UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Campus de São José do Rio Preto

Gustavo Fernando Balbino

Perfilamento Humano Através Do Uso De Algoritmos Computacionais Para Processamento Da Voz

São José do Rio Preto
2022

Gustavo Fernando Balbino

Perfilamento Humano Através Do Uso De Algoritmos Computacionais Para Processamento Da Voz

Trabalho de Conclusão de Curso (TCC) apresentado como parte dos requisitos para obtenção do título de Bacharel em Ciência da Computação, junto ao Conselho de Curso de Bacharelado em Ciência da Computação, do Instituto de Biociências, Letras e Ciências Exatas da Universidade Estadual Paulista "Júlio de Mesquita Filho", Câmpus de São José do Rio Preto.

Orientador: Prof. Dr. Rodrigo Capobianco Guido

São José do Rio Preto
2022

B172p

Balbino, Gustavo Fernando
Perfilamento humano através do uso de algoritmos computacionais para processamento da voz / Gustavo Fernando Balbino. -- São José do Rio Preto, 2022
52 p. : il., tabs.

Trabalho de conclusão de curso (Bacharelado - Ciência da Computação) - Universidade Estadual Paulista (Unesp), Instituto de Biociências Letras e Ciências Exatas, São José do Rio Preto
Orientador: Rodrigo Capobianco Guido

1. Voz. 2. Fala. 3. Perfilamento. 4. Processamento de Sinais. 5. Características. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Biblioteca do Instituto de Biociências Letras e Ciências Exatas, São José do Rio Preto. Dados fornecidos pelo autor(a).

Essa ficha não pode ser modificada.

Gustavo Fernando Balbino

Perfilamento Humano Através Do Uso De Algoritmos Computacionais Para Processamento Da Voz

Trabalho de Conclusão de Curso (TCC) apresentado como parte dos requisitos para obtenção do título de Bacharel em Ciência da Computação, junto ao Conselho de Curso de Bacharelado em Ciência da Computação, do Instituto de Biociências, Letras e Ciências Exatas da Universidade Estadual Paulista "Júlio de Mesquita Filho", Câmpus de São José do Rio Preto.

Comissão Examinadora

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

Profa. Dra. Renata Spolon Lobato
UNESP – Câmpus de São José do Rio Preto

Prof. Dr. Aleardo Manacero Júnior
UNESP – Câmpus de São José do Rio Preto

São José do Rio Preto
2022

Dedico à minha família, base do meu caminho até aqui.

Agradecimentos

Agradeço, primeiramente a Deus e à Sua Palavra que me trouxeram à Verdade e possibilitou-me a força necessária para seguir em frente. Agradeço à minha família: meu pai, Fernando, e minha mãe, Glaucia, além do meu querido irmãozinho Lukas, que me forneceram a estrutura essencial para que eu pudesse chegar até aqui. Agradeço à minha namorada, e futura esposa, Lara, que me alegrou nos momentos mais difíceis, sempre com um sorriso em seu rosto. Agradeço aos amigos mais próximos que fiz durante minha graduação: Eric, Everton, Filipe, Luis Marcello e Vinícius, os quais vou me recordar sempre com muito carinho. Agradeço ao meu orientador: Prof. Dr. Rodrigo, sem seu conhecimento, este projeto nunca teria sido possível. Agradeço a todos os professores que passaram pela minha vida universitária, compartilhando comigo parte do seu inestimável conhecimento.

Eu descobri em mim mesmo desejos os quais nada nesta Terra pode satisfazer. A única explicação lógica é que eu fui feito para outro mundo.

- C. S. Lewis

Resumo

Perfilamento humano através da voz pode ser definido como o processo de estimar características pessoais de um indivíduo utilizando sua voz. Um ato muito comum entre humanos é construir a imagem de alguém somente ouvindo sua fala. Com o avanço da tecnologia, tornou-se possível o emprego de métodos para realizar tais predições de maneira computacional, sendo exequível a dedução parâmetros físicos, geográficos, emocionais, médicos, sociais, etc. de indivíduos, por meio de sinais digitais produzidos por sua voz, com o intermédio técnicas de aprendizado de máquina aplicadas a classificadores.

