UNIVERSIDADE FEDERAL FLUMINENSE
ESCOLA DE ENGENHARIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA E DE TELECOMUNICAÇÕES

VINÍCIUS FLORES CARDOSO

Comparação das técnicas MFCC, PNCC e ZCPA na identificação de patologias relacionadas à voz, usando Redes Neurais Artificiais

NITERÓI

2023

UNIVERSIDADE FEDERAL FLUMINENSE
ESCOLA DE ENGENHARIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA E DE TELECOMUNICAÇÕES

VINÍCIUS FLORES CARDOSO

Comparação das técnicas MFCC, PNCC e ZCPA na identificação de patologias relacionadas à voz, usando Redes Neurais Artificiais

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Engenharia Elétrica e de Telecomunicações da Universidade Federal Fluminense como requisito parcial para a obtenção do título de Mestre em Engenharia Elétrica e de Telecomunicações. Área de concentração: Sinais e Sistemas de Comunicações Móveis.

Orientador:
Prof. Dr. Edson Luiz Cataldo Ferreira

NITERÓI

2023

Ficha catalográfica automática - SDC/BEE
Gerada com informações fornecidas pelo autor

C268c Cardoso, Vinícius Flores
Comparação das técnicas MFCC, PNCC e ZCPA na identificação de patologias relacionadas à voz, usando Redes Neurais Artificiais / Vinícius Flores Cardoso. - 2023.
78 f.

Orientador: Edson Luiz Cataldo Ferreira.
Dissertação (mestrado)-Universidade Federal Fluminense, Escola de Engenharia, Niterói, 2023.

1. Voz. 2. Inteligência artificial. 3. Sistema de telecomunicação. 4. Produção intelectual. I. Ferreira, Edson Luiz Cataldo, orientador. II. Universidade Federal Fluminense. Escola de Engenharia.III. Título.

CDD - XXX

Bibliotecário responsável: Debora do Nascimento - CRB7/6368

VINÍCUS FLORES CARDOSO

Comparação das técnicas MFCC, PNCC e ZCPA na identificação de patologias relacionadas à voz, usando Redes Neurais Artificiais

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Engenharia Elétrica e de Telecomunicações da Universidade Federal Fluminense como requisito parcial para a obtenção do título de Mestre em Engenharia Elétrica e de Telecomunicações. Área de concentração: Sinais e Sistemas de Comunicações Móveis.

Aprovada em 14 de dezembro de 2023.

BANCA EXAMINADORA

Prof. Dr. Edson Luiz Cataldo Ferreira – Orientador
Universidade Federal Fluminense – UFF

Prof. Dr. Leonardo Forero Mendoza
Universidade do Estado do Rio de Janeiro – UERJ

Prof. Dr. Pedro Vladimir Gonzalez Castellano
Universidade Federal Fluminense – UFF

Niterói
2023

À familia

Agradecimentos

À minha grande família, pais, tios, padrinhos, sobrinhos, primos e afilhados por sempre estarem perto quando necessário.

À minha nova família, Thais, por ser um exemplo de força e uma parceira de vida; Pedro, por ser um garoto de um coração gigante e meu filho Benício (Beni), que ainda nem chegou, mas que já me faz transbordar de felicidade.

Ao meu orientador, Edson, por todo conhecimento transmitido, paciência e oportunidade de poder aprender cada vez mais.

Ao professor Leonardo Forero Mendoza, pelas orientações e oportunidade de grande aprendizado.

Ao professor Jorge Petrucio Viana, por abrir a minha mente para grandes possibilidades.

Aos colegas e amigos do Colégio Estadual Baltazar Bernardino e Colégio Salesiano Santa Rosa, por fazerem os meus dias mais felizes e agradáveis.

À Secretaria de Estado de Educação do Rio de Janeiro, por reconhecer a importância da minha formação na transformação da realidade educacional dos alunos.

Por fim, agradeço aos meus amigos. Por estarem ao meu lado para combater o bom combate por aquilo que acredito.

Resumo

A comunicação humana é um fenômeno complexo e multifacetado, no qual a voz desempenha um papel central. Este estudo considera a voz humana e suas patologias, explorando como as modernas tecnologias de inteligência artificial podem auxiliar no diagnóstico e no melhor entendimento dessas condições. A voz não é apenas um meio de comunicação, mas também uma janela para a saúde física e emocional do indivíduo, refletindo aspectos fisiológicos, psicológicos e sociais. Esta pesquisa aborda a complexidade da voz humana e as dificuldades inerentes ao diagnóstico de patologias vocais, que frequentemente apresentam desafios na área da saúde.

O estudo é motivado pela necessidade de métodos de diagnóstico mais precisos, objetivos e não invasivos para patologias das cordas vocais, especialmente em profissionais que dependem fortemente de suas vozes, como professores. Ao focar em profissionais que utilizam intensivamente a voz, a pesquisa reconhece e aborda as condições desafiadoras sob as quais esses indivíduos operam, incluindo ambientes ruidosos e estressantes que aumentam o risco de problemas vocais.

Essa pesquisa emprega tecnologias avançadas de inteligência artificial para uma análise detalhada das características vocais, utilizando métodos como *MFCC (Mel-Frequency Cesptrum Coefficients)*, *PNCC (Power-Normalized Cepstral Coefficients)* e *ZCPA (Zero-Crossings with Peak Amplitudes)*. Estes métodos são explorados para avaliar sua eficácia na detecção de patologias vocais, considerando diferentes condições ambientais e tipos de patologias. Além disso, diferentes redes neurais - *DNN (Deep Neural Networks)*, *CNN (Convolutional Neural Networks)*, *LSTM (Long Short-Term Memory)* e *BiLSTM (Bidirectional Long Short-Term Memory)* - são utilizadas para avaliar a classificação de vozes patológicas e saudáveis, oferecendo avanços significativos a respeito do potencial dessas tecnologias no campo da saúde vocal.

Os resultados demonstram que o método MFCC se destacou pela sua alta eficiência na classificação de patologias vocais e vozes saudáveis, alcançando taxas de acerto notáveis, especialmente na detecção de casos como cistos, edema de Reinke, nódulos e paralisia. Por outro lado, o método PNCC, embora tenha identificado uma proporção considerável de casos patológicos, mostrou-se discutível devido a uma maior incidência de falsos negativos e positivos. Quanto ao ZCPA, mostrou-se menos consistente em comparação com os outros dois métodos, indicando a necessidade de refinamentos adicionais e mais testes para aprimorar sua aplicação em diagnósticos vocais.

Palavras-chave: Identificação de patologias, Inteligência artificial, Saúde vocal.

Abstract

Human communication is a complex and multifaceted phenomenon where voice plays a central role. This study considers the human voice and its pathologies, exploring how modern artificial intelligence technologies can assist in diagnosing and better understanding these conditions. The voice is not only a means of communication, but also a window into the physical and emotional health of the individual, reflecting physiological, psychological, and social aspects. This research addresses the complexity of the human voice and the difficulties inherent in the diagnosis of vocal pathologies, which often present challenges in the health area.

The study is motivated by the need for more accurate, objective and non-invasive diagnostic methods for vocal cord pathologies, especially in professionals who rely heavily on their voices, such as teachers. By focusing on professionals who use their voice intensively, the research recognizes and addresses the challenging conditions under which these individuals operate, including noisy and stressful environments that increase the risk of voice problems.

This research employs advanced artificial intelligence technologies for a detailed analysis of vocal characteristics, using methods such as *MFCC (Mel-Frequency) Cesptrum Coefficients)*, *PNCC (Power-Normalized Cepstral Coefficients)* and *ZCPA (Zero-Crossings with Peak Amplitudes)*. These methods are explored to evaluate their effectiveness in the detection of vocal pathologies, considering different environmental conditions and types of pathologies. In addition, different neural networks - *DNN (Deep Neural Networks)*, *CNN (Convolutional Neural Networks)*, *LSTM (Long Short-Term Memory)* and *BiLSTM (Bidirectional Long Short-Term Memory)* - are used to evaluate the classification of pathological and healthy voices, offering significant advances regarding the potential of these technologies in the field of vocal health.

The results show that the MFCC method stood out for its high efficiency in the classification of vocal pathologies and healthy voices, achieving remarkable success rates, especially in the detection of cases such as cysts, Reinke's edema, nodules and paralysis. On the other hand, the PNCC method, although it identified a considerable proportion of pathological cases, proved to be debatable due to a higher incidence of false negatives and positives. As for the ZCPA, it proved to be less consistent compared to the other two methods, indicating the need for further refinements and more tests to improve its application in vocal diagnostics.

Keywords: Pathologies identification, Artificial intelligence, Vocal health.

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

BiLST	Bidirectional Long Short-Term Memory
CNN	Convolutional Neural Networks
DNN	Deep Neural Networks
IA	Inteligência Artificial
LSTM	Long Short-Term Memory
MFCC	Mel-Frequency Cepstral Coefficients
PNCC	Power-Normalized Cepstral Coefficients
ZCPA	Zero-Crossings with Peak Amplitudes

Sumário

Capítulo 1

Introdução

A voz humana desempenha um papel fundamental na sociedade, sendo um instrumento vital de comunicação e expressão pessoal. Ela transcende a mera função de transmitir palavras, carregando em si nuances emocionais e identitárias que são tão únicas quanto impressões digitais. Na comunicação diária, a voz não apenas transmite informações, mas também emoções, intenções e características pessoais, tornando-se uma ferramenta poderosa para a interação social [1]. Cada voz possui características distintas - timbre, tom, ritmo, e entonação - que refletem não apenas a fisiologia individual, mas também aspectos culturais, educacionais e psicológicos. Esta identidade vocal é tão marcante que, mesmo sem ver a pessoa, frequentemente reconhecemos quem fala apenas pelo som.

Um aspecto crítico que surge é a prevalência de patologias das cordas vocais, especialmente em profissionais que as utilizam intensivamente, como professores. Estes profissionais estão frequentemente expostos a condições que predispõem a problemas vocais, como uso prolongado da voz, ambientes com alta sonoridade e estresse. As patologias não apenas afetam a eficácia profissional, mas também a qualidade de vida [2].

Na análise da saúde vocal, os diagnósticos de patologias podem ser em algum momento subjetivos, como na medicina de forma geral, sujeitos a um "julgamento enviesado ou ruidoso", como discutido em [3]. Algumas vezes, esses diagnósticos dependem excessivamente da interpretação subjetiva, o que pode levar a avaliações imprecisas.

Com os avanços da inteligência artificial (IA) aplicadas em vários setores da sociedade, inclusive na área da saúde, a análise das características vocais por meio dessa tecnologia se tornou uma ferramenta valiosa. Várias pesquisas buscam melhorar a precisão no diagnóstico de doenças e condições de saúde. Este progresso na IA oferece um suporte significativo para os profissionais de saúde, melhorando a eficácia no diagnóstico e na

tomada de decisões terapêuticas.

A IA pode ser empregada para automatizar a classificação da voz em patológica ou saudável. Para essa tarefa, os métodos requerem extração inicial de características que representam a voz em estudo. No que diz respeito à extração de características vocais, este estudo emprega os métodos *MFCC* (*Mel-Frequency Cesptrum Coefficients*), *PNCC* (*Power-Normalized Cepstral Coefficients*) e *ZCPA* (*Zero-Crossings with Peak Amplitudes*), que fornecem uma análise objetiva e detalhada das características de interesse.

A literatura sugere que, em sistemas de reconhecimento de voz com bases de dados faladas, ou seja, pronúncia de palavras e presença de ruídos, o método ZCPA frequentemente supera o MFCC, conforme demonstrado nos estudos de [4], [5] e [6]. Além disso, o método PNCC também é considerado mais robusto que o MFCC em situações com ruído, conforme indicado em [7] e [8]. No entanto, em ambientes sem ruído, o MFCC tende a ser mais eficaz que o ZCPA. Quanto à comparação entre PNCC e MFCC nesses contextos, eles apresentam desempenho bastante similar. Em estudos voltados para a identificação de patologias nas cordas vocais, como apresentado em [9], [10], [11] e [12], observa-se a eficácia do método MFCC na extração de características da voz, que são utilizadas em redes neurais artificiais para classificar vozes patológicas e saudáveis.

Nossa pesquisa tem como objetivo empregar os métodos mencionados para registrar as características de vozes que apresentam patologias, tais como nódulos, paralisia, edema de Reinke e cistos, além de vozes saudáveis. Isso será feito analisando fragmentos de 0,5 e 0,7 segundos da emissão sustentada das vogais /a/ e /e/. Os testes foram realizados em vozes sem ruído e com adição de ruído branco.

Aliada aos métodos de extração de características da voz, vamos investigar como as diferentes redes neurais se comportam no contexto de classificação da saúde vocal. Para essa pesquisa, utilizamos as redes: DNN (*Deep Neural Networks*), CNN (*Convolutional Neural Networks*), LSTM (*Long Short-Term Memory*) e BiLSTM (*Bidirectional Long Short-Term Memory*), utilizadas em [13], [14], [15], [16] e [17], onde são empregadas em uma ampla gama de aplicações, refletindo a capacidade de aprender e reconhecer padrões complexos em dados.

Portanto, integrar essas tecnologias avançadas na prática clínica não apenas melhora a precisão do diagnóstico das patologias vocais, mas também contribui para a objetividade e redução dos vieses e ruídos no julgamento clínico. Essa questão torna-se particularmente importante em ambientes com alta sonoridade e desafiantes, a exemplo das salas de aula, onde é comum professores enfrentarem problemas vocais. A integração de mé-

todos avançados de análise vocal com o conhecimento clínico especializado aponta para uma era renovada no diagnóstico e tratamento de patologias das cordas vocais.

Este trabalho visa contribuir para a pesquisa na área de identificação de patologias em vozes, utilizando ferramentas automáticas de classificação no apoio à decisão. No segundo capítulo, exploramos a anatomia da voz humana e examinamos as patologias focadas neste estudo, incluindo nódulos, paralisia vocal, edema de Reinke e cistos. O terceiro capítulo é dedicado à descrição dos processos de extração de características vocais, empregando técnicas como MFCC, PNCC e ZCPA. No quarto capítulo, nos aprofundamos nas redes neurais aplicadas na análise vocal, abrangendo DNN, CNN, LSTM e BiLSTM, junto com suas respectivas configurações. O quinto capítulo detalha a metodologia adotada, incluindo a seleção do banco de vozes e as configurações dos métodos de extração e das redes neurais. No sexto capítulo, discutimos os resultados obtidos, apresentando tabelas de matriz de confusão para ilustrar a eficácia dos métodos e das redes neurais utilizadas. Por fim, a conclusão sintetiza os achados, oferecendo uma análise abrangente dos resultados.

Capítulo 2

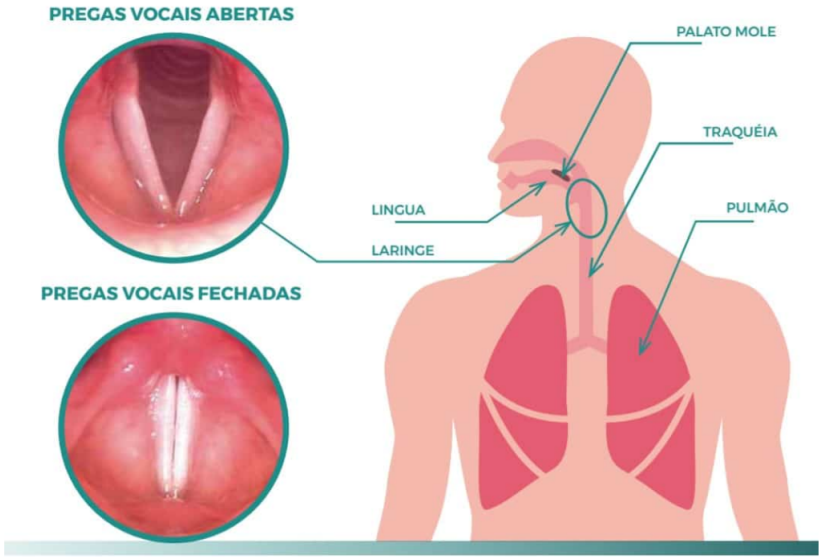
Fisiologia da voz humana

A voz é um elemento fundamental na sociedade, pois serve como uma ferramenta primordial de comunicação e expressão humana. Mais do que simplesmente um meio de transmitir palavras, a voz também desempenha um papel crucial na expressão da identidade individual. Cada pessoa tem uma voz única, que pode refletir sua personalidade, emoções e experiências de vida. Através da voz, podemos expressar alegria, tristeza, raiva, surpresa e uma variedade de outras emoções.

O ponto de partida da criação da voz é o pulmão, que fornece o ar necessário como fonte de energia para a produção de som. Ao expirar, o ar passa pela traqueia e chega até as pregas vocais (ou cordas vocais), localizadas na laringe. As cordas vocais são duas pregas musculares que vibram quando o ar passa entre elas, produzindo o som primário da voz. Após a geração desse som básico, ele é modulado pelo trato vocal, que inclui a garganta, a cavidade oral, os seios nasais e o nariz. Essas estruturas atuam como um amplificador, modificando a qualidade do som ao alterar sua ressonância. A língua, os lábios e os dentes são especialmente importantes na articulação, ou seja, na formação das palavras. Eles ajudam a formar os diferentes sons da fala, modificando o fluxo de ar e a ressonância da cavidade oral. Na figura 2.1, podemos observar o processo fisiológico da voz.

Uma variedade de fatores como o uso excessivo ou inadequado da voz, infecções, lesões, doenças neurológicas, tumores e condições congênitas podem gerar algum tipo de patologia. Independentemente da causa, essas patologias geralmente resultam em alterações na vibração das cordas vocais, interferindo na capacidade das cordas vocais de vibrar [19] de maneira eficaz e uniforme, alterando a frequência (espectro da frequência), intensidade e qualidade do timbre.

