterizados por hipofunção, que é a falta de tensionamento nas pregras vocais, resultando em uma voz ofegante. A disfonia psicogênica tem como causa transtornos de ordem psicológica, afetando pessoas introvertidas, vulneráveis, e propensas ao estresse e à depressão, e que experienciam conflitos familiares ou profissionais.

2.1.2 Disfonias Orgânicas

São alterações vocais decorrentes de diversos fatores alheios ao próprio uso da voz. São normalmente causadas por mudanças nas estruturas dos órgãos envolvidos na produção vocal, assim como por doenças neurológicas (BEHLAU, 2001).

De acordo com a pesquisa realizada por Cohen *et al*. (2012), a Laringite Aguda é a disfonia orgânica com maior número de diagnósticos entre pessoas que procuram tratamento médico devido a problemas de voz, sendo 42,1% dos diagnósticos concedidos. Segundo Caserta (2015), esta disfonia se caracteriza por um inchaço na região da laringe, e sua principal etiologia é uma infecção no trato respiratório superior causada por algum vírus, como por exemplo o influenza, conforme pode ser visualizado na Figura 2 (B), em contraste com a parte (A) desta mesma figura, a qual apresenta uma laringe saudável. A laringite também pode ser causada por infecções bacterianas respiratórias, como apresentado na Figura 2 (C), e, em alguns casos, até mesmo por fungos, como apresentado na Figura 2 (D).

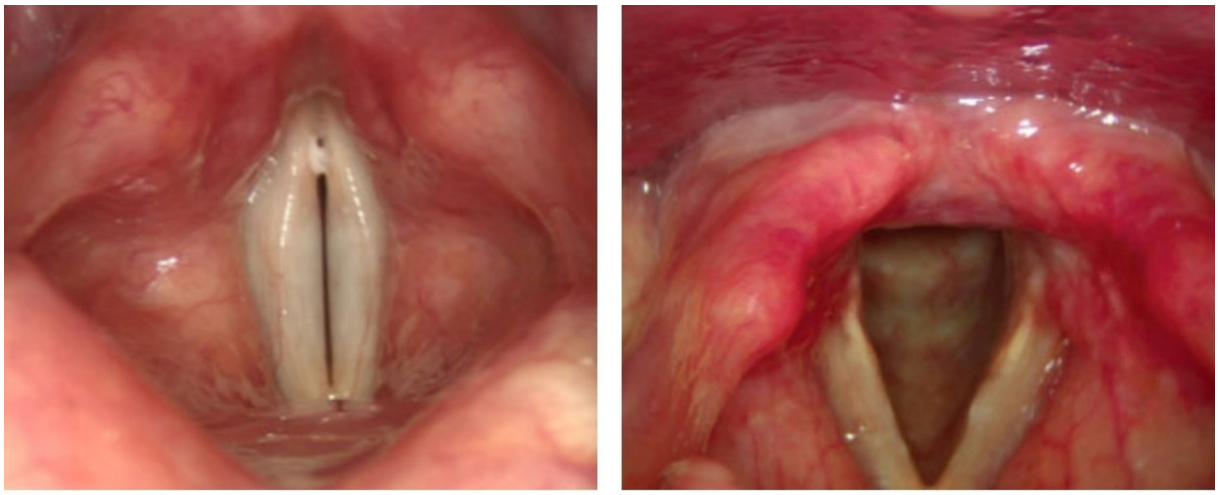
Figura 2 – Exemplos de laringe saudável (A), e apresentando Laringite Aguda, nas formas (B) viral, (C) bacteriana, e (D) fúngica.



Fonte: Adaptado de Fleischer *et al*. (2020) [Parte A], e adaptado do Centro Médico da Universidade de Nebraska (2006b) [Partes B, C, D]

Para Van-Houtte *et al*. (2010), a Doença do Refluxo Gastroesofágico (DRGE) também está entre as principais causas de disfonias orgânicas na população adulta. Segundo Kellerman e Kintanar (2017), acontece quando ácidos estomacais se deslocam pelo esôfago, provocando sensação de azia, até atingir a faringe, fazendo com que parte destes ácidos se desloquem pelo canal respiratório até a região das pregas vocais, resultando no aparecimento de inchaços e feridas nesta região. A Figura 3 (A) exibe uma laringe considerada saudável, enquanto a Figura 3 (B) exibe uma laringe de um paciente com DRGE, na qual é possível observar os danos causados pelos ácidos estomacais.

Figura 3 – Exemplo de laringe saudável (A), e apresentando DRGE (B).



(A) Laringe Saudável (B) Laringe de paciente com DRGE

Fonte: Adaptado da Associação Britânica de Voz (2009)

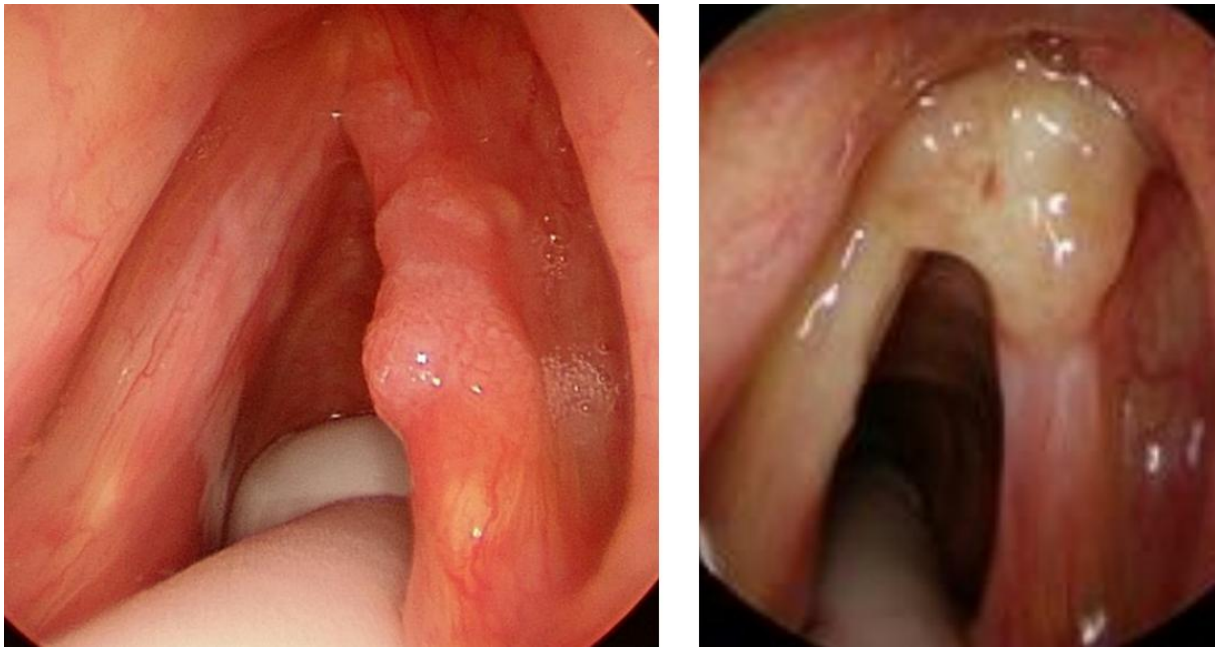
Outros exemplos de disfonias consideradas orgânicas são tumores, carcinomas, paralisia ou paresia do nervo laríngeo-recorrente, granulomas e lesões provocados por processo de intubação, e ainda transtornos neurológicos, como o mau de Parkinson (BEHLAU, 2001; SALONI *et al.*, 2014). A Figura 4 apresenta exemplos de tumores laríngeos malignos.

2.1.3 Disfonias Organofuncionais

São, de acordo com Behlau (2001), alterações vocais causadas por lesões benignas nas estruturas dos órgãos envolvidos na produção vocal, que se devem, na maioria dos casos, a um agravamento de alguma disfonia funcional não tratada.

Dentre os principais tipos de disfonias organofuncionais, de acordo com Van-Houtte *et al*. (2010), está o aparecimento de nódulos nas pregas vocais, especialmente em crianças e adolescentes, segundo Tuzuner *et al*. ((TUZUNER *et al.*, 2017)) devido a comportamentos

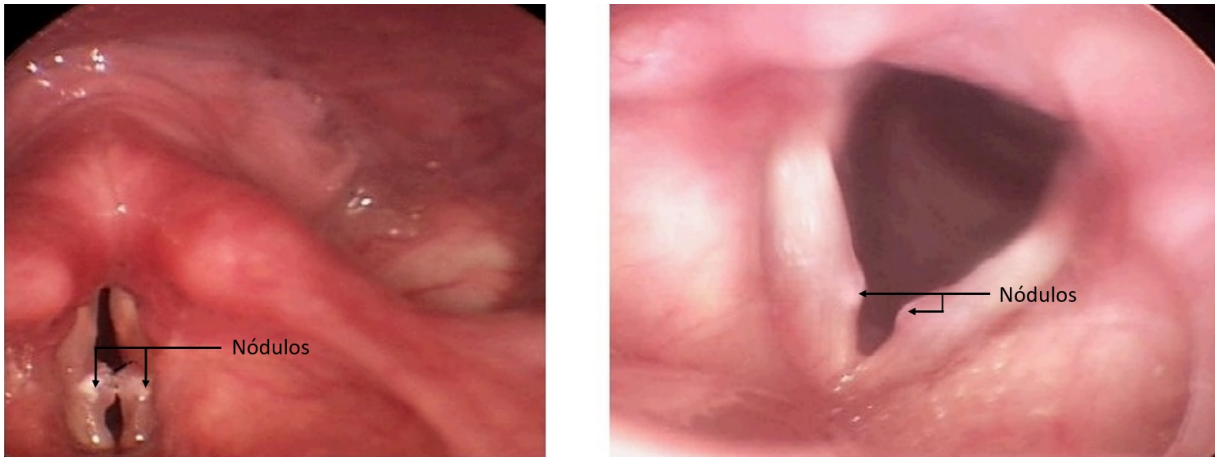
Figura 4 – Exemplos de laringe apresentando tumores malignos.



Fonte: Adaptado de Mannelli *et al*. (2020)

vocais inadequados, como gritos frequentes. Para a população adulta, o aparecimento de nódulos nas pregas vocais pode estar relacionado a fatores profissionais, como no caso de professores, e também a comportamentos sociais associados a pessoas extrovertidas, impulsivas, ou agressivas (KARKOS; MCCORMICK, 2009). Na Figura 5, pode-se observar exemplos de nódulos vocais.

Figura 5 – Fotografias de laringes com nódulos vocais bilaterais.



Fonte: Adaptado do Centro Médico da Universidade de Nebraska (2006a)

Ainda segundo Van-Houtte *et al*. (2010), outros tipos comuns de disfonias organofuncionais são o edema de Reinke, mostrado na Figura 6 (A), e caracterizado pelo inchaço de uma ou ambas pregas vocais apresentando acúmulo de fluido em sua parte interna, e também a formação de cistos e pólipos na região laríngea, apresentados na Figura 6 partes (B) e (C) respectivamente,

podendo tais condições terem como causa comportamentos vocais inadequados, fatores profissionais, e uso de cigarros, dentre outros (TAVALUC; TAN-GELLER, 2019; SALONI *et al.*, 2014).

Figura 6 – Exemplos de disfonias organofuncionais produzidas por (A) edema de Reinke unilateral de grau 3, (B) cisto unilateral, e (C) pólipo unilateral.

