Universidade Estadual Paulista
Instituto de Biociências, Letras e Ciências Exatas
Departamento de Ciência da Computação e Estatística

Douglas Eduardo da Silva

Impacto das Disfonias de Curta Duração na Autenticação Biométrica por Voz

São José do Rio Preto - SP
2024

Douglas Eduardo da Silva

Impacto das Disfonias de Curta Duração na Autenticação Biométrica por Voz

Monografia apresentada ao Programa de graduação em Ciência da Computação da UNESP para obtenção do título de Bacharel.

Orientador: Prof. Dr. Eng. Rodrigo Capobianco Guido

São José do Rio Preto - SP

2024

Douglas Eduardo da Silva

Impacto das Disfonias de Curta Duração na Autenticação Biométrica por Voz

Monografia apresentada ao Programa de graduação em Ciência da Computação da UNESP para obtenção do título de Bacharel.

Orientador: Prof. Dr. Eng. Rodrigo Capobianco Guido

Comissão Examinadora

Prof. Dr. Eng. Rodrigo Capobianco Guido
UNESP – Câmpus de São José do Rio Preto
Orientador

Prof. Dr. Lucas Correia Ribas
UNESP – Câmpus de São José do Rio Preto

Prof. Dr. Diego Renan Bruno
UNESP – Câmpus de São José do Rio Preto

São José do Rio Preto
2024

S586i

Silva, Douglas Eduardo da

Impacto das disfonias de curta duração na autenticação biométrica por voz / Douglas Eduardo da Silva. -- , 2024

51 p. : il., tabs.

Trabalho de conclusão de curso (-) - Universidade Estadual Paulista (UNESP), Instituto de Biociências Letras e Ciências Exatas, São José do Rio Preto,

Orientador: Rodrigo Capobianco Guido

1. Processamento de Sinais. 2. Biometria. 3. Distúrbios da voz. I. Título.

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

Dedico à minha família, que alicerçou minha jornada até aqui.

Agradecimentos

Agradeço à Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), que fomentou o estudo aqui relatado com concessão de bolsa de iniciação científica. Agradeço a Deus que, com a sua graça, me possibilitou realizar todas estas coisas. Agradeço à minha família que me deu todo o suporte e amparo necessários para seguir em frente. Agradeço aos amigos mais próximos que fiz durante a minha graduação, cuja a parceria foi essencial em todas as etapas. Agradeço ao meu orientador, Prof. Dr. Eng. Rodrigo, por todo o empenho em me ajudar a desenvolver este projeto, sem sua colaboração não seria possível fazê-lo.

"Comece fazendo o que é necessário, depois o que é possível, e de repente você estará fazendo o impossível."

São Francisco de Assis

Resumo

As alterações involuntárias e acusticamente perceptíveis na fala, conhecidas como disfonias, podem comprometer a precisão da identificação biométrica de indivíduos por meio da voz. Considerando a inovação do tema, este projeto teve como objetivo investigar as implicações das disfonias de curta duração no processo de Autenticação Biométrica por Voz (ABL), visando à adaptação de algoritmos que abordem essa questão. Com o apoio de uma base de dados de vozes obtida por meio de parcerias, foram desenvolvidos algoritmos para a extração de características representativas dos sinais de voz em dois momentos distintos: durante o período de alteração vocal e após a recuperação, utilizando tanto uma abordagem baseada na seleção manual de características quanto uma abordagem automática. Os algoritmos foram avaliados para validar sua eficiência na distinção das classes dos sinais. Os resultados indicaram que a extração automática de características, por meio de *autoencoder*, mostrou-se uma abordagem promissora para lidar com as disfonias de curta duração e melhorar o desempenho da ABL.

Palavras-chave: Processamento de Sinais. Verificação de locutores. Disfonias de curta duração.

Abstract

The involuntary and acoustically perceptible alterations in speech, known as dysphonias, can compromise the accuracy of biometric identification of individuals through voice. Considering the innovative nature of the subject, this project aimed to investigate the implications of short-term dysphonias in the Voice Biometric Authentication (VBA) process, focusing on the adaptation of algorithms to address this issue. Supported by a voice database obtained through partnerships, algorithms were developed to extract representative features from voice signals at two distinct moments: during the vocal alteration period and after recovery, using both a manually selected feature-based approach and an automatic approach. The algorithms were evaluated to validate their efficiency in distinguishing signal classes. The results indicated that automatic feature extraction, using an *autoencoder*, proved to be a promising approach to address short-term dysphonias and improve the performance of VBA.

Keywords: Signal processing. Speaker Verification. Short-term Dysphonias.

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas

ABLs Autenticação Biométrica de Locutores

ASV Verificação automática de Locutores

AUC Área Sob a Curva ROC

DsCD Disfonias de Curta Duração

EEG Sinais de eletroencefalia

MFCC *Mel Frequency Cepstral Coefficients*

PCA Análise de Componentes Principais

STFT *Short Time Fourier Transform*

SVM *Support Vector Machine*

WPT *Wavelet Pack Transform*

ZCR *Zero Crossing Rate*

Sumário

Capítulo 1

Introdução

1.1 Considerações iniciais

Os resultados obtidos em estudos prévios na área de processamento digital de sinais de voz, apoiados pela Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) e pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), desenvolvidos neste instituto, evidenciam a relevância das pesquisas relacionadas às ABLs. Publicações científicas, como os artigos de HOMAYOON (2011) e NEUSTEIN; PATIL (2012), reforçam essa importância. Além disso, essa realidade é demonstrada pelo número significativo de artigos científicos submetidos e publicados em periódicos de alto impacto, como o *IEEE Signal Processing Magazine*, *Elsevier Neurocomputing* e *Elsevier Computer Speech and Language*.

Observa-se também uma outra linha de pesquisa que tem capturado a atenção da comunidade científica internacional no contexto da análise de sinais de voz, conforme evidenciado por diversas publicações recentes, como as de GUPTA et al. (2021), CASTELLANA et al. (2018) e HEMMERLING; SKALSKI; GAJDA (2016): os algoritmos acústico-computacionais para análise de vozes disfônicas, que envolvem alterações acústicas perceptíveis na fala. Alguns desses estudos estão empregando técnicas avançadas de reconhecimento de padrões, como as redes neurais profundas (DENG; YU et al., 2014a; GOODFELLOW, 2016; DENG, 2018), exemplificados por referências como HARAR et al. (2017) e ZHANG et al. (2017), além da

lógica paraconsistente, como discutido por FONSECA et al. (2017).

1.2 Motivação

No contexto que abrange a interseção dessas subáreas, surge um desafio significativo que tem ganhado destaque apesar dos avanços modestos na sua resolução: a autenticação de locutores afetados por distúrbios na voz, que alteram em diferentes graus a qualidade acústica da fala, sejam eles de origem orgânica, funcional ou mista (HUCHE; ALLALI, 2005a, 2005b). Este cenário delimita um campo de investigação intrigante e carente de atenção: permitir que sistemas biométricos de autenticação de locutores em condições normais também sejam eficazes quando esses indivíduos estão temporariamente afetados por disfonias, como gripes, rouquidões e resfriados. Esse tema constitui o foco do projeto de Auxílio à Pesquisa Regular FAPESP intitulado "Aprimorando os Sistemas de Autenticação Biométrica por Voz: Robustez Mediante Disfonias de Curta Duração", sob a supervisão do orientador deste estudo, atualmente em desenvolvimento.

A produção limitada existente sobre ABLs afetados por DsCD, conforme indicado nos registros do *Web of Science*, envolve principalmente experimentos iniciais e considerações preliminares (TULL; RUTLEDGE, 1996b; TULL; RUTLEDGE; LARSON, 1996; TULL; RUTLEDGE, 1996a; PRAVENA; DHIVYA et al., 2012; LANSFORD; BERISHA; UTIANSKI, 2016; HUCKVALE; BEKE, 2017), não oferecendo soluções definitivas para o problema em questão. Além disso, conforme mencionado no artigo jornalístico referenciado (LINDSAY, 2014), os algoritmos para ABLs precisam ser capazes de identificar de maneira eficaz indivíduos afetados por disfonias, especialmente as mais leves e de curta duração, pois essas condições não alteram a estrutura básica do trato vocal do locutor utilizada para a caracterização vocal. Assim, este trabalho se caracteriza e se justifica como um suporte essencial ao projeto principal de Auxílio à Pesquisa mencionado.

1.3 Objetivos

O propósito central deste estudo, alinhado ao projeto de Auxílio à Pesquisa Regular FAPESP intitulado “Aprimorando os Sistemas de Autenticação Biométrica por Voz: Robustez Mediante Disfonias de Curta Duração”, englobou os seguintes objetivos específicos:

Realizar uma revisão bibliográfica detalhada sobre o estado da arte em ABLs, complementada por diálogos com profissionais da saúde para uma compreensão mais profunda das DsCD;

Estabelecer parcerias, visando consolidar uma base de dados abrangente de sinais vocais. Essa base de dados incluiu gravações de fala de indivíduos em dois contextos distintos: durante episódios de DsCD, como rouquidões e resfriados, e após a recuperação da saúde;

Extrair características distintivas dos sinais de voz de cada locutor em ambos os cenários, ou seja, durante e após a presença de DsCD, utilizando abordagens de extração manual de características e aprendizado de características por meio de *autoencoders*.

1.4 Organização do trabalho

O texto vindouro do presente trabalho está organizado da seguinte forma:

- No Capítulo 2 apresenta-se uma a fundamentação teórica dos principais conceitos empregados para o desenvolvimento deste projeto e uma série de trabalhados publicados envolvendo a área de reconhecimento de locutores, mostrando como são inúmeras as possibilidades de se realizar essa tarefa.

- No Capítulo 3 apresenta-se, com detalhes, as etapas realizadas para desenvolver o trabalho proposto e de que forma os conceitos discutidos no capítulo anterior foram utilizados.