Figura 2.1: Fisiologia da voz



Fonte: [18]

2.1 Respiração

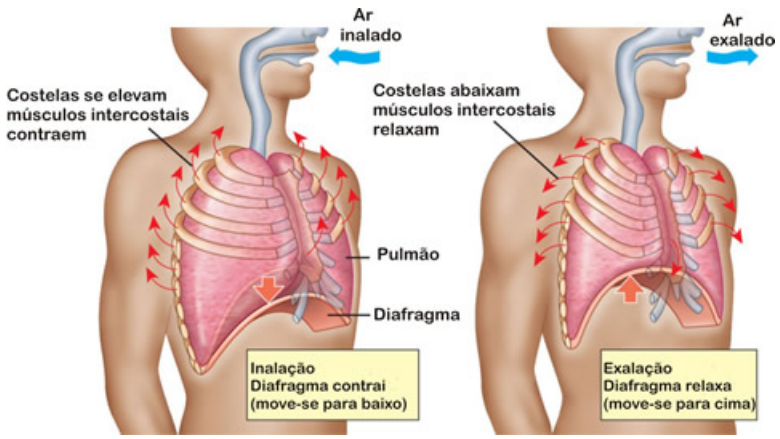
A respiração é a primeira etapa na produção da voz e é essencial para que ela ocorra de forma eficaz. Inicialmente ocorre a inalação com o ar sendo levado para dentro dos pulmões. Este ar é então utilizado para produzir som. Os músculos do diafragma e os músculos intercostais expandem o tórax, reduzindo a pressão interna e permitindo que o ar entre nos pulmões. Este processo de inspiração é crucial para fornecer o fluxo de ar necessário para a fonação [20].

Após a inalação, o ar é expelido dos pulmões durante a exalação. Este fluxo de ar ascendente passa pela laringe, onde estão localizadas as cordas vocais. Na figura 2.2, apresentamos o processo de respiração.

2.2 Fonação

A fonação ocorre principalmente na laringe, que está localizada na parte superior da traqueia. A laringe contém as cordas vocais, duas bandas musculares que vibram quando o ar passa entre elas. O ajuste da tensão e do espaçamento dessas cordas vocais é crucial para controlar a altura e a qualidade do som produzido. Quando as cordas vocais estão juntas e tensas, elas vibram mais rapidamente, produzindo sons mais agudos. Se estão

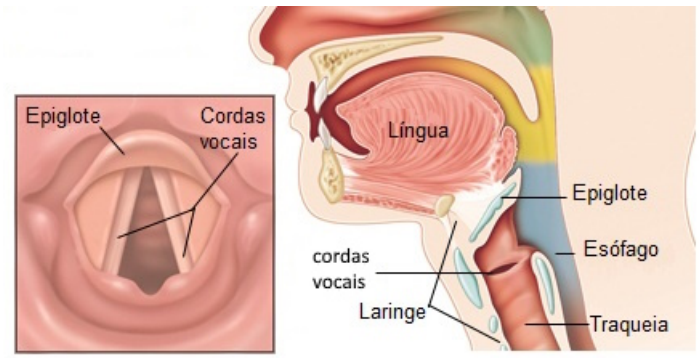
Figura 2.2: Processo de respiração



Fonte: [21]

mais relaxadas e afastadas, os sons são mais graves. Na figura 2.3, exibimos a anatomia da laringe e uma vista aproximada das cordas vocais.

Figura 2.3: Anatomia da laringe



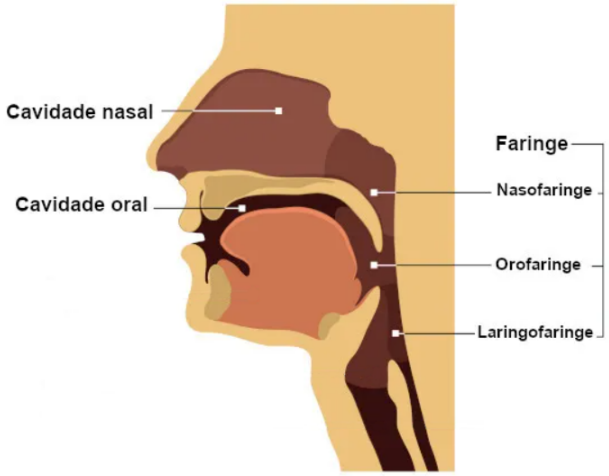
Fonte: Adaptado de [22]

O processo de fonação é controlado pelo sistema nervoso, que coordena a atividade dos músculos da laringe [23]. Esses músculos controlam não só a tensão das cordas vocais, mas também a abertura da glote, o espaço entre as cordas vocais. Quando a glote está fechada, o ar dos pulmões cria uma pressão nas cordas vocais, fazendo-as vibrar e gerando som. Este som então viaja pelas cavidades de ressonância do corpo, como a boca, o nariz e os seios paranasais, onde é modificado para produzir a voz como a conhecemos.

2.3 Ressonância

As ondas sonoras produzidas pelas cordas vocais são amplificadas e modificadas pelas cavidades ressonantes num processo denominado ressonância. Esta ocorre principalmente nas cavidades do trato vocal superior, que incluem a faringe, a cavidade nasal e a cavidade oral, tal como observamos na figura 2.4.

Figura 2.4: Ressonância



Fonte: Adaptado de [24]

Cada cavidade tem seu próprio conjunto de frequências naturais nas quais ressoa mais eficientemente. Estas frequências são conhecidas como formantes [25]. Quando as ondas sonoras interagem com essas cavidades, certas frequências são amplificadas enquanto outras são atenuadas. Este processo de filtragem seletiva é o que dá à voz humana sua qualidade única e permite a variedade de sons e timbres que podemos produzir.

2.4 Articulação

Outro elemento importante na produção do som é a articulação, que é definida como a maneira como moldamos o som produzido na laringe para formar palavras e frases compreensíveis. É realizada principalmente pela boca, incluindo os lábios, os dentes, o céu da boca (palato), a língua e a mandíbula, como podemos observar na figura 2.5. Esses componentes trabalham juntos para modificar as características do som, como sua qualidade, volume e tonalidade, e para formar os diferentes fonemas da fala.

Os sons que compõem a língua falada passam por todo o processo da fisiologia da voz, como: respiração, fonação, ressonância até chegar a articulação. E essa última etapa é de

Figura 2.5: Articulação

Fonte: [26]

grande importância para a formação das vogais e consoantes.

Para a articulação das vogais, a posição da língua e o formato do trato vocal são cruciais. As vogais são produzidas com um fluxo de ar relativamente livre e são diferenciadas pela forma como a boca é configurada e pela posição da língua. Por exemplo, para a vogal "a", a boca fica aberta e a língua fica baixa, enquanto para a vogal "i", a boca se fecha um pouco e a língua se eleva. O palato mole (ou véu palatino) e os lábios também desempenham um papel importante na formação das vogais, modificando a ressonância e o timbre do som.

As consoantes, por outro lado, são produzidas com algum grau de obstrução no trato vocal. Dependendo da consoante, esta obstrução pode ocorrer em diferentes locais, como os lábios (labiais), os dentes (dentais), o palato (palatais) ou a garganta (guturais). Além disso, a maneira como o ar é liberado ou bloqueado cria diferentes tipos de consoantes, como as oclusivas (p, b, t, d, k, g), nas quais o fluxo de ar é completamente interrompido, ou as fricativas (f, v, s, z), nas quais o ar passa por uma estreita abertura, causando um som de fricção.

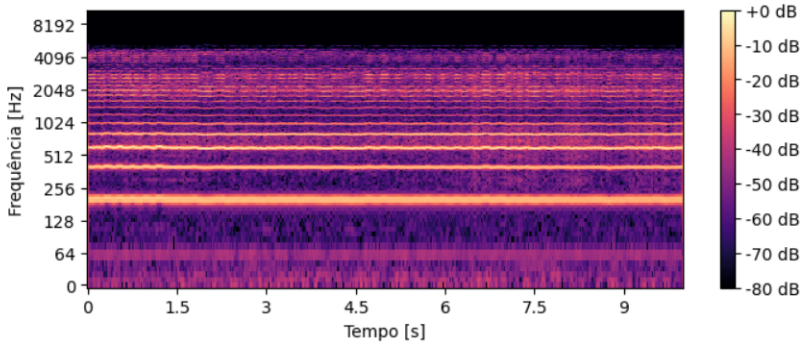
2.5 Patologias das cordas vocais

As patologias das cordas vocais referem-se a uma variedade de condições que afetam a saúde e o funcionamento das cordas vocais, estruturas essenciais para a produção da voz. Essas patologias podem variar em tipo e gravidade, e podem afetar a fala, a respiração e a qualidade da voz de uma pessoa.

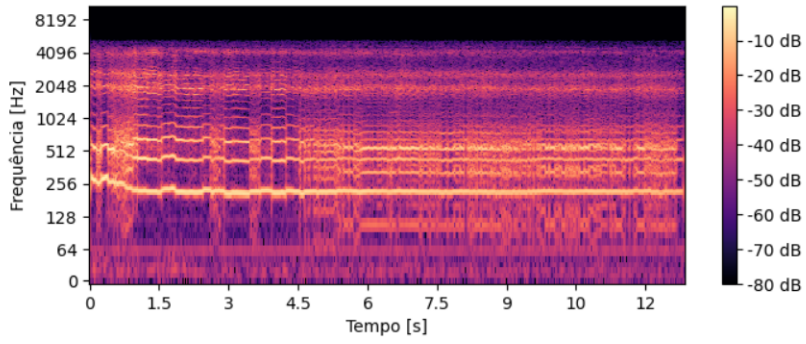
Normalmente, as cordas vocais vibram de maneira uniforme e simétrica, criando ondas

sonoras regulares que produzem uma voz clara e distintamente modulada. No entanto, os distúbios das cordas vocais perturbam esse processo. O espectro de frequência da voz, que é representação visual das diferentes frequências que compõem a voz, também é afetado. As patologias podem limitar a capacidade das cordas vocais de vibrar em certas frequências, resultando em uma redução da amplitude tonal. Em casos de paralisia vocal, onde uma ou ambas as cordas vocais não se movem adequadamente, o impacto no espectro de frequência pode ser ainda mais pronunciado. A paralisia pode levar a uma incapacidade de alcançar tons mais altos ou mais baixos, além de dificultar o controle do volume da voz. Isso ocorre porque a tensão e o comprimento das cordas vocais, essenciais para produzir diferentes frequências, são comprometidos. Na figura 2.6, apresentamos o espectrograma de uma voz saudável (a) e uma voz com paralisia (b).

Figura 2.6: Espectrograma de voz saudável (a) e com paralisia (b)



(a)



(b)

Outra questão relevante é a fadiga vocal. Indivíduos com patologias das cordas vocais frequentemente experimentam uma rápida fadiga vocal devido ao esforço adicional necessário para falar. Isso pode resultar em uma diminuição da capacidade de sustentar a voz em determinadas frequências por períodos prolongados. Neste trabalho, analisamos quatro tipos de distúrbios que acometem as pregas vocais: nódulo, paralisia, edema de Reinke e cisto.

Nódulo

Os nódulos vocais, frequentemente pequenos e sólidos, surgem nas cordas vocais devido ao uso excessivo ou impróprio da voz, resultando em lesões superficiais. Com o passar do tempo, essas lesões podem evoluir para nódulos, similares a calosidades, localizados nos pontos de maior contato e atrito durante a vibração vocal. Podem ser singulares ou múltiplos e são inofensivos.

Essas protuberâncias afetam a habilidade das cordas vocais de vibrarem harmoniosamente e livremente, impactando a qualidade vocal. Sintomas típicos incluem rouquidão, desconforto ou sensação de tensão na garganta, dificuldade em sustentar notas ao cantar, perda do alcance vocal usual ou clareza reduzida.

Geralmente, a ocorrência de nódulos vocais é vinculada a profissões ou a atividades que demandam uso constante e vigoroso da voz, como a de cantores, professores, locutores e atores. Contudo, podem também aparecer em indivíduos que não utilizam a voz profissionalmente, mas possuem hábitos vocais inadequados, como falar alto ou frequentemente, ou em casos de refluxo gastroesofágico, que irrita as cordas vocais [27].

Paralisia

A paralisia afeta as cordas vocais quando os nervos responsáveis por seu controle sofrem danos ou interrupções. Isso impede o movimento adequado das cordas vocais, que normalmente abrem e fecham suavemente, facilitando a respiração e a produção de voz. A paralisia de uma ou ambas as cordas vocais compromete essa mobilidade, podendo resultar em problemas como dificuldades na fala, na respiração e na deglutição.

Diversos fatores podem causar a paralisia das cordas vocais, incluindo danos aos nervos que as inervam. Esses danos podem ser provocados por cirurgias no pescoço ou no tórax, lesões no pescoço, tumores, infecções virais ou doenças neurológicas, como o acidente vascular cerebral (AVC) [28]. Os sintomas variam conforme a paralisia afeta uma ou ambas as cordas vocais. Com apenas uma corda vocal afetada, a voz pode se tornar rouca, fraca ou áspera, e pode haver dificuldade para falar alto ou por longos períodos. Já a paralisia de ambas as cordas vocais pode dificultar a respiração devido ao estreitamento da abertura entre elas, restringindo o fluxo de ar.

Edema de Reinke

O edema de Reinke é uma patologia que afeta a camada mais externa da lâmina

própria nas cordas vocais, o chamado espaço de Reinke [29]. Essencial para a geração da voz, esta região abriga fibras colágenas e elásticas essenciais para as vibrações vocais. Neste distúrbio, um acúmulo incomum de um líquido gelatinoso no espaço de Reinke provoca o inchaço das cordas vocais, resultando em sintomas vocais como uma voz mais grave, rouquidão e, em situações extremas, dificuldades respiratórias. A alteração na voz surge porque o inchaço faz com que as cordas vocais vibrem de maneira anormal, modificando a sonoridade.

Diversos fatores podem causar o edema de Reinke, com destaque para o tabagismo crônico, uso excessivo da voz, refluxo gastroesofágico e, em certos casos, hipotireoidismo. O tabagismo é particularmente danoso, já que os elementos químicos do cigarro provocam irritação e inflamação contínua nas cordas vocais.

Cisto

Cisto nas cordas vocais são pequenas lesões esféricas, repletas de líquido ou substância semi-sólidas. Tal condição pode ser desencadeada por múltiplas causas, incluindo irritação crônica, uso excessivo da voz (como em cantores ou professores), e predisposições genéticas [30], que levam ao bloqueio de uma glândula secretora nas cordas vocais.

Os sintomas associados a um cisto nas cordas vocais abrangem rouquidão, alterações no timbre vocal, dificuldades para projetar a voz e, em alguns casos, perda total da voz. É possível, contudo, que cistos não produzam sintomas evidentes, sendo identificados apenas em exames realizados por outros motivos.

Capítulo 3

Extração de características da voz

A extração de características da voz é o processo de converter os sinais de áudio da fala em representações numéricas que podem ser usadas para análises posteriores, como reconhecimento de fala, identificação de locutor, detecção de emoções, entre outras aplicações. Essas características são derivadas das propriedades acústicas da fala e são úteis para compreender e processar a informação contida na voz. Neste trabalho, vamos utilizar três modelos de extração de características: *Mel-Frequency Cepstral Coefficients (MFCC)*, *Power-Normalized Cepstral Coefficients (PNCC)* e *Zero-Crossing with Peak Amplitude (ZCPA).*

3.1 Pré-processamento

O pré-processamento em sinal de voz refere-se às etapas iniciais de tratamento de sinais de áudio, que ocorrem anteriormente a aplicação de algoritmos de processamento ou análise mais avançados. Essas etapas são essenciais para melhorar a qualidade e a utilidade dos sinais de áudio. As etapas de pré-processamento aplicadas foram: pré-ênfase e janelameto.

3.1.1 Pré-ênfase

No processamento de fala, é comum ocorrer uma queda no espectro do sinal devido a várias razões. A queda no espectro de sinal significa que as frequências mais altas no sinal de áudio são registradas com menor intensidade do que as frequências mais baixas, o que pode resultar em uma qualidade de áudio indesejada. A pré-ênfase (filtragem de pré-ênfase) é uma técnica usada para resolver esse problema. Ela envolve o aumento das

amplitudes das frequências mais altas em relação às frequências mais baixas no sinal de áudio, tendo como objetivo compensar a queda no espectro de sinal, tornando o sinal de áudio mais perceptualmente equilibrado em termos de intensidade de frequências. Esse procedimento leva a melhora da qualidade do áudio e torna o sinal mais adequado para análise e processamento subsequentes. A função de transferência típica de um filtro de pré-ênfase é da forma:

$$H(z) = 1 - \alpha z^{(-1)}, \qquad 0,9 \leq \alpha \leq 1 \tag{3.1}$$

Nessa equação, $z^{(-1)}$ representa um atraso de uma amostra no sinal, e α é um coeficiente que determina a quantidade de ênfase a ser aplicada às altas frequências.

3.1.2 Janelamento

O processo de janelamento é uma etapa fundamental no processamento de sinais de voz. O objetivo desta etapa é dividir um sinal contínuo em segmentos menores, conhecidos como "janelas" ou "frames", para que possamos aplicar técnicas de processamento de sinais em intervalos de tempo curtos e bem definidos. Isso é importante para análise espectral, redução de ruído, extração de características e outras tarefas de processamento de sinais. O janelamento de Hamming é comumente usado em processamento de sinais para reduzir o vazamento espectral e melhorar a qualidade da análise espectral de sinais em domínio de frequência e é definida pela seguinte função:

$$\omega(n) = \begin{cases} 0.54 - 0.46 \cdot cos\left(\frac{2\pi n}{N-1}\right), & 0 \leq n \leq N-1 \\ 0 \qquad\qquad\qquad\qquad\quad , & \text{c.c.} \end{cases} \tag{3.2}$$

Onde:

- $\omega(n)$ é o valor da janela na posição n.
- N é o tamanho da janela em amostras.
- cos é a função cosseno.

3.2 Transformada Discreta de Fourier (Discrete Fourier Transforms - DFT)

A transformada discreta de Fourier é uma técnica usada na análise de sinais de áudio, incluindo a voz, para extrair informações sobre as características espectrais ao longo do tempo. Após a realização do janelamento (subseção 3.1.2), é aplicada uma transformada rápida de Fourier (*Fast Fourier Transform - FFT*) em cada segmento. A FFT é um conjunto de algoritmos eficientes para calcular a transformada discreta de Fourier.