Fonte: sataloff *et al*. (2014) [Parte A], e adaptado do Centro Médico da Universidade de Nebraska (2006a)

2.2 ESPECTROGRAMAS DE VOZ

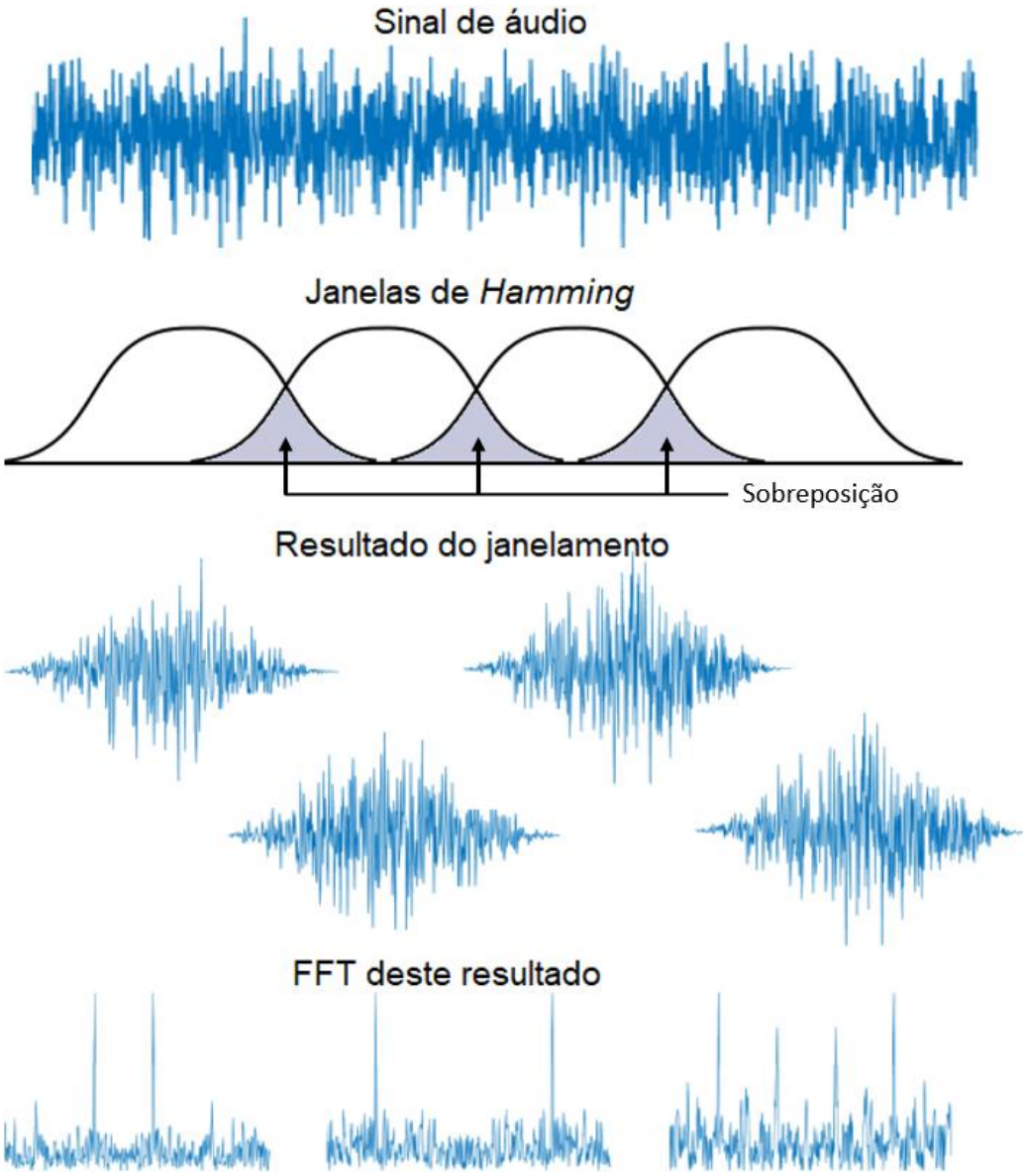
Uma das representações mais comuns de sinais é pela sua forma de onda, que consiste em um gráfico contendo a amplitude deste sinal em função do tempo (TOHYAMA, 2020). Para sinais de áudio, esta representação evidencia a amplitude dos sons captados pelo dispositivo de gravação, o que pode ser usado para detectar os momentos em que uma pessoa fala. Todavia, esta representação não expõe informações sobre o conteúdo espectral destes sinais, como a presença de harmônicos e sub-harmônicos. Ainda de acordo com Tohyama (2020), estas informações podem ser visualizadas a partir de representações que utilizam o domínio da frequência, ao invés do tempo, e são normalmente obtidas aplicando a Transformada Rápida de Fourier, do inglês *Fast Fourier Transform* (FFT).

Uma vez que o som produzido pelas pregas vocais e modulado pelo trato vocal é abundante em componentes espectrais que variam de acordo com o tempo, conforme explicado por Wu *et al*. (2018), espectrogramas de voz fornecem uma representação adequada destes sinais, por apresentarem componentes nos domínios do tempo e da frequência concomitantemente.

Para a produção de um espectrograma de voz, o sinal vocal é subdividido em fragmentos de poucos milissegundos a partir de alguma técnica de janelamento, como a de *hamming*, normalmente apresentando algum nível de sobreposição entre fragmentos adjacentes, e cada um deles tem seus componentes espectrais calculados a partir do módulo de sua transformada de

Fourier de tempo discreto. A Figura 7 apresenta estas etapas de forma gráfica.

Figura 7 – Exemplo de janelamento e aplicação da FFT a um sinal para a obtenção de seu espectrograma.



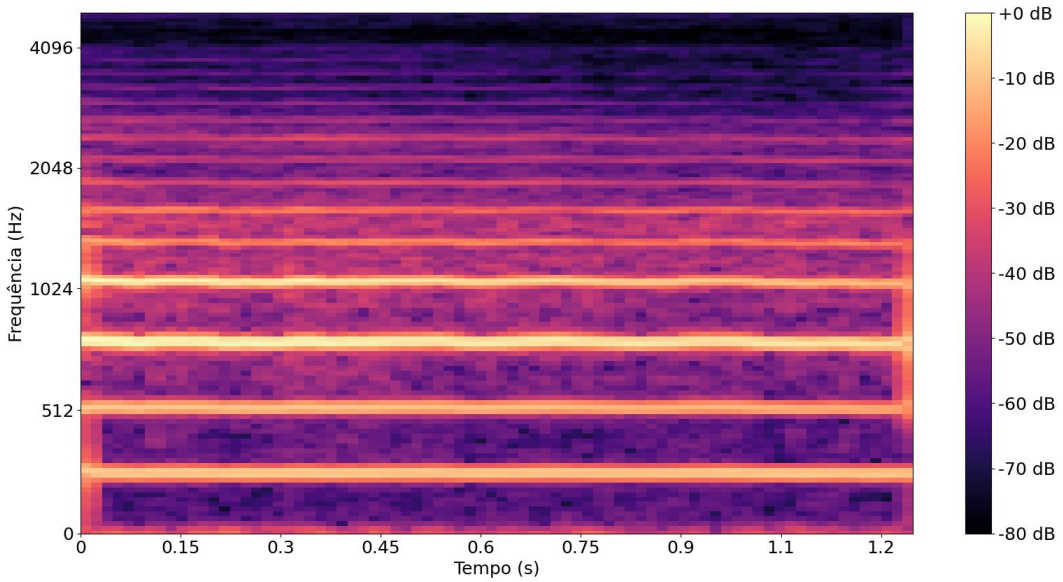
Fonte: Adaptado de Jeon *et al.* (2020)

Após este cálculo, um gráfico tridimensional é produzido, de forma que o eixo das abscissas represente unidades de tempo, o eixo das ordenadas represente unidades de frequência, e suas cores representem a contribuição atribuída a cada frequência do sinal na respectiva unidade de tempo (ALMEIDA *et al.*, 2009; WU *et al.*, 2018).

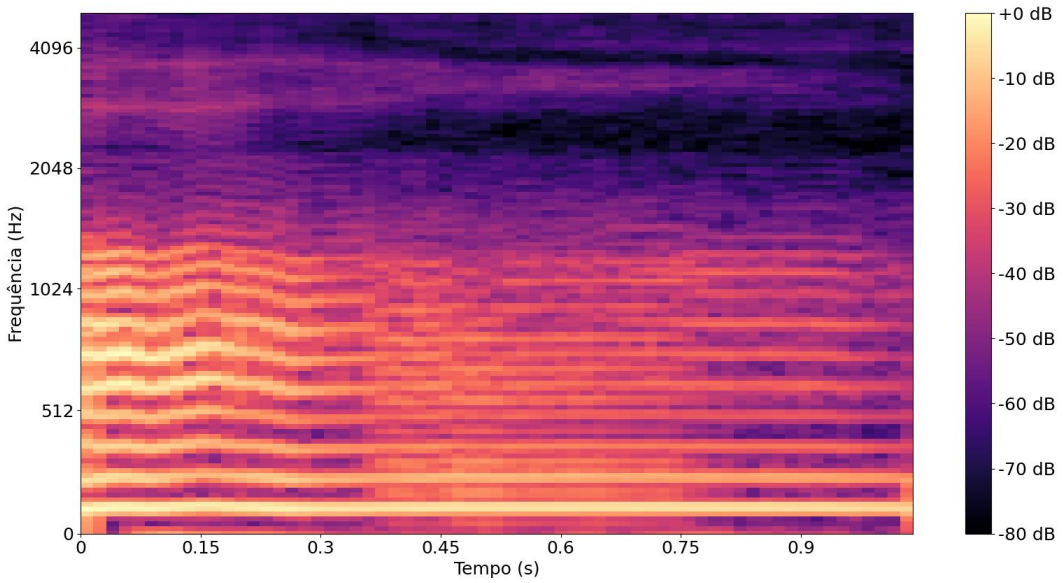
A Figura 8 (A) contém o espectrograma de uma voz considerada saudável, enquanto a Figura 8 (B) contém o de uma voz que apresenta disfonia. Ao compará-los, pode-se perceber as distorções vocais provocadas pela sua presença. Além disso, ressalta-se que o eixo das ordenadas destes espectrogramas, que representa o domínio da frequência, está apresentado em escala de

mel, a qual melhor se aproxima da forma como frequências sonoras são percebidas por seres humanos (DING *et al.*, 2021).

Figura 8 – Exemplos de espectrogramas de voz saudável e com disfonia.



(A) Paciente com voz saudável



(B) Paciente apresentando disfonia

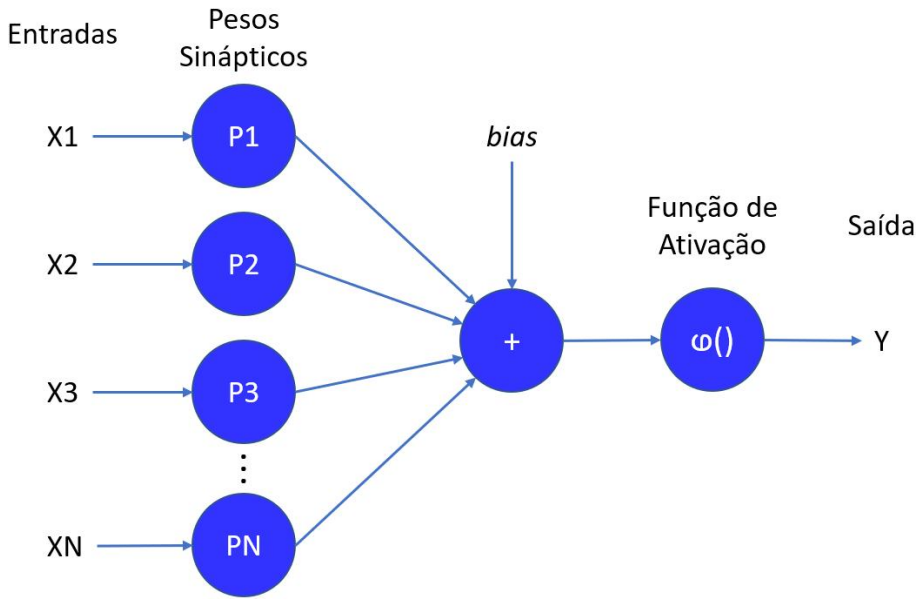
Fonte: Autoria Própria.

2.3 APRENDIZADO PROFUNDO E REDES NEURAIS CONVOLUCIONAIS

O aprendizado profundo consiste em uma técnica na qual, por meio de um processo iterativo, um algoritmo identifica padrões em exemplos de entrada, como séries temporais, textos, imagens, sons, entre outros, e os correlacionam para obter uma saída desejada. De acordo com

Krose e Smagt (2011), para isso, são comumente utilizadas Redes Neurais Artificiais (RNA), que consistem em uma estrutura contendo diversos neurônios artificiais interconectados, sendo cada um deles uma representação matemática de um neurônio biológico, denominada *Perceptron*, e apresentada na Figura 9.

Figura 9 – Representação de um neurônio artificial.