- No Capítulo 4 apresenta-se os testes e os resultados obtidos com o trabalho desenvolvido.

- No Capítulo 5 apresenta-se a conclusão, discutindo os resultados obtidos e propostas para trabalhos futuros.

Capítulo 2

Revisão Bibliográfica

2.1 Definição dos Conceitos

2.1.1 Reconhecimento de locutores

Biometria de voz, também podendo ser referida como reconhecimento de voz, compreende o processo de identificação, verificação, classificação e, por extensão, segmentação, rastreamento e detecção de locutores. De forma geral, é um termo genérico usado para referenciar qualquer procedimento que envolva o conhecimento da identidade de uma pessoa baseado em sua voz.

Primeiramente, vale destacar que um sistema de reconhecimento de voz busca capturar e representar as características únicas do trato vocal de um indivíduo. Essa representação pode assumir duas abordagens principais: uma modelagem matemática, que representa o sistema fisiológico responsável pela produção da fala, ou uma abordagem estatística, que busca capturar padrões de saída semelhantes aos encontrados no trato vocal humano.

Uma vez estabelecido o modelo e sua associação com um indivíduo, o sistema torna-se capaz de avaliar novas amostras de voz, determinando a probabilidade de terem sido geradas por esse modelo específico em comparação com outros modelos disponíveis no sistema. Essa metodologia fundamental constitui a base de todas as aplicações de reconhecimento de voz.

(HOMAYOON, 2011).

2.1.2 Métodos de análise para identificação de padrões

A identificação de padrões em sinais, como os de voz, envolve a aplicação de técnicas analíticas que extraem características relevantes para distinguir ou correlacionar diferentes condições do emissor do sinal. Tais métodos baseiam-se, em grande parte, na análise de parâmetros associados à energia do sinal e suas propriedades espectrais. Por meio dessa análise, é possível captar informações essenciais sobre a consistência e a variabilidade da fonte geradora do sinal, permitindo a detecção de mudanças sutis em diferentes contextos ou estados.

Em análises de sinais de voz, técnicas específicas são utilizadas para examinar aspectos como a quantidade de energia acumulada em diferentes partes do sinal e a frequência com que a forma de onda cruza determinados limiares. Essas abordagens permitem explorar tanto a regularidade na produção do sinal quanto a variação de padrões ao longo do tempo. Além disso, os parâmetros críticos definidos durante essas análises servem como fundamentos para a extração de características, pois orientam a seleção de atributos que serão posteriormente utilizados para treinamento de modelos de aprendizado de máquina.

2.1.3 Extração de características

A extração de características é uma etapa essencial em sistemas de autenticação biométrica por voz. Esse processo visa transformar os sinais de voz brutos em um conjunto de parâmetros ou atributos que capturem as informações mais relevantes para a identificação ou verificação da identidade de um indivíduo.

Uma ampla busca na literatura faz crer que existem principalmente duas formas empregadas para aprendizado automático de características, conforme segue:

2.1.4 *Handcraft Extraction*

A extração artesanal de características consiste na situação em que o engenheiro do sistema escolhe as características apropriadas a serem extraídas do sinal sob análise (GUIDO, 2018). Desse modo, trata-se de um processo que busca identificar e selecionar manualmente as características relevantes entre os dados - neste caso, mais especificamente, no sinal - para uso em modelos de aprendizado de máquina ou análises estatísticas.

Uma aplicação comum desse processo é a Classificação por *Template*, onde as características extraídas manualmente são comparadas com modelos predefinidos para classificar novas amostras.

Classificação por *Template*

A classificação por *Template* é um método de classificação que utiliza modelos representativos, ou "*templates*", para comparar novos exemplos com as classes existentes. Nesse processo, mede-se a similaridade entre um dado de entrada e os *templates* de cada classe, e o novo exemplo é classificado com base no grau de correspondência com o *template* mais semelhante.

2.1.5 *Feature Learning*

Nos últimos anos, as técnicas derivadas da investigação em aprendizado de máquina têm exercido influência significativa sobre diversos campos de processamento de sinais e informações, abrangendo tanto os contextos convencionais quanto os emergentes, e englobando elementos essenciais do aprendizado de máquina e da inteligência artificial (DENG; YU et al., 2014b).

Aprendizado de Máquina

Nesse sentido, o aprendizado de máquina pode ser aplicado aos projetos de biometria de voz, colaborando com a identificação de características semelhantes em diferentes sinais, pois trata-se de uma área que utiliza algoritmos para adquirir representações dos dados, visando capturar relações complexas entre as informações (DENG; YU et al., 2014b).

Notavelmente, o aprendizado de características que se utiliza de recursos de aprendizado de máquina é denominado *feature learning*. Esse conceito se concentra na identificação automática e na representação das características mais relevantes e discriminativas dos dados. Ao utilizar esse método, objetiva-se permitir que algoritmos aprendam diretamente a partir dos dados, sem a necessidade de uma extração manual ou prévia das características. Em geral, tal forma de aprendizado de características baseia-se em uma estrutura de rede neural conhecida como *autoencoder* (DENG; YU et al., 2014b).

As características extraídas por meio de um *autoencoder* podem, então, ser utilizadas como insumos para o treinamento de modelos classificadores, com o objetivo de prever as classes dos dados de entrada de maneira mais precisa e eficiente.

2.1.6 Algoritmos Classificadores

Algoritmos classificadores são técnicas de aprendizado de máquina projetadas para prever a classe de um dado com base em seus atributos. Existem diversas abordagens para classificação, cada uma com fundamentos matemáticos e estatísticos distintos. O *Support Vector Machine* (SVM), por exemplo, busca encontrar o hiperplano ótimo que separa as classes, maximizando a margem entre elas e garantindo uma melhor generalização do modelo. Outro algoritmo amplamente utilizado é o *Random Forest*, que cria múltiplas árvores de decisão a partir de amostras aleatórias dos dados e combina suas previsões para melhorar a precisão e reduzir o risco de sobreajuste. Já a Regressão Logística é um classificador linear que utiliza uma função

para modelar a probabilidade de uma variável binária, sendo eficaz em problemas onde as classes são separáveis de forma linear, além de ser simples de entender e implementar.

2.1.7 *Data augmentation*

A técnica de *Data augmentation* é amplamente utilizada em aprendizado de máquina para aumentar a quantidade e a diversidade dos dados disponíveis, por meio da criação de novas amostras sintéticas a partir de dados originais. Essa abordagem é especialmente valiosa em tarefas de classificação, pois permite que os modelos sejam treinados com um conjunto de dados mais variado, melhorando sua capacidade de generalização e reduzindo o risco do chamado *overfitting* que é um ajuste excessivo aos dados de treinamento.

2.1.8 Curva ROC

A curva ROC (*Receiver Operating Characteristic*) é uma representação gráfica amplamente utilizada para avaliar o desempenho de modelos de classificação binária. Ela ilustra a relação entre a taxa de verdadeiros positivos (TPR, do inglês *True Positive Rate*) e a taxa de falsos positivos (FPR, *False Positive Rate*) para diferentes limiares de decisão. Esses limiares determinam a partir de qual valor o modelo classifica uma instância como pertencente à classe positiva ou negativa.

A curva ROC permite visualizar o equilíbrio entre a sensibilidade (capacidade de identificar corretamente os positivos) e a especificidade (capacidade de identificar corretamente os negativos) de um modelo. Um modelo ideal teria um TPR alto e um FPR baixo, resultando em uma curva próxima ao canto superior esquerdo do gráfico.

Além da curva ROC, a Área Sob a Curva (AUC, do inglês *Area Under the Curve*) é uma

métrica quantitativa derivada da curva ROC e serve como um indicador do desempenho geral do modelo. O valor da AUC varia entre 0 e 1, onde AUC = 1 indica um modelo com desempenho perfeito, e AUC = 0.5 representa um modelo com desempenho equivalente ao de uma classificação aleatória. Valores de AUC menores que 0.5 indicam que o modelo está classificando pior do que um palpite aleatório.

2.2 Estado da Arte

Na área de autentição biométrica por voz, vários estudos ainda estão sendo produzidos de forma a permitir e aprimorar mecanismos de reconhecimento de locutores. Alguns desses trabalhos motivam o desenvolvimento deste projeto.

BHATTACHARYYA et al. (2009), em uma abordagem mais voltada para a segurança da informação, examinaram várias abordagens de autenticação biométrica, com ênfase na biometria por voz. O estudo produzido examina as vantagens e desvantagens de diversas abordagens para aumentar a precisão e a segurança dos sistemas, incluindo reconhecimento baseado em características vocais distintas e técnicas de processamento de sinais. Além disso, são discutidas as tendências futuras e as maneiras pelas quais essas tecnologias podem ser melhoradas para enfrentar os desafios de segurança digital atuais.

MENG; ALTAF; JUANG (2020) apresentaram avanços significativos na autenticação ativa de voz, que consiste em uma abordagem inovadora para verificar continuamente a identidade de um locutor usando sinais de voz muito curtos. Para tal, utilizaram-se de técnicas como adaptação de modelo e treinamento de erro mínimo de verificação (MVE), o estudo alcançou taxas de erro igual em janelas de 3-4% em um conjunto de dados de 25 locutores, com cada decisão de autenticação baseada em apenas 1 segundo de voz. Esses resultados superam métodos tradicionais, representando um ganho absoluto de 3.88% em relação aos modelos i-vector em um conjunto de dados de referência, demonstrando a viabilidade e eficácia da autenticação contínua de voz em tempo real.