A fórmula da DFT é dada por:

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{-j\frac{2\pi}{N}kn} \tag{3.3}$$

Onde:

- $X(k)$ é a representação no domínio da frequência do sinal no índice k.
- $x(n)$ é o valor do sinal de voz no índice de tempo n.
- N é o número de pontos na transformada.
- j é a unidade imaginária.
- $\frac{2\pi}{N}$ é a frequência fundamental.

A DFT é usada para transformar um sinal de áudio no domínio do tempo em seu equivalente no domínio da frequência, o que resulta em um espectro de magnitude para cada quadro. Este espectro é calculado tomando o quadrado do módulo da transformada discreta de Fourier.

3.3 Coeficientes Cepstrais de Frequência Mel (Mel-Frequency Cesptrum Coefficients - MFCC)

MFCC é um processo fundamental na área de processamento de sinais de fala e reconhecimento de fala. Essa técnica tem sido amplamente utilizada para representar de forma compacta e informativa as características acústicas de um sinal de fala, tornando-o adequado para análise, reconhecimento de fala, identificação de locutor, sistemas de diálogo e classificação por algoritmos de aprendizado de máquina. O processo de extração

de coeficientes envolve as seguintes etapas: pré-processamento (pré-ênfase e janelamento), transformada de Fourier, bancos de filtros Mel, logaritmo dos coenficientes de Mel e transformada Cosseno Discreta.

A ideia do método MFCC é simular a forma como o sistema auditivo humano percebe diferentes frequências, através da escala Mel. A escala Mel é um componente fundamental no cálculo do MFCC e é usada para mapear as frequências lineares (geralmente em Hertz) em frequências Mel.

3.3.1 Escala Mel e Banco de filtros Mel

A escala Mel é uma escala perceptual de frequência que simula a maneira como os humanos percebem diferentes frequências sonoras. Em termos simples, a escala Mel é uma representação da frequência que leva em consideração a sensibilidade do ouvido humano a diferentes frequências. A fórmula para converter uma frequência f (em Hertz) para a frequência m (em Mel) e a sua inversa são, respectivamente:

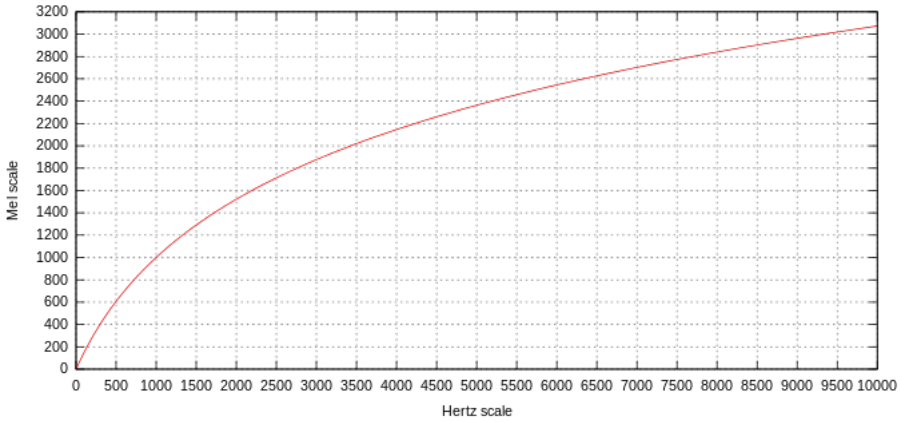
$$m = M(f) = 1127 \times ln\left(1 + \frac{f}{700}\right) \tag{3.4}$$

$$f = M^{-1}(m) = 700\left(e^{\frac{m}{1127}} - 1\right) \tag{3.5}$$

Na figura 3.1, podemos ver a relação entre as escalas Hz e Mel.

Os filtros Mel são aplicados à transformada de Fourier para capturar as características perceptualmente relevantes da frequência, mapeando o espectro de um sinal de áudio em uma representação mais perceptual, que se assemelha à resposta do ouvido humano. A criação dos filtros Mel é feita a partir dos pontos Mel usando funções triangulares [7], onde cada filtro é centrado em um ponto Mel e possui uma forma triangular. Os coeficientes dos filtros Mel são calculados com base nas equações:

Figura 3.1: Escala Hz x escala Mel



Fonte: [31]

$$H_m(k) = \begin{cases} 0, & \text{se } k < k_{m-1} \\ \frac{k-k_{m-1}}{k_m-k_{m-1}}, & \text{se } k_{m-1} \leq k < k_m \\ \frac{k_{m+1}-k}{k_{m+1}-k_m}, & \text{se } k_m \leq k < k_{m+1} \\ 0, & \text{se } k \geq k < k_{m+1} \end{cases} \tag{3.6}$$

Onde:

- k é a frequência em Mel.
- $k_{(m-1)}$ e $k_{(m+1)}$ são as frequências dos filtros vizinhos.
- $k(m)$ é a frequência central do filtro triangular.

Depois de aplicar o banco de filtros de Mel, os coeficientes de Mel são calculados, obtendo-se uma série de valores que representam a energia em cada faixa de frequência do banco de filtros. Como os filtros Mel são projetados para simular a resposta do ouvido humano às diferentes frequências, eles são distribuídos de maneira não linear ao longo da escala de frequência, e por esse motivo utiliza-se o logaritmo posteriormente. A fórmula geral para calcular os coeficientes de Mel, em que M é o número de coeficientes de Mel desejados, é a seguinte:

$$Mi = \sum_{k=1}^{N} P(k) \cdot H_m(k, f_i) \tag{3.7}$$

Onde:

- M_i é o í-ésimo coeficiente de Mel.
- N é o número de pontos no espectro de potência.
- $P(k)$ é o espectro de potência na frequência k.
- $H_m(k, f_i)$ é a função triangular.

3.3.2 Logaritmo dos coeficientes de Mel

O próximo passo, após calcular os coeficientes de Mel, é aplicar o logaritmo aos respectivos coeficientes. O logaritmo é uma operação matemática que torna a representação dos coeficientes mais sensível às diferenças em baixos níveis de energia e menos sensível a diferenças em altos níveis de energia. O motivo para isso é que o ouvido humano percebe o som de maneira aproximadamente logarítmica em relação à intensidade. Portanto, aplicar o logaritmo aos coeficientes de Mel ajuda a aproximar a resposta auditiva humana e torna a representação mais adequada para a análise da fala. A transformação logarítmica aplicada nos coeficientes de Mel é dada pela fórmula:

$$Ci = log(M_i) \tag{3.8}$$

Onde:

- Ci é o i-ésimo coeficiente de Mel transformado.
- M_i é o í-ésimo coeficiente de Mel extraído na fórmula 3.7.

3.3.3 Transformada Cosseno Discreta (Discrete Cosine Transform - DCT)

A DCT é aplicada aos logaritmos das energias dos filtros Mel para obter os coeficientes MFCC. Ela converte um sinal ou uma matriz de valores em uma representação no domínio das frequências, permitindo a compactação de informações, eliminação de redundâncias e retenção das características mais importantes do sinal. A fórmula da DCT para um sinal unidimensional é dada por:

$$DCT(u) = C(u) \cdot \sum_{x=0}^{N-1} f(x) \cdot cos\left(\frac{(2x+1)u\pi}{2N}\right), \quad \text{para } u = 0, 1, \ldots, N-1 \tag{3.9}$$

Onde:

- $DCT(u)$ é o coeficiente DCT na posição u;
- $C(u)$ é um coeficiente de escala que é igual a $\frac{1}{\sqrt{2}}$ para u =0 e 1 para $u > 0$;
- $f(x)$ é o valor do sinal na posição x;
- N é o número total de amostras no sinal.

A DCT é amplamente utilizada em várias aplicações, incluindo compressão de imagem (como o formato JPEG), compressão de áudio (como o formato MP3) e em muitas outras áreas de processamento de sinais.

3.3.4 Normalização

A normalização de dados é a etapa final no processamento de fala e em tarefas de reconhecimento de padrões. Como os coeficientes extraídos serão analisados posteriormente por algoritmos de aprendizado de máquina, a normalização melhora a capacidade do modelo de generalizar, ajuda a reduzir o risco de *overfitting*; outro benefício é a redução da variabilidade dos dados, tornando-os comparáveis em diferentes situações, há também a melhora na convergência e redução de problemas numéricos, tornando, assim, as características mais interpretáveis e facilitando a análise dos resultados do modelo, uma vez que as unidades das características são consistentes.

3.4 Coeficientes Cepstrais Normalizados pela Potência (Power-Normalized Cepstral Coefficients - PNCC)

O método de extração de características PNCC é uma técnica comumente usada no processamento de sinais de áudio para representar e analisar informações relevantes de espectros de frequência. São particularmente úteis em tarefas de reconhecimento de fala, classificação de áudio e outras aplicações de processamento de sinais de áudio, onde a representação precisa das características do espectro de frequência é essencial.

Os PNCCs são uma extensão dos MFCCs (Mel-Frequency Cepstral Coefficients) e têm como objetivo principal capturar características perceptivas do áudio, semelhantes à forma como o ouvido humano processa o som, tendo a vantagem de ter mais robustez em ambientes ruidosos [32].

Os PNCCs são obtidos a partir de uma transformação da magnitude do espectro de potência por meio da aplicação de filtros perceptivos lineares. Esses filtros são projetados para simular a sensibilidade do ouvido humano a diferentes frequências. Em seguida, são utilizados para calcular os coeficientes cepstrais, que representam as características relevantes do sinal de áudio.

A obtenção de coeficientes segue um conjunto de passos que incluem as fases a seguir: pré-processamento (pré-ênfase e janelamento), transformada de Fourier, bancos de filtros gammatone, função de potenciação e transformada Cosseno Discreta.

3.4.1 Filtros gammatone

Os filtros gammatone são uma forma de modelar a resposta em frequência do ouvido humano, são baseados na escala de Bandas Retangulares Equivalentes (ERB)[33], que representam bem a resposta impulsional da membrana basilar. Eles foram projetados para simular a forma como a cóclea, a parte do ouvido que traduz o som em sinais nervosos, processa os sinais de áudio. Os filtros gammatone capturam a essência de como as diferentes frequências são filtradas antes de serem percebidas pelo cérebro.

Em termos de processamento de sinais, um filtro gammatone é um filtro passa-banda que pode ser caracterizado por sua função de transferência ou sua resposta ao impulso. Podendo ser representada como:

$$g(t) = a \cdot t^{n-1} \cdot e^{-2\pi bt} \cdot cos(2\pi f_c t + \phi), \qquad \text{para } t \geq 0 \tag{3.10}$$

Onde:

- a é a amplitude;
- t é o tempo;
- n é a ordem do filtro;
- b é a largura de banda retangular equivalente ERB;
- f_c é a frequência central do filtro;
- ϕ é a fase inicial do filtro.

A fórmula mais comumente utilizada para calcular a ERB é baseada na pesquisa de Moore e Glasberg (1983), que desenvolveram uma aproximação para a largura de

banda crítica do ouvido humano. A fórmula proposta para calcular a ERB em função da frequência é:

$$ERB(f) = 24.7 + (4.32f + 1) \tag{3.11}$$

Onde:

- $ERB(f)$ é a largura de banda retangular equivalente em Hertz,
- f é a frequência em kiloHertz.

Após cada frame do sinal ser passado através de um banco de filtros gammatone, o resultado é uma representação do sinal de fala em termos de intensidade em várias bandas de frequência crítica, sendo, então, aplicada uma função de potenciação - operação não linear - para ajudar a diferenciar entre componentes de sinal que são importantes para a percepção e aqueles que não são, como ruídos de fundo ou componentes irrelevantes.

3.5 Cruzamento por Zero com Amplitude de Pico (Zero-Crossings with Peak Amplitudes - ZCPA)

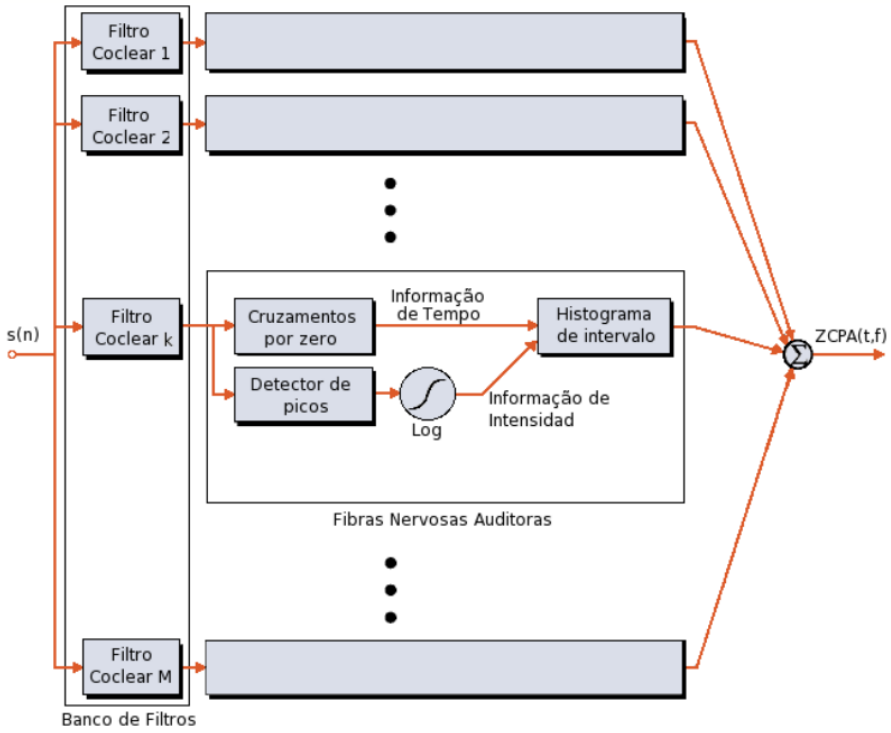
O método ZCPA é uma técnica de extração de características em processamento de sinais de voz, com uma abordagem robusta na análise de sinais acústicos, especialmente no reconhecimento de voz e de locutores em ambientes ruidosos. Esta técnica se concentra em identificar e analisar dois componentes principais de um sinal de voz: a taxa de cruzamento por zero (Zero Crossing Rate - ZCR) e a amplitude de pico (Peak Amplitude).

No início do processo de extração do ZCPA, o sinal vocal é inicialmente submetido a um estágio de pré-processamento. Subsequentemente, o sinal passa por um conjunto de filtros, que fraciona o sinal em diversas faixas de frequência, resultando em um gráfico de frequência baseado nos pontos de cruzamento de zero. O acréscimo nesse histograma é determinado pelo logaritmo da maior intensidade registrada em cada intervalo. Finalmente, sobre este histograma, aplica-se a DCT para obter os coeficientes cepstrais ZCPA. Inspirado pela forma como o ouvido humano processa sons, este método se concentra em simular a capacidade da cóclea, uma parte essencial do ouvido interno, de decompor sons complexos em componentes de frequência mais simples.

3.5.1 Banco de filtros

Na base do banco de filtros cocleares, está a ideia de capturar as características essenciais dos sons da maneira como são percebidos pelo ouvido humano. Este processo envolve a transformação do sinal de áudio em uma série de canais ou faixas de frequência, cada um correspondendo a uma parte diferente da cóclea. Essa segmentação é crucial porque o ouvido humano não percebe todas as frequências da mesma maneira; algumas são mais salientes do que outras, dependendo de sua intensidade e contexto. Na figura 3.2, exibimos o método ZCPA, onde os M filtros cocleares representam o deslocamento mecânico da membrana basilar [34].

Figura 3.2: Método ZCPA



Fonte: [35]

As k bandas são disposta segundo a escala Bark pela equação,

$$f_{Bark} = 13 \arctan\left(\frac{76f}{1000}\right) + 3.5 \arctan\left(\frac{f}{7500}\right)^2 \tag{3.12}$$

e f_{Bark} é a frequência perceptual em Bark e f é a frequência em Hertz.

3.5.2 Taxa de cruzamento por zeros

A taxa de cruzamento por zeros (*Zero Crossing Rate - ZCR*) tem como objetivo verificar a quantidade de vezes que um sinal passa pelo valor zero (muda de sinal) em um determinado período de tempo ou amostras. No contexto de áudio, pode ser usado para identificar características de um som, onde sinais de voz com ZCR alto geralmente indicam ruído, enquanto que sinais com ZCR baixo indicam vozes sem ruído [6].

Após a saída do banco de filtros, é feito um processamento de cruzamento positivo por zero para obter o pico máximo ($p_k i$) entre cruzamentos sucessivos e o inverso do tamanho do intervalo ($f_k i$).

$$p_k(i) = \max_{z_k(i) \leq n < z_k(i+1)} \{s_k(n)\} \tag{3.13}$$

$$f_k(i) = \frac{1}{z_k(i+1) - z_k(i)} \tag{3.14}$$

Onde $z_k(i)$ e $z_k(i+1)$ são cruzamentos sucessivos, e $s_k(n)$ é a saída de cada sub-banda do banco de filtros.

3.5.3 Histograma de frequências

Para cada característica de cruzamento de zeros com detecção de picos, é construído um histograma. Cada pico de amplitude extraído passa por uma transformação não-linear para realçar certas características do sinal, como picos proeminentes ou padrões únicos de frequência. O objetivo principal é quantificar a distribuição das frequências presentes no sinal de voz. Isto é feito dividindo o espectro de frequência em vários intervalos, conhecidos como *"bins"* [4]. Cada *bin* representa um intervalo de frequência específico. A escolha do número e o tamanho desses *bins* afeta diretamente a resolução e a precisão da análise espectral. A atribuição de um peso a cada *bins* é determinada pela soma das transformações não-lineares [5] dos picos de amplitude que caem dentro do intervalo de frequência correspondente ao *bins*. A aplicação dessas transformações não-lineares e a subsequente ponderação dos bins resultam em um histograma que reflete a distribuição de energia ao longo do espectro de frequência do sinal de voz. Com isso, é calculado o DCT do histograma, gerando os coeficientes ZCPA.