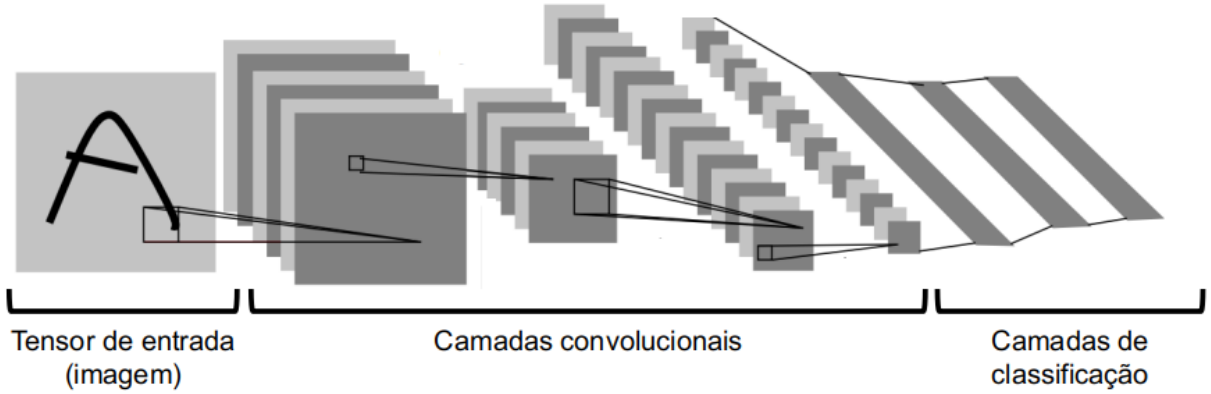
Fonte: Autoria Própria.

RNAs tradicionais empregam uma topologia do tipo *Feed-Forward*, a qual é composta por três ou mais camadas de neurônios artificiais, sendo a primeira conectada ao vetor de entrada de dados e a última conectada ao vetor de saídas da RNA. As camadas intermediárias são denominadas "escondidas", pelo fato de não estarem diretamente conectadas a um destes vetores. Uma RNA é dita "rasa"quando possui apenas uma camada escondida, e "profunda"quando possui duas ou mais destas camadas.

Apesar de redes *Feed-Forward* apresentarem boa performance em tarefas de classificação, não são adaptadas para lidar com grandes quantidades de informações de entrada, como em análise de áudio e classificação de imagens, pelo fato de todos os neurônios de uma camada estarem conectados a todos os outros da camada adjacente. Por conta disso, LeCun *et al*. (1998) conceberam uma nova topologia, denominada Rede Neural Convolucional (CNN), a qual é apresentada na Figura 10. Uma CNN é formada por várias camadas de extração de características, que são filtros compostos por neurônios artificiais aplicados a um tensor de entrada por meio de convolução, além de uma ou mais camadas de classificação, compostas por neurônios artificiais totalmente conectados. Dessa forma, a utilização de camadas convolucionais diminui o número

de parâmetros treináveis da rede, tornando-a otimizada.

Figura 10 – Rede Neural Convolucional.



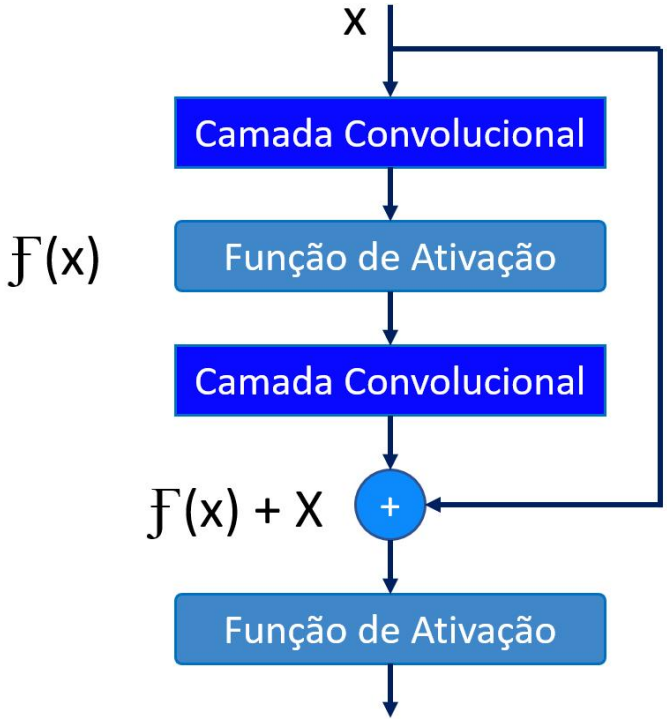
Fonte: Adaptado de LeCun *et al.* (1998).

Como complemento às camadas convolucionais, outras estruturas são utilizadas por CNNs para performar a extração de características. Um exemplo é a camada de *Pooling*, a qual realiza uma sub-amostragem do seu tensor de entrada, e torna a CNN mais robusta a pequenas translações em sua entrada (GOODFELLOW *et al.*, 2016). Outro exemplo é a utilização de blocos residuais, desenvolvidos por He *et al*. (2015), nos quais uma conexão especial é utilizada para somar o tensor de entrada de uma camada convolucional ao resultado de sua camada adjacente, conforme representado na Figura 11, o que tem um efeito prático de evitar o sobreajuste da CNN, e permite a adição de um número maior de parâmetros treináveis.

O treinamento de RNAs que possuem alguma camada escondida passou a ser possível a partir do trabalho de Rumelhart *et al*. (1986), que apresentou o algoritmo de *Retropropagação de Erros*, o qual utiliza as derivadas parciais das funções matemáticas representadas pelos neurônios da RNA em relação ao seu erro de predição, para realizar ajustes incrementais nos pesos sinápticos de cada neurônio. A partir deste algoritmo, foram desenvolvidos novos incrementos e adaptações, com as finalidades de evitar que o processo de otimização fique preso a mínimos locais, e de diminuir o tempo necessário ao treinamento da RNA (RUDER, 2016).

Dentre eles, de acordo com Zhou *et al*. (2020), um dos algoritmos mais notórios é o Gradiente-Descendente Estocástico, do inglês *Stochastic Gradient Descent* (SGD), o qual, para cada etapa de retropropagação, utiliza um pequeno grupo de amostras escolhidas de forma aleatória dentre os dados de treinamento, como forma de reduzir custos computacionais. Ainda, algumas de suas implementações incorporam um termo denominado *momentum*, que utiliza frações da última atualização de pesos sinápticos dos neurônios para definir a magnitude da próxima, o que aumenta sua robustez a mínimos locais.

Figura 11 – Bloco residual utilizado em CNNs.



Fonte: Adaptado de He *et al.* (2015)

2.4 TRABALHOS RELACIONADOS

Diversas metodologias foram desenvolvidas até o momento para realizar a detecção automática de disfonias, com a utilização de bases de dados previamente estabelecidas na literatura. Dentre elas, destacam-se a base de dados da Enfermaria de Olhos e Ouvidos de Massachusetts, do inglês *Massachusetts Eye and Ear Infirmary* (MEEI), e a Base de dados de voz de Saarbruecken, do inglês *Saarbruecken Voice Database* (SVD), sendo esta última consideravelmente maior em número de amostras e tipos de classificação disponíveis, além de possuir maior consistência de dados, o que por consequência a torna mais desafiadora para a aplicação de tais metodologias.

Martínez et al. (2012) realizam uma comparação entre resultados obtidos utilizando a MEEI e a SVD na detecção binária de disfonias. Para isso, foi utilizado um classificador de Modelo de Misturas Gaussianas, do inglês *Gaussian Mixture Model* (GMM), que tinha como entrada uma série de características extraídas dos sinais de voz, tais como Coeficientes Cepstrais de Frequência de Mel, do inglês *Mel-Frequency Cepstral Coefficients* (MFCC), e Taxa de Ruído Normalizada, do inglês *Normalized Noise Ratio* (NNE). Dentre as várias abordagens realizadas, os autores obtiveram acurácia máxima de 94,8% para os dados da MEEI, e 79,4% para os da

SVD, aplicando um processo de validação cruzada de resultados. No entanto, para a SVD, os melhores resultados foram obtidos considerando todas as vogais e intonações presentes na base de dados, de modo que gravações de um mesmo paciente poderiam estar contidas nos conjuntos de treino e teste simultaneamente. De acordo com os próprios autores, isto pode tornar tais resultados enviesados. Quando considerado apenas intonações da vogal /a/ sustentada, foi obtida acurácia de 71,8%.

No trabalho de Hammami *et al*. (2016), é realizada a extração da frequência de *pitch* de sinais vocais, além de suas três primeiras frequências de ressonância, para serem utilizadas como entrada de um classificador por Máquina de Vetores de Suporte, do inglês *Support Vector Machine* (SVM). Para a avaliação deste método, foi utilizado um subconjunto da SVD composto por apenas 160 vozes, sendo 80 consideradas saudáveis, e 80 com a presença de disfonias. Como resultado foi obtida uma acurácia de 86%.

Além de realizar a detecção binária de disfonias, Muhammad *et al*. (2017) realizam uma etapa de classificação, considerando disfonias causadas por cistos, paralisias das pregas vocais, e pólipos. Para isso, foram utilizados subconjuntos da MEEI e da SVD, além de uma base de dados desenvolvida pelos próprios autores, denominada *Arabic Voice Pathology Database* (AVPD), totalizando 414 amostras de sinais vocais. Para a extração de características, foi aplicada uma técnica de filtragem interativo-adaptativa a cada um destes sinais, para extrair seu respectivo sinal de excitação glotal, o qual teve sua derivada de primeira ordem transformada em espectrograma para que fossem computados seus Padrões Derivados Entrelaçados, do inglês *Interlaced Derivative Patterns* (IDP). Com isso, uma etapa de classificação foi aplicada utilizando SVM. Após aplicar um processo de validação cruzada, foi obtida uma acurácia de detecção de 88,5%, e de classificação de 90,3%.

Ao invés de realizar uma longa etapa de extração de características dos sinais vocais, Wu *et al*. (2018) propõem a utilização de espectrogramas, em formato de imagens, a serem classificados por uma CNN cuja arquitetura foi desenvolvida pelos próprios autores. Desse modo, foi realizada a detecção binária de disfonias, com a utilização de um subconjunto da SVD contendo 964 amostras. Neste trabalho, não foi mencionada a realização de validação cruzada dos resultados, e foram obtidas acurácias de 66% para o subconjunto de validação, e 77% para o de teste.

Esta abordagem também é explorada no trabalho de Belenko *et al*. (2020), porém com a utilização da técnica de Rede Convolucional de Crença Profunda, do inglês *Convolutional Deep*

Belief Network (CDBN), a qual provê um melhor conjunto de pesos sinápticos iniciais para o treinamento da CNN. Assim, dentre os testes realizados, foi obtida uma acurácia máxima de 73%.

Ding *et al*. (2021) realizaram a extração do Logaritmo dos Coeficientes Espectrais de Frequência de Mel, do inglês *Log Mel-Frequency Spectral Coefficients* (MFSC), além de suas derivadas de primeira e segunda ordem, como entrada de uma CNN com arquitetura ResNet, a qual possui complexidade e capacidade de generalização consideravelmente superiores às arquiteturas descritas nos trabalhos anteriores. Além disso, foram adicionados a esta CNN estruturas chamadas de "Atenção Residual", que têm potencial de aprimoramento de sua performance. Com isso, foi obtida uma acurácia máxima de 81,6% para detecção de disfonias, utilizando a base de dados SVD, e 82,2% utilizando a SVD para treino, e uma base de dados desenvolvida pelos próprios autores para testes. Ainda, foi avaliada a performance de classificação entre 4 das disfonias melhores representadas na SVD, tendo obtido acurácias entre 62,7% e 28,0%, de acordo com a classe avaliada.

Oliveira *et al*. (2020) utilizaram a transformada de Wavelet para extração de características dos sinais vocais, as quais foram utilizadas em um classificador do tipo *Random Forest*, para detecção de vozes com disfonia. Para a obtenção de resultados, foram utilizadas duas bases de dados: A SVD, e o Banco de Dados de Avaliação Avançada da Função de Voz, do inglês *Advanced Voice Function Assessment Database* (AVFAD), sendo esta última relativamente recente e com quantidade significativa de amostras, sendo 709 no total. Dentre os testes realizados, foram obtidas acurácias máximas de 83,16% para vozes do AVFAD, e 78,89% para as da SVD.

3 MATERIAL E MÉTODOS

Este capítulo contém uma descrição dos materiais utilizados neste trabalho, assim como a metodologia desenvolvida.

3.1 BASES DE DADOS

Dentre as bases de dados disponíveis na literatura, a SVD e a AVFAD foram escolhidas para a realização deste trabalho, por possuírem a maior quantidade de amostras, bem como de tipos distintos de disfonia, além de estarem disponíveis para uso gratuito.