HUNDAL; HAMDE (2017) compararam várias técnicas de extração de características para sistemas de autenticação por voz, destacando suas vantagens e desvantagens. O estudo pôde afimar os seguintes tópicos: A técnica de MFCC replica o sistema auditivo humano ao converter frequências de Hz para escala Mel, embora amplifique ruídos devido à sua escala logarítmica. A STFT é rápida e oferece boa imunidade a ruídos, mas não fornece resolução tempo-frequência adequada. A Transformada Wavelet, com sua decomposição em múltiplos níveis, fornece uma análise detalhada do sinal, mas é computacionalmente mais complexa. A Codificação Linear Preditiva (LPC) é eficiente em termos de velocidade de computação, mas falha ao diferenciar palavras com sons de vogais semelhantes. Os resultados mostram que a precisão das técnicas varia de 83,75% a 87,5%, com diferentes tempos de processamento.

ALMAADEED; AGGOUN; AMIRA (2015), discutiram técnicas avançadas para identificação de falantes, destacando a combinação de múltiplas redes neurais com análise de wavelet. Essa abordagem inovadora supera métodos convencionais como Modelos de Mistura Gaussiana (GMM), Redes Neurais de Retropropagação (BPNN) e PCA tanto em termos de velocidade de identificação quanto de precisão. Testes realizados no banco de dados GRID mostram uma precisão de 97,5% com um tempo de identificação rápido de 50 ms usando WPT para extração de características. Essa técnica promete aplicabilidade direta em dispositivos industriais para segurança e autenticação, estabelecendo um novo padrão para sistemas de identificação de falantes em tempo real.

SHAYAMUNDA; RAMOTSOELA; HANCKE (2020), desenvolveram um sistema autônomo de reconhecimento de voz para autenticação de usuários em aplicações industriais, utilizando o Odroid XU4. Foi adotado um modelo dependente de texto, que exige menos amostras de treinamento e apresenta uma melhor taxa de reconhecimento. O sistema implementa várias técnicas para melhorar a precisão do reconhecimento, incluindo detecção de atividade vocal (VAD), remoção de silêncio e redução de ruído. Para a verificação de usuários, foram empregados modelos de mistura Gaussiana (GMM) com coeficientes MFCC extraídos das amostras de áudio. O sistema alcançou uma taxa de reconhecimento de 79% e um tempo médio de autenticação de 4,67 segundos, mostrando-se eficaz para uso em ambientes industriais, apesar de algumas limitações em condições de baixa razão sina-ruído.

LUPU; CIOBAN (2009), apresentaram uma solução inovadora e econômica para a autenticação de locutores e diagnóstico de formas de ondas biomédicas (EEG, ECG) utilizando a codificação TESPAR (Time Encoded Signal Processing and Recognition). Diferente das abordagens convencionais que empregam métodos complexos de processamento de sinais no domínio da frequência e demandam altos recursos computacionais, a técnica TESPAR é implementada em uma plataforma de microcontrolador ATMEL, oferecendo uma alternativa de baixa complexidade e custo reduzido. A metodologia TESPAR se destaca por necessitar de requisitos computacionais duas ordens de magnitude menores em comparação com outros métodos tradicionais, como espectrogramas, distorção temporal dinâmica, quantização vetorial, redes neurais e modelos ocultos de Markov. Os experimentos conduzidos no estudo demonstraram a eficácia do sistema, analisando variações nas matrizes TESPAR-S de acordo com diferentes modos de pronúncia e tipos de senha, alcançando melhores resultados em enunciações cooperativas com velocidade normal. Esta abordagem representa um avanço significativo na área de autenticação biométrica por voz, possibilitando implementações eficientes em plataformas de potência média como microcontroladores.

LAMEL; GAUVAIN (2000), abordaram avanços significativos no reconhecimento de locutor baseado em fala, explorando o impacto do conteúdo linguístico nos dados de treinamento e teste. Utilizando técnicas como treinamento multi-estilo e treinamento específico por tipo, o estudo investiga como diferentes tipos de material de fala, como sequências de dígitos, sentenças SEPT e sentenças Le Monde, influenciam o desempenho de identificação e verificação de locutores. Os resultados mostram que o treinamento específico por tipo tende a proporcionar desempenho superior quando o teste é realizado com o mesmo tipo de dados de treinamento, destacando a importância de um conteúdo linguístico correspondente entre treinamento e teste para otimizar a precisão do reconhecimento de locutor. Além disso, o estudo explora a dificuldade adicional do reconhecimento de locutor em condições de fala espontânea, revelando maiores taxas de erro devido à variação no estilo de fala e na diversidade de contextos fonéticos.

BRYDINSKYI et al. (2024), investigaram diversas técnicas avançadas de autenticação biométrica de voz para a verificação de falantes. Eles utilizaram PCA para a redução de dimensionalidade e a Análise Discriminante Linear (LDA) para maximizar a separabilidade entre

classes. Como característica fundamental para a representação da voz, os autores empregaram MFCC. Para aumentar a robustez do sistema, métodos de normalização de pontuação, como Z-norm e T-norm, foram aplicados. Na etapa de classificação, foram utilizadas SVM. Além disso, os autores integraram modelos de mistura Gaussiana (GMM) e i-vectors, criando um sistema híbrido que combina diferentes abordagens para alcançar alta precisão na verificação de falantes. Esses esforços resultaram em um sistema robusto e eficiente para autenticação biométrica de voz.

HAN et al. (2021), exploraram a verificação de locutor independente de texto utilizando EEG para melhorar a robustez do sistema em ambientes ruidosos. Utilizando um modelo de ponta baseado em aprendizado profundo, especificamente uma rede neural recorrente (RNN) com células LSTM ou GRU, o estudo combina características de EEG com MFCC para a verificação de locutor. A pesquisa destaca que características de EEG são menos suscetíveis ao ruído de fundo, aumentando assim a precisão da verificação de locutor em condições adversas. Para avaliação, foram construídos bancos de dados sincronizados de fala e EEG com diferentes níveis de ruído de fundo (40 dB e 65 dB), e os resultados demonstraram que a inclusão de características de EEG reduz a taxa de erro igual (EER), indicando uma melhoria significativa na robustez do sistema de verificação de locutor.

RAMOJI; KRISHNAN; GANAPATHY (2020), abordaram ASV, apresentando um modelo de ponta que integra tanto a extração de embeddings x-vector quanto o backend NPLDA (Discriminative Neural Probabilistic Linear Discriminant Analysis) em uma estrutura neural de ponta a ponta. Inicialmente, o sistema utiliza um modelo x-vector baseado em redes neurais convolutivas (TDNN) para extrair embeddings fixas de falas de duração variável. Em seguida, esses embeddings são processados por um modelo NPLDA que foi otimizado através de uma função de custo discriminativa, resultando em melhorias significativas sobre os modelos tradicionais de PLDA (Gaussian Probabilistic Linear Discriminant Analysis). O modelo proposto é treinado utilizando pares de enunciados de falantes com características MFCC, e a otimização conjunta de ambas as etapas permite um desempenho superior nas avaliações dos datasets NIST SRE 2018 e 2019, mostrando uma robustez considerável em condições de teste com diferentes durações de áudio e níveis de ruído.

Os experimentos e resultados apresentados nos estudos supracitados motivam e sustentam diversas técnicas empregadas neste estudo. Eles fornecem diretrizes sobre os métodos atualmente utilizados na área de processamento de sinais para autenticação biométrica por voz, destacando suas aplicações variadas e os resultados obtidos.

Capítulo 3

Detalhamento do Trabalho Proposto

3.1 Desenvolvimento

Neste capítulo, são explorados os aspectos metodológicos adotados na concepção e desenvolvimento do trabalho proposto. Após a explanação detalhada no capítulo anterior sobre os principais fundamentos técnicos pertinentes ao projeto e uma análise sucinta do panorama atual da área, o conteúdo subsequente pode ser abordado com maior ênfase. O capítulo está estruturado em seções que especificam cada fase do projeto realizado, oferecendo uma visão abrangente das abordagens e procedimentos adotados.

3.1.1 Base de dados

Com o auxílio de parcerias, foi possível construir uma base de dados composta por cerca de trinta gravações de voz, variando entre diferentes locutores em termos de idade, gênero, altura e peso corporal. A base foi organizada de forma que cada locutor tivesse sua voz gravada em dois estágios distintos. No primeiro estágio, a voz apresentava características disfônicas, normalmente causadas por patologias temporárias, como resfriados. No segundo estágio, a voz foi registrada

após a recuperação do locutor, sem sinais de disfonia.

Como o processo de coleta de dados não oferecia riscos aos voluntários, não houve necessidade de emissão do parecer do comitê de ética. A maioria dos áudios foi captada e enviada por meio do aplicativo de mensagens *WhatsApp*, sendo posteriormente convertida para o formato *WAV* para a análise.

No total, foram obtidos 30 arquivos de áudio, nos quais os locutores verbalizaram uma contagem de 1 a 10. Destes, 15 arquivos pertencem à classe "normal"(sem disfonia) e os outros 15 à classe "patológica"(com disfonia), com os tamanhos dos arquivos em formato *WAV* variando de 400 a 800 kB. Cada arquivo de áudio possui uma taxa de amostragem de 16.000 amostras por segundo. Os arquivos de áudio variam de 5 a 9 segundos, com uma média de 7,4 segundos e um desvio padrão de 1,356 segundos.

3.1.2 Pré-processamento

Como parte fundamental para o processamento posterior dos sinais presentes na base de dados, realizou-se um pré-processamento, visando adequar os dados de cada sinal para serem processados, conforme segue.

Primeiramente, dados os arquivos de entrada, digitalizados na taxa de 16000 amostras por segundo com quantização de 16 bits, e utilizando bibliotecas disponíveis para a linguagem de programação *Phyton*, realizou-se a extração de dados brutos dos arquivos de voz, com o armazenamento de tais dados de cada sinal em um respectivo vetor de valores inteiros.

Na sequência, normalizaram-se os sinais em termos de amplitude, sendo que, para isso, identificou-se a maior amplitude em módulo entre os quadros amostrados de cada sinal e dividiu-se todas as demais amplitudes por este valor.