3.5.4 Coeficientes Delta e Delta-Delta

Na esfera do reconhecimento vocal e análise da fala, a incorporação destes coeficientes contribui significativamente para aperfeiçoar a exatidão e a eficiência dos sistemas. Ao analisar a voz, não só as características "estáticas" são importantes, mas também as características "dinâmicas" (características que mudam ao longo do tempo), o que é capturado pelos coeficientes delta (primeira derivada temporal das características espectrais do sinal de voz) dado pela fórmula,

$$\Delta_t = \frac{\sum_{n=1}^{N} n \cdot (c_{t+n} - c_{t-n})}{2 \sum_{n=1}^{N} n^2} \tag{3.15}$$

onde Δ_t é a diferencial em função do tempo, em termos dos coeficientes estáticos (c_{t+n} até c_{t-n}). E N é a quantidade de amostras requeridas para determinar os coeficientes dinâmicos.

O delta-delta são os coeficientes da segunda derivada dos coeficientes delta (coeficientes de aceleração)[36], onde os resultados são adquiridos ao replicar a derivada sobre os achados da primeira derivação. Fundamentalmente, enquanto os coeficientes delta dão uma ideia sobre a velocidade de variação, os delta-delta oferecem uma perspectiva sobre a aceleração dessa variação.

Capítulo 4

Redes Neurais Artificiais (RNAs)

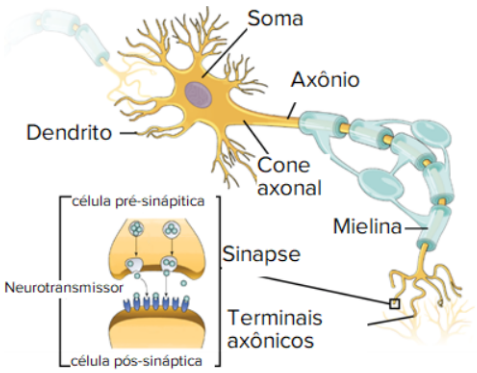
As Redes neurais artificiais são modelos computacionais inspirados pelo funcionamento do cérebro humano e seu sistema nervoso [37]. Elas são compostas por unidades interconectadas, chamadas de neurônios artificiais ou nodos, que simulam de forma simplificada um neurônio humano (neurônio biológico). Os neurônios artificiais processam e transmitem informações através de conexões ponderadas, semelhantes às sinapses no cérebro.

4.1 Neurônio biológio e neurônio artificial

Os neurônios biológicos são as unidades fundamentais do cérebro e do sistema nervoso, responsáveis por receber, processar e transmitir informações através de sinais elétricos e químicos. Cada neurônio possui um corpo celular, que contém o núcleo e organelas que mantêm as funções celulares; dendritos, que são extensões curtas que recebem sinais de outros neurônios; e um axônio, uma extensão longa que transmite sinais para outras células. Os sinais são recebidos nos dendritos como impulsos elétricos. Estes impulsos viajam pelo corpo celular e são transmitidos para fora do neurônio através do axônio. Na extremidade do axônio, o neurônio forma sinapses com outras células, que podem ser outros neurônios ou células musculares, por exemplo. Quando um impulso elétrico atinge uma sinapse, ele provoca a liberação de neurotransmissores, que são substâncias químicas que atravessam a fenda sináptica e transmitem o sinal para a próxima célula. Essa transmissão pode resultar em uma variedade de respostas, como a geração de um novo impulso elétrico, a ativação de uma resposta muscular ou a modulação de funções metabólicas.

Na figura 4.1, podemos ver um neurônio biológico.

Figura 4.1: Neurônio biológico

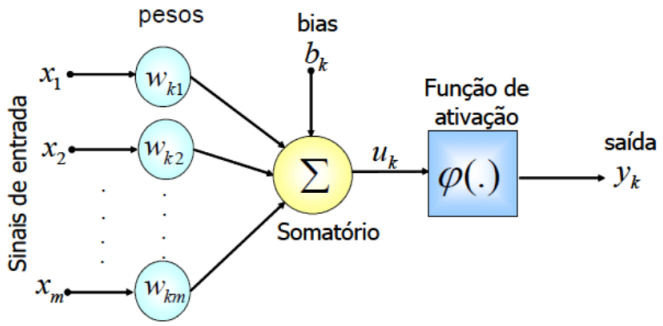


Fonte: [38]

Um neurônio artificial é uma construção conceitual que emula a função de um neurônio biológico. Ele foi idealizado como parte de uma rede neural artificial e serve como unidade básica de processamento dentro dessa rede. Inspirado no funcionamento dos neurônios biológicos, um neurônio artificial recebe vários sinais de entrada (que podem ser dados brutos ou saídas de outros neurônios artificiais), multiplicando cada entrada por um peso (que representa a força da conexão sináptica) e somando esses valores [39]. Essa soma ponderada é então transmitida por uma função de ativação, que determina se e como o sinal deve prosseguir. A função de ativação é fundamental porque introduz não-linearidade ao processo, permitindo que a rede neural aprenda e execute tarefas complexas.

Na figura 4.2, exibimos um modelo de um neurônio artificial.

Figura 4.2: Neurônio artificial



Fonte: [39]

O neurônio artificial é formado pelos seguintes componentes: x são as entradas da rede; os valores de w são pesos sinápticos, que são parâmetros que o modelo ajusta durante o treinamento para fazer previsões mais precisas; b_k é o viés (bias), que é adicionado para ajustar a saída do neurônio junto com os pesos ponderados das entradas [40]; u_k é

a combinação linear, ou seja, o resultado do somatório (Σ) dos processos anteriores. A função de ativação é representada por $\varphi(.)$ e y_k é a saída do neurônio. A fórmula da saída (y_k) de um neurônio k é dada pela equação 4.1:

$$y_k = \varphi \cdot u_k \tag{4.1}$$

sendo a fórmula da combinação linear (u_k) representada pela equação 4.2:

$$u_k = \left(\sum_{i=1}^{n} w_i x_i + b \right) \tag{4.2}$$

As RNAs são usadas em aprendizado de máquina e inteligência artificial para resolver uma variedade de tarefas, como classificação, regressão, reconhecimento de padrões, processamento de linguagem natural, entre outros [17]. Existem vários tipos de redes neurais artificiais, cada uma com sua própria arquitetura e aplicação específica.

Neste trabalho, usamos as redes neurais feedforward (feed forward network - FFN), convolucionais (convolutional neural network - CNN) e recorrentes (recurrent neural network - RNN)

4.2 Configurações das redes neurais

4.2.1 Função de ativação

Uma função de ativação em uma rede neural é uma função matemática, que é fundamental para a habilidade do modelo de lidar com complexidade e aprender não-linearidades. É ela que permite um neurônio ser ativado ou não, isto é, se a informação que o neurônio está processando é relevante para a tarefa em mãos ou deve ser ignorada. As funções de ativação mais comumente usadas são:

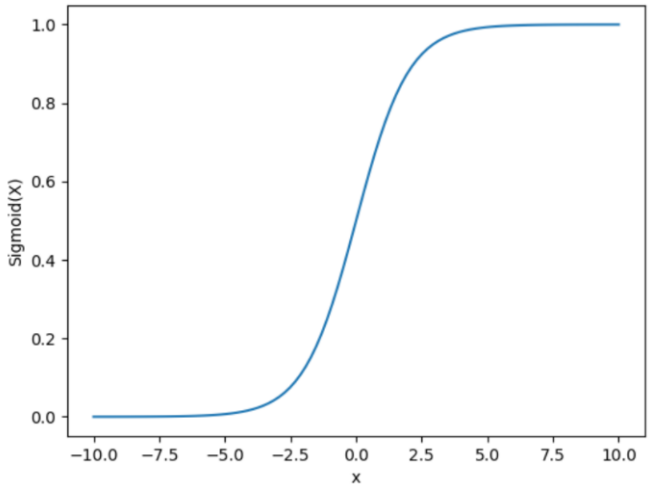
Sigmoid

É uma função importante porque ajuda a converter valores de entrada em valores entre 0 e 1 [41], o que é útil, especialmente em problemas de classificação binária. A função de ativação sigmoid é expressa pela fórmula 4.3:

$$\varphi(x) = \frac{1}{1 + e^{-x}} \tag{4.3}$$

Nesta fórmula, e é a base do logaritmo natural e x é o valor de entrada para a função. Quando x aumenta, $\varphi(x)$ se aproxima de 1, e quando x diminui, $\varphi(x)$ se aproxima de 0. Na figura 4.3, podemos observar o comportamento da função.

Figura 4.3: Gráfico da função sigmoid



A função sigmoide é muito utilizada em camadas de saída de redes neurais que realizam classificação binária, onde a saída é interpretada como a probabilidade de pertencer a uma das duas classes (0 ou 1) e também na regulação do fluxo de informações dentro das células de memória da rede LSTM, decidindo quais informações devem ser esquecidas ou lembradas.

Softmax

A função de ativação softmax é especialmente importante em tarefas de classificação multiclasse, ou seja, classificar dados em diferentes categorias, sendo uma generalização da função sigmoide para casos não-binários. É uma função matemática utilizada em redes neurais para converter um conjunto de valores em uma distribuição de probabilidade, tal que os valores produzidos pertencem ao intervalo [0, 1], onde sua soma é igual a 1 [42]. Ou seja, num problema com i classes, por exemplo, a função softmax vai produzir i valores, que somam 1, onde cada valor representa a probabilidade da instância pertencer a uma das i possíveis classes.

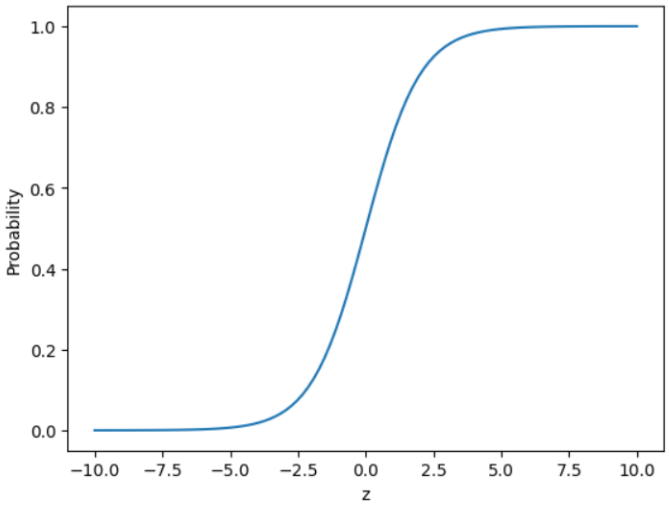
A fórmula da função de ativação softmax é dada pela equação 4.4.

$$\varphi(z)_i = \frac{e^{z_i}}{\sum_{j=1}^{N} e^{z_j}} \tag{4.4}$$

Onde e é a base do logaritmo natural, z é o vetor de entrada para a classe i vetor de entrada z e $\sum_{j=1}^{N} e^{zj}$ é a soma de todos os valores exponenciais para todas as classes possíveis j.

Na figura 4.4, apresentamos o gráfico da função softmax, onde podemos observar o valor z (para cada classe i), à medida que esse aumenta, a probabilidade da classe correspondente se aproxima de 1, indicando uma maior confiança na classificação dessa classe.

Figura 4.4: Gráfico da função softmax



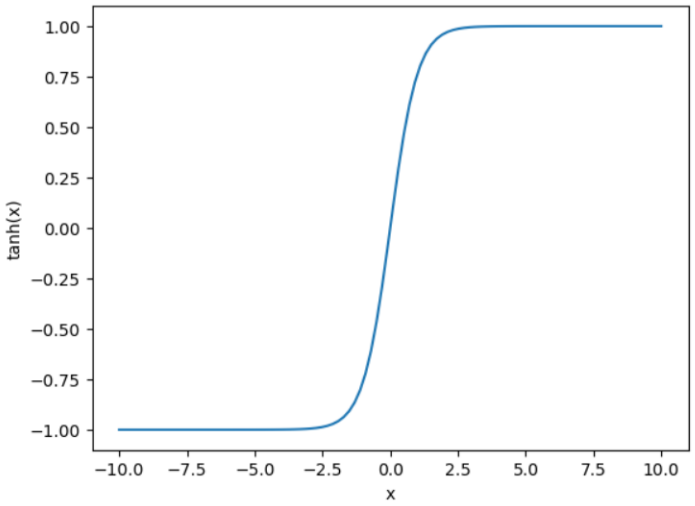
Tanh (Tangente Hiperbólica)

Essa função é comumente usada em redes neurais devido à sua propriedade de simetria em relação à origem e a sua capacidade de mapear valores negativos para valores negativos próximos a -1, e valores positivos para valores positivos próximos a 1 [43]. A normalização dos valores de saída em torno de zero pode melhorar a eficiência do treinamento. A fórmula da função tanh é definida como pela equação 4.5:

$$\varphi(x) = \frac{e^{x} - e^{x}}{e^{x} - e^{x}} \tag{4.5}$$

Onde e é a base do logaritmo natural e x é a entrada para a função. Na figura 4.5, exibimos o gráfico da função tanh.

Figura 4.5: Gráfico da função tanh



O gráfico é centrado em torno do valor 0 no eixo y. A função aumenta suavemente de -1 para 1 à medida que o valor de entrada x aumenta de valores negativos para positivos.

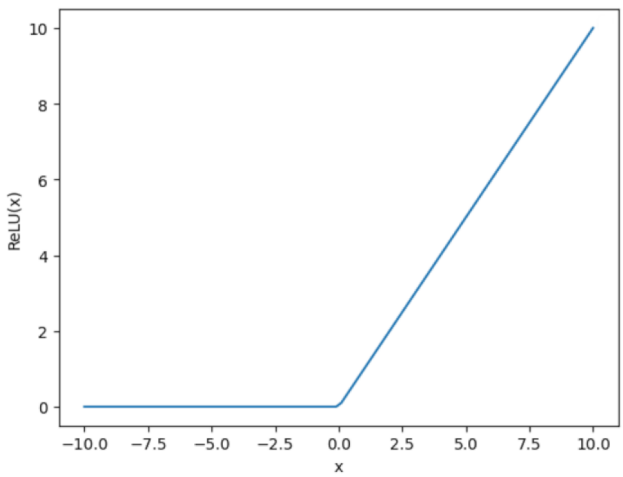
ReLU (Unidade Linear Retificada)

A função de ativação ReLU (Rectified Linear Unit) é especialmente comum em redes neurais profundas. Ela é definida pela fórmula 4.6:

$$\varphi(x) = max(0, x) \tag{4.6}$$

Isso significa que, para qualquer valor de entrada x, a função ReLU retorna 0 para $x \leq 0$ e retorna x para $x > 0$. Em termos simples, ela "ativa" um neurônio apenas se o sinal de entrada é positivo, como podemos observar o gráfico na figura 4.6.

Figura 4.6: Gráfico da função ReLU

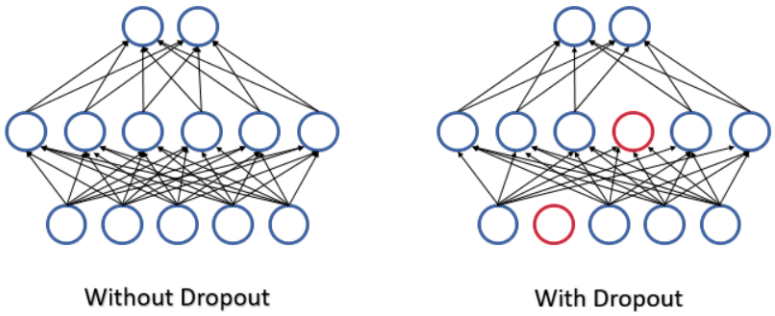


O comportamento linear para entradas positivas e nulo para entradas negativas é o que define a ReLU.

4.2.2 Dropout

O *dropout* é uma técnica de regularização em redes neurais utilizada para melhorar o desempenho e a capacidade de generalização, especialmente em contextos de aprendizado profundo, onde é amplamente utilizada para combater o problema de *overfitting*. Em uma rede neural, *overfitting* ocorre quando o modelo aprende padrões específicos do conjunto de dados de treino, mas falha em generalizar esses aprendizados para dados novos ou não vistos anteriormente. Para resolver esse tipo de problema, o *dropout* funciona "desligando" aleatoriamente um conjunto de neurônios durante o treinamento de uma rede neural [44]. Na figura 4.7, podemos visualizar uma rede neural antes e depois da utilização do *dropout*.

Figura 4.7: Rede neural sem dropout e com dropout



Fonte: [45]

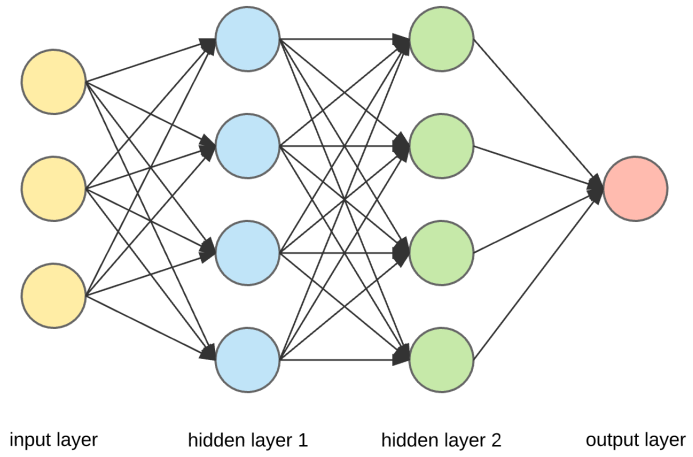
A técnica de *dropout* é eficaz, pois ao ignorar uma porção de neurônios temporariamente em cada etapa de treinamento, consegue impedir que unidades individuais se especializem demais em respostas específicas. Em outras palavras, ao desativar aleatoriamente certas unidades, a rede é forçada a não depender excessivamente de qualquer conjunto específico de neurônios. Isso ajuda a rede a aprender representações mais robustas, não depndendo de uma configuração específica de neurônios para fornecer uma resposta.