O presente projeto explora esse campo de estudos, e busca desenvolver uma aplicação, com base em técnicas de processamento de sinais digitais, extração de características e classificadores, que seja capaz de realizar estimativas acerca dos seguintes atributos físicos e fisiológicos dos seres humanos: altura, peso, faixa etária e sexo, utilizando-se somente de gravações digitais de áudio contendo uma pronúncia de texto pré-definido.

Palavras-chave: Voz, Fala, Perfilamento, Processamento de Sinais, Características.

Abstract

Human profiling from voice can be defined as the procedure of making estimates about personal characteristics from a individual using their voice to do so. It is fairly common among people to construct an image, in their minds, of someone they have never seen from hearing their voice alone. As the field of computer science evolves, it became feasible to employ computational methods for predicting a number of parameters of the speaker, namely, physical attributes, geographic aspects, emotions, health, social affairs, etc., through digital audio data recorded from their voice, and by applying digital signal processing techniques supported bey machine learning.

The work aims to explore this area of the Information Theory field, and intend to present a software application that is able to, through a voice recording containing an utterance of a pre-defined script, predict the following physical parameters about the speaker: height, weight, age, and gender.

Keywords: Voice, Speech, Profiling, Digital Signal Processing, Features.

Lista de Figuras

Lista de Tabelas

Sumário

Lista de Abreviações

ANN Artificial Neural Network

AoI Attributes of Interest

AUC Area Under the Curve

DLT Deviation from Linear Trend

EMA Erro Médio Absoluto

FN False Negative

FP False Positive

FPQ Frequency Perturbation Quotient

HMM Hidden Markov Model

MLP Multi-Layer Perceptron

PCM Pulse-code modulation

PVI Pitch Variation Index

RAP Relative Average Perturbation

RMS Root Mean Squared

ROC Receiving Operating Characteristic

SO Sistema Operacional

SVM Support Vector Machine

TEO Teager Energy Operator

TN True Negative

TP True Positive

VoIP Voice over Internet Protocol

WAV Waveform Audio File Format

ZCR Zero Crossing Rate

Capítulo 1

Introdução

1.1 Considerações iniciais

O termo "perfilamento humano" é definido na literatura como um processo de dedução de características humanas, geralmente de natureza física e/ou demográfica. Tal processo pode encontrar utilidade em diferentes âmbitos como aplicações forenses, segurança e cumprimento da lei, assistência médica, *marketing*, entre outros. Existe um número de métodos que podem ser empregados para se alcançar esse objetivo.

O uso da voz tem sido apresentado como uma novidade nesse campo de estudos. Através da análise do sinal da fala, utilizando-se de artifícios computacionais, é possível determinar características pessoais do locutor como gênero, idade, etnia, porte físico, e outros aspectos.

Neste contexto, Khan (2019) indica os principais desafios do perfilamento humano através da voz como: *(i)* encontrar uma representação que compreenda a relação entre a voz e o locutor, *(ii)* tornar tal representação flexível ao ruído inerente do mundo real e *(iii)* assegurar que essa representação é generalizável em diferentes condições e características do locutor.

1.2 Motivação

Os trabalhos anteriores na área de perfilamento humano através da voz apresentam-na não somente como um campo de estudos relativamente novo, mas também de grande interesse científico. As pesquisas realizadas durante o desenvolvimento do presente projeto revelaram a relevância da área para a sociedade através de suas utilidades, como identificação de pessoas, aplicações forenses e caracterização de doenças, bem como os desafios ainda a serem superados.