Dado que muitos sinais adquiridos experimentalmente, como os de voz pertencentes ao experimento em questão, podem conter um componente de corrente contínua, conhecido como componente DC, que pode obscurecer a análise espectral e dificultar a detecção de padrões e

características importantes, procedeu-se à remoção desse componente nos sinais em questão. Em função da definição da Transformada Discreta de Fourier, comumente usada para análise espectral, a remoção do componente DC pode ser realizada simplesmente calculando-se a média das amplitudes do sinal em questão e subtraindo-se, posteriormente, esse valor de todas as amostras do mesmo.

A partir disso, aplicaram-se as técnicas para identificar padrões de interesse baseadas em energia e em taxa de cruzamento zero, utilizando-se dos métodos descritos a seguir.

3.1.3 Métodos de análise dos sinais para identificar padrões de interesse

Os métodos utilizados para identificar as caraterísticas comuns entre as vozes dos mesmos locutores em diferentes condições de saúde vocal baseiam-se, essencialmente, na análise dos níveis de energia presentes no sinal e nos componentes espectrais dos mesmos, sendo de autoria do orientador deste projeto de pesquisa, conforme registrado nas referências (GUIDO, 2016a, 2016b). Assim, implementaram-se os algoritmos seguintes para cada um dos sinais componentes da base:

- Método A_3: De acordo com GUIDO (2016a), este método consiste em determinar os comprimentos proporcionais do sinal sob análise que são necessários para atingir proporções pré-definidas de energia em relação à energia total nele contida. Assim, esse método é ideal para inspecionar a constância na ação da entidade física responsável pela geração do sinal, que no caso, é o conjunto vocal do locutor, formado pelos pulmões e tratos vocal e nasal.

  A variável *C* presente no algoritmo que implementa este método (GUIDO, 2016a, p. 268-269), é definida como um nível crítico de energia, ou seja, um valor percentual entre 0 e 100. Nesse sentido, para cada um dos sinais analisados, aplicou-se o algoritmo implementando este método em duas situações experimentais distintas, isto é, com *C* assumindo, respectivamente, os valores 1 e 10 com o objetivo de analisar o esforço de

sinal para aumentar a energia em 1 ou 10 por cento, respectivamente.

- Método B_3: Definido por GUIDO (2016b), este método é bastante similar ao A_3, entretanto consiste em determinar os comprimentos proporcionais do sinal sob análise que são necessários para alcançar porcentagens pré-definidas da ZCR, isto é, o número de vezes que a forma de onda do sinal cruza a amplitude zero. Assim, o método B_3 também se caracteriza como sendo ideal para inspecionar a constância na frequência da entidade física responsável por gerar o sinal. A variável *C* presente no algoritmo (GUIDO, 2016b, p. 254-255) é, desta vez, definida como sendo o nível crítico de ZCR, sendo um valor entre 0 e 100. Ou seja, este método analisa quanto de frequência do sinal é necessária para se alcançar C% do ZCR total. Assim, para cada um dos sinais analisados, aplicou-se o algoritmo implementando este método em duas situações distintas, com *C* assumindo os valores 1 e 10.

Combinando ambos os métodos, os quais encontram-se ilustrados nas Figuras 3.1 e 3.2, foram gerados vetores de características para cada sinal de modo que, para cada sinal foram gerados dois vetores configurando a variável C e concatenando vetores da seguinte forma:

- **Vetor 1**: Concatenação do vetor gerado pelo Método A_3 para C=1 com o vetor gerado pelo Método B_3 para C=10
- **Vetor 2**: Concatenação do vetor gerado pelo Método A_3 para C=10 com o vetor gerado pelo Método B_3 para C=1

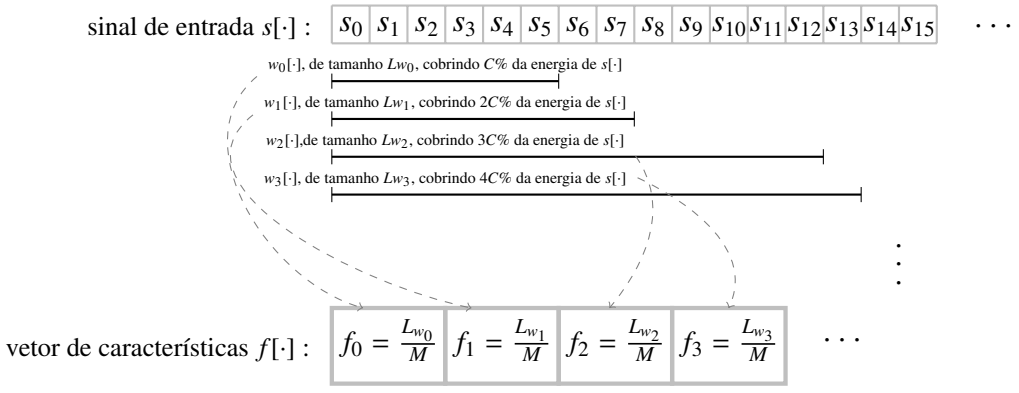


Figura 3.1 – Exemplo 1D para A_3, onde L_{w_i} representa o comprimento da janela $w_i[\cdot]$, para $i = 0, 1, 2, 3, ..., T-1$ (GUIDO, 2016a).

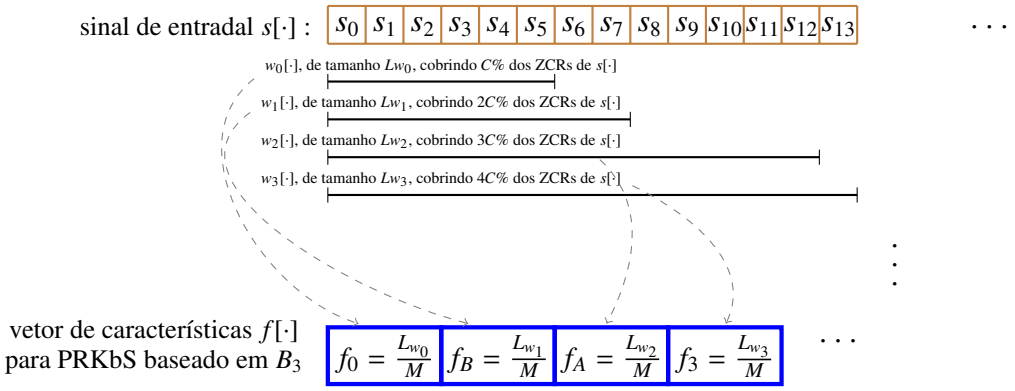


Figura 3.2 – Exemplo 1D para B_3, no qual L_{w_i} representa o comprimento da janela $w_i[\cdot]$, para $i = 0, 1, 2, 3, ..., T-1$ (GUIDO, 2016b).

3.1.4 Implementação de rotinas para extração de características

Classificação por *Template*

Inicialmente, adotando uma abordagem centrada na extração manual de características, definiu-se a distância entre vetores como uma característica significativa para a identificação do tipo de voz. Para isso, utilizou-se uma técnica de classificação por *template*, da qual extraiu-se métricas para validar essa definição.

Os vetores de características foram rotulados como normais ou patológicos, e cada uma dessas classes foi dividida em dois conjuntos: um para treino e outro para teste, cada um contendo 50% da classe total.

Para determinar se um vetor do conjunto de teste pertence à classe normal ou patológica, calcula-se a distância euclidiana entre esse vetor e todos os vetores de cada classe do conjunto de treino. Em seguida, identifica-se a distância mínima e classifica-se o vetor como pertencente à classe cuja distância mínima foi encontrada.

A figura 3.3 ilustra um exemplo de classificação por *template* baseada na distância de forma espacial.

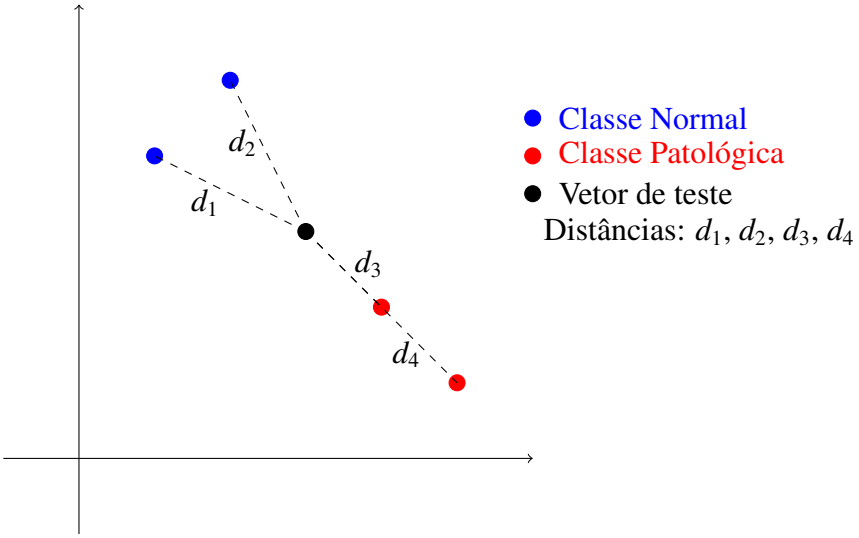


Figura 3.3 – Exemplo espacial de classificação por *template*

Data Augmentation

Em um segundo momento, utilizou-se um método de extração de características baseado no uso de *autoencoder*. Entretanto, esse método requer que haja um vasto conjunto de dados para treinamento para obter-se boa capacidade de generalização. Dessa forma, foi necessário aplicar o processo de *Data Augmentation* para gerar variações sintéticas dos vetores de características. Para tal, utilizou-se da geração de ruído gaussiano da seguinte forma:

1. Para cada amostra no conjunto original, foi gerado um ruído aleatório, seguindo uma distribuição normal (gaussiana) com média 0 e desvio padrão 0.1. Em outras palavras, para cada amostra original, são gerados n números aleatórios com média 0 e desvio padrão 0.1, onde n é o tamanho do vetor original de amostra.