4.3 Rede neural densa (Dense neural network - DNN)

Uma DNN (rede neural densa) é uma das arquiteturas mais tradicionais e fundamentais no campo de aprendizado profundo (*deep learning*). Esta arquitetura é caracterizada

pela sua estrutura simples, onde cada neurônio de uma camada está conectado a todos os neurônios da camada subsequente, estas conexões são chamadas de “densas” porque formam uma rede densamente interligada sem lacunas entre os neurônios em camadas sucessivas [13]. Essas redes são um componente chave no campo do aprendizado de máquina e desempenham um papel vital em uma variedade de tarefas, incluindo classificação, regressão e até mesmo como blocos de construção em sistemas mais complexos, como redes neurais convolucionais (CNNs) e redes neurais recorrentes (RNNs). Na figura 4.8, podemos ver a estrutura de uma DNN.

Figura 4.8: Estrutura DNN



Fonte: [46]

Para entender o funcionamento de uma Rede Neural Densa, é essencial explorar sua estrutura e funcionamento interno. As DNNs são redes neurais do tipo feedforward, ou em português “alimentação direta”, são a forma mais simples e direta de redes neurais artificiais. Numa FFN, a informação move-se apenas em uma direção — para a frente — desde a entrada, que recebe os sinais, camadas ocultas que processam esses sinais, e uma camada de saída que fornece a predição ou classificação final. Não há ciclos ou laços nas conexões, cada nó da camada anterior está conectado a cada nó da próxima camada. Cada camada é composta de unidades, ou neurônios, que são basicamente funções matemáticas que recebem entradas, aplicam pesos (que são aprendidos durante o treinamento da rede), somam um bias (também aprendido durante o treinamento) e, finalmente, aplicam uma função de ativação não-linear.

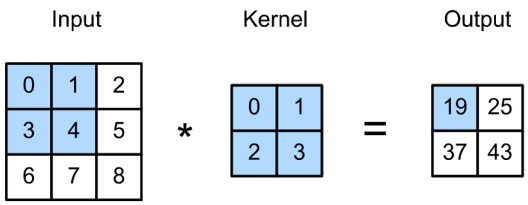
4.4 Rede neural convolucional (Convolutional neural network - CNN)

Uma CNN (rede neural convolucional) é um tipo de rede neural artificial profundamente interligada que é amplamente utilizada no processamento de imagens e visão computacional. Inspiradas pela complexa organização do córtex visual dos mamíferos, essas redes foram estruturadas de maneira a imitar a forma como os seres humanos percebem e reconhecem elementos visuais. Este tipo de rede é composto por múltiplas camadas que, juntas, são capazes de captar e processar uma hierarquia de características visuais (padrões espaciais e temporais em dados), o que as torna eficientes na resolução de problemas relacionados à visão computacional, como reconhecimento de imagens, classificação, detecção de objetos e análise de vídeo [14].

O funcionamento de uma CNN inicia-se na aplicação de suas camadas convolucionais, onde filtros (kernel) são empregados para realizar a operação de convolução sobre a imagem de entrada. Os filtros, pequenas matrizes de pesos aprendíveis, movem-se sobre a imagem e executam uma soma ponderada dos valores dos pixels [47], resultando em mapas de características que captam aspectos específicos, como bordas ou texturas. À medida que a informação flui pela rede, os filtros das camadas subsequentes conseguem reconhecer padrões cada vez mais abstratos, devido à combinação das características identificadas anteriormente.

A Figura 4.9, mostra um processo de convolução aplicada na matriz de entrada (*input*) que representa a imagem, um *kernel* de ordem 2 (2 linhas x 2 colunas) e a saída (*output*) que é uma matriz de ordem 2.

Figura 4.9: Convolução



Fonte: [48]

Logo após a convolução, normalmente são aplicadas as camadas de *pooling*, que têm a função de reduzir a dimensão espacial dos mapas de características. Essa redução é crucial não apenas para diminuir a carga computacional e o número de parâmetros da rede, mas também para aumentar a robustez das características detectadas, tornando-as

mais tolerantes a variações e a distorções na imagem de entrada.

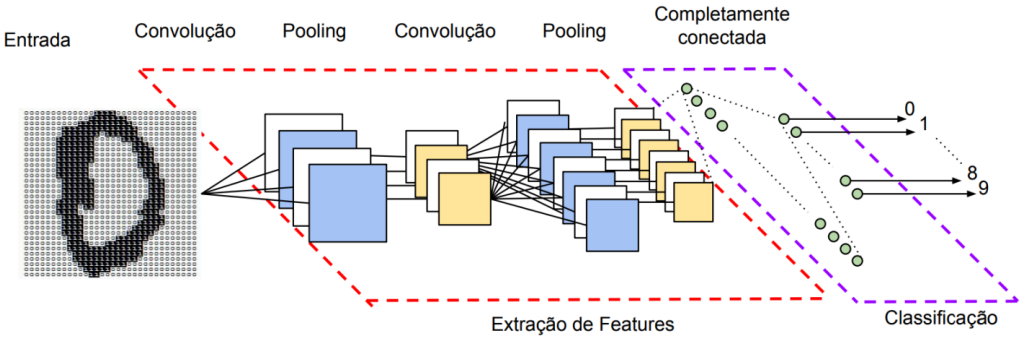
Ao longo do treinamento e do ajuste das CNNs, técnicas de normalização podem ser utilizadas para estabilizar a aprendizagem, equilibrando a distribuição dos valores de entrada para cada neurônio.

Concluindo a arquitetura das CNNs, encontram-se as camadas totalmente conectadas, que desempenham o papel de classificadores. Nesse estágio, todas as características extraídas e combinadas pelas camadas anteriores são utilizadas para determinar a saída final da rede, que pode ser, por exemplo, a classificação de uma imagem em uma categoria específica. Esses dados passam então por uma estrutura de FFN para a tarefa de classificação.

A adoção de funções de ativação não-lineares é um aspecto adicional que contribui para a eficácia das CNNs. Estas funções são empregadas para permitir que a rede aprenda relações complexas e não-lineares entre as entradas e as saídas.

Na figura 4.10, podemos ver a estrutura de uma CNN.

Figura 4.10: Estrutura CNN



Fonte: [49]

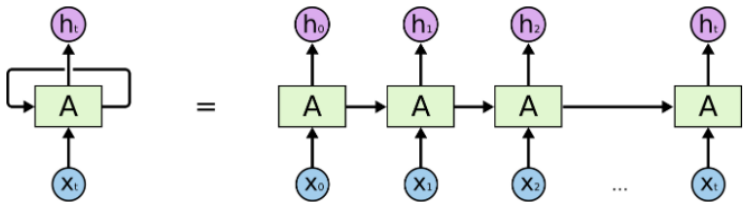
O treinamento de uma CNN é geralmente realizado por meio de retropropagação e do algoritmo de gradiente descendente, processos que ajustam os pesos e os filtros da rede com o objetivo de minimizar uma função de perda predeterminada. Esse ajuste é feito de forma iterativa, utilizando grandes conjuntos de dados anotados, em um processo que requer uma quantidade substancial de poder computacional.

4.5 Memória de longo e curto prazo (Long short-term memory - LSTM)

As LSTMs são uma espécie especial de redes neurais recorrentes (RNNs). As Redes Neurais Recorrentes são uma classe de redes neurais que possuem a capacidade de processar sequências de dados, como séries temporais ou linguagem natural. Isso é possível porque, ao contrário das redes neurais feedforward, onde as informações se movem em uma única direção, as RNNs são projetadas para processar sequências de dados, onde a saída de um passo de tempo é condicionada não apenas pela entrada atual, mas também por uma representação dos passos anteriores em repetição [50]. Essa memória temporária é útil quando o contexto é importante para a realização da tarefa, como em previsão de sequência ou modelagem de linguagem.

Na figura 4.11, mostramos uma parte da rede neural A, onde os valores de x são as entradas e o valores de h, as saídas.

Figura 4.11: Uma célula RNN



Fonte: [51]

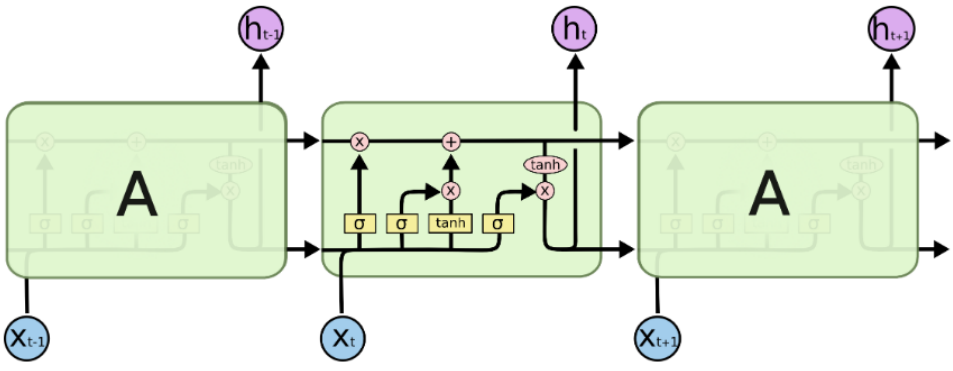
No entanto, as RNNs comuns enfrentam uma dificuldade significativa conhecida como o problema do desaparecimento do gradiente. Durante o treinamento, quando os gradientes são propagados para trás através do tempo para ajustar os pesos da rede (um processo conhecido como backpropagation through time, ou BPTT), ocorrendo a dissipação do gradiente. Isso dificulta a aprendizagem de dependências de longo prazo dentro das sequências, pois o modelo se torna incapaz de conectar eventos distantes no tempo.

A arquitetura LSTM foi introduzida por Hochreiter e Schmidhuber em 1997, para

combater o problema do desaparecimento do gradiente. As células LSTM possuem uma estrutura mais complexa do que as unidades padrão de RNN, contendo três portões (gates): o portão de esquecimento (forget gate), o portão de entrada (input gate) e o portão de saída (output gate) [15]. Esses portões regulam o fluxo de informações dentro e fora da célula, permitindo à LSTM adicionar ou remover informações de seu estado celular, que é uma espécie de "memória" da rede. A porta de esquecimento decide qual informação será descartada do estado celular. A porta de entrada atualiza o estado celular com novas informações, enquanto a porta de saída decide o que será transferido para a próxima etapa temporal. A interação dessas portas permite que as células LSTMs mantenham informações relevantes ao longo do tempo e descartem informações desnecessárias, resolvendo efetivamente o problema de manter dependências de longo prazo.

Na figura 4.12, apresentamos o funcionamento da cadeia de uma rede neural A. Onde x_{t-1}, x_t, x_{t+1}, h_{t-1}, h_t e h_{t+1} são respectivamente as entradas e saídas anteriores, atual e posteriores. Os portões (gates) são funções sigmoides (σ) que estabelecem valores no intervalo [0, 1], que tem como atribuição manter ou descartar as informações necessárias ou desnecessárias, respectivamente, para a tomada de decisão. A função tangente hiperbólica ($tanh$) tem como objetivo controlar o fluxo de informações e o estado de memória, limitando os valores no intervalo [-1, 1].

Figura 4.12: Mecanismo interno do LSTM



Fonte: [51]

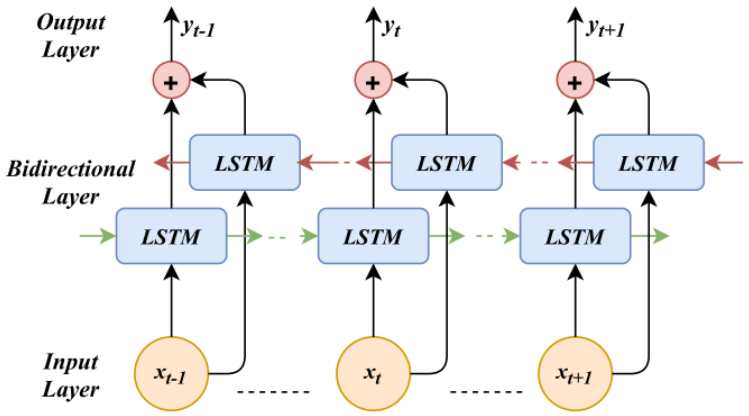
A importância das LSTMs no campo da aprendizagem de máquina e do processamento de sequências de dados é vasta. Elas têm sido usadas com sucesso em aplicações como reconhecimento de fala, tradução automática, geração de texto, previsão de séries temporais e muito mais. Sua habilidade para lidar com sequências de entrada e saída de comprimentos variáveis e sua robustez para lidar com dependências de longo alcance as tornam uma escolha popular para problemas complexos em sequências de dados.

4.6 Memória de longo e curto prazo bidirecional (Long short-term memory bidirectional - Bidirectional LSTM)

As redes neurais recorrentes bidirecionais (BiLSTMs) são uma variação das LSTMs que podem melhorar o desempenho do modelo em certas tarefas. Elas fazem isso treinando duas LSTMs em vez de uma, ou seja, é adicionda mais uma camada LSTM. Uma LSTM processa a sequência de entrada na ordem do tempo, enquanto a outra processa a sequência de entrada na ordem inversa do tempo. Isso permite que a rede capture informações do passado (processamento da sequência na ordem do tempo) e informações futuras (processamento da sequência na ordem inversa do tempo). As informações dessas duas LSTMs são então combinadas para fazer a previsão final. Isso pode ser particularmente útil em tarefas onde o contexto futuro é tão importante quanto o contexto passado [16]. Em seguida, combinamos as saídas de ambas as camadas LSTM de várias maneiras, como média, soma, multiplicação ou concatenação.

Na figura 4.13, exibimos o modelo de uma rede neural BiLSTM, onde x_{t-1}, x_t, x_{t+1}, y_{t-1}, y_t e y_{t+1} são respectivamente as entradas e saídas anteriores, atual e posteriores. A setas verdes indicam a sequência na ordem do tempo e as setas vermelhas na ordem inversa do tempo.

Figura 4.13: Modelo BiLSTM



Fonte: [52]

Em suma, as redes neurais recorrentes bidirecionais representam um avanço significativo na modelagem de sequências temporais, permitindo que os modelos capturem informações tanto do passado quanto do futuro para fazer previsões mais precisas. Elas são amplamente utilizadas em uma variedade de aplicações, incluindo tradução automática, reconhecimento de fala e análise de sentimentos.

Capítulo 5

Metodologia

Neste trabalho, realizamos uma análise detalhada das características vocais, essencial para identificar patologias relacionadas à voz. Utilizando um banco de dados de vozes que inclui amostras com nódulo, paralisia, edema de Reinke, cisto e também vozes saudáveis, aplicamos métodos avançados de processamento de sinal para extrair características distintas. Os métodos escolhidos são o MFCC, PNCC e ZCPA, cada um com suas peculiaridades e eficácia na captura das nuances das vozes.

Os coeficientes extraídos são então alimentados em diferentes arquiteturas de redes neurais: DNN, CNN, LSTM e BiLSTM. Essas redes neurais são treinadas e validadas para identificar se uma voz possui patologia ou é considerada saudável. Este capítulo detalha o banco de dados, os métodos de extração de características e as arquiteturas de rede neural empregadas.

5.1 Banco de dados

O banco de dados utilizado para as análises foram extraídas da base de dados do professor Edson Cataldo e do site do *Saarbruecken Voice Database (SVD)* [53], que chamaremos de base 1 e 2 respectivamente.

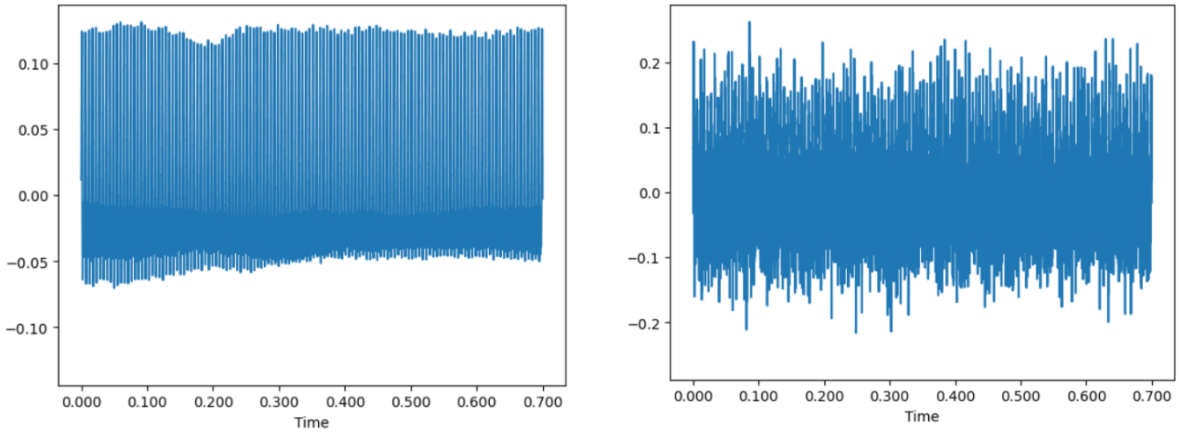
A base 1 é composta por 11 vozes saudáveis (6 masculinas e 5 femininas), 12 vozes com nódulo (1 masculina e 11 femininas) e 8 vozes com paralisia (4 masculinas e 4 femininas).

A base 2 é constituída por 687 vozes saudáveis (259 masculinas e 428 femininas), 6 vozes com cisto (1 masculina e 5 femininas), 68 vozes com edema de Reinke (7 masculinas e 61 femininas). Do total de vozes com edema de Reinke, 34 possuíam mais de uma patologia, como é o caso de um sinal de voz que foi diagnosticado com edema de Reinke,

laringite e voz senil. Não houve separação das vozes masculinas e femininas, e o banco de vozes (banco 1 e 2) foi agrupado por tipo de patologia e as vozes contêm pronúncias sustentadas da vogal /a/ e /e/, em alguns casos as pronúncias no tom normal, alto e baixo.