1.3 Objetivos

O objetivo do presente trabalho é desenvolver um algoritmo computacional capaz de determinar o perfil humano de uma pessoa através de uma gravação digital de sua fala como dado de entrada. Desse modo, estabeleceu-se que o algoritmo deve conseguir identificar as seguintes características físicas dos indivíduos: **altura, peso, idade e sexo**. O estudo pretende também realizar uma análise comparativa entre dois algoritmos classificadores: um simples, baseado em distância euclidiana, e outro mais robusto, baseado em *Randon Decision Forest*, verificando a eficiência de cada um nos limites do presente projeto.

1.4 Organização da Monografia

O trabalho foi organizado em capítulos, visto que este é extenso e multidisciplinar, trazendo conceitos de matemática, processamento de sinais, algoritmos computacionais e trato vocal. A organização do conteúdo subsequente dar-se-á da seguinte forma:

- Segundo capítulo: Revisão bibliográfica compreendendo os conceitos fundamentais para o entendimento acerca do projeto desenvolvido;
- Terceiro capítulo: Descrição dos métodos aplicados para se alcançar os objetivos do projeto, bem como o detalhamento de sua implementação, empregando o conteúdo

especificado no capítulo anterior;

- Quarto capítulo: Apresentação dos testes e resultados obtidos com o trabalho elaborado;

- Quinto capítulo: Conclusões e propostas para trabalhos futuros.

Capítulo 2

Fundamentação Teórica

Este capítulo aborda toda a fundamentação teórica que serve como base para o entendimento do trabalho, e também para dar o devido respaldo ao conteúdo abstrato que é empregado no desenvolvimento do algoritmo detalhado no capítulo seguinte.

Inicialmente são apresentadas as definições dos conceitos fundamentais para o entendimento do projeto desenvolvido e seus objetivos. Em seguida, é feita uma breve exposição de algumas obras produzidas na área, de forma a situar o leitor acerca do estado da arte da área de estudos explorada no presente trabalho.

2.1 Definição dos Conceitos

2.1.1 Uso da voz no perfilamento humano

A expressão **Perfilamento humano através da voz** refere-se ao procedimento que tem por objetivo determinar características e informações (chamados **parâmetros**) de cunho pessoal de um indivíduo, utilizando-se apenas de sua voz. É importante ressaltar uma distinção entre os vocábulos "voz" e "fala". A primeira remete ao som que é produzido pelo trato vocal do ser humano, a medida que a última se trata do sinal produzido quando entoa-se a voz em algo de significado. (SINGH, 2019)

Perfilamento pode ser definido como a dedução de todos os tipos de parâmetros, nesse caso através do uso da voz. Esse espaço é vasto e pode ser distribuído em um número de categorias, como os exemplos abaixo:

- **Comportamentais:** domínio, liderança, outros comportamentos;
- **Demográficos:** Etnia, origens, nível educacional;
- **Ambientais:** Localização no momento da fala, aparelhos usados para captação da voz, redondeza do locutor;
- **Clínicos:** Patologias, efeitos de medicações, saúde física;
- **Físicos:** Altura, peso, estrutura do rosto;
- **Fisiológicos:** Idade, nível hormonal, ritmo cardíaco;
- **Psicológicos:** Personalidade, emoções;
- **Sociais:** Classe econômica, renda, profissão.