2. O ruído gerado foi somado diretamente à amostra original, ou seja, foi aplicado a cada característica (*feature*) da amostra, de forma que o vetor resultante manteve a mesma dimensão da amostra original, mas com valores levemente modificados, criando-se, assim, uma nova amostra (um novo vetor) modificada.

3. Esse processo foi repetido 50 vezes para cada amostra, gerando diferentes variações da mesma.

Teoricamente, se tem que dada a amostra de dados $x \in \mathbb{R}^n$, com n características, para gerar uma nova amostra aumentada $\tilde{x}$, foi adicionado um ruído gaussiano ϵ em cada característica da amostra original, onde $\epsilon \sim \mathcal{N}(0, 0.1)$. A nova amostra $\tilde{x}$ é então dada por:

$$\tilde{x} = x + \epsilon$$

A figura 3.4 exemplifica de forma visual o resultado desse procedimento do ponto de vista espacial.

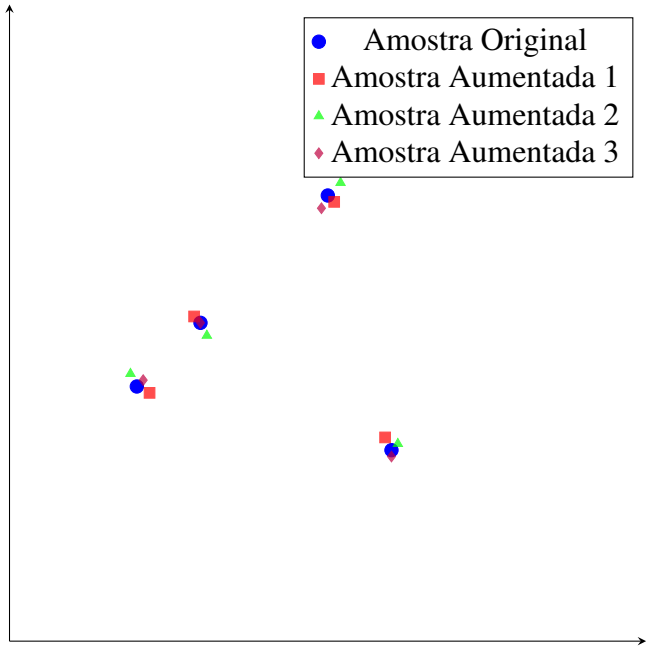


Figura 3.4 – Exemplo de *data augmentation* no conjunto de características

Feature Learning

Em seguida, foram empregadas técnicas de extração de características utilizando um *auto-encoder*, que recebe um conjunto de dados e gera uma representação compacta destes dados, buscando, a partir dessa representação, retornar ao conjunto original, como ilustrado na Figura 3.5. Nesse processo, identificam-se as características mais relevantes para a representação dos dados. A implementação do autoencoder foi realizada conforme descrito a seguir.

Construiu-se um *autoencoder* com duas camadas principais: uma camada de codificação (encoder) e uma camada de decodificação (decoder). A entrada do modelo consiste em vetores

de tamanhos variáveis, com o número de neurônios na camada de entrada igual ao número de características de cada vetor. A camada de codificação tem 12 neurônios, definidos com base na quantidade de dados disponíveis no conjunto de treinamento, e utiliza a função de ativação ReLU (*Rectified Linear Unit*). A função ReLU foi escolhida por sua capacidade de introduzir não linearidade, permitindo ao modelo aprender representações complexas e não-lineares dos dados.

A camada de decodificação possui um número de neurônios igual ao da camada de entrada, ou seja, o mesmo número de características presentes nos dados originais, com a função de ativação *sigmoide*. A função sigmoide foi escolhida para a camada de saída para mapear as saídas entre 0 e 1, o que é adequado para dados normalizados, como no caso deste modelo.

Nesse sentido, a arquitetura do autoencoder pode ser descrita da seguinte forma:

- **Camada de Entrada**: Número de neurônios igual ao número de características dos dados de entrada.
- **Camada de Codificação**: 12 neurônios, com a função de ativação ReLU.
- **Camada de Decodificação**: Número de neurônios igual ao número de características de entrada, com a função de ativação *sigmoide*.

O *autoencoder* foi compilado utilizando o otimizador Adam, que é eficiente para otimização em redes neurais profundas, ajustando automaticamente as taxas de aprendizado durante o treinamento. A função de perda escolhida foi o erro médio quadrático (MSE), que mede a diferença entre as saídas previstas e os valores reais, sendo comumente usada em modelos de reconstrução de dados.

O treinamento do modelo foi realizado por 5 épocas, onde o número de passagens completas pelo conjunto de dados de treinamento foi definido como 5. O tamanho de lote (`batch_size`) foi estabelecido como 1, o que significa que o modelo atualiza seus parâmetros após processar cada amostra individualmente. Essa configuração foi escolhida para permitir que o modelo se ajuste rapidamente a novas informações, dado que as atualizações dos parâmetros ocorrem com maior frequência.

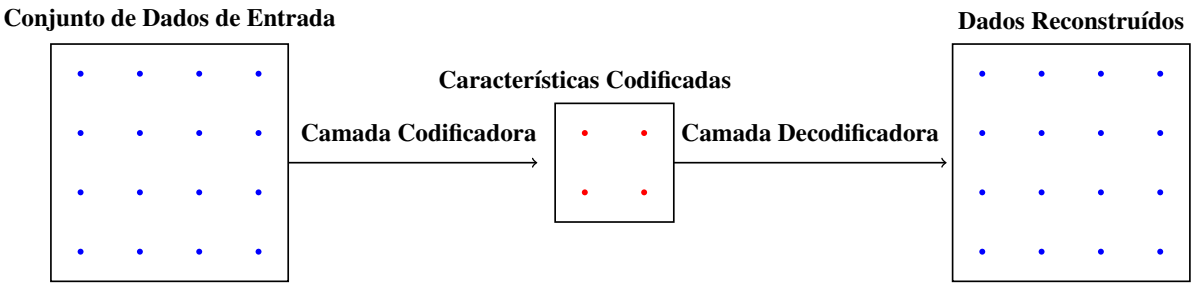


Figura 3.5 – Funcionamento de um *autoencoder*

Classificadores

As características codificadas geradas pelo *autoencoder* foram utilizadas como entrada para os modelos de *Support Vector Machine* (SVM), *Random Forest* e Regressão Logística, a fim de coletar métricas que possam avaliar o desempenho de seu uso na tarefa de distinguir as classes.

O modelo SVM foi implementado utilizando a classe SVC da biblioteca *sklearn* oferecida para a linguagem Python, com o parâmetro *probability=True* ativado. Essa configuração permite que o modelo retorne as probabilidades das classes durante a predição, o que é crucial para a avaliação do seu desempenho. O *kernel* utilizado para o SVM é linear na camada de saída, o que implica que a separação das classes é realizada por um hiperplano linear, maximizando a margem entre as classes. O treinamento do SVM é realizado utilizando o método *fit*, onde o modelo é ajustado às características codificadas do conjunto de treinamento (*X_train_balanced*) e suas respectivas classes (*y_train_balanced*). Durante esse processo, a função de custo do SVM é minimizada, buscando o hiperplano que maximiza a margem entre as classes, conforme ilustrado na Figura 3.6.

O modelo *Random Forest* foi implementado utilizando a classe *RandomForestClassifier* da biblioteca *sklearn*. Cada árvore de decisão foi treinada utilizando uma amostra aleatória dos dados de treinamento, e em cada divisão da árvore é considerada uma seleção aleatória das características. Esse processo, denominado *bagging* (*Bootstrap Aggregating*), visa reduzir a variância e melhorar a robustez do modelo. No caso do *Random Forest* adotado, o número de árvores é configurado com *n_estimators=100*, ou seja, o modelo usou 100 árvores de decisão para tomar a decisão final.

O treinamento do *Random Forest* foi realizado utilizando o método *fit*, no qual o modelo é ajustado aos dados de treinamento balanceados (X_train_balanced) e suas respectivas classes

(*y_train_balanced*). O treinamento busca otimizar o processo de criação das árvores de decisão, garantindo que cada árvore faça boas divisões para distinguir as classes. Após o treinamento, o modelo realiza a predição sobre o conjunto de teste (*X_encoded_test*) utilizando o método *predict*.

O modelo Regressão Logística foi implementado utilizando a classe *LogisticRegression* da biblioteca *sklearn*. O modelo utilizou a *função de custo log-verossimilhança*, que é minimizada durante o treinamento, ajustando os coeficientes da combinação linear das características de entrada para otimizar as predições. O treinamento foi realizado utilizando o método *fit*, no qual o modelo é ajustado aos dados de treinamento balanceados (*X_train_balanced*) e suas respectivas classes (*y_train_balanced*). O parâmetro *max_iter=1000* garante que o modelo tenha iterações suficientes para convergir para uma solução ótima, caso necessário.

Após o treinamento, o modelo realiza predições sobre o conjunto de teste (*X_encoded_test*) utilizando o método *predict*, fornecendo a classe prevista para cada instância. Para a avaliação do desempenho, também retorna as probabilidades associadas a cada classe por meio do método *predict_proba*, que fornece a confiança do modelo em suas predições.