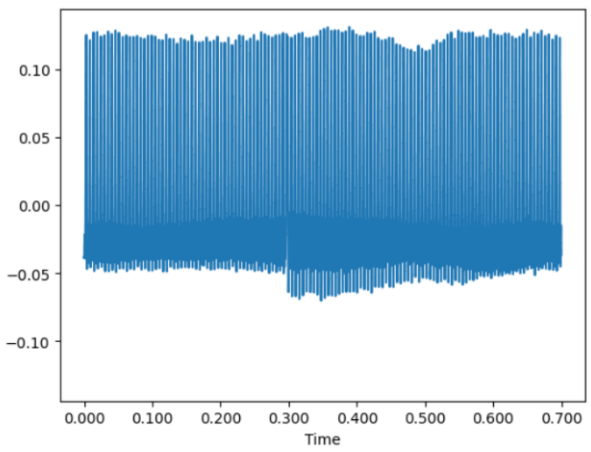
As redes neurais precisam de um grande número de dados para serem capazes de fazer generalizações e, com isso, ter um bom desempenho no reconhecimento de objetos. Por esse motivo, utilizaremos técnicas computacionais para o aumento de dados. Os sinais das vozes foram fragmentados em 0,5 e 0,7 segundos, adicionamos ruído branco e efetuamos um deslocamento temporal da parte inicial do áudio para pontos específicos. Com isso, obtivemos sinais de áudios originais sem alteração, com ruído branco ou deslocado, e com ruído branco e deslocado. Nas figuras 5.1 a, b e 5.2, mostramos respectivamente o *waveform* de um áudio original, com ruído branco e com deslocamento temporal.

Figura 5.1: *Waveform* original (a) e com ruído (b)



(a) (b)

Figura 5.2: *Waveform* deslocado



Para adicionar o ruído branco e fazer o deslocamento dos áudios, utilizamos a biblioteca *NumPy* no *Python*. A função *numpy.random.randn()* gera um *array* de números aleatórios, seguindo uma distribuição normal (Gaussiana) [54], onde um fator de multiplicação aumenta ou diminui o nível do ruído. Os valores utilizados como fator de multiplicação foram: 0.05 (mais ruído), 0.03, 0.01, 0.009 e 0.007 (menos ruído). E a função *np.roll(data, x)* [55] desloca os elementos de um array x posições para frente. Os valores de deslocamento foram respectivamente 1000, 3500 e 6000.

Após a conclusão do aumento do banco de dados, totalizamos 1188 vozes saudáveis, 1017 vozes com nódulo e 765 vozes com paralisia do banco 1 e 1359 vozes saudáveis, 261 vozes com cisto, 1152 vozes com edema de Reink do banco 2.

Para avaliar a eficácia dos métodos de extração em vozes com ruído, introduzimos ruído branco com um fator multiplicador de 0.20, significativamente mais alto do que o aplicado nos estudos iniciais. Nos testes, combinamos vozes com nódulos e paralisia, agrupando-as sob a categoria de patologias, e as comparamos com vozes saudáveis.

5.2 Extração de características das vozes

Os procedimentos adotados neste estudo visam extrair os coeficientes cepstrais do sinal de voz. Todos os processos para extrair características foram executados usando *Python*. O método MFCC foi realizado utilizando a biblioteca *Librosa*, através da função *librosa.feature.mfcc()*, onde foram extraídos 40 coeficientes. Para extrair características pelo método PNCC, empregou-se a função *spafe.features.pncc()* da biblioteca *Spafe*. Nesse processo, 32 coeficientes foram obtidos para áudios de 0,5 segundos, e 45 coeficientes para áudios de 0,7 segundos. O método ZCPA foi implementado por [5], utilizando a biblioteca *Scipy* através da fução *signal.filtfilt()*, onde foram extraídos 66 coeficientes.

5.3 Redes Neurais

Para desenvolver e validar nossos modelos de redes neurais, utilizamos um conjunto de dados abordado na seção 5.1. Este conjunto de dados foi dividido em três segmentos distintos para treinamento, validação e teste do modelo. Alocamos 70% dos dados para treinamento, o que permitiu ao modelo aprender e ajustar-se às características das vozes patológicas e saudáveis. A seguir, 10% dos dados foram destinados à validação. Por fim, os restantes 20% dos dados foram utilizados para teste.

O código foi desenvolvido em *Python*, utilizando bibliotecas específicas para redes neurais: *NumPy*, *Tensorflow*, *Keras* e *Sklearn*.

A rede DNN foi construída com uma camada densa de 32 neurônios e função de ativação $ReLU$, seguida de uma camada de saída com 2 neurônios, adequada para classificação binária, usando a função de ativação *softmax*.

Na CNN, foram implementados dois blocos convolucionais. O primeiro tem uma camada convolucional unidimensional com 64 filtros, *kernel* de tamanho 10 e função de ativação $ReLU$, além de um *dropout* de 40% e uma camada de max-pooling de tamanho 4. O segundo bloco, também unidimensional, possui 128 filtros, *kernel* de tamanho 10, *padding* = *"same"*, garantindo que a saída tenha o mesmo tamanho que a entrada, função de ativação $ReLU$, seguido por um *dropout* de 40% e max-pooling de tamanho 4. Essa rede também inclui uma camada densa de 64 neurônios com 40% de dropout e uma camada de saída com 2 neurônios e função de ativação *softmax*.

A LSTM conta com três camadas, iniciando com duas camadas LSTM de 100 unidades cada, seguidas de uma camada de saída com dois neurônios e função de ativação *softmax*, intercaladas por dois *dropouts* de 30%.

Por fim, a Bi_LSTM possui duas camadas LSTM bidirecionais de 100 unidades cada, com um *dropout* de 30% entre elas, e uma camada de saída com dois neurônios e ativação *softmax*.

Na compilação dos modelos, utilizamos a função de perda *categorical_crossentropy*, que é comumente usada para problemas de classificação multiclasse. A métrica usada foi *accuracy* e o otimizador, *adam*.

Capítulo 6

Resultados

Na avaliação de modelos de classificação, é fundamental compreender a terminologia usada para descrever os resultados das previsões. Estes termos não apenas definem o tipo de acerto ou erro cometido pelo modelo, mas também são essenciais para a interpretação da sua performance em cenários práticos. A seguir, apresenta-se a descrição dos principais termos utilizados nesse trabalho.

VP (verdadeiro positivo) corresponde às instâncias em que o modelo corretamente prevê a classe Positiva. FN (falso negativo) descreve as situações onde o modelo erra ao classificar um verdadeiro caso da classe Positiva como Negativo. Em contraste, FP (falso positivo) ocorre quando o modelo equivocadamente classifica um caso da classe Negativa como Positivo. VN (verdadeiro negativo) é a correta identificação de um caso Negativo pelo modelo. Estabelecemos que a classe positiva indica a presença de uma patologia vocal, enquanto a classe negativa representa a condição de uma voz saudável.

Neste capítulo, exploramos a eficácia dos modelos de redes neurais em classificar corretamente patologias de voz, utilizando o conjunto de métricas de avaliação de desempenho: acurácia, precisão, *recall* e o *f1-Score*, tal como apresentamos nas equações 6.1, 6.2, 6.3 e 6.4, respectivamente.

$$A = \frac{VP + VN}{VP + VN + FP + FN} \tag{6.1}$$

$$P = \frac{VP}{VP + FP} \tag{6.2}$$

$$R = \frac{VP}{VP + FN} \tag{6.3}$$

$$f1 = \frac{2 \cdot P \cdot R}{P + R} \tag{6.4}$$

A acurácia nos dá uma visão geral da capacidade do modelo em fazer previsões corretas (tanto positivas quanto negativas), mas isso não revela o quadro completo. Ao aprofundarmos na precisão, analisamos a taxa de verdadeiros positivos sobre todas as previsões positivas, o que é crucial quando o custo de um falso positivo é alto. O *recall*, por outro lado, se concentra na proporção de positivos reais identificados, sendo vital em cenários onde não se pode deixar de detectar um caso real. O *f1-Score* é uma métrica que harmoniza precisão e recall, fornecendo um indicador global de desempenho balanceado. Especialmente útil quando precisamos de um equilíbrio entre a identificação correta de casos positivos (patologias de voz) e a minimização de alarmes falsos, um *f1-Score* alto indica que o modelo é confiável tanto em sua assertividade quanto em sua sensibilidade.

6.1 Validação dos métodos

Abordaremos neste tópico a validação dos estudos previamente mencionados. Na introdução, foram abordadas técnicas de extração de características para o reconhecimento de falantes, destacando-se a robustez dos métodos PNCC e ZCPA na identificação correta de palavras pronunciadas em ambientes ruidosos. Adicionalmente, ressaltou-se a eficiência do método MFCC na distinção entre vozes patológicas e saudáveis, especialmente em situações de emissão de vogais sustentadas. A validação que será executada tem o objetivo de distinguir entre vozes saudáveis e vozes saudáveis que estão sob a influência de ruídos em diversos níveis de acordo com a tabela 6.1. Conforme descrito no capítulo 5, a base de dados é composta por segmentos de 0,5 e 0,7 segundos das vogais /a/ e /e/ sustentadas.

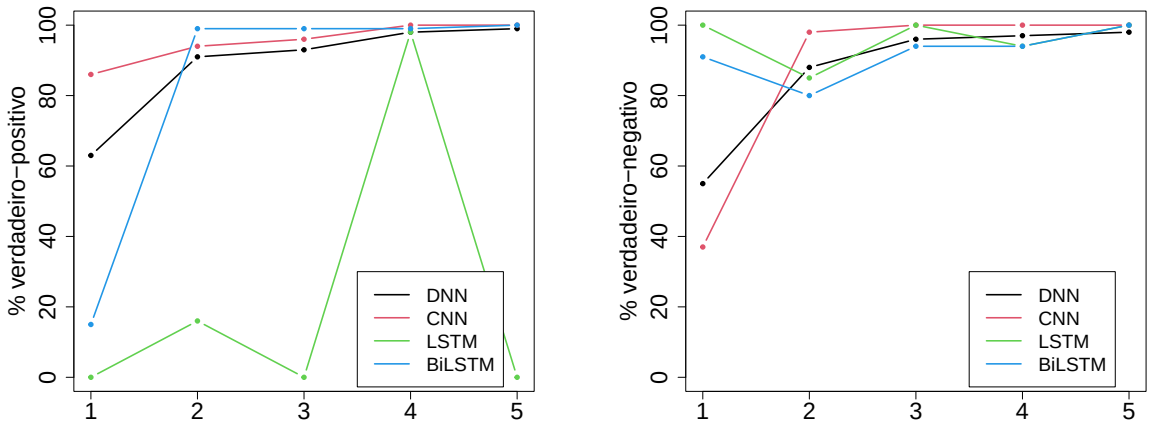
Tabela 6.1: Níveis de ruído correspondentes a cada fator

Fator	Nível do ruído
0.01	Nível 1 - Ruído baixo
0.05	Nível 2 - Ruído médio
0.20	Nível 3 - Ruído alto
0.50	Nível 4 - Ruído muito alto
2.00	Nível 5 - Só ruído

Ao examinar os resultados obtidos pelo método MFCC, constatamos que todas as redes neurais (DNN, CNN, LSTM, BiLSTM) analisadas alcançaram uma taxa de classificação de 100% para vozes com todos os níveis de ruídos e vozes sem ruído.

Na análise feita no método PNCC, a figura 6.1 ilustra no eixo horizontal os diferentes níveis de ruído, visando aferir a eficácia na detecção de vozes saudáveis tanto em condições de ruído quanto sem ruído. No eixo vertical, é mensurada a porcentagem de acurácia para os casos de verdadeiro positivo (voz com ruído) e verdadeiro negativo (voz sem ruído). Foi notado um aumento progressivo, conforme o nível de ruído aumentava, na precisão da distinção entre vozes afetadas por ruído e as não afetadas, com as taxas de acerto para ambos os verdadeiros, convergindo para faixas entre 98% e 100%. Destaca-se, no entanto, que a rede neural LSTM apresentou um desvio do padrão geral observado.

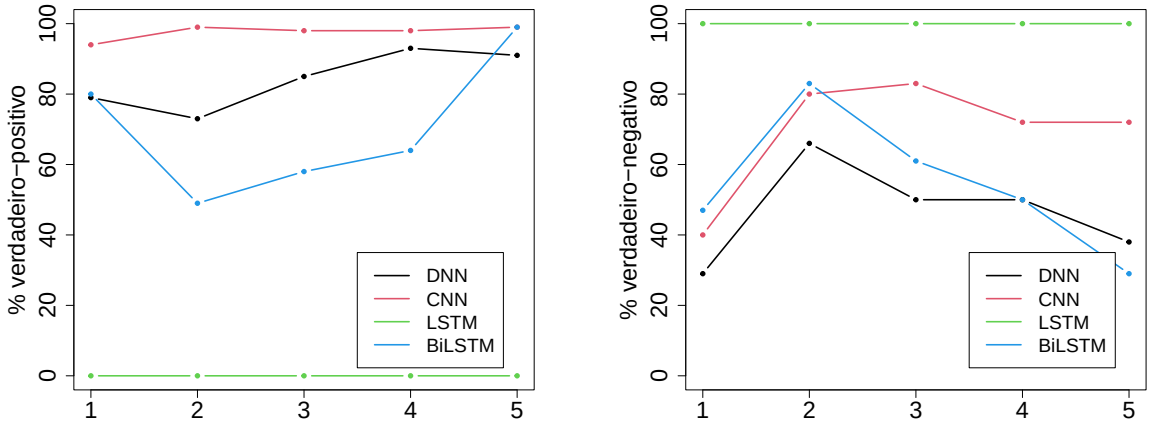
Figura 6.1: PNCC, voz com ruídos (a) e voz sem ruído (b)



(a) (b)

Analisando o método ZCPA, conforme apresentado na figura 6.2, nota-se que um incremento no nível de ruído resulta em um aumento na taxa de acertos de verdadeiro positivo, com as porcentagens mais elevadas, situando-se entre os níveis de ruído 4 e 5. Além disso, destaca-se a discrepância encontrada nos resultados da rede neural LSTM. No caso de verdadeiro negativo, as redes neurais atingem um ápice de maior acerto no nível de ruído 2.

Figura 6.2: ZCPA, voz com ruídos (a) e voz sem ruído (b)



(a) (b)

6.2 Resultados - MFCC

Nesta seção, abordamos os resultados da pesquisa que se concentra na utilização de redes neurais para a classificação de patologias vocais. Através do método MFCC, os modelos de redes neurais: DNN, CNN, LSTM e BiLSTM, foram considerados na tarefa de classificação.

De acordo com a tabela 6.2, os melhores resultados foram obtidos para identificação da patologia cisto, onde todos os modelos alcançaram um ótimo desempenho, com 100% de verdadeiro positivo e verdadeiro negativo e 0% de falso negativo e falso positivo. Isso significa que cada um dos modelos foi capaz de identificar todos os casos de cisto e todos os casos saudáveis sem cometer erros.

Tabela 6.2: Matriz de confusão dos modelos para cisto e saudável usando MFCC

	MFCC (cisto x saudável)			
	DNN	CNN	LSTM	BiLSTM
Verdadeiro-Positivo	100%	100%	100%	100%
Falso-Negativo	0%	0%	0%	0%
Falso-Positivo	0%	0%	0%	0%
Verdadeiro-Negativo	100%	100%	100%	100%

Os resultados da tabela 6.3 mostram que, para a classificação de edema de Reinke, o BiLSTM exibiu acurácia perfeita, enquanto os outros modelos tiveram pequenas taxas de falso negativo, com o CNN sendo o menos preciso. Todos os modelos foram extremamente eficazes em evitar falsos positivos, indicando uma alta especificidade na detecção de casos saudáveis.

Tabela 6.3: Matriz de confusão dos modelos para edema de Reinke e saudável usando MFCC

	MFCC (reinke x saudável)			
	DNN	CNN	LSTM	BiLSTM
Verdadeiro-Positivo	99%	96%	98%	100%
Falso-Negativo	1%	4%	2%	0%
Falso-Positivo	0%	0%	0%	0%
Verdadeiro-Negativo	100%	100%	100%	100%

Podemos verificar na tabela 6.4 que os modelos DNN, LSTM e BiLSTM apresentaram resultados semelhantes, com 99% de Verdadeiro-Positivo e 1% de Falso-Negativo, indicando alta sensibilidade na detecção de nódulos. O CNN, por sua vez, diferenciou-se ao alcançar 100% de Verdadeiro-Positivo, sem erros de Falso-Negativo, mas com uma taxa ligeiramente inferior de Verdadeiro-Negativo (98%). O LSTM se destacou com 100% de Verdadeiro-Negativo, mostrando-se extremamente preciso na classificação de casos saudáveis. Falsos-Positivos foram raros, com o LSTM não apresentando nenhum e os outros modelos com uma taxa de 1% a 2%. Em resumo, o CNN mostrou-se mais eficaz na identificação de nódulos, enquanto o LSTM foi superior na correta identificação de vozes saudáveis.

Tabela 6.4: Matriz de confusão dos modelos para nódulo e saudável usando MFCC

	MFCC (nódulo x saudável)			
	DNN	CNN	LSTM	BiLSTM
Verdadeiro-Positivo	99%	100%	99%	99%
Falso-Negativo	1%	0%	1%	1%
Falso-Positivo	1%	2%	0%	1%
Verdadeiro-Negativo	99%	98%	100%	99%

A tabela 6.5 mostra que os modelos DNN, CNN e BiLSTM tiveram um desempenho quase perfeito na classificação de casos de paralisia e saudáveis usando MFCC, com 99% de verdadeiro positivo e 1% de Falso-Negativo, além de manterem 100% de verdadeiro negativo, indicando uma classificação correta de todos os casos saudáveis. O modelo LSTM teve uma leve queda na taxa de verdadeiro positivo (98%) e uma taxa um pouco maior de Falso-Negativo (2%), mas ainda assim manteve 100% de verdadeiro negativo.

Nenhum dos modelos produziu falsos positivos, evidenciando uma alta especificidade na detecção de condições saudáveis.

Tabela 6.5: Matriz de confusão dos modelos para paralisia e saudável usando MFCC

	MFCC (paralisia x saudável)			
	DNN	CNN	LSTM	BiLSTM
Verdadeiro-Positivo	99%	99%	98%	99%
Falso-Negativo	1%	1%	2%	1%
Falso-Positivo	0%	0%	0%	0%
Verdadeiro-Negativo	100%	100%	100%	100%

Em conclusão, os resultados indicam que as redes neurais, são extremamente eficazes na classificação de patologias vocais usando o método MFCC. A alta taxa de acertos (98% a 100%) em quase todos os modelos e condições testadas demonstra o potencial significativo dessas técnicas avançadas de inteligência artificial no campo do diagnóstico vocal.