Desenvolvida por Barry e Pützer (2007), a SVD é composta por gravações de 2037 indivíduos, sendo destes 687 considerados saudáveis, e 1350 classificados entre 71 tipos de disfonias. Todas as seções de gravação foram registradas no mesmo ambiente, com a utilização de um microfone com resolução de 16 bits e taxa de amostragem de 50kHz, e possuem os seguintes dados:

- Gravações das pronúncias das vogais */a/*, */u/*, e */i/*, nos tons baixo, normal, alto, e crescente-decrescente;
- Gravação da pronúncia da sentença em língua alemã "Guten Morgen, wie geht es Ihnen?", que pode ser traduzida como "Bom dia, como você está?";
- Arquivo com informações sobre o paciente: idade, sexo, tipo de disfonia (se houver), e comentários médicos.

Ressalta-se que esta base de dados possui boa representatividade de vozes quanto à faixa etária dos indivíduos que a compõem, sendo que a menor idade encontrada foi de 4 anos, e a maior foi de 94 anos, e números semelhantes de indivíduos entre 18 e 70 anos. Nota-se também que, aproximadamente 50% destes indivíduos correspondem ao sexo masculino, enquanto os outros 50% correspondem ao feminino.

Já a AVFAD, desenvolvida por Jesus *et al*. (2017), possui gravações de 709 indivíduos, sendo destes 363 considerados saudáveis, e 346 classificados entre 26 tipos de disfonia, de acordo com o manual desenvolvido por Behlau e Gasparini (2007). Neste caso, as seções de gravação

foram produzidas utilizando um microfone com resolução de 16 bits e taxa de amostragem de 48kHz, e possuem os seguintes dados:

- Gravações das pronúncias das vogais */a/*, */u/*, e */i/*, em tom normal, sendo cada vogal pronunciada três vezes em um mesmo áudio;
- Gravações das pronúncias de seis sentenças da versão em português da CAPE-V, sendo cada uma pronunciada três vezes em um mesmo áudio. São elas: "A Marta e o avô vivem naquele casarão rosa velho", "Sofia saiu cedo da sala", "A asa do avião andava avariada", "Agora é hora de acabar", "A minha mãe mandou-me embora", e "O Tiago comeu quatro peras";
- Gravação da pronúncia da versão em português do texto *The North Wind and The Sun*, do fabulista Æsop (LIBRARY-OF-CONGRESS, 2009);
- Gravação da pronúncia de uma fala livre;
- Arquivo com informações sobre o paciente: idade, sexo, tipo de disfonia, entre outros.

De forma semelhante à SVD, esta base de dados possui boa representatividade de vozes quanto à faixa etária de seus indivíduos, sendo que neste caso, a menor idade encontrada foi de 18 anos, e a maior foi de 93 anos, com números semelhantes de indivíduos entre 25 e 70 anos. Já para a distribuição de acordo com o sexo dos indivíduos, nota-se que 70% dos indivíduos correspondem ao sexo feminino, e apenas 30% ao sexo masculino.

3.2 RECURSOS COMPUTACIONAIS

Um computador foi disponibilizado pela UTFPR para auxiliar no desenvolvimento deste trabalho. Este computador foi equipado com um processador *Intel Core i7* de nona geração, 16GB de memória de acesso randômico, e uma placa gráfica *Nvidia GeForce GTX 1650*, necessária para o treinamento de RNAs. Assim, códigos foram desenvolvidos em linguagem de programação Python com o auxílio da *framework* PyTorch, por ser uma das mais utilizadas para tarefas de aprendizado de máquina (OTT *et al.*, 2020).

3.3 METODOLOGIA

Após a revisão da literatura e a definição das bases de dados a serem utilizadas, seus conteúdos foram analisados para determinar o tipo de classificação a ser adotado. Com isso, foi constatado que, para ambas SVD e AVFAD, mais de 50% de todas as amostras de indivíduos com disfonia pertencem a uma das cinco classificações mais recorrentes da referida base de dados, enquanto que as classificações menos recorrentes podem estar representadas por poucos ou apenas 1 indivíduo, o que faz com que não possam ser identificadas por algoritmos de aprendizado de máquina.

Por conta disso, trabalhos anteriores, como os de Muhammad *et al*. (2017) e Ding *et al*. (2021), realizaram suas etapas de classificação considerando apenas três e quatro tipos de disfonia com considerável representatividade, respectivamente. Já para o presente trabalho, disfonias foram categorizadas de acordo com o critério apresentado em Behlau (2001), o qual foi descrito na seção 2.1 deste texto.

A partir deste critério, foi possível categorizar 46 dentre os 71 tipos de disfonia presentes na SVD, e 23 dentre os 26 presentes na AVFAD, tendo os demais sido desconsiderados por não serem considerados disfonias propriamente ditas, como por exemplo o envelhecimento vocal, distúrbios de fala, ou sintomas de outras disfonias, como a diplofonia. As tabelas 1 e 2 apresentam os tipos de disfonias considerados e suas devidas classificações, para as bases de dados SVD e AVFAD, respectivamente.

Tabela 1 – Tipos de disfonias considerados para a SVD.

Categoria	Disfonias consideradas
Funcional	Disfonia, Disfonia Funcional, Disfonia Hiperfuncional, Disfonia Hipofuncional, Disfonia Hipotônica, Falsete Mutacional, Disfonia Psicogênica, Microfonia Psicogênica, Rinofonia Aberta, Rinofonia Fechada, Rinofonia Mista
Orgânica	Carcinoma nas Pregas Vocais, Carcinoma de Epiglote, Carcinoma In Situ, Cisto Cervical Mediano, Cisto Valecular, Condroma, Cordectomia, Laringocele, Monocordite, Papiloma, Paquidermia de Contato, Paralisia do Nervo Recorrente, Tumor Hipofaríngeo, Tumor Laríngeo, Tumor Nasofaríngeo, Granuloma de Intubação, Lesão de Intubação, Esclerose Lateral Amiotrófica, Paralisia Bulbar, Hiperostose Esquelética Idiopática Difusa, Doença do Refluxo Gastroesofágico, Síndrome de Down, Doença de Parkinson, Distúrbio do Movimento Laríngeo Central, Laringite, Lesão do Nervo Laríngeo Superior, Neuralgia do Nervo Laríngeo Superior, Disfonia Espasmódica, Sinéquia Vocal
Organofuncional	Cisto, Edema de Reinke, Granuloma, Nódulo de fonação, Pólipo nas Pregas Vocais, Leucoplasia

Fonte: Autoria própria.

Uma vez que as gravações de indivíduos com disfonia da SVD contém comentários médicos associados, estes foram analisados como forma de eliminar inconsistências. Com isso,

Tabela 2 – Tipos de disfonias considerados para a AVFAD.

Categoria	Disfonias consideradas
Funcional	Sulco das Pregas Vocais, Puberfonia, Disfonia de Tensão Muscular Primária, Disfonia Adaptativa
Orgânica	Varizes e Ectasia das Pregas Vocais, Laringite Aguda, Refluxo Laringofaríngeo, Trauma da Mucosa Laríngea, Transtorno Depressivo Maior (Recorrente), Paralisia Unilateral do Nervo Laríngeo Recorrente, Paresia do Nervo Laríngeo Recorrente (Unilateral ou Bilateral), Paralisia Bilateral Periférica do Nervo laríngeo recorrente, Esclerose Lateral Amiotrófica , Doença de Parkinson
Organofuncional	Nódulos de Prega Vocal, Pólipos de Prega Vocal, Cisto Subepitelial de Prega Vocal, Lesão Reativa de Prega Vocal, Edema de Reinke, Cicatriz de Prega Vocal Própria, Granuloma de Prega Vocal Não-Relacionado à Intubação, Leucoplasia, Hemorragia de Prega Vocal

Fonte: Autoria própria.

foram identificadas condições incompatíveis com o tipo de categorização adotado, resultando na desconsideração dos pacientes que as possuem. São elas:

- Amostras classificadas como disfonia funcional, em que, de acordo com os comentários associados, o paciente também possui disfonia orgânica ou organofuncional;
- Amostras classificadas como disfonia orgânica, em que, de acordo com os comentários associados, o paciente também possui disfonia organofuncional, ou foi previamente submetido a múltiplas seções de fonoterapia para acobertar sua disfonia;
- Amostras classificadas como disfonia funcional ou organofuncional, em que, de acordo com os comentários associados, o paciente encontra-se em fase de recuperação após cirurgia laríngea.

Entretanto, ressalta-se que uma parte significativa de tais comentários médicos não possuíam clareza ou quantidade de informação suficiente para realizar esta distinção, porém, nestes casos as respectivas amostras foram consideradas válidas, para que houvesse dados suficientes para avaliar os algoritmos de classificação. A tabela 3 apresenta o número de amostras consideradas e descartadas, para cada categoria de disfonia:

Tabela 3 – Gravações mantidas e descartadas da SVD, de acordo com suas categorizações.

Categoria	Número total de gravações	Gravações mantidas	Gravações descartadas
Funcional	528	414	114
Orgânica	618	459	159
Organofuncional	175	112	63
Saudável	687	687	0

Fonte: Autoria própria.

No caso da AVFAD, as gravações não possuíam comentários médicos associados, logo todas foram consideradas válidas. Entretanto, como pode ser visualizado na tabela 4, o número

de gravações de pacientes com disfonias funcionais é muito baixo, sendo inferior em mais de dez vezes ao de disfonias organofuncionais, que corresponde à segunda categoria menos representada, e por isso esta categoria foi desconsiderada.

Tabela 4 – Número de gravações da AVFAD para cada categoria de disfonia.

Categoria	Número total de gravações
Funcional	14
Orgânica	171
Organofuncional	141
Saudável	363

Fonte: Autoria própria.

Ao comparar as tabelas 3 e 4, percebe-se que a SVD possui uma maior quantidade de gravações para cada categoria, com exceção da Organofuncional. Porém, em cada gravação da AVFAD, a pronúncia da referida vogal sustentada era realizada três vezes, com uma duração média superior a dois segundos para cada uma delas. Assim, para aumentar a quantidade de dados, cada uma das três pronúncias de cada gravação foi separada como uma gravação distinta. Além isso, as que possuíam duração superior a 2,4 segundos foram divididas em duas partes, com exceção das gravações de pacientes saudáveis, por estarem em maior número.

A tabela 5 apresenta o número total de gravações utilizadas de cada base de dados, após a realização das etapas de seleção e aumento de dados.

Tabela 5 – Número total de gravações utilizadas, para cada base de dados.

Categoria	Número de gravações (SVD)	Número de gravações (AVFAD)
Funcional	414	-
Orgânica	459	913
Organofuncional	112	713
Saudável	687	1089

Fonte: Autoria própria.

Conforme apresentado na seção 2.2 deste trabalho, a utilização de espectrogramas evidencia características vocais que podem facilitar a detecção de disfonias, aprimorando assim a eficácia de métodos computacionais desenvolvidos para este fim. Para que fossem gerados os espectrogramas, um recorte de 1 segundo foi aplicado a cada gravação, com a adoção de *padding* reflexivo nas que possuíam duração menor, e então cada gravação foi normalizada em amplitude, de forma que seu menor e maior valores fossem “-1” e “+1”, respectivamente.

Após isso, foram gerados seus espectrogramas utilizando, da mesma forma que Mohammed *et al*. (2020), um janelamento do tipo *hamming* com duração de 64 milissegundos, e 50% de sobreposição. Em seguida, suas representações de frequência foram redimensionadas para a escala de mel, utilizando 64 bandas, e frequências superiores a 5kHz foram removidas por não

conterem características relevantes a este contexto. Além disso, valores e nomenclaturas dos seus eixos foram removidos, pois não seriam úteis à CNN.