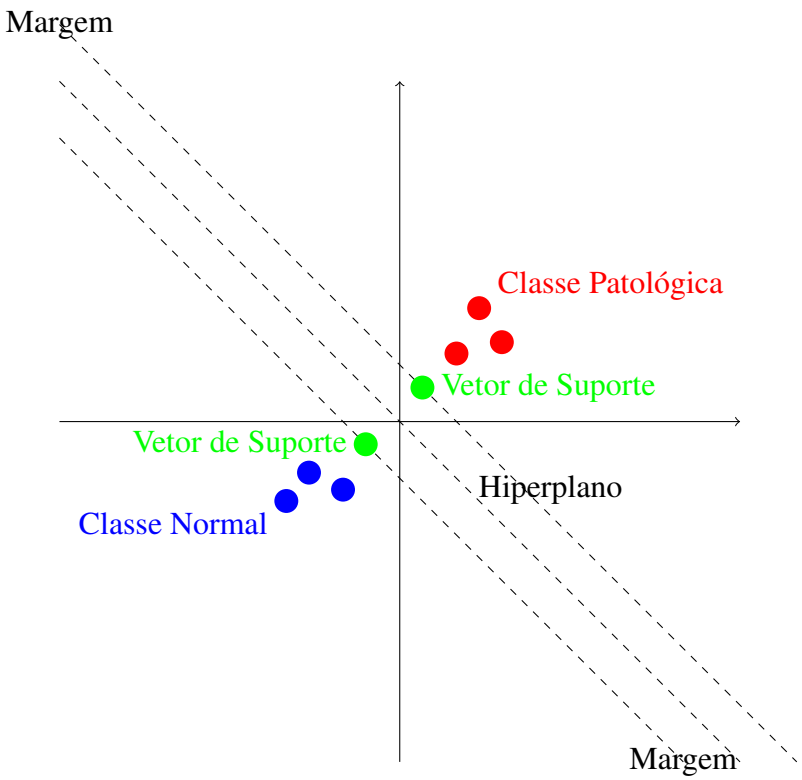


Figura 3.6 – Representação do modelo SVM.

De modo geral, o processo de *Feature Learning* adotado pode ser resumido pelo fluxograma contido na figura 3.7.

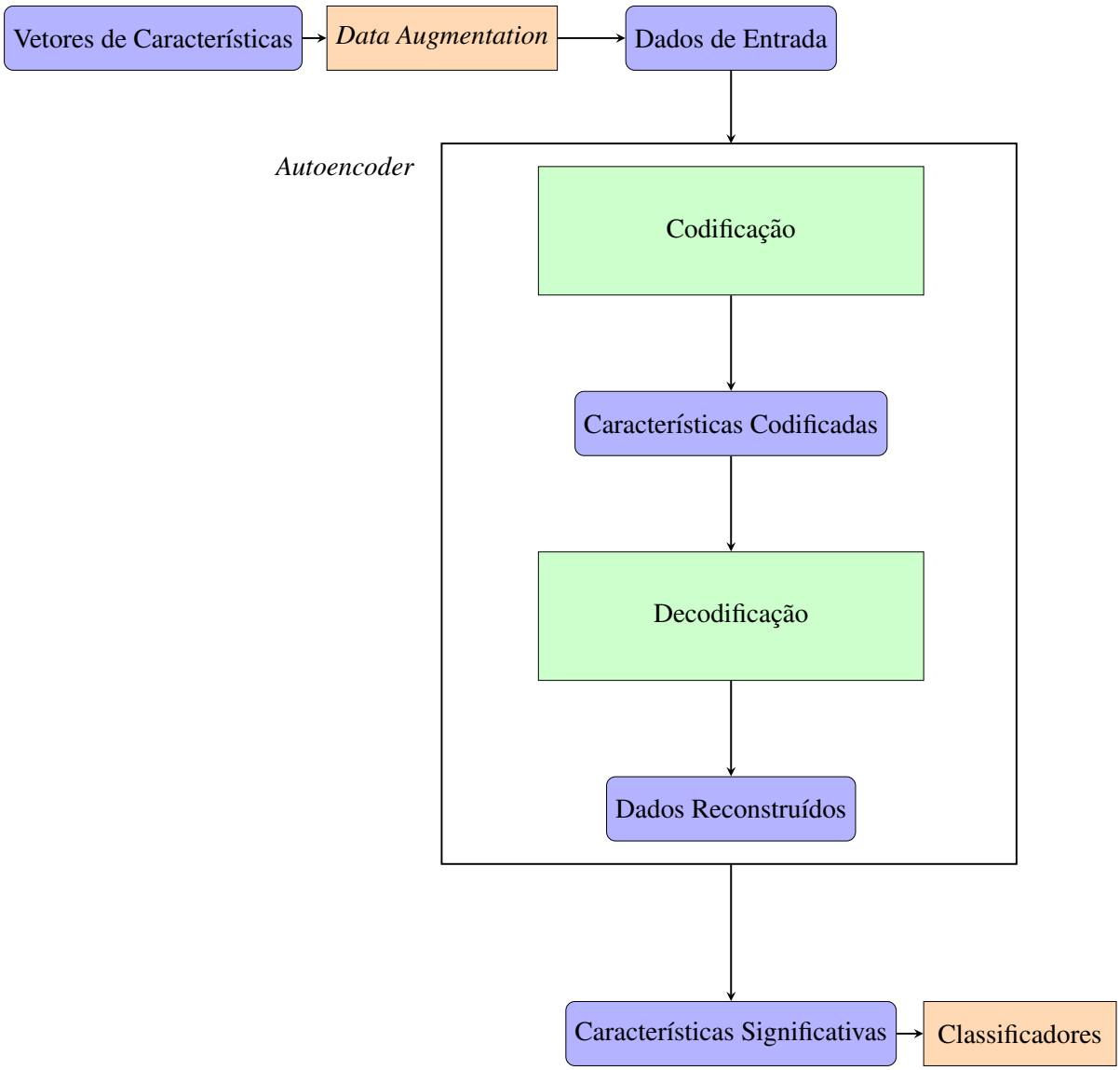


Figura 3.7 – Fluxograma do Processo de *Feature Learning*.

Métricas

Para avaliar e comparar a classificação por *Template* com *Feature Learning*, bem como identificar se são eficientes no processo de extração de características foram calculas duas métricas principais, conforme descrito a seguir.

Acurácia

A acurácia é uma métrica de classificação que mede a proporção de previsões corretas em

relação ao total de previsões feitas. Ela foi calculada seguindo a fórmula:

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

Curva ROC

Como mencionado no Capítulo 2, a Curva ROC (*Receiver Operating Characteristic*) mostra a relação entre a taxa de verdadeiros positivos (sensibilidade) e a taxa de falsos positivos para diferentes limiares de classificação. Sendo assim, para avaliar os modelos, a curva ROC foi esboçada seguindo estas fórmulas:

- Eixo Y: Taxa de Verdadeiros Positivos (TVP) ou Sensibilidade: a fração de exemplos positivos corretamente classificados.

$$\text{Sensibilidade} = \frac{\text{VP}}{\text{VP} + \text{FN}}$$

- Eixo X: Taxa de Falsos Positivos (TFP): a fração de exemplos negativos classificados incorretamente como positivos.

$$\text{TFP} = \frac{\text{FP}}{\text{FP} + \text{VN}}$$

Graficamente, a curva ROC é exemplificada pela Figura 3.8

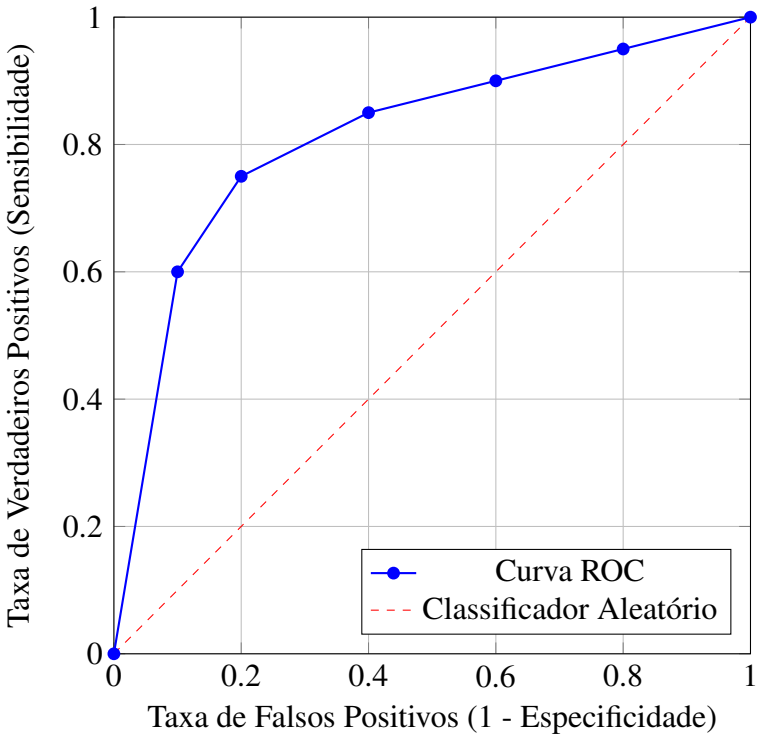


Figura 3.8 – Exemplo de Curva ROC

Área sob a Curva ROC

A Área sob a Curva ROC (AUC) foi a principal métrica utilizada para analisar a eficácia dos métodos testados neste estudo, bem como compará-los entre si. Para seu cálculo, utiliza-se a integral da curva ROC, então, pode-se defini-la da seguinte forma:

$$\text{AUC} = \int_0^1 \text{Sensibilidade}(t)\, d(1 - \text{Especificidade}(t))$$

Capítulo 4

Testes e Resultados

Este capítulo apresenta os resultados obtidos com a implementação dos métodos proposto neste trabalho. Além disso, são detalhadas as especificações da plataforma de testes e discutidos os resultados obtidos.

4.1 Plataforma de Desenvolvimento e Testes

Todos os algoritmos foram desenvolvidos na linguagem Python e testados no ambiente de desenvolvimento Google Colab. Este ambiente, amplamente utilizado em projetos de aprendizado de máquina e ciência de dados, oferece um serviço de Jupyter Notebook hospedado, fornecendo acesso a recursos como CPU e memória virtualizados.