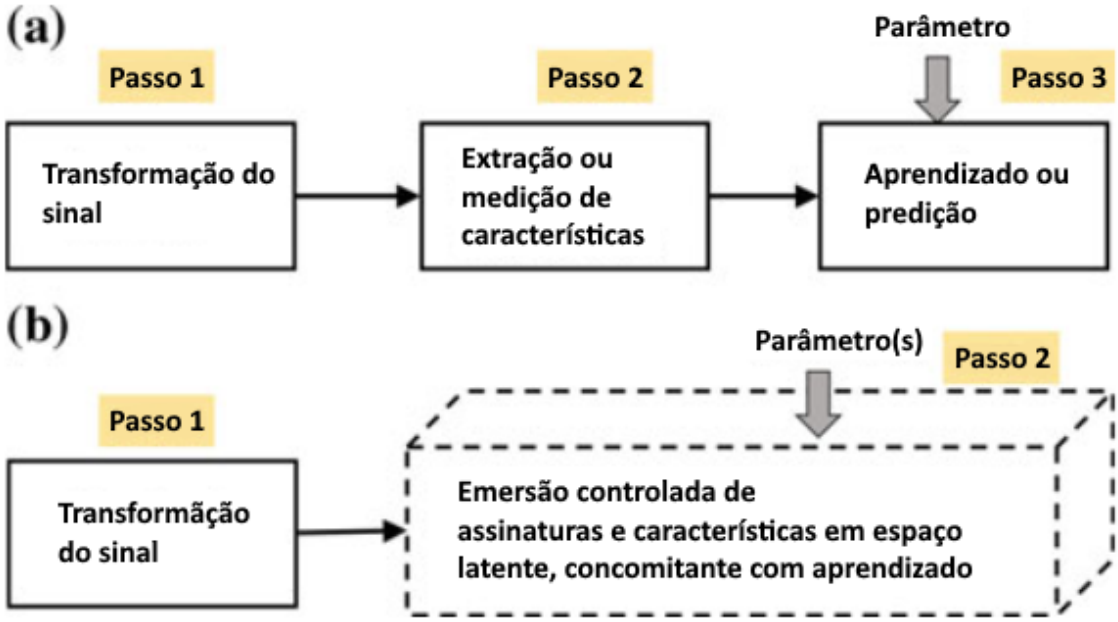
Uma variedade de estudos (ZEMLIN, 1982) (MALINOW et al., 1982) (AGHA, 2005) (KEANE, 1999) (NAKARAT et al., 2014) (HOLLIEN, 2013) sugere que a voz está relacionada a muitos outros parâmetro humanos, a saber: estrutura facial, estrutura corporal, benevolência, competência, dominância, personalidade e outros. Alguns aspectos de natureza ambiental do indivíduo também foram correlacionados a voz.

É comum que os humanos, mesmo que inconscientemente, façam julgamentos a respeito dos outros a partir de suas vozes, assim como conseguem facilmente fazer previsões sobre as condições ao redor do locutor (se está falando de um telefone, ou se a gravação da voz foi feita de uma proximidade física). Porém, tais julgamentos estão condicionados às limitações humanas, como sua capacidade de audição, habilidades interpretativas, e estado mental, o que causa, na maioria da vezes, análises imprecisas sobre o indivíduo.

Os métodos computacionais, todavia, estão isentos dessas limitações. Uma vez que o processo é "terceirizado" para uma máquina, a maior parte das ambiguidades e imprecisões humanas caem por terra. No entanto, vale ressaltar que desafios associados estão presentes, e estes tem um papel importante para as pesquisas na área, uma vez que ainda não estão bem sedimentados. Alguns exemplos: **variabilidade inerente na voz humana**, **disfarces da voz**, **variabilidade dos parâmetros de perfilamento**, **encontrar assinaturas dos parâmetros de perfilamento**

O processo de perfilamento computacional geralmente segue duas abordagens: **orientado a conhecimento** e **orientado aos dados**. No primeiro caso, as características que são conhecidas por se correlacionar com os parâmetros a serem deduzidos são extraídas, quantificadas e, então, mapeadas para os parâmetros dentro de um intervalo que seja estatisticamente relevante. O último utiliza técnicas de processamento de dados de sinais e está ilustrado na Figura 2.1.

Figura 2.1: Abordagens (a) *top-down* e (b) *bottom-up* do modelo orientado a dados



Fonte: adaptado de (SINGH, 2019).