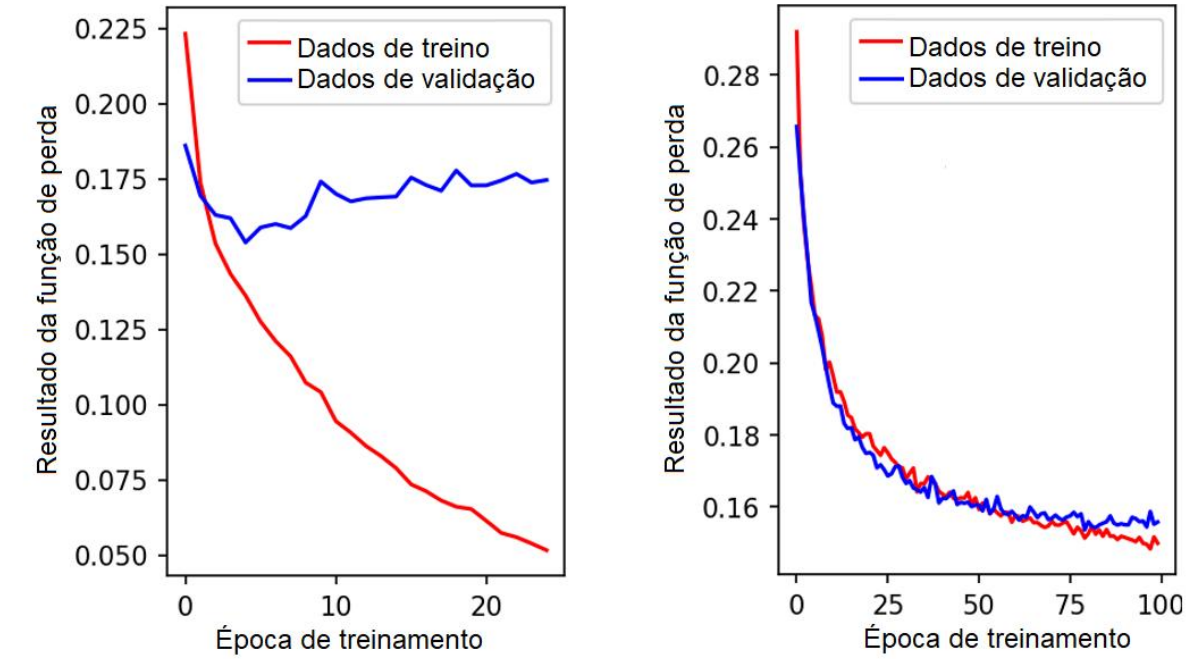
A partir da análise de estudos que utilizam CNNs, como os de Han *et al*. (2018) e Hussain *et al*. (2019), percebe-se que a utilização de arquiteturas pré-existentes de CNN previamente treinadas em grandes bases de dados pode facilitar seu processo de treino, e aprimorar consideravelmente seus resultados.

Para este trabalho, foi selecionada a arquitetura EfficientNet B0, desenvolvida por Tan e Le (2020), pré-treinada na base de dados Imagenet, por apresentar um bom desempenho com um número reduzido de parâmetros treináveis. Assim, optou-se pela utilização de esquema de cores RGB na confecção dos espectrogramas, para que não houvesse a necessidade de remodelar a camada de entrada da CNN, e para adaptá-la à resolução das imagens de entrada e ao número de classes empregados neste trabalho, sua camada totalmente conectada original foi substituída por uma de *pooling* adaptativo, para que fosse criado um vetor de características de tamanho fixo de 1000 neurônios, e uma nova camada totalmente conectada para realizar a classificação, adotando o uso do *dropout* para melhorar seu desempenho.

Após a definição da arquitetura de CNN a ser utilizada, foram realizados testes preliminares de classificação de disfonias, com o intuito de obter parâmetros de treinamento adequados. Com isso, foi definida a utilização do algoritmo de otimização SGD, com taxa de aprendizado de 0,0001, função de perda do tipo Erro Quadrático Médio (EQM), *dropout* de 20%, treinamento simultâneo de todas as camadas da CNN, e como critério de parada de treinamento a não-redução da perda de validação por 20 épocas. Ainda que esta configuração torne o processo de treinamento moroso, ela demonstrou melhores resultados práticos e maior robustez à ocorrência de sobreajuste. A Figura 12 apresenta dois processos de treino utilizados para exemplificação, realizados em uma mesma base de dados: Na parte (A), é utilizado o algoritmo de otimização Adamax, e na parte (B) é utilizado o algoritmo SGD, sendo mantidos os demais parâmetros constantes. Nota-se, para o primeiro caso, que a CNN incorre em sobreajuste já nas primeiras épocas de treinamento, o que não ocorre no segundo caso. A figura 15, contida no apêndice A, é uma captura de tela do código desenvolvido contendo uma lista de todos os parâmetros utilizados para realizar os treinamentos.

A extração de resultados foi conduzida de forma independente para cada base de dados, e realizada em duas etapas: Na primeira, foi avaliada a capacidade da CNN de distinguir as vozes pertencentes a cada categoria de disfonia das consideradas saudáveis, e na segunda, foi avaliada

Figura 12 – Exemplo de teste preliminar utilizado na escolha dos parâmetros de treinamento da CNN.



(A) Treinamento utilizando algoritmo Adamax **(B) Treinamento utilizando algoritmo SGD**

Fonte: Autoria própria.

a capacidade da CNN de reconhecer todas as categorias de disfonia simultaneamente.

Os sinais da vogal /a/ sustentada foram escolhidos para a avaliação de ambas bases de dados, uma vez que esta vogal apresenta melhores resultados de acordo com os trabalhos anteriormente apresentados (MUHAMMAD *et al.*, 2017; BELENKO *et al.*, 2020). Para a SVD, resultados foram produzidos para esta vogal considerando os tons baixo, médio, alto, bem como a junção destes três. No caso da AVFAD, todas as emissões da vogal /a/ sustentadas foram realizadas em tom médio, logo, apenas este pôde ser utilizado.

Além disso, para garantir a qualidade dos resultados apresentados, as bases de dados foram divididas em três subconjuntos, sendo eles: 70% dos dados para treinamento, 15% para validação, e 15% para teste. Esta divisão foi realizada de forma a garantir que todas as gravações de um mesmo indivíduo estejam alocadas no mesmo subconjunto, e também foi aplicado um processo de validação cruzada de cinco etapas, para assegurar a qualidade dos resultados apresentados. Por fim, a avaliação dos resultados foi feita a partir das métricas de precisão (equação 1), revocação (equação 2), F1 Score (equação 3), e acurácia (equação 4).

$$P(i) = \frac{VP(i)}{VP(i) + FP(i)} \tag{1}$$

$$R(i) = \frac{VP(i)}{VP(i) + FN(i)} \tag{2}$$

$$F1(i) = \frac{2P(i)R(i)}{P(i) + R(i)} \tag{3}$$

$$AC = \frac{CP}{NS} \tag{4}$$

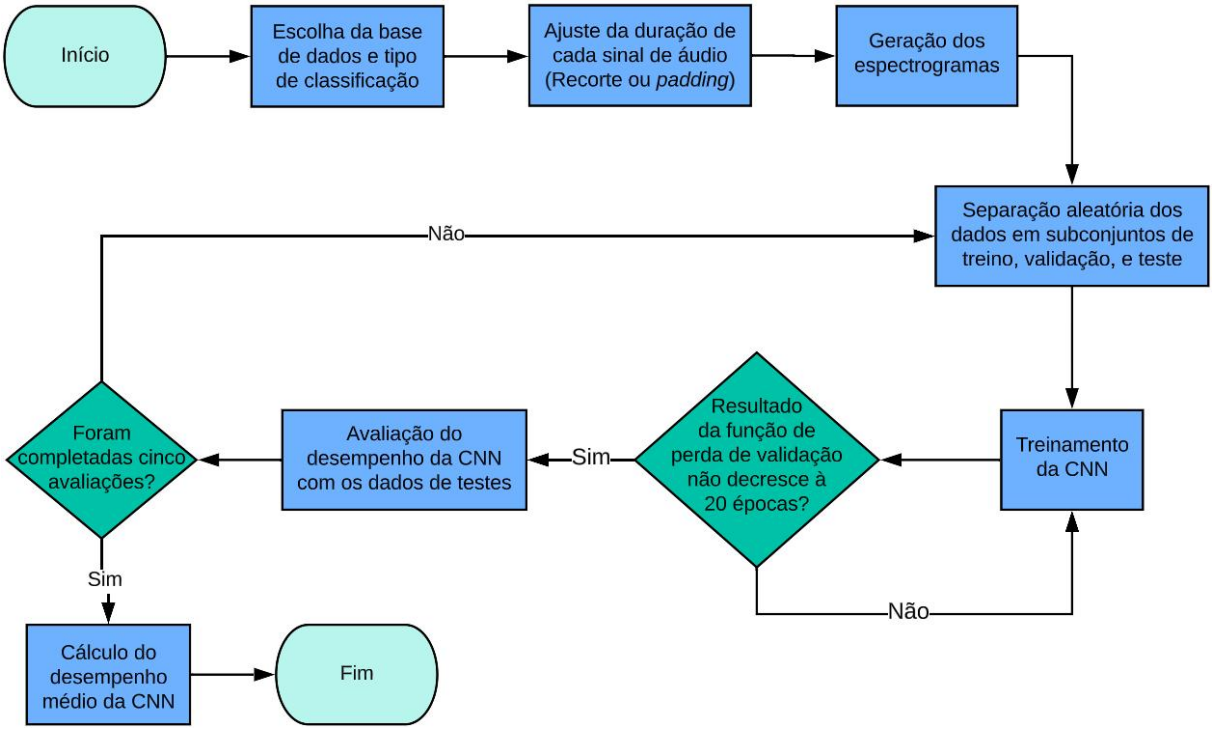
Onde:

- i refere-se a classe analisada;
- P refere-se a métrica de precisão;
- R refere-se a métrica de revocação;
- $F1$ refere-se a métrica de F1 Score;
- AC refere-se a métrica de acurácia;
- VP é o número de predições corretamente classificadas como positivas para a classe analisada;
- FP é o número de predições incorretamente classificadas como positivas para a classe analisada;
- FN é o número de predições incorretamente classificadas como negativas para a classe analisada.
- CP é o número total de predições classificadas corretamente;
- NS é o número total de espectrogramas.

4 RESULTADOS E DISCUSSÃO

A Figura 13 apresenta o fluxograma do algoritmo desenvolvido para avaliar os diferentes treinamentos da CNN e coletar seus resultados, aplicando o processo de validação cruzada, e a Figura 14 apresenta exemplos dos espectrogramas gerados durante a execução deste algoritmo. Além disso, o Apêndice A contém capturas de telas de processos de treinamento e seus resultados.

Figura 13 – Fluxograma do algoritmo desenvolvido para os treinamentos da CNN.

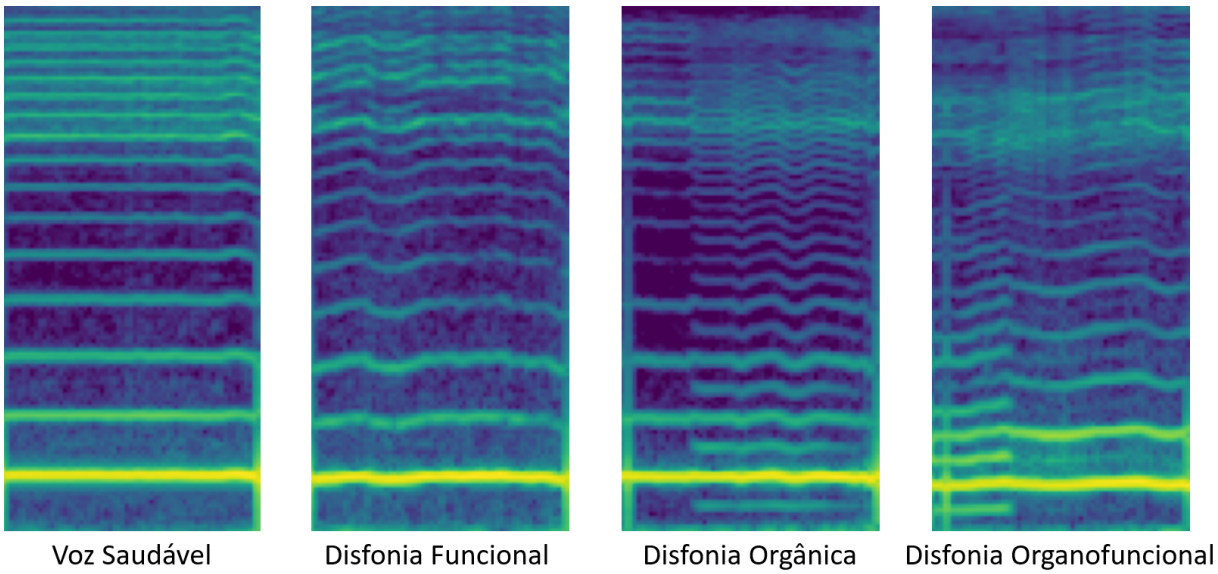


Fonte: Autoria própria.