4.2 Classificação por *Template*

O algoritmo foi executado em três lotes distintos de 50, 100 e 200 iterações. Em cada lote, a acurácia e a AUC (Área Sob a Curva ROC), foram calculadas, fornecendo uma visão

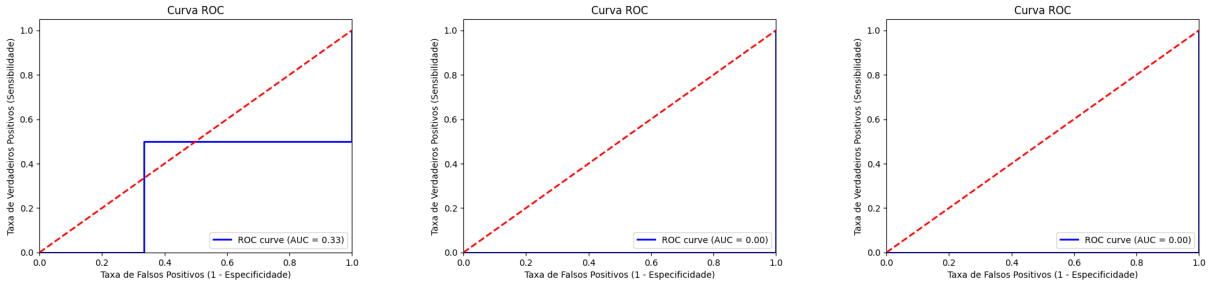
desempenho do modelo em diferentes cenários. A curva ROC média foi traçada, o que permite avaliar a capacidade de separação das classes em cada lote de iterações.

Os resultados médios obtidos para as diferentes métricas ao fim de cada lote das iterações estão apresentados na Tabela 4.1.

Número de Iterações	Acurácia (%)	AUC
50	46	0.03
100	40	0.05
200	41	0.05

Tabela 4.1 – Resultados de Acurácia e AUC em Classificação por *Template*

As Figuras 4.1(a), 4.1(b) e 4.1(c) apresentam as curvas ROC correspondentes ao desempenho do modelo após 50, 100 e 200 iterações, respectivamente.



(a) Curva ROC média após 50 iterações

(b) Curva ROC média após 100 iterações

(c) Curva ROC média após 200 iterações

Figura 4.1 – Curva ROC do método de Classificação por *Template*

Observou-se que o classificador em análise apresentou um desempenho insatisfatório na tarefa de distinguir corretamente entre as classes (normal e patológica) dos vetores de características. Esse desempenho insatisfatório é claramente evidenciado pela AUC, que se manteve extremamente próxima de 0.0 em todos os cenários testados, independentemente do número de iterações. Tal valor indica que o modelo foi incapaz de separar adequadamente as classes, agindo de maneira praticamente aleatória. Além disso, a acurácia, consistentemente abaixo de 50%, confirma a predominância de erros, com o modelo classificando erroneamente a maioria dos exemplos. Esses resultados sugerem que a abordagem adotada, que faz uso da distância euclidiana como principal característica para a diferenciação entre os sinais, não foi capaz de capturar de forma eficiente as particularidades dos dados, resultando em uma classificação inadequada.

4.3 *Feature Learning*

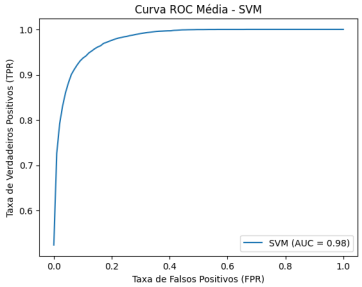
Assim como no modelo de classificação por *template*, o processo de *Feature Learning* foi realizado em blocos de iterações de 50, 100 e 200. Em cada um desses blocos, os dados foram aleatoriamente aumentados, e um *autoencoder* foi treinado para extrair características significativas, que foram então utilizadas para treinar os modelos classificadores. Ao final de cada bloco, foram calculadas a Acurácia e a AUC, permitindo a avaliação da eficácia dos modelos em distinguir as classes com base na saída do *autoencoder*.

Os resultados médios obtidos para as diferentes métricas ao longo das iterações estão apresentados na Tabela 4.2.

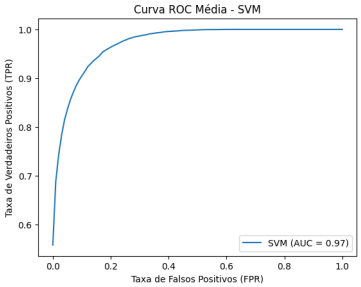
Número de Iterações	**Acurácia (%)**	**AUC**
50	SVM: 92 Random Forest: 88 Logistic Regression: 80	SVM: 0.98 Random Forest: 0.96 Logistic Regression: 0.89
100	SVM: 90 Random Forest: 88 Logistic Regression: 80	SVM: 0.97 Random Forest: 0.95 Logistic Regression: 0.89
200	SVM: 90 Random Forest: 87 Logistic Regression: 80	SVM: 0.95 Random Forest: 0.95 Logistic Regression: 0.89

Tabela 4.2 – Resultados de Acurácia e AUC em *Feature Learning*

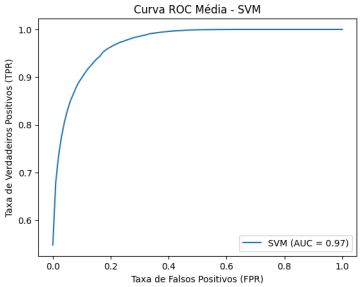
A Figura 4.2 ilustra as curvas ROC que refletem o desempenho de cada modelo após a realização de 50, 100 e 200 iterações, respectivamente.



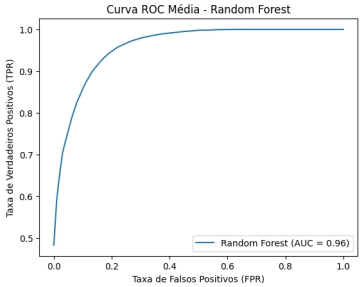
(a) Curva ROC média após 50 iterações para o modelo SVM



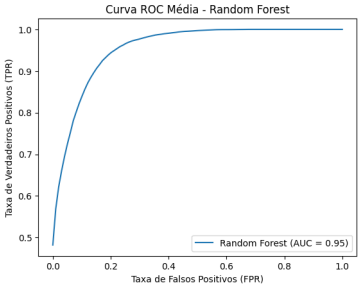
(b) Curva ROC média após 100 iterações para o modelo SVM



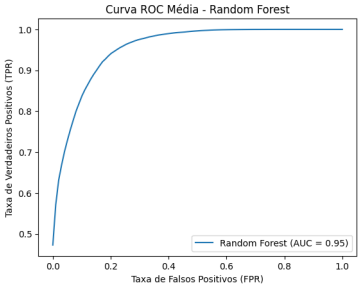
(c) Curva ROC média após 200 iterações para o modelo SVM



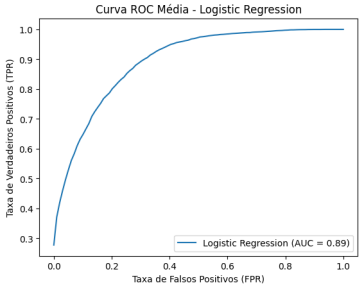
(d) Curva ROC média após 50 iterações para o modelo *Random Forest*



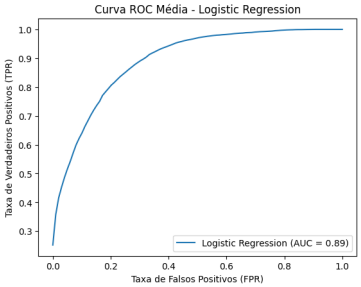
(e) Curva ROC média após 100 iterações para o modelo *Random Forest*



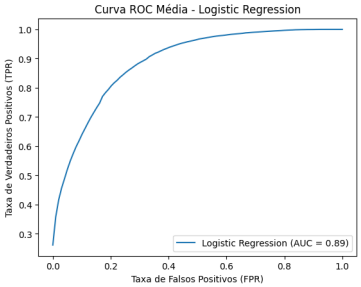
(f) Curva ROC média após 200 iterações para o modelo *Random Forest*



(g) Curva ROC média após 50 iterações para o modelo *Logistic Regression*



(h) Curva ROC média após 100 iterações para o modelo *Logistic Regression*



(i) Curva ROC média após 200 iterações para o modelo *Logistic Regression*

Figura 4.2 – Curvas ROC do método de *Feature Learning* para diferentes iterações e condições.

Observou-se que a extração de características por meio do *autoencoder* foi capaz de identificar de forma eficiente as informações relevantes dos vetores de características dos sinais de voz. Isso ficou evidente pelo desempenho consistente e elevado dos modelos de classificação, que foram treinado com essas características extraídas. A alta AUC média, independentemente da quantidade de iterações e do modelo classificador adotado, reforça a capacidade do modelo em distinguir corretamente entre as classes, indicando uma boa separação entre os dados normais e patológicos. Além disso, a acurácia, que se manteve muito próxima ou acima de 90% nos

cenários analisados, demonstra uma predominância de acertos nas predições dos modelos. Esses resultados apontam que o uso do *Feature Learning* com *autoencoder* foi eficiente na extração de características, capturando as particularidades dos dados e resultando em uma classificação adequada e de bom desempenho.

Capítulo 5

Conclusões

Os estudos realizados neste trabalho têm como objetivo a extração de características distintivas dos sinais de voz de cada locutor antes e após a presença de DsCD, avaliando duas abordagens: uma de extração manual de características, utilizando a distância euclidiana, e outra de aprendizado automático de características por meio de *autoencoders*. A importância dessas abordagens se deve à necessidade de algoritmos em sistemas de Autenticação Biométrica de Locutores (ABLs) serem capazes de identificar com eficácia indivíduos afetados por disfonias, especialmente as mais leves e de curta duração, já que essas condições não alteram significativamente a estrutura do trato vocal, utilizada na caracterização vocal.