2.1.2 Formato WAV

O formato WAV, forma curta de WAVEform, é um formato de áudio criado pelas empresas Microsoft e IBM, em 1991, para armazenamento de arquivos de áudio em computadores pessoais. Embora um arquivo WAV possa conter áudio comprimido, a forma mais comum de se utilizar o formato é para guardar áudio sem compressão, ou seja, sem perdas de dados, que pode ser obtido através da Modulação por Código de Pulsos (*Pulse-code modulation* - PCM, em inglês), que se trata de um método para representação digital de amostras de áudios analógicos.

Seu uso é comum em sistemas de processamento de voz pelo fato de possibilitar o armazenamento de áudio sem perda de qualidade, e também por ser um formato padrão, permitindo que sejam salvos e processados facilmente em qualquer dispositivo.

2.1.3 Extração de características

Características são uma parte muito importante para qualquer sistema que envolva o processamento a voz para se alcançar algum objetivo. Uma definição bem feita de tais características é o ponto chave para a identificação de aspectos humanos. A fala é um sinal contínuo de tamanho variável que contém informação. Dessa forma é possível extrair características de natureza global e local. As globais, também conhecidas como "de longo termo", estão associadas às estatísticas "brutas" como média, mediana, valores máximos e mínimos, e desvio padrão do sinal. Já as locais estão mais ligadas às variações temporais, onde geralmente tenta-se aproximar de um estado estacionário.

Mais especificamente, é possível classificar as características relevantes para o estudo de perfilamento humano através da voz em alguns conjuntos descritos abaixo:

- **Características temporais:** Determináveis no domínio temporal, como energia do sinal, taxa de cruzamento com zero (ZCR), volume, ritmo de fala, etc.

- **Características espectrais:** Determináveis no domínio da frequência, como tom ou

frequência fundamental, harmonia, fluxo espectral, etc.

- **Características compostas:** Necessitam de medição tanto em termos de tempo quanto de frequência, como ritmo, melodia, etc.
- **Características qualitativas:** Propriedades de caráter humano como nasalidade, respiração, irregularidade, etc.

2.1.4 *Cepstrum*

Sanchez et al. apud Deng & O'Shaughnessy descreve o espectro da fala como sendo $S = E \cdot H$, onde E representa a excitação e H representa a ressonância do trato vocal. A análise *cepstral* é gerada a partir do seguinte:

$$S = E \cdot H \Rightarrow log(S) = log(E \cdot H) \Rightarrow log(S) = log(E) + log(H)$$

Uma das aplicações da análise *cepstral* está na detecção de período de *pitch* da fala, que pode ser definido como o intervalo de tempo entre dois fechamentos glóticos que controlam o fluxo de ar de E. (SANCHEZ et al., 2009)

2.1.5 Entropia

O termo entropia é objeto de estudo dos mais variados campos do conhecimento, como física, termodinâmica, astrofísica, genética e também tecnologia da informação. Baseado no trabalho de Shannon (1948), entropia pode ser definida como a medida de imprevisibilidade do conteúdo da informação, a qual se iguala a zero quando o resultado é completamente previsível. Sua fórmula matemática é dada por:

$$H = \sum_{i=0}^{K-1} p_i \cdot log_{\beta}\left(\frac{1}{p_i}\right)$$

onde p_i é a probabilidade do i-ésimo dado distinto de um conjunto $s[.] = \{s_0, s_1, s_2, \ldots, s_{M-1}\}$ de tamanho M com K símbolos distintos em cada. Sua expressão é dada por:

$$p_i = \frac{(\alpha)_i}{M} = \frac{1}{(M/(\alpha)_i)},$$

Quando comparada com métodos de energia e ZCR, a entropia pode revelar características do sinal de voz que são mais difíceis de serem encontradas baseadas nesses anteriores. (GUIDO, 2018)

2.1.6 Algoritmos classificadores

Um algoritmo classificador, em geral, pode ser descrito como uma função que pondera ou "pesa" as características de entrada, de forma que a saída separa-se em uma classe de valores positivos e uma de valores negativos. O processo de treinamento de classificadores consiste na identificação das funções (ou pesos) que fornecem a melhor e mais precisa separação das duas classes.