As tabelas 6 e 7 apresentam os resultados obtidos para a tarefa de distinção binária entre vozes saudáveis e não-saudáveis, para cada categoria analisada. Para a tabela 6, estão destacados os melhores resultados obtidos para as diferentes categorizações, de acordo com o tom em que foram realizadas as pronúncias. Ressalta-se que foram aplicados processos de treinamento distintos para cada tom e categoria de disfonia.

Para os resultados obtidos com a SVD, a partir da análise da tabela 6, nota-se que a utilização simultânea de todos os tons da vogal sustentada apresenta maior desempenho, para todas as categorias de disfonia, de forma análoga aos resultados obtidos por Martínez et al. (2012), e pode ser atribuído à utilização de uma maior quantidade de dados para treinamento e testes. Considerando os resultados obtidos utilizando-se um tom único para a vogal sustentada, o

Figura 14 – Exemplos de espectrogramas gerados pelo algoritmo desenvolvido.



Fonte: Autoria própria.

Tabela 6 – Resultados para a distinção entre vozes saudáveis e não-saudáveis, para cada categoria de disfonia (SVD)

Categoria	Tom	Acurácia	Precisão (Saudável)	Revocação (Saudável)	Precisão (Disfonia)	Revocação (Disfonia)	F1 Score
Funcional	Baixo	62,2	60,4	65,7	58,4	**52,6**	**55,3**
	Médio	67,5	69,5	**86,0**	60,7	36,8	45,8
	Alto	62,9	68,7	74,7	50,8	43,2	46,7
	Todos	**69,0**	**71,1**	84,8	**62,8**	42,6	50,8
Orgânica	Baixo	72,9	72,7	88,3	74,2	49,4	59,3
	Médio	75,4	**77,2**	84,3	72,2	**62,1**	**66,8**
	Alto	72,2	73,6	84,1	70,5	54,1	61,2
	Todos	**76,1**	75,7	**88,8**	**77,3**	57,2	65,7
Organofuncional	Baixo	66,0	70,2	77,1	59,0	48,2	53,0
	Médio	68,9	**75,4**	77,9	55,3	52,5	53,9
	Alto	67,0	74,8	78,5	59,6	**55,0**	**57,2**
	Todos	**72,2**	**75,4**	**87,0**	**61,7**	42,7	50,5

Fonte: Autoria própria.

tom médio permitiu um maior desempenho.

Para ambas bases de dados, a CNN apresentou maior desempenho na detecção de disfonias orgânicas, o que possivelmente deve-se ao fato desta categoria ser composta de disfonias causadas por mudanças estruturais nos órgãos do aparelho fonador, que são melhor evidenciadas nos espectrogramas. Para as disfonias organofuncionais, que também apresentam tais mudanças estruturais porém apenas de forma benigna, a CNN apresentou resultados medianos. Já para as disfonias funcionais, que normalmente são distorções vocais não-acentuadas em comparação com as outras categorias, a CNN teve dificuldades em encontrar características que as representassem

Tabela 7 – Resultados para a distinção entre vozes saudáveis e não-saudáveis, para cada categoria de disfonia (AVFAD)

Categoria	Acurácia	Precisão (Saudável)	Revocação (Saudável)	Precisão (Disfonia)	Revocação (Disfonia)	F1 Score
Orgânica	82,8	83,3	85,8	83,1	79,4	81,2
Organofuncional	77,3	77,4	86,2	77,8	65,8	71,3

Fonte: Autoria própria.

nos espectrogramas, visto a baixa assertividade obtida.

As tabelas 8 e 9 apresentam os resultados obtidos para o reconhecimento simultâneo das categorias de disfonia. Nelas, as notações *Acc, P, R,* e *F1* denotam as métricas de *acurácia, precisão, revocação,* e *F1 Score*, respectivamente, e as notações *S, F, O,* e *OF* representam as categorias *Saudável, Funcional, Orgânica,* e *Organofuncional,* respectivamente. Para os dados da SVD, estão destacados os melhores resultados obtidos de acordo com o tom em que foram realizadas as pronúncias.

Tabela 8 – Resultados para a classificação entre as categorias de disfonias (SVD). As nomenclaturas "Acc", "P","R", e "F1" denotam as métricas de acurácia, precisão, revocação, e F1 Score, respectivamente, e as nomenclaturas "(S)", "(F)", "(O)", e "(OF)" denotam as classes de indivíduos saudáveis, e disfonias Funcional, Orgânica, e Organofuncional, respectivamente.

Tom	Acc	P(S)	R(S)	F1(S)	P(F)	R(F)	F1(F)	P(O)	R(O)	F1(O)	P(OF)	R(OF)	F1(OF)
Baixo	49,5	54,3	65,1	59,2	**29,1**	3,60	6,41	50,1	**51,6**	50,8	37,0	32,7	34,7
Médio	52,9	56,0	70,1	62,3	14,1	**10,9**	**12,3**	42,1	38,5	40,2	**44,7**	**37,9**	**41,0**
Alto	48,8	52,6	72,0	60,8	16,2	2,70	4,63	**55,4**	46,8	50,7	33,0	24,6	28,2
Todos	**53,2**	**57,4**	**81,0**	**67,2**	25,0	0,50	0,98	54,7	51,2	**52,9**	36,3	23,5	28,5

Fonte: Autoria própria.

Tabela 9 – Resultados para a classificação entre as categorias de disfonias (AVFAD). A nomenclatura "F1" denota a métrica de F1 Score, e as nomenclaturas "(S)", "(O)", e "(OF)" denotam as classes de indivíduos saudáveis, e disfonias Orgânica e Organofuncional, respectivamente.

Acurácia	Precisão (S)	Revocação (S)	F1 (S)	Precisão (O)	Revocação (O)	F1 (O)	Precisão (OF)	Revocação (OF)	F1 (OF)
59,8	75,1	78,5	76,8	52,2	56,0	54,0	38,3	30,8	34,1

Fonte: Autoria própria.

Da mesma forma como na etapa de detecção binária de disfonias, a utilização simultânea de todos os tons produziu melhores resultados para a SVD, sendo estes levemente superiores aos obtidos para o tom médio. Além disso, nota-se que a CNN obteve êxito em reconhecer disfonias orgânicas dessa base de dados, porém não foi capaz de reconhecer disfonias funcionais, classificando-as incorretamente como saudáveis devido ao baixo grau de distorção vocal dos indivíduos que as possuem. Para o reconhecimento de disfonias organofuncionais, mesmo estando pouco representadas nesta base de dados, os resultados obtidos foram superiores aos

de disfonias funcionais, indicando que a CNN também é capaz de reconhecer esta categoria de disfonia, porém de forma limitada.

Para a AVFAD, a CNN obteve desempenho levemente superior ao da SVD no reconhecimento de disfonias orgânicas e organofuncionais. Contudo, neste caso, não houve classificação de disfonias funcionais, e as demais categorias foram representadas por menores quantidades de disfonias, o que facilita a tarefa de classificação, como também pode ser observado na etapa de distinção binária de disfonias.

A tabela 10 apresenta uma comparação entre os resultados do presente trabalho com os de trabalhos anteriores, para a distinção binária entre vozes saudáveis e com disfonia utilizando a SVD.

Tabela 10 – Comparação de resultados entre este trabalho e trabalhos anteriores. As nomenclaturas "(O)" e "(OF)" denotam as classes de disfonia orgânica e organofuncional, respectivamente.

Autores	Disfonias	Quantidade de Disfonias	Maior Acurácia (SVD)
Martínez *et al* (2012)	Todos os tipos	71	79,4%
Wu *et al* (2018)	(O, OF)	6	77%
Belenko *et al* (2020)	(O, OF)	6	73%
Ding *et al* (2021)	Todos os tipos	71	81,6%
Presente trabalho	(O)	29	76,1%
Presente trabalho	(OF)	6	72,2%

Fonte: Autoria própria.

Ao analisá-la, pode-se inferir que a metodologia utilizada para a extração de características e classificação dos sinais vocais produziu resultados satisfatórios, sendo superiores aos de Belenko *et al* (2020) e próximos aos de Wu *et al* (2018), mesmo empregando uma maior variedade de tipos de disfonias. Além disso, para a detecção de disfonias organofuncionais, que ocorreu de forma não-balanceada devido à baixa quantidade de amostras disponíveis, a acurácia máxima obtida foi de 72,2%, comparável à obtida por Belenko *et al* (2020), de 73%.

Apesar do trabalho de Martínez *et al* (2012) ter atingido uma maior performance de detecção de disfonias, conforme explicado na seção 2.4 deste texto, seus resultados podem estar enviesados devido à possibilidade de pacientes poderem ter suas vozes representadas em dados de treinamento e teste simultaneamente. Além disso, o trabalho de Ding *et al* (2021), que também apresentou performance superior, foi conduzido utilizando uma arquitetura mais complexa de CNN, a qual necessita de melhores recursos de *hardware* para seu treinamento.

Para a AVFAD, este trabalho obteve acurácia máxima de classificação de 82,8%, a qual é comparável à obtida por Oliveira *et al*, (2020), de 83,16%, sendo este o único trabalho deste tipo encontrado na literatura que utiliza essa mesma base de dados. Já para as etapas de categorização

de disfonias, não foram encontrados trabalhos similares na literatura, impossibilitando assim a efetuação de análises comparativas.

5 CONCLUSÕES

O presente trabalho propõe uma nova metodologia para a classificação automática de disfonias, em que estas são agrupadas em categorias de acordo com sua etiologia, de modo a extrair uma informação mais detalhada que uma simples distinção entre vozes saudáveis e com disfonia, porém mantendo um número reduzido de classificações possíveis, sem a necessidade de desconsiderar disfonias com baixa representatividade entre os dados.

Os resultados aqui apresentados demonstram que a escolha realizada para os métodos de extração de características e classificação dos sinais vocais proporcionou um desempenho de classificação semelhante ao encontrado no estado da arte atual, permitindo a identificação de disfonias orgânicas e organofuncionais com 76,1% e 72,2% de acurácia, respectivamente.

Assim, observa-se que a adoção do critério de etiologia pode ser aplicada para separar disfonias organofuncionais, que acontecem de forma benigna e em geral podem ser amenizadas a partir de fonoterapia, das disfonias orgânicas, que podem ter causas malignas e exigir tratamentos complexos. Portanto, apesar da dificuldade do método em reconhecer disfonias funcionais, sua utilização atende aos objetivos propostos.

5.1 TRABALHOS FUTUROS

Uma vez que a CNN não foi capaz de distinguir disfonias funcionais de modo satisfatórios, novos métodos de extração de características podem ser empregados em trabalhos futuros, assim como possíveis novos critérios de agrupamento de disfonias que tenham maior relação com as distorções por elas provocadas nos sinais vocais.

Além disso, métodos podem ser empregados para realizar classificações quanto à qualidade dos sinais vocais, ao invés de tipos de disfonias, de forma semelhante aos métodos perceptivo-auditivos de diagnóstico, como o CAPE-V. Para isso, seria necessário a disponibilização de profissionais para classificar gravações das bases de dados de acordo com tal método, porém de um ponto de vista computacional, uma vez que as classificações seriam diretamente relacionadas aos níveis de distorção vocal dos pacientes, há potencial para uma melhora significativa na assertividade dos classificadores.

REFERÊNCIAS

ALMEIDA, Nathalee C de; BARROS, Jannayna D; SOARES, Heliana B; BRESOLIN, Adriano de A; GUERREIRO, Ana Maria G; BRANDAO, Glaucio B. A new computational tool for voice analysis based on fft, wavelet transform, and spectrogram. *In*: AMERICAN SOCIETY OF MECHANICAL ENGINEERS. **Summer Bioengineering Conference**. *[S.l.]*, 2009. v. 48913, p. 805–806.

ALTMAN, Kenneth W.; ATKINSON, Cory; LAZARUS, Cathy. Current and emerging concepts in muscle tension dysphonia: A 30-month review. **Journal of Voice**, v. 19, n. 2, p. 261–267, 2005. ISSN 0892-1997. Disponível em: https://www.sciencedirect.com/science/article/pii/S0892199704000505.

BARRY, WJ; PÜTZER, M. Saarbrucken voice database. **Institute of Phonetics**, University of Saarland, 2007. Disponível em: http://stimmdb.coli.uni-saarland.de/.

BEHLAU, Mara. **Voz: O Livro do Especialista**. *[S.l.]*: Thieme Revinter, 2001. ISBN 8573095253.