Diante disso, comparou-se o desempenho de ambos os métodos: a classificação baseada na distância euclidiana e o aprendizado com *autoencoders*. O primeiro método demonstrou-se ineficiente, incapaz de distinguir corretamente entre as classes normais e patológicas. Por outro lado, a abordagem baseada em *autoencoders* mostrou-se bastante eficiente, apresentando alta taxa de acertos na classificação. Esses resultados sugerem que o *Feature Learning* via *autoencoders* será uma ferramenta promissora no processo de autenticação biométrica de locutores afetados por DsCD, contribuindo para uma identificação mais precisa em casos de disfonia.

Diversos fatores podem ter influenciado os resultados obtidos neste estudo, sendo o principal deles a limitação do tamanho da base de amostras. Embora tenham sido feitos esforços para torná-la o mais diversificada possível, a base ainda carecia de maior amplitude, o que pode ter impactado a representatividade dos dados e, consequentemente, o desempenho dos modelos

testados. Um aspecto fora do convencional foi a aplicação de técnicas de *Data Augmentation* diretamente nos vetores de características, em vez de nos sinais de áudio, como é mais comum. Embora essa abordagem ousada tenha fugido do padrão, ela contribuiu de forma significativa para melhorar os resultados dos treinamentos, gerando métricas positivas. No entanto, o ideal seria contar com uma base de dados maior e mais variada de forma natural, sem depender de vetores sintéticos. Para trabalhos futuros, recomenda-se a utilização de uma base de sinais ampliada e diversificada, o que aumentaria a robustez das análises e resultaria em classificações mais generalizáveis e consistentes.

5.0.1 Nota Pessoal

De maneira geral, o projeto teve um impacto significativo no meu desenvolvimento pessoal e no meu amadurecimento. Ele me proporcionou a oportunidade de aprofundar meus conhecimentos em áreas específicas do processamento de sinais, o que, sem dúvida, resultou em um grande enriquecimento do meu aprendizado. As principais dificuldades surgiram na fase de coleta de dados, especialmente em função da necessidade de capturar a voz de um mesmo indivíduo em diferentes estágios vocais, desde o estado rouco até o retorno à normalidade. Esse processo, além de demorado, foi desafiador devido à dificuldade de encontrar voluntários dispostos a seguir os critérios necessários para a pesquisa. Esses fatores tornaram a criação de uma base de dados robusta mais trabalhosa e lenta.

Apesar desses obstáculos, acredito que os resultados obtidos com este trabalho irão contribuir para o projeto maior vinculado ao Auxílio à Pesquisa Regular FAPESP, intitulado “Aprimorando os Sistemas de Autenticação Biométrica por Voz: Robustez Mediante Disfonias de Curta Duração”, que era seu principal objetivo. Além disso, espero que este estudo tenha agregado, ainda que modestamente, à área de pesquisa, inspirando novos estudos e avanços. Pretendo, num futuro próximo, dar continuidade ao que foi iniciado nesse projeto, em um programa de pós-graduação.

Referências Bibliográficas

ALMAADEED, N.; AGGOUN, A.; AMIRA, A. Speaker identification using multimodal neural networks and wavelet analysis. *Iet Biometrics*, Wiley Online Library, v. 4, n. 1, p. 18–28, 2015.

BHATTACHARYYA, D. et al. Biometric authentication techniques and its future possibilities. In: IEEE. *2009 Second International Conference on Computer and Electrical Engineering*. [S.l.], 2009. v. 2, p. 652–655.

BRYDINSKYI, V. et al. Comparison of modern deep learning models for speaker verification. *Applied Sciences*, MDPI, v. 14, n. 4, p. 1329, 2024.

CASTELLANA, A. et al. Discriminating pathological voice from healthy voice using cepstral peak prominence smoothed distribution in sustained vowel. *IEEE Transactions on Instrumentation and Measurement*, IEEE, v. 67, n. 3, p. 646–654, 2018.

DENG, L. *Deep learning in natural language processing*. [S.l.]: Springer, 2018.

DENG, L.; YU, D. et al. Deep learning: methods and applications. *Foundations and trends® in signal processing*, Now Publishers, Inc., v. 7, n. 3–4, p. 197–387, 2014.

DENG, L.; YU, D. et al. Deep learning: methods and applications. *Foundations and trends® in signal processing*, Now Publishers, Inc., v. 7, n. 3–4, p. 197–387, 2014.

FONSECA, E. S. et al. Linear prediction and discrete wavelet transform to identify pathology in voice signals. In: IEEE. *2017 Signal Processing Symposium (SPSympo)*. [S.l.], 2017. p. 1–4.

GOODFELLOW, I. *Deep learning*. [S.l.]: MIT press, 2016.

GUIDO, R. C. A tutorial on signal energy and its applications. *Neurocomputing*, Elsevier, v. 179, p. 264–282, 2016.

GUIDO, R. C. Zcr-aided neurocomputing: a study with applications. *Knowledge-based Systems*, Elsevier, v. 105, p. 248–269, 2016.

GUIDO, R. C. A tutorial review on entropy-based handcrafted feature extraction for information fusion. *Information Fusion*, Elsevier, v. 41, p. 161–175, 2018.

GUPTA, S. et al. Residual neural network precisely quantifies dysarthria severity-level based on short-duration speech segments. *Neural Networks*, Elsevier, v. 139, p. 105–117, 2021.

HAN, Y. et al. Robust end-to-end speaker verification using eeg. In: IEEE. *2020 28th European Signal Processing Conference (EUSIPCO)*. [S.l.], 2021. p. 1170–1174.

HARAR, P. et al. Voice pathology detection using deep learning: a preliminary study. In: IEEE. *2017 international conference and workshop on bioinspired intelligence (IWOBI)*. [S.l.], 2017. p. 1–4.

HEMMERLING, D.; SKALSKI, A.; GAJDA, J. Voice data mining for laryngeal pathology assessment. *Computers in biology and medicine*, Elsevier, v. 69, p. 270–276, 2016.

HOMAYOON, B. Fundamentals of speaker recognition. *Springer Science Business Media*, 2011.

HUCHE, F. L.; ALLALI, A. A voz: patologia vocal de origem funcional. In: *A voz: patologia vocal de origem funcional*. [S.l.: s.n.], 2005. p. 187–187.

HUCHE, F. L.; ALLALI, A. A voz: patologia vocal de origem funcional. In: *A voz: patologia vocal de origem funcional*. [S.l.: s.n.], 2005. p. 187–187.

HUCKVALE, M. A.; BEKE, A. It sounds like you have a cold! testing voice features for the interspeech 2017 computational paralinguistics cold challenge. In: INTERNATIONAL SPEECH COMMUNICATION ASSOCIATION (ISCA). [S.l.], 2017.

HUNDAL, J. K.; HAMDE, S. Some feature extraction techniques for voice based authentication system. In: IEEE. *2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI)*. [S.l.], 2017. p. 419–421.

LAMEL, L.; GAUVAIN, J.-L. Speaker verification over the telephone. *Speech Communication*, Elsevier, v. 31, n. 2-3, p. 141–154, 2000.

LANSFORD, K. L.; BERISHA, V.; UTIANSKI, R. L. Modeling listener perception of speaker similarity in dysarthria. *The Journal of the Acoustical Society of America*, AIP Publishing, v. 139, n. 6, p. EL209–EL215, 2016.

LINDSAY, M. You're the voice: the science behind speaker recognition tech. *The Conversation*, 2014. Acessado em Outubro de 2024. Disponível em: ⟨http://theconversation.com/youre-the-voice-the-science-behind-speaker-recognition-tech-31579⟩.

LUPU, E.; CIOBAN, M. Voice biometric system. In: SPRINGER. *International Conference on Advancements of Medicine and Health Care through Technology: 23–26 September, 2009, Cluj-Napoca, Romania*. [S.l.], 2009. p. 239–242.

MENG, Z.; ALTAF, M. U. B.; JUANG, B.-H. F. Active voice authentication. *Digital Signal Processing*, Elsevier, v. 101, p. 102672, 2020.

NEUSTEIN, A.; PATIL, H. A. *Forensic speaker recognition*. [S.l.]: Springer, 2012. v. 1.

PRAVENA, D.; DHIVYA, S. et al. Pathological voice recognition for vocal fold disease. *International journal of computer applications*, Citeseer, v. 47, n. 13, 2012.

RAMOJI, S.; KRISHNAN, P.; GANAPATHY, S. Neural plda modeling for end-to-end speaker verification. *arXiv preprint arXiv:2008.04527*, 2020.

SHAYAMUNDA, C.; RAMOTSOELA, T.; HANCKE, G. P. Biometric authentication system for industrial applications using speaker recognition. In: IEEE. *IECON 2020 The 46th Annual Conference of the IEEE Industrial Electronics Society*. [S.l.], 2020. p. 4459–4464.

TULL, R. G.; RUTLEDGE, J. ‘cold speech’for automatic speaker recognition. In: *Acoustical Society of America 131st Meeting Lay Language Papers*. [S.l.: s.n.], 1996.

TULL, R. G.; RUTLEDGE, J. C. Analysis of “cold-affected”speech for inclusion in speaker recognition systems. *The Journal of the Acoustical Society of America*, Acoustical Society of America, v. 99, n. 4_Supplement, p. 2549–2574, 1996.

TULL, R. G.; RUTLEDGE, J. C.; LARSON, C. R. *Cepstral analysis of “cold-speech”for speaker recognition: a second look*. Tese (Doutorado) — Acoustical Society of America, 1996.

ZHANG, X. et al. Pathological voice recognition by deep neural network. In: IEEE. *2017 4th International Conference on Systems and Informatics (ICSAI)*. [S.l.], 2017. p. 464–468.