A Máquina de Vetor de Suporte (em inglês, *Support Vector Machine* - SVM), Redes Neurais Artificiais (em inglês, *Artificial Neural Networks* - ANN), Florestas de Decisão Aleatória (em inglês, *Random Decision Forest*, e *Perceptron* de Múltiplas Camadas (em inglês, *Multi-Layer Perceptron* - MLP) fazem parte das abordagens computacionais de algoritmos classificadores mais recentes, capazes de gerar divisões mais complexas entre as classes de dados. Vale ressaltar que cada um tem suas vantagens e desvantagens.

No presente trabalho, um número de classificadores é devidamente utilizado e treinado, a fim de comparar seus resultados no estudo proposto.

Support Vector Machine

Support Vector Machine (SVM) é um algoritmo de classificador de aprendizado supervisionado, que também tem aplicações para regressão. É capaz de realizar classificação linear e não-linear. Uma classificação linear é implementada através da função *Kernel*, enquanto que em uma não-linear, as funções *Kernel* são polinomiais, polinomiais complexos, funções de base radial Gaussiana, e funções tangentes hiperbólicas. A escolha apropriada de tais

Kernels e seus parâmetros é imprescindível para uma boa execução do algoritmo. (GOVE; FAYTONG, 2012)

O principal conceito por trás do SVM é a separação de características uma das outras em planos diferentes, formando uma espécie de hiperplano entre duas classes. O SVM então seleciona amostras de ambas as classes, chamados de vetores de suporte, os quais estão o mais próximo possível do hiperplano. O objetivo é maximizar a margem entre os vetores de suporte e o hiperplano, de modo a encontrar as fronteiras de decisão mais genéricas. (MISRA; LI, 2020)

Perceptron

Multi-Layer Perceptron (MLP) é um algoritmo classificador de aprendizado supervisionado, que possui a partir de 3 camadas, as quais são arranjadas da seguinte forma: uma camada de entrada de dados e uma de saída, com uma ou mais camadas não-lineares entre elas. A camada de entrada recebe os primeiros valores e os repassam para a primeira camada do meio, que transforma tais valores em uma somatória linear com um fator multiplicador para cada resultado e, em sequência, é passada por uma função não linear. A camada de saída recebe os valores da última camada do meio, e os transforma em um resultado. (MIA et al., 2015)

Após a obtenção do resultado, é calculada a diferença entre este e o resultado esperado, gerando um erro. A partir de tal erro, é calculado um parâmetro que será a taxa de mudança dos nós, ou neurônios. Após o término do processo, o caminho contrário é feito, alterando o valor dos pesos de cada neurônio baseado no parâmetro calculado anteriormente, de forma a tentar minimizar o erro da saída. O procedimento é então reiniciado até que se obtenha o melhor resultado. (RUDER, 2016)

Random Decision Forest

Árvores de Decisão é um algoritmo classificador, também de aprendizado supervisionado, que similarmente pode ser utilizado para regressão. Apresenta uma estrutura simples e eficiente, tendo destaque com um amplo campo de aplicação. Este modelo é comumente referenciado pelo seu esquema de dividir e conquistar, uma vez que seu funcionamento se baseia em nós que guardam informações e se subdividem em nós "filhos" a partir de uma regra pré-definida. (SILVA, 2005)

Esse modelo de classificação se baseia em receber treinamentos que já possuem uma caracterização, dessa forma, o algoritmo pode ser aplicado em um conjunto de teste, que possuem registros com rótulos de classes ainda desconhecidas, possibilitando a criação de novas(BENTO; KUX; KÖRTING, 2019). Seu funcionamento começa pela escolha das regras, que nada mais são do que perguntas de resposta positiva ou negativa, definindo-se, assim, qual a melhor divisão do nó pai em seus dois filhos. O procedimento é repetido até que seja impossível continuar, obtendo-se um resultado. É realizada então uma análise que buscará o caminho na árvore com a menor complexidade e menor taxa de erro. (SAFAVIAN; LANDGREBE, 1991)