BEHLAU, Mara; ZAMBON, Fabiana; GUERRIERI, Ana Cláudia; ROY, Nelson. Epidemiology of voice disorders in teachers and nonteachers in brazil: Prevalence and adverse effects. **Journal of Voice**, v. 26, n. 5, p. 665.e9–665.e18, 2012. ISSN 0892-1997. Disponível em: https://www.sciencedirect.com/science/article/pii/S0892199711001664.

BEHLAU MARA AND GASPARINI, Gisele. Classification manual for voice disorders-i - cmvd-i. *In*: **Revista da Sociedade Brasileira de Fonoaudiologia**. *[s.n.]*, 2007. Disponível em: https://doi.org/10.1590/S1516-80342007000100014>.

BELENKO, Mikhail; BURYM, Nikita; BALAKSHIN, Pavel. Detection of defective speech using convolutional neural networks. *In*: **CEUR Workshop Proceedings**. *[S.l.: s.n.]*, 2020.

CASERTA, Mary T. Acute laryngitis. **Mandell, Douglas, and Bennett's principles and practice of infectious diseases**, Elsevier, p. 760, 2015.

COHEN, Seth M; KIM, Jaewhan; ROY, Nelson; ASCHE, Carl; COUREY, Mark. Prevalence and causes of dysphonia in a large treatment-seeking population. **The Laryngoscope**, Wiley Online Library, v. 122, n. 2, p. 343–348, 2012.

CRISTOFOLINI, Carla. **Produção e Articulação dos Sons da Fala**. Universidade Federal de Santa Catarina, 2013. Disponível em: https://moodle.ufsc.br/pluginfile.php/905248/mod_resource/content/1/Sons_da_fala_letras.pdf.

DING, Huijun; GU, Zixiong; DAI, Peng; ZHOU, Zhou; WANG, Lu; WU, Xiaoxiao. Deep connected attention (dca) resnet for robust voice pathology detection and classification. **Biomedical Signal Processing and Control**, v. 70, p. 102973, 2021. ISSN 1746-8094. Disponível em: https://www.sciencedirect.com/science/article/pii/S174680942100570X.

FLEISCHER, Susanne; PFLUG, Christina; HESS, Markus. Dipping and rotating: two maneuvers to achieve maximum magnification during indirect transnasal laryngoscopy. **European Archives of Oto-Rhino-Laryngology**, v. 277, 05 2020.

GHAZANFAR, Asif A; RENDALL, Drew. Evolution of human vocal production. **Current biology**, Elsevier, v. 18, n. 11, p. R457–R460, 2008.

GOODFELLOW, Ian; BENGIO, Yoshua; COURVILLE, Aaron. **Deep Learning**. *[S.l.]*: MIT Press, 2016. http://www.deeplearningbook.org.

HAMMAMI, Imen; SALHI, Lotfi; LABIDI, Salam. Pathological voices detection using support vector machine. *In*: **2016 2nd International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)**. *[S.l.: s.n.]*, 2016. p. 662–666.

HAN, Dongmei; LIU, Qigang; FAN, Weiguo. A new image classification method using cnn transfer learning and web data augmentation. **Expert Systems with Applications**, v. 95, p. 43–56, 2018. ISSN 0957-4174. Disponível em: http://www.sciencedirect.com/science/article/pii/S0957417417307844.

HAUX, Reinhold. Medical informatics: Past, present, future. **International Journal of Medical Informatics**, v. 79, n. 9, p. 599–610, 2010. ISSN 1386-5056. Disponível em: https://www.sciencedirect.com/science/article/pii/S1386505610001140.

HE, Kaiming; ZHANG, Xiangyu; REN, Shaoqing; SUN, Jian. **Deep Residual Learning for Image Recognition**. arXiv, 2015. Disponível em: https://arxiv.org/abs/1512.03385.

HUSSAIN, Mahbub; BIRD, Jordan J.; FARIA, Diego R. A study on cnn transfer learning for image classification. *In*: LOTFI, Ahmad; BOUCHACHIA, Hamid; GEGOV, Alexander; LANGENSIEPEN, Caroline; MCGINNITY, Martin (Ed.). **Advances in Computational Intelligence Systems**. Cham: Springer International Publishing, 2019. p. 191–202. ISBN 978-3-319-97982-3.

JEON, Hohyub; JUNG, Yongchul; LEE, Seongjoo; JUNG, Yunho. Area-efficient short-time fourier transform processor for time–frequency analysis of non-stationary signals. **Applied Sciences**, v. 10, n. 20, 2020. ISSN 2076-3417. Disponível em: https://www.mdpi.com/2076-3417/10/20/7208.

JESUS, L.; BELO, I.; MACHADO, J.; HALL, A. The advanced voice function assessment databases (avfad): Tools for voice clinicians and speech engineering research. *In*: **Advances in Speech-language Pathology**. *[s.n.]*, 2017. Disponível em: http://acsa.web.ua.pt/.

KARKOS, Petros D; MCCORMICK, Maxwell. The etiology of vocal fold nodules in adults. **Current opinion in otolaryngology & head and neck surgery**, LWW, v. 17, n. 6, p. 420–423, 2009.

KELLERMAN, Rick; KINTANAR, Thomas. Gastroesophageal reflux disease. **Primary Care: Clinics in Office Practice**, v. 44, n. 4, p. 561–573, 2017. ISSN 0095-4543. Gastroenterology. Disponível em: https://www.sciencedirect.com/science/article/pii/S0095454317300957.

KOSZTYŁA-HOJNA, Bożena; MOSKAL, Diana; SITNIK, Anna Łobaczuk; KRASZEWSKA, Anna; ZDROJKOWSKI, Maciej; BISZEWSKA, Jolanta; SKORUPA, Małgorzata. Psychogenic voice disorders. **Polish Journal of Otolaryngology**, v. 72, n. 4, p. 26–34, 2018. ISSN 0030-6657. Disponível em: https://otolaryngologypl.com/resources/html/article/details?id=186441&language=en.

KROSE, Ben; SMAGT, Patrick van der. **An introduction to neural networks**. *[S.l.: s.n.]*, 2011.

LECUN, Yann; BOTTOU, Léon; BENGIO, Yoshua; HAFFNER, Patrick. Gradient-based learning applied to document recognition. **Proceedings of the IEEE**, Ieee, v. 86, n. 11, p. 2278–2324, 1998.

LIBRARY-OF-CONGRESS. **About the *Æsop for Children***. 2009. Disponível em: https://www.read.gov/aesop/143.html.

LYBERG-ÅHLANDER, Viveka; RYDELL, Roland; FREDLUND, Peeter; MAGNUSSON, Cecilia; WILéN, Staffan. Prevalence of voice disorders in the general population, based on the stockholm public health cohort. **Journal of Voice**, v. 33, n. 6, p. 900–905, 2019. ISSN 0892-1997. Disponível em: https://www.sciencedirect.com/science/article/pii/S0892199718301759.

MANNELLI, Giuditta; COMINI, Lara Valentina; SANTORO, Roberto; BETTIOL, Alessandra; VANNACCI, Alfredo; DESIDERI, Isacco; BONOMO, Pierluigi; PIAZZA, Cesare. T1 glottic cancer: Does anterior commissure involvement worsen prognosis? **Cancers**, v. 12, n. 6, 2020. ISSN 2072-6694. Disponível em: https://www.mdpi.com/2072-6694/12/6/1485.

MARTÍNEZ, David; LLEIDA, Eduardo; ORTEGA, Alfonso; MIGUEL, Antonio; VILLALBA, Jesús. Voice pathology detection on the saarbrücken voice database with calibration and fusion of scores using multifocal toolkit. *In*: TOLEDANO, Doroteo Torre; GIMÉNEZ, Alfonso Ortega; TEIXEIRA, António; RODRÍGUEZ, Joaquín González; GÓMEZ, Luis Hernández; HERNÁNDEZ, Rubén San Segundo; CASTRO, Daniel Ramos (Ed.). **Advances in Speech and Language Technologies for Iberian Languages**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012. p. 99–109. ISBN 978-3-642-35292-8.

MEHTA, VK; DEB, PS; RAO, D SUBBA. Application of computer techniques in medicine. **Medical Journal Armed Forces India**, v. 50, n. 3, p. 215–218, 1994. ISSN 0377-1237. Disponível em: https://www.sciencedirect.com/science/article/pii/S0377123717310651.

MOHAMMED, Mazin Abed; ABDULKAREEM, Karrar Hameed; MOSTAFA, Salama A.; GHANI, Mohd Khanapi Abd; MAASHI, Mashael S.; GARCIA-ZAPIRAIN, Begonya; OLEAGORDIA, Ibon; ALHAKAMI, Hosam; AL-DHIEF, Fahad Taha. Voice pathology detection and classification using convolutional neural network model. **Applied Sciences**, v. 10, n. 11, 2020. ISSN 2076-3417. Disponível em: https://www.mdpi.com/2076-3417/10/11/3723.

MUHAMMAD, Ghulam; ALSULAIMAN, Mansour; ALI, Zulfiqar; MESALLAM, Tamer A; FARAHAT, Mohamed; MALKI, Khalid H; AL-NASHERI, Ahmed; BENCHERIF, Mohamed A. Voice pathology detection using interlaced derivative pattern on glottal source excitation. **Biomedical signal processing and control**, Elsevier, v. 31, p. 156–164, 2017.

OLIVEIRA, Brígida F. C.; MAGALHãES, Deborah M. V.; FERREIRA, Daniel S.; MEDEIROS, Fátima N. S. Combined sustained vowels improve the performance of the haar wavelet for pathological voice characterization. *In*: **2020 International Conference on Systems, Signals and Image Processing (IWSSIP)**. *[S.l.: s.n.]*, 2020. p. 381–386.

OMORI, Koichi. Diagnosis of voice disorders. **JMAJ**, v. 54, n. 4, p. 248–253, 2011.

OTT, Jordan; PRITCHARD, Mike; BEST, Natalie; LINSTEAD, Erik; CURCIC, Milan; BALDI, Pierre. A fortran-keras deep learning bridge for scientific computing. *In*: . *[S.l.: s.n.]*, 2020.

PEREIRA, Gabriela da Cunha; LEMOS, Isadora de Oliveira; GADENZ, Camila Dalbosco; CASSOL, Mauriceia. Effects of voice therapy on muscle tension dysphonia: A systematic literature review. **Journal of Voice**, v. 32, n. 5, p. 546–552, 2018. ISSN 0892-1997. Disponível em: https://www.sciencedirect.com/science/article/pii/S0892199717300668.

RAVì, Daniele; WONG, Charence; DELIGIANNI, Fani; BERTHELOT, Melissa; ANDREU-PEREZ, Javier; LO, Benny; YANG, Guang-Zhong. Deep learning for health informatics. **IEEE Journal of Biomedical and Health Informatics**, v. 21, n. 1, p. 4–21, 2017.

ROSEN, Deborah Caputo; SATALOFF, Johnathan Brandon; SATALOFF, Robert Thayer. **Psychology of voice disorders**. *[S.l.]*: Plural Publishing, 2020.

ROY, Nelson; MERRILL, Ray M.; GRAY, Steven D.; SMITH, Elaine M. Voice disorders in the general population: Prevalence, risk factors, and occupational impact. **The Laryngoscope**, v. 115, n. 11, p. 1988–1995, 2005. Disponível em: https://onlinelibrary.wiley.com/doi/abs/10.1097/01.mlg.0000179174.32345.41.

RUDER, Sebastian. An overview of gradient descent optimization algorithms. **arXiv preprint arXiv:1609.04747**, 2016.

RUMELHART, David E.; HINTON, Geoffrey E.; WILLIAMS, Ronald J. Learning representations by back-propagating errors. **Nature**, v. 323, p. 533–536, 1986.

SALONI; SHARMA, Rajender K; GUPTA, Anil K. Disease detection using voice analysis: a review. **Int. J. of Medical Engineering and Informatics**, v. 6, p. 189 – 209, 01 2014.

SATALOFF, Robert T; CHOWDHURY, Farhad; HAWKSHAW, Mary J; JOGLEKAR, Shruti; PORTNOY, Joel E. **Voz: O Livro do Especialista**. *[S.l.]*: Jaypee Brothers Medical Publishers (P) Ltd., 2014. ISBN 9789350906521.

SIMPSON, C Blake; FLEMING, Daniel J. Medical and vocal history in the evaluation of dysphonia. **Otolaryngologic Clinics of North America**, Elsevier, v. 33, n. 4, p. 719–729, 2000.