6.3 Resultados - PNCC

Nesta parte do estudo, exploramos os resultados obtidos na classificação de patologias vocais utilizando redes neurais via o método PNCC.

A tabela 6.6 indica uma variação significativa na eficácia dos modelos de aprendizado de máquina. O CNN e o BiLSTM mostraram um desempenho robusto em detectar casos de cisto com 86% de verdadeiro positivo, embora o CNN tenha uma taxa maior de falso positivo (45%) em comparação ao BiLSTM (40%). O DNN teve um desempenho notavelmente mais baixo com apenas 37% de Verdadeiro-Positivo. Surpreendentemente, o LSTM não conseguiu identificar corretamente nenhum caso de cisto (0% de Verdadeiro-Positivo), mas teve um desempenho perfeito na classificação de casos saudáveis (100% de verdadeiro negativo). Os resultados indicam que a escolha do modelo pode ter um impacto substancial na precisão do diagnóstico, com o LSTM sendo extremamente específico, mas não sensível, e o CNN e o BiLSTM oferecendo um equilíbrio melhor entre sensibilidade e especificidade.

Na tarefa de classificação de edema de Reinke versus tecidos saudáveis apresentado na 6.7, o modelo LSTM obteve a maior taxa de Verdadeiro-Positivo com 79%, sugerindo uma boa sensibilidade na identificação correta de casos patológicos. No entanto, esse modelo também registrou a mais alta taxa de falso positivo, 59%, indicando uma propensão significativa para classificar incorretamente tecidos saudáveis como patológicos. O CNN

Tabela 6.6: Matriz de confusão dos modelos para cisto e saudável usando PNCC

	PNCC (cisto x saudável)			
	DNN	CNN	LSTM	BiLSTM
Verdadeiro-Positivo	37%	86%	0%	86%
Falso-Negativo	63%	14%	100%	14%
Falso-Positivo	33%	45%	0%	40%
Verdadeiro-Negativo	77%	55%	100%	60%

apresentou um desempenho mais equilibrado com 58% de verdadeiro positivo e a menor taxa de falso positivo, 15%, juntamente com a melhor taxa de verdadeiro negativo, 85%, evidenciando uma precisão superior em discriminar casos saudáveis. Por outro lado, o DNN mostrou a menor taxa de verdadeiro positivo, 47%, e o BiLSTM teve uma performance razoável com 70% de verdadeiro positivo, mas com taxas de falso negativo e falso positivo que indicam uma necessidade de ajustes para melhorar sua precisão diagnóstica. De forma geral, o método CNN obteve o melhor desempenho com acurácia de 72%, enquanto os modelos DNN, LSTM e BiLSTM resultaram em acurácias de 62%, 59% e 68% respectivamente.

Tabela 6.7: Matriz de confusão dos modelos para edema de Reinke e saudável usando PNCC

	PNCC (reinke x saudável)			
	DNN	CNN	LSTM	BiLSTM
Verdadeiro-Positivo	47%	58%	79%	70%
Falso-Negativo	53%	42%	21%	30%
Falso-Positivo	25%	15%	59%	25%
Verdadeiro-Negativo	75%	85%	41%	65%

A tabela 6.8 revela que o modelo LSTM se sobressai com 88% de Verdadeiro-Positivo, indicando alta sensibilidade, mas com uma taxa de Falso-Positivo relativamente alta de 35%. O CNN, apesar de ter apenas 57% de Verdadeiro-Positivo, apresenta a melhor especificidade com 82% de Verdadeiro-Negativo e a menor taxa de Falso-Positivo (18%). O DNN mostra um equilíbrio com 70% de Verdadeiro-Positivo e 78% de Verdadeiro-Negativo, enquanto o BiLSTM, com 64% de Verdadeiro-Positivo e 80% de Verdadeiro-Negativo, reflete uma precisão moderada tanto na identificação de casos de nódulo quanto na exclusão de não-nódulos. A variação nos resultados destaca a complexidade na escolha do modelo ideal, onde a precisão em identificar corretamente as patologias deve ser cuidadosamente balanceada com a capacidade de evitar alarmes falsos.

Tabela 6.8: Matriz de confusão dos modelos para nódulo e saudável usando PNCC

	PNCC (nódulo x saudável)			
	DNN	CNN	LSTM	BiLSTM
Verdadeiro-Positivo	70%	57%	88%	64%
Falso-Negativo	30%	43%	12%	36%
Falso-Positivo	22%	18%	35%	20%
Verdadeiro-Negativo	78%	82%	65%	80%

O modelo LSTM, apresentado na tabela 6.9, se destaca com a maior taxa de Verdadeiro-Positivo (76%) e uma baixa taxa de Falso-Negativo (24%), sugerindo uma boa capacidade de identificar corretamente casos de paralisia. Em contraste, o CNN teve a menor taxa de Verdadeiro-Positivo (34%), mas a maior de vVerdadeiro negativo (98%), indicando uma alta especificidade, mas uma sensibilidade limitada. O DNN apresentou um desempenho moderado em Verdadeiro positivo (57%) e uma especificidade razoável (81%), enquanto o BiLSTM mostrou um equilíbrio entre verdadeiro positivo (61%) e Verdadeiro-Negativo (80%). O baixo falso positivo do CNN (2%) é notável, apontando para sua forte capacidade em corretamente excluir casos saudáveis.

Tabela 6.9: Matriz de confusão dos modelos para paralisia e saudável usando PNCC

	PNCC (paralisia x saudável)			
	DNN	CNN	LSTM	BiLSTM
Verdadeiro-Positivo	57%	34%	76%	61%
Falso-Negativo	43%	66%	24%	39%
Falso-Positivo	19%	2%	28%	20%
Verdadeiro-Negativo	81%	98%	72%	80%

Os resultados mostram que diferentes modelos têm desempenhos variados. O CNN e o BiLSTM são mais eficazes na detecção de cistos, mas o CNN tem uma taxa de falso positivo maior. O LSTM é perfeito em identificar vozes saudáveis, mas fraco em detectar patologias, dando uma visão (acurácia) geral de 55% do modelo. Na classificação de edema de Reinke, o LSTM tem a maior taxa de verdadeiro positivo. Em casos de nódulos, o LSTM tem alta sensibilidade, mas também muitos falsos positivos. Na identificação de paralisia, o LSTM é eficaz, mas o CNN tem baixa taxa de verdadeiro positivo e alta de verdadeiro negativo, indicando alta especificidade. Esses resultados sublinham a importância da escolha cuidadosa do modelo para o diagnóstico preciso em patologias vocais e que o método PNCC, aliado com as redes neurais verificadas, não é tão eficiente quanto o MFCC.

6.4 Resultados - ZCPA

Nesta seção da pesquisa, investigamos os desempenhos alcançados nas avaliações com as redes neurais empregando o método ZCPA.

Nas tabelas 6.10, 6.11, 6.12 e 6.13, observamos que os altos índices de falso negativo sugerem uma tendência desses modelos em classificar incorretamente casos de patologia como saudáveis. Em contraste, a alta taxa de verdadeiro negativo em todos os modelos indica uma forte capacidade de identificar corretamente indivíduos saudáveis. Considerando as informações apresentadas nas tabelas, para obter uma perspectiva mais precisa dos modelos, isto é, a proximidade com que os modelos correspondem à realidade, destacamos que a acurácia varia de 51% a 75%.

Tabela 6.10: Matriz de confusão dos modelos para cisto e saudável usando ZCPA

	ZCPA (cisto x saudável)			
	DNN	CNN	LSTM	BiLSTM
Verdadeiro-Positivo	14%	51%	0%	0%
Falso-Negativo	86%	49%	100%	100%
Falso-Positivo	6%	44%	0%	0%
Verdadeiro-Negativo	94%	66%	100%	100%

Tabela 6.11: Matriz de confusão dos modelos para edema de Reinke e saudável usando ZCPA

	ZCPA (reinke x saudável)			
	DNN	CNN	LSTM	BiLSTM
Verdadeiro-Positivo	47%	11%	0%	0%
Falso-Negativo	53%	89%	100%	100%
Falso-Positivo	42%	9%	0%	0%
Verdadeiro-Negativo	58%	91%	100%	100%

Tabela 6.12: Matriz de confusão dos modelos para nódulo e saudável usando ZCPA

	ZCPA (nódulo x saudável)			
	DNN	CNN	LSTM	BiLSTM
Verdadeiro-Positivo	34%	14%	0%	22%
Falso-Negativo	66%	86%	100%	78%
Falso-Positivo	26%	11%	0%	24%
Verdadeiro-Negativo	74%	89%	100%	76%

Tabela 6.13: Matriz de confusão dos modelos para paralisia e saudável usando ZCPA

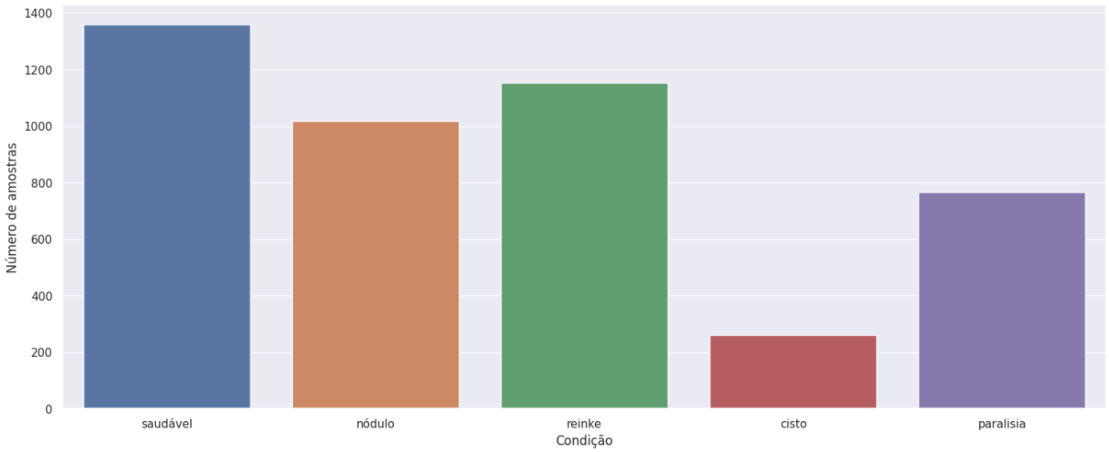
	ZCPA (paralisia x saudável)			
	DNN	CNN	LSTM	BiLSTM
Verdadeiro-Positivo	1%	18%	0%	0%
Falso-Negativo	99%	82%	100%	100%
Falso-Positivo	0%	4%	0%	0%
Verdadeiro-Negativo	100%	96%	100%	100%

6.5 Identificação de vozes com diversas patologias e vozes saudáveis

Considerando os resultados promissores obtidos pelo método MFCC, e os resultados atingidos pelos métodos PNCC e ZCPA em combinação com as redes neurais DNN, CNN, LSTM e BiLSTM na distinção entre vozes com um tipo de patologia e vozes saudáveis, optamos por avaliar a eficácia dessas abordagens ao integrar todas as vozes com patologia - nódulos, paralisia, edema de Reinke e cistos - com amostras de vozes saudáveis. Nesta seção, abordaremos os resultados através das métricas de avaliação de desempenho: acurácia, precisão, *recall* e *f1-Score* de cada modelo de rede neural.

Na figura 6.3, apresentamos um gráfico de barras com o número de amostras que foram utilizadas neste teste, sendo 1359 vozes saudáveis, 1152 vozes com edema de Reinke, 1017 vozes com nódulo, 765 vozes com paralisia e 261 vozes com cisto.

Figura 6.3: Número de amostras de vozes saudáveis e com patologias



6.5.1 Resultados - MFCC

Os resultados do modelo de classificação vocal apresentados na tabela 6.14 exibe alta acurácia, com destaque para o diagnóstico perfeito de nódulos e paralisia. Embora a

identificação de cistos e edema de Reinke não tenha atingido 100%, as altas taxas de acurácia obtidas indicam a confiabilidade do método.

Tabela 6.14: Classificação da rede DNN na detecção de patologias usando MFCC

Condição	**Precisão**	**Recall**	**f1-Score**
Cisto	0.93	0.88	0.90
Nódulo	1.00	1.00	1.00
Paralisia	1.00	1.00	1.00
Reinke	0.96	0.98	0.97
Saudável	1.00	0.99	1.00
Acurácia = 0.99			

O modelo CNN da tabela 6.15 de classificação vocal apresenta uma acurácia geral de 0,97, um pouco abaixo da apresentada no modelo DNN. A precisão na identificação de nódulos e paralisia continua se destacando. A condição de cisto, apesar de ter uma alta precisão, mostrou um recall mais baixo, sugerindo uma tendência do modelo em não detectar todos os casos reais. O desempenho na condição de edema de Reinke e na identificação de vozes saudáveis continua se mostrando eficiente.

Tabela 6.15: Classificação da rede CNN na detecção de patologias usando MFCC

Condição	**Precisão**	**Recall**	**f1-Score**
Cisto	0.97	0.67	0.79
Nódulo	1.00	1.00	1.00
Paralisia	1.00	1.00	1.00
Reinke	0.92	0.98	0.95
Saudável	0.98	0.99	0.99
Acurácia = 0.97			

O modelo LSTM, apresentado na tabela 6.16, se mantém com uma alta precisão para nódulo, paralisia, edema de Reinke e vozes saudáveis e obteve um aumento considerável no recall, mostrando a melhora na detecção de cisto. A performance geral do modelo é acentuada pelo alto valor da acurácia.

Tabela 6.16: Classificação da rede LSTM na detecção de patologias usando MFCC

Condição	**Precisão**	**Recall**	**f1-Score**
Cisto	0.93	0.93	0.93
Nódulo	1.00	1.00	1.00
Paralisia	1.00	1.00	1.00
Reinke	0.98	0.96	0.97
Saudável	0.99	1.00	0.99
Acurácia = 0.99			

Por fim, na tabela 6.17, o modelo de classificação mantém uma eficácia notável com uma acurácia de 99%, evidenciando uma precisão e recall perfeitos na identificação de nódulos, paralisia e vozes saudáveis. O desempenho na detecção de cistos e edema de Reinke também foi elevado.

Tabela 6.17: Classificação da rede BiLSTM na detecção de patologias usando MFCC

Condição	**Precisão**	**Recall**	**f1-Score**
Cisto	0.95	0.91	0.93
Nódulo	1.00	1.00	1.00
Paralisia	1.00	1.00	1.00
Reinke	0.98	0.99	0.98
Saudável	1.00	1.00	1.00
Acurácia = 0.99			

6.5.2 Resultados - PNCC

Os resultados dos modelos de classificação vocal para o método PNCC apresentados na tabelas 6.18, 6.19, 6.20 e 6.21, apresentam uma acurácia de 0.47 a 0.65, indicando a necessidade de aprimoramento no método de extração de características para a identificação de vozes com patologias e vozes saudáveis.

Tabela 6.18: Classificação da rede DNN na detecção de patologias usando PNCC

Condição	**Precisão**	**Recall**	**f1-Score**
Cisto	0.00	0.00	0.00
Nódulo	0.50	0.70	0.58
Paralisia	0.44	0.08	0.13
Reinke	0.48	0.58	0.52
Saudável	0.44	0.52	0.48
Acurácia = 0.47			

Tabela 6.19: Classificação da rede CNN na detecção de patologias usando PNCC

Condição	**Precisão**	**Recall**	**f1-Score**
Cisto	1.00	0.09	0.16
Nódulo	0.60	0.96	0.74
Paralisia	0.88	0.09	0.17
Reinke	0.61	0.66	0.63
Saudável	0.65	0.77	0.70
Acurácia = 0.63			

Tabela 6.20: Classificação da rede LSTM na detecção de patologias usando PNCC

Condição	**Precisão**	**Recall**	**f1-Score**
Cisto	0.44	0.25	0.31
Nódulo	0.74	0.88	0.80
Paralisia	0.80	0.44	0.57
Reinke	0.55	0.64	0.59
Saudável	0.63	0.67	0.65
Acurácia = 0.64			

Tabela 6.21: Classificação da rede BiLSTM na detecção de patologias usando PNCC

Condição	**Precisão**	**Recall**	**f1-Score**
Cisto	0.47	0.30	0.37
Nódulo	0.76	0.77	0.76
Paralisia	0.65	0.62	0.64
Reinke	0.57	0.71	0.64
Saudável	0.69	0.62	0.65
Acurácia = 0.65			

6.5.3 Resultados - ZCPA

As tabelas 6.22, 6.23, 6.24 e 6.25, exibem os resultados dos modelos de classificação vocal usando o método ZCPA com uma acurácia variando entre 0.30 e 0.50. Isso sugere a necessidade de melhorias no método de extração de características para a identificação de vozes com patologias e vozes saudáveis.