SULICA, Lucian. Laryngoscopy, stroboscopy and other tools for the evaluation of voice disorders. **Otolaryngologic Clinics of North America**, v. 46, n. 1, p. 21–30, 2013. ISSN 0030-6665. Office Procedures in Laryngology. Disponível em: https://www.sciencedirect.com/science/article/pii/S0030666512001351.

TAN, Mingxing; LE, Quoc V. Efficientnet: Rethinking model scaling for convolutional neural networks. *In*: 36^{th} **International Conference on Machine Learning**. *[s.n.]*, 2020. Disponível em: https://arxiv.org/abs/1905.11946.

TAVALUC, Raluca; TAN-GELLER, Melin. Reinke's edema. **Otolaryngologic Clinics of North America**, v. 52, n. 4, p. 627–635, 2019. ISSN 0030-6665. Advancements in Clinical Laryngology. Disponível em: https://www.sciencedirect.com/science/article/pii/S0030666519300477.

THE BRITISH VOICE ASSOCIATION, Voice Care. **Reflux and your Voice**. 2009. Disponível em: https://www.britishvoiceassociation.org.uk/voicecare_reflux-and-voice.htm.

TOHYAMA, Mikio. Chapter 3 - modulation waveform and masking effect. *In*: TOHYAMA, Mikio (Ed.). **Acoustic Signals and Hearing**. Academic Press, 2020. p. 47–62. ISBN 978-0-12-816391-7. Disponível em: https://www.sciencedirect.com/science/article/pii/B9780128163917000115.

TUZUNER, Arzu; DEMIRCI, Sule; OGUZ, Haldun; OZCAN, Kursat Murat. Pediatric vocal fold nodule etiology: what are its usual causes in children? **Journal of Voice**, Elsevier, v. 31, n. 4, p. 506–e19, 2017.

UNIVERSITY OF NEBRASKA, Medical Center. **Benign Lesions**. 2006. Disponível em: https://app1.unmc.edu/medicine/heywood/laryngealdisease/Data/benignlesions.htm.

UNIVERSITY OF NEBRASKA, Medical Center. **Infectious Laryngitis**. 2006. Disponível em: https://app1.unmc.edu/medicine/heywood/laryngealdisease/Data/infection.htm.

VAN-HOUTTE, Evelyne; VAN-LIERDE, Kristiane; D'HAESELEER, Evelien; CLAEYS, Sofie. The prevalence of laryngeal pathology in a treatment-seeking population with dysphonia. **The Laryngoscope**, Wiley Online Library, v. 120, n. 2, p. 306–312, 2010.

WU, Huiyi; SORAGHAN, John; LOWIT, Anja; CATERINA, Gaetano Di. Convolutional neural networks for pathological voice detection. *In*: IEEE. **2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)**. *[S.l.]*, 2018. p. 1–4.

ZHOU, Pan; FENG, Jiashi; MA, Chao; XIONG, Caiming; HOI, Steven; E, Weinan. Towards theoretically understanding why sgd generalizes better than adam in deep learning. arXiv, 2020. Disponível em: https://arxiv.org/abs/2010.05627.

ZRAICK, Richard I; KEMPSTER, Gail B; CONNOR, Nadine P; THIBEAULT, Susan; KLABEN, Bernice K; BURSAC, Zoran; THRUSH, Carol R; GLAZE, Leslie E. Establishing validity of the consensus auditory-perceptual evaluation of voice (cape-v). ASHA, 2011.

APÊNDICES

APÊNDICE A – CAPTURAS DE TELA REFERENTES AOS PROCESSOS DE TREINAMENTO DA CNN

A Figura 15 é uma captura de tela do código-fonte desenvolvido, apresentando todos os parâmetros utilizados em relação à base de dados, confecção de espectrogramas, processo de treinamento, e arquitetura da CNN, para que seja realizado o processo de treinamento das CNNs.

Figura 15 – Parâmetros de treinamento da CNN.

```
# Base de dados e espectrogramas:
remake       = True           # Re-computa os espectrogramas
data_dir     = os.getcwd()    # Diretório onde se localiza a base de dados
dataset_name = "SVD_ALL_A"    # Nome da base de dados

sampling_rate = 50000         # Taxa de amostragem: 50kHz
audio_size    = 50000         # 50.000 amostras -> 1 segundo
window_size   = 0.064         # Janela de 64ms
hop_div       = 2             # 50% de sobreposição
top_db        = 80            # dB máximo: 80
crop_mode     = 'fill_crop'   # Padding reflexivo para áudios menores que o especificados, e "cropping" para áudios maiores
freq_min      = 0             # Frequência mínima do espectrograma
freq_max      = 5000          # Frequência máxima do espectrograma
is_rgb        = True          # Uso de imagens RGB ou em escala de cinza
spec_y_type   = 'mel'         # Uso de espectrogramas em escala de mel
n_mels        = 64            # Uso de 64 bandas de mel. Utilizado apenas se <spec_y_type = 'mel'>

# Parâmetros de treinamento:
train_perc = 0.70             # 70% dos dados para treinamento
valid_perc = 0.15             # 15% dos dados para validação
                              # Logos, os demais 15% serão utilizados para testes

max_epochs             = 500  # Número máximo de épocas de treino
freeze_epochs          = 15   # Número de épocas em que a parte convolucional da CNN é congelada,
                              # para a realização de "Fine-tunning" em CNNs pré-treinadas

n_epochs_to_stop_train = 20   # Interrompe o treinamento após a não-diminuição da perda de validação por ao menos 20 épocas

learning_rate    = 0.0001           # Taxa de aprendizado
batch_size       = 4                # Tamanho da batelada
use_CrossEntropy = False            # Escolha entre as funções de perda "Entropia Cruzada Categórica" e "Erro Quadrático Médio"
Optimizer        = torch.optim.SGD  # SGD Optimizer

# Parâmetros da arquitetura da CNN
architecture    = 'EfficientNetB0'  # Arquitetura da CNN
pretrained      = True              # Utilização de rede pré-treinada
dropout         = 0.20              # Utilização de dropout
```

Fonte: Autoria própria.

A Figura 16 apresenta um exemplo da saída de informações do *script* desenvolvido neste trabalho, contendo a perda e a acurácia médias para cada época de treino de uma CNN, bem como a atuação do critério de parada de treinamento, conforme descrito na seção 3.3 deste texto.

A Figura 17 apresenta de forma gráfica os resultados mostrados anteriormente na Figura 16, após o término do processo de treinamento. A partir desta representação, é possível observar que este processo de treino específico ocorreu de forma satisfatória, e não houve sobreajuste dos pesos sinápticos da CNN.

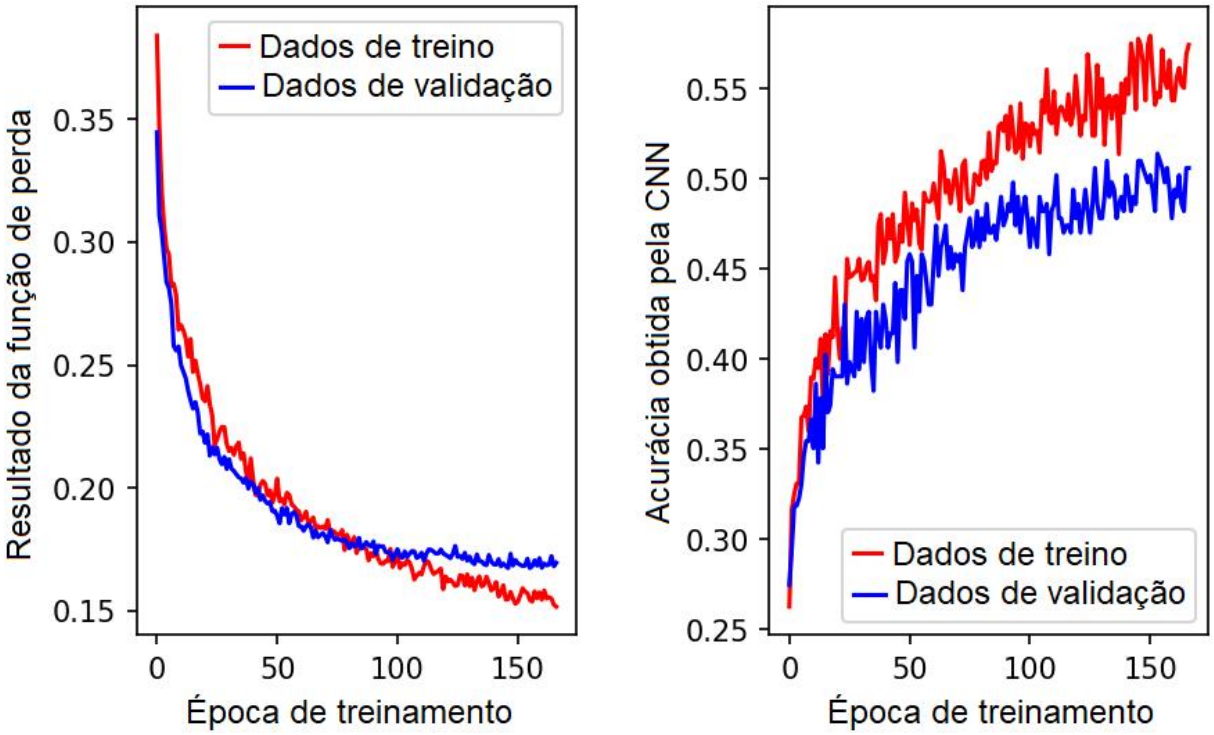
Figura 16 – Exemplo do monitoramento do processo de treino da CNN.

Epoch	Train Loss	Valid Loss	Train Acc	Valid Acc
001	0.3840	0.3446	0.2628	0.2749
002	0.3470	0.3110	0.3166	0.2948
003	0.3207	0.3042	0.3251	0.3187
004	0.3064	0.2936	0.3311	0.3187
005	0.2973	0.2837	0.3311	0.3227
006	0.2950	0.2812	0.3677	0.3307
007	0.2821	0.2750	0.3686	0.3466
008	0.2830	0.2577	0.3737	0.3546
009	0.2783	0.2558	0.3601	0.3546
010	0.2643	0.2575	0.3899	0.3665
011	0.2662	0.2499	0.3891	0.3506
012	0.2641	0.2473	0.4002	0.3865
013	0.2608	0.2446	0.3951	0.3426
014	0.2534	0.2395	0.4113	0.3785
015	0.2606	0.2356	0.3729	0.3506
016	0.2471	0.2323	0.4138	0.4024
017	0.2518	0.2349	0.3916	0.3705
018	0.2465	0.2306	0.4155	0.3745
019	0.2421	0.2221	0.4121	0.3944
020	0.2364	0.2230	0.4454	0.3904
		...		
152	0.1562	0.1687	0.5580	0.4940
153	0.1591	0.1682	0.5410	0.4821
154	0.1577	0.1684	0.5478	0.5139
155	0.1572	0.1711	0.5452	0.5100
156	0.1544	0.1674	0.5717	0.5060
157	0.1537	0.1682	0.5546	0.4980
158	0.1579	0.1705	0.5503	0.5060
159	0.1551	0.1683	0.5657	0.4940
160	0.1575	0.1714	0.5435	0.4781
161	0.1545	0.1675	0.5435	0.4940
162	0.1582	0.1692	0.5563	0.4900
163	0.1550	0.1685	0.5614	0.5020
164	0.1555	0.1687	0.5520	0.4861
165	0.1550	0.1722	0.5503	0.4821
166	0.1524	0.1681	0.5691	0.5060
167	0.1517	0.1695	0.5742	0.5060
Stopping training to avoid overfitting!				

Fonte: Autoria própria.

Por fim, a Figura 18 apresenta os resultados obtidos para este processo de treinamento de exemplo. Para cada processo de validação cruzada, devem ser armazenados estes resultados de cinco processos de treinamento distintos, para que seja calculado o valor médio de cada métrica.

Figura 17 – Gráfico de exemplo da evolução da perda e da acurácia da CNN ao longo das épocas de treinamento.



Fonte: Autoria própria.

Figura 18 – Exemplo de resultados de teste da CNN.

```
Accuracy = 55.0201
Precision: [0.582, 0.0, 0.344, 0.578]  |  Recall: [0.864, 0.0, 0.177, 0.544]
Confusion Matrix:
[[89.  0.  8.  6.]
 [ 7.  0.  1.  8.]
 [38.  0. 11. 13.]
 [19.  0. 12. 37.]]
Classes: ['Sem Disfonia', 'Disfonia Organofuncional', 'Disfonia Funcional', 'Disfonia Orgânica']
```

Fonte: Autoria própria.