Tabela 6.22: Classificação da rede DNN na detecção de patologias usando ZCPA

Condição	**Precisão**	**Recall**	**f1-Score**
Cisto	0.00	0.00	0.00
Nódulo	0.54	0.86	0.66
Paralisia	0.00	0.00	0.00
Reinke	0.37	0.27	0.31
Saudável	0.46	0.70	0.55
Acurácia = 0.47			

Tabela 6.23: Classificação da rede CNN na detecção de patologias usando ZCPA

Condição	**Precisão**	**Recall**	**f1-Score**
Cisto	0.67	0.04	0.07
Nódulo	0.57	0.87	0.69
Paralisia	0.54	0.20	0.30
Reinke	0.40	0.35	0.37
Saudável	0.49	0.63	0.55
Acurácia = 0.50			

Tabela 6.24: Classificação da rede LSTM na detecção de patologias usando ZCPA

Condição	**Precisão**	**Recall**	**f1-Score**
Cisto	0.00	0.00	0.00
Nódulo	0.00	0.00	0.00
Paralisia	0.00	0.00	0.00
Reinke	0.00	0.00	0.00
Saudável	0.30	1.00	0.46
Acurácia = 0.30			

Tabela 6.25: Classificação da rede BiLSTM na detecção de patologias usando ZCPA

Condição	**Precisão**	**Recall**	**f1-Score**
Cisto	0.15	0.05	0.08
Nódulo	0.55	0.71	0.62
Paralisia	0.47	0.30	0.36
Reinke	0.41	0.46	0.43
Saudável	0.51	0.52	0.52
Acurácia = 0.48			

Capítulo 7

Conclusões e trabalhos futuros

7.1 Conclusões

Na busca por avanços na identificação de patologias da voz, a utilização de métodos de extração de características e redes neurais tem se mostrado uma área promissora de pesquisa. Estes métodos são fundamentais para a análise e classificação de vozes, permitindo distinguir entre condições patológicas e vozes saudáveis com maior precisão. As redes neurais, em particular, oferecem uma abordagem sofisticada, capaz de aprender e adaptar-se a complexidades inerentes aos dados vocais. Esta capacidade as torna ferramentas valiosas na detecção e diagnóstico de patologias vocais, um campo onde a precisão é crucial tanto para o tratamento adequado quanto para a tranquilidade dos pacientes.

O método MFCC (Mel-Frequency Cepstral Coefficients) se destacou por sua alta eficiência na classificação de patologias vocais utilizando redes neurais. Em testes realizados, os modelos DNN, CNN, LSTM e BiLSTM apresentaram resultados notáveis. Eles conseguiram detectar todos os casos de cisto (100%) e mostraram uma taxa de acerto entre as redes neurais de 96% e 100% para diagnósticos de edema de Reinke, nódulo e paralisia. Em relação à detecção de casos saudáveis, os modelos conseguiram identificar corretamente 100% dos casos na maioria das condições. No entanto, ao comparar vozes com nódulos e vozes saudáveis, a eficiência oscilou entre 98% e 100%. Esses resultados indicam que o MFCC é uma ferramenta altamente confiável e precisa para o diagnóstico de patologias vocais, com uma notável ausência de falsos negativos e falsos positivos, o que reforça sua eficácia. Portanto, o MFCC se estabelece como uma opção promissora para aplicações práticas na área de saúde vocal, especialmente em condições ideais de análise.

Em contraste com o MFCC, o método PNCC (Power-Normalized Cepstral Coefficients) registrou um desempenho mais variado e moderado. Ao serem aplicados em modelos

de redes neurais, os resultados obtidos com o PNCC mostraram uma eficácia flutuante. A taxa de identificação de patologias vocais variou significativamente, alcançando um pico de 86% e um mínimo de 0%. Além disso, os modelos apresentaram uma alta incidência de falsos negativos e falsos positivos, com taxas variando de 12% a 100%. Esses resultados indicam que, apesar de sua capacidade de identificar uma proporção considerável de casos patológicos, o PNCC frequentemente falha ao classificar corretamente casos saudáveis, confundindo-os com patológicos. Essa tendência diminui a confiabilidade do PNCC como método de diagnóstico preciso de patologias vocais. Isso sugere a necessidade de melhorias no método ou a combinação com outras técnicas para aprimorar sua precisão e confiabilidade em diagnósticos vocais.

Finalmente, o método ZCPA (Zero-Crossing Rate of the Power-Amplified speech signal) mostrou-se o menos eficaz entre os três. Há a necessidade de maiores testes e ajustes, pois os modelos de redes neurais tiveram altos índices de falso negativos. Embora os modelos tenham sido eficazes em identificar casos saudáveis (100% de verdadeiro-negativo), a incapacidade de detectar condições patológicas compromete seriamente a utilidade do ZCPA para diagnósticos vocais.

É importante notar que, conforme a literatura já citada durante o trabalho, os métodos PNCC (Power-Normalized Cepstral Coefficients) e ZCPA (Zero-Crossing Rate of the Power-Amplified speech signal) são geralmente considerados mais robustos para a identificação de vozes em condições de ruído, superando frequentemente o desempenho do método MFCC (Mel-Frequency Cepstral Coefficients). No entanto, no contexto específico da nossa pesquisa e com base no banco de dados utilizado, conforme detalhado no capítulo sobre metodologia, observamos um desempenho superior do método MFCC em todos os casos testados. Esta constatação sugere que, embora os métodos PNCC e ZCPA possam ter vantagens teóricas em certas condições, o MFCC demonstrou ser mais eficaz nas condições específicas e nos parâmetros do nosso experimento.

7.2 Trabalhos futuros

Os resultados obtidos nesta pesquisa destacam a importância de refinar os métodos de extração de características vocais, particularmente o PNCC e o ZCPA, para melhorar a detecção de patologias e avaliar a saúde vocal. Adicionalmente, surge a perspectiva de explorar a combinação de diferentes modelos de redes neurais, com o intuito de alcançar um desempenho ainda mais elevado e consistente em uma variedade de testes.

Referências

[1] BEHLAU, M.; PONTES, P.; MORETI, F. *Higiene vocal: cuidando da voz*. Rio de Janeiro: Thieme Revinter Publicações LTDA, 2018.

[2] SILVA, S. S. L. da. Principais patologias laríngeas em professores. *Distúrbios da Comunicação*, v. 30, n. 4, p. 767–775, 2018.

[3] KAHNEMAN, D.; SIBONY, O.; SUNSTEIN, C. R. *Ruído: uma falha no julgamento humano*. Objetiva: 1ª edição, 2021.

[4] SILVA, D. G. d.; CUADROS, C. D. R.; ABRAHAM, A. Reconhecimento robusto de locutor baseado nos atributos zcpac. *in: Simpósio Brasileiro de Telecomunicações*, v. 3, 2007.

[5] SILVA, K. R. F. d. *Reconhecimento de locutor em ambientes ruidosos: uma comparação entre os métodos de extração de características MFCC e ZCPA, 2023. 98 f.* Niterói, RJ, 2023. Trabalho de Conclusão de Curso (Graduação) - Engenharia de Telecomunicações, Universidade Federal Fluminense.

[6] SILVA, V. G. R. d. *Análise do sinal de fala para reconhecimento de emoções utilizando representação semântica. 2022. 95 f.* São Cristóvão, SE, 2022. Dissertação (Mestrado em Engenharia Elétrica), Universidade Federal de Sergipe.

[7] SIQUEIRA, J. K. *Reconhecimento de Voz Contínua com Atributos MFCC, SSCH e PNCC, Wavelet Denoising e Redes Neurais. 2011. 85 f.* Rio de Janeiro, RJ, 2013. Dissertação (Mestrado em Engenharia Elétrica), Pontifícia Universidade Católica do Rio de Janeiro.

[8] SIQUEIRA, J. K.; ALCAIM, A. Comparação dos atributos mfcc, ssch e pncc para reconhecimento robusto de voz contínua. *Proc. XXIX Simpósio Brasileiro de Telecomunicações*, 2011.

[9] RIBEIRO, G.; GOMES, R.; COSTA, S.; COSTA, W. Análise mel-cepstral na discriminação de patologias laríngeas. In: *XXIV congresso brasileiro de engenharia biomédica*. Uberlândia-MG: CBEB, 2014.

[10] VIEIRA, V. J.; COSTA, S. C.; COSTA, W. C. d. A.; CORREIA, S. E.; ARAÚJO, J. M. F. R. de. Avaliaçao de desempenho na classificaçao de patologias larıngeas por análise lpc de sinais de voz e redes neurais mlp. In: *Anais do XIII Congresso Brasileiro de Inteligência Computacional*. Porto de Galinhas, PE: SBIC, 2013. p. 1–6.

[11] ALI, A.; GANAR, S. Intelligent pathological voice detection. *Int. J. Innov. Res. Technol*, v. 5, n. 5, p. 92–95, 2018.

[12] CORDEIRO, H.; FONSECA, J.; GUIMARÃES, I.; MENESES, C. Voice pathologies identification speech signals, features and classifiers evaluation. In: IEEE. *2015 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA)*. Poznan, Poland, 2015. p. 81–86.

[13] NAZARI, F.; YAN, W. Convolutional versus dense neural networks: Comparing the two neural networks performance in predicting building operational energy use based on the building shape. *arXiv preprint arXiv:2108.12929*, 2021.

[14] ALBAWI, S.; MOHAMMED, T. A.; AL-ZAWI, S. Understanding of a convolutional neural network. In: IEEE. *2017 international conference on engineering and technology (ICET)*. [S.l.], 2017. p. 1–6.

[15] LE, X.-H.; HO, H. V.; LEE, G.; JUNG, S. Application of long short-term memory (lstm) neural network for flood forecasting. *Water*, MDPI, v. 11, n. 7, p. 1387, 2019.

[16] FAN, B.; XIE, L.; YANG, S.; WANG, L.; SOONG, F. K. A deep bidirectional lstm approach for video-realistic talking head. *Multimedia Tools and Applications*, Springer, v. 75, p. 5287–5309, 2016.

[17] GOMES, D. T. *Redes Neurais Recorrentes Para Previsão de Séries Temporais de Memórias Curta e Longa, 2005. 153 f.* Campinas, SP, 2005. Dissertação (Mestrado), Instituto de Matemática, Estatística e Computação Científica. Universidade Estadual de Campinas.

[18] PLANETA Música. Disponível em: <https://blog.planetamusica.net/como-funciona-a-voz-conheca-o-instrumento-musical-do-nosso-corpo/>. Acesso em: 20, nov 2023.

[19] OLIVEIRA, I. e. S. G. d. *Vibração das pregas vocais em diferentes frequências: análise teórico-experimental. 2017. 118 f.* Belo Horizonte, MG, 2017.

[20] WALDOW, M. L. d. C. *Estratégias respiratórias e seus efeitos na qualidade da voz cantada: um estudo acústico e perceptivo. 2015. 111 f.* São Paulo, 2015. Dissertação (Mestrado em Linguistica Aplicada e Estudos da Linguagem), Pontifícia Universidade Católica de São Paulo, PUC-SP.

[21] Só Biologia. Disponível em: <https://www.sobiologia.com.br/conteudos/ FisiologiaAnimal/respiracao6.php>. Acesso em: 20, nov 2023.

[22] SITTA, F. M. E. *Fonoonline Blog*. Disponível em: <https://ericasitta.wordpress.com/2015/04/16/conheca-a-laringe/>. Acesso em: 20, nov 2023.

[23] BEHLAU, M. *O livro do especialista*. Rio de Janeiro: Revinter, 2004, V.1.

[24] MUNDO educação. Disponível em: <https://mundoeducacao.uol.com.br/biologia/faringe.htm/>. Acesso em: 20, nov 2023.

[25] FUKUYAMA, E. E. Análise acústica da voz captada na faringe próximo à fonte glótica através de microfone acoplado ao fibrolaringoscópio. *Revista Brasileira de Otorrinolaringologia*, SciELO Brasil, v. 67, p. 776–786, 2001.

[26] MUSIC Station. Disponível em: <https://mundoeducacao.uol.com.br/biologia/faringe.htm/>. Acesso em: 20, nov 2023.

[27] BRAGA, J. N.; OLIVEIRA, D. S. F. de; ATHERINO, C. C. T.; SCHOTT, T. C. A.; SILVA, J. C. Nódulos vocais: análise anátomo-funcional. *Revista CEFAC*, Instituto Cefac, São Paulo, v. 8, n. 2, p. 223–229, 2006.

[28] GARCIA, M. d. M.; MAGALHÃES, F. P.; DADALTO, G. B.; MOURA, M. V. T. d. Avaliação por imagem da paralisia de pregas vocais. *Radiologia Brasileira*, SciELO Brasil, Belo Horizonte, MG, v. 42, p. 321–326, 2009.

[29] APARECIDA, C.; FINGER, L. S.; ROSA, J. d. C.; BRANCALIONI, A. R. Lesões organofuncionais do tipo nódulos, pólipos e edema de reinke. *Revista CEFAC*, SciELO Brasil, Belo Horizonte, MG, v. 13, p. 735–748, 2011.

[30] NEGREIROS, B. C. P. *Cisto em prega vocal. 1997. 28.* São Paulo, 1997. Trabalho de Conclusão de Curso (Especialização) - Centro de Especialização em Fonoaudiologia Clínica-CEFAC.

[31] PNGWING. Disponível em: <https://www.pngwing.com/pt/free-png-tqhfx>. Acesso em: 20, nov 2023.

[32] AL-KALTAKCHI, M. T. S.; TAHA, H. A. A.-R.; SHEHAB, M. A.; ABDULLAH, M. A. M. Comparison of feature extraction and normalization methods for speaker recognition using grid-audiovisual database. *Indonesian Journal of Electrical Engineering and Computer Science*, v. 18, p. 782–789, 2020.

[33] KIM, C.; STERN, R. M. Power-normalized cepstral coefficients (pncc) for robust speech recognition. *Ieee transactions on audio, speech, and language processing*, v. 24, p. 1315–1329, 2016.

[34] ALMEIDA, C. R. *Extratores de características acústicas inspirados no sistema periférico auditivo. 2014. 56 f.* São Cristóvão, SE, 2014. Dissertação (Mestrado em Engenharia Elétrica), Universidade Federal de Sergipe.

[35] CUADROS, C. D. R. *Reconhecimento de voz e de locutor em ambientes ruidosos: comparação das técnicas MFCC e ZCPA, 2007. 121 f.* Niterói, RJ, 2007. Dissertação (Mestrado em Engenharia de Telecomunicações), Universidade Federal Fluminense.

[36] GORDILLO, C. D. A. *Reconhecimento de Voz Contínua Combinando os Atributos MFCC e PNCC com Métodos de Robustez SS, WD, MAP e FRN. 2013. 101 f.* Rio de Janeiro, RJ, 2013. Dissertação (Mestrado em Engenharia Elétrica), Pontifícia Universidade Católica do Rio de Janeiro.

[37] DUARTE, A. A. *Wavelets e redes neurais aplicadas à estimação de volume de tráfego de veículos , 2001. 79 f.* Salvador, BA, 2001. Dissertação (Mestrado em Engenharia Elétrica), Universidade Federal da Bahia.

[38] ACADEMY, K. *O neurônio, estrutura e função.* Disponível em: <https://pt.khanacademy.org/science/6-ano/vida-e-evoluo-os-sistemas-do-corpo-humano/os-neuronios/a/o-neuronio-estrutura-e-funcao>. Acesso em: 20, nov 2023.

[39] SOARES, P. L. B.; SILVA, J. P. da. Aplicação de redes neurais artificiais em conjunto com o método vetorial da propagação de feixes na análise de um acoplador direcional baseado em fibra Ótica. *Revista Brasileira de Computação Aplicada*, v. 3, n. 2, p. 58–72, 2011.

[40] BOCHIE, K.; GILBERT, M. da S.; GANTERT, L.; BARBOSA, M. d. S. M.; MEDEIROS, D. S. V. de; CAMPISTA, M. E. M. Aprendizado profundo em redes desafiadoras: Conceitos e aplicações. *Sociedade Brasileira de Computação*, 2020.

[41] FLECK, L.; TAVARES, M. H. F.; EYNG, E.; HELMANN, A. C.; ANDRADE, M. A. d. M. Redes neurais artificiais: princípios básicos. *Revista Eletrônica Científica Inovação e Tecnologia*, v. 1, n. 13, p. 47–57, 2016.

[42] NUNES, J. A. C. *Additive Margin Softmax e funções Sinc para Reconhecimento de Locutor, 2020. 17 f.* Recife, PE, 2020. Dissertação (Mestre em Ciência da Computação), Programa de Pó-graduação em Ciência da Computação. Universidade Federal de Pernambuco.

[43] RIZZO, I. V.; CANATO, R. L. C. Inteligência artificial: funções de ativação. *Revista Prospectus*, v. 2, n. 2, p. 51–65, 2020.

[44] WANG, S.; MANNING, C. Fast dropout training. In: PMLR. *international conference on machine learning*. Stanford University, 2013. p. 118–126.

[45] MQL5. *Redes Neurais de Maneira Fácil (Parte 12): Dropout*. Disponível em: <https://www.mql5.com/pt/articles/9112>. Acesso em: 20, nov 2023.

[46] DERTAT, A. *Applied Deep Learning - Part 1: Artificial Neural Networks*. Disponível em: <https://towardsdatascience.com/applied-deep-learning-part-1-artificial-neural-networks-d7834f67a4f6>. Acesso em: 20, nov 2023.

[47] WU, J. Introduction to convolutional neural networks. *National Key Lab for Novel Software Technology. Nanjing University. China*, v. 5, n. 23, p. 495, 2017.

[48] LEARNING, D. into D. *A Operação de Correlação Cruzada*. Disponível em: <https://pt.d2l.ai/chapter_convolutional-neural-networks/conv-layer.html>. Acesso em: 20, nov 2023.

[49] NÓBREGA, J. X. *Uso de rede neural convolucional na identificação de hipertensão arterial*. Campo Grande, MS, 2022. Dissertação (Mestrado em Computação Aplicada), Universidade Federal do Mato Grosso do Sul.

[50] STAUDEMEYER, R. C.; MORRIS, E. R. Understanding lstm. a tutorial into long short-term memory recurrent neural networks. *arXiv preprint arXiv:1909.09586*, 2019.

[51] OLAH, C. *Understanding LSTM Networks, 2015*. Disponível em: <https://colah.github.io/posts/2015-08-Understanding-LSTMs/>. Acesso em: 20, nov 2023.

[52] IHIANLE, I.; NWAJANA, A.; EBENUWA, S.; OTUKA, R.; OWA, K.; ORISATOKI, M. A deep learning approach for human activities recognition from multimodal sensing devices. *IEEE Access*, v. 8, p. 179028–179038, 10 2020.

[53] SAARBRUECKEN VOICE DATABASE. Disponível em: <https://stimmdb.coli.uni-saarland.de/index.php4>. Acesso em: 20, nov 2023.

[54] NUMPY. Disponível em <https://numpy.org/doc/stable/ reference/random/generated/numpy.random.randn.html/>. Acesso em: 20, nov 2023.

[55] NUMPY. Disponível em: <https://numpy.org/doc/stable /reference/generated/numpy.roll/>. Acesso em: 20, nov 2023.