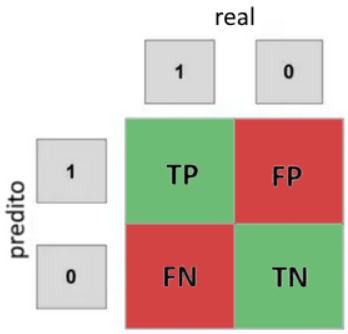
Matriz de Confusão

Matrizes de confusão são estruturas utilizadas para demonstrar o desempenho de um algoritmo de classificação. Cada linha representa o resultado de classificação do algoritmo,e cada coluna representa o valor real da classe de entrada. Seus elementos representam os valores de verdadeiro positivo (o algoritmo previu corretamente a classe), verdadeiro negativo (previu corretamente a classe de um contra exemplo), falso negativo (previu-se erroneamente a classe) e falso positivo (previu-se erroneamente a classe do contra exemplo).

A ordem da matriz é determinada pela quantidade de classes que serão usadas pelo classificador. Na Figura 2.2 está ilustrado como uma matriz de confusão de ordem 2 é formada. Uma matriz de confusão cuja soma da diagonal principal resulte no maior valor possível, em

comparação com outras, é considerada a melhor matriz. De maneira análoga, a matriz com a soma da diagonal principal resultando no menor valor, em comparação, é chamada pior matriz. Ambas são usadas para representar o melhor e o pior caso de classificação do algoritmo, respectivamente.

Figura 2.2: Matriz de confusão 2x2



Fonte: elaborado pelo autor.

Curva ROC

A curva de Características Operacionais Recebidas (do inglês, *Receiving operating characteristic* - ROC) se trata de um gráfico cujo objetivo é ilustrar o quão satisfatório um classificador binário (que classifica em no máximo duas possíveis classes) se comporta. Em suma, a curva apresenta a progressão de matrizes de confusão obtidas durante a execução do algoritmo de classificação, traçando a taxa de verdadeiros positivos em contraste com a taxa de falsos positivos. Quanto mais distante do centro do gráfico, em direção acima, a curva se encontra, melhor o desempenho do classificador.

A área sob a curva ROC é uma medida que pode ser obtida para facilitar a análise da própria curva. Seu valor é um número entre 0 e 1. Quanto mais próximo de 1, melhor o classificador está performando. Mais próximo de 0.5 significa que não está sendo possível verificar uma clara distinção entre as classes. Ao se aproximar de 0, indica que o classificador está falhando em determinar as classes corretamente.

2.2 Estado da Arte

Nesta seção, serão apresentados alguns trabalhos desenvolvidos no campo de pesquisas de perfilamento humano através da voz. Dentre as obras selecionadas encontram-se artigos científicos, monografias e teses de mestrado. Cada um é discutido brevemente a seguir, onde são demonstrados seus objetivos e resultados, a fim de elucidar sua contribuição científica.

Em sua tese de mestrado, Khan (2019) defende que os principais desafios desta área de estudo estão em encontrar uma representação, ou modelo, que seja capaz de compreender a complexa relação existente entre a voz e os parâmetros do locutor, e que seja resiliente aos ruídos do ambiente. Mas isso não é tudo: o modelo ideal também precisa ser generalizável para quaisquer condições da gravação de voz e características do indivíduo em questão. A fim de encontrar uma solução para tais problemas, o autor propõe um modelo, utilizando técnicas de redes neurais de ponta, combinando características obtidas através de métodos orientados a conhecimento, com abordagens orientadas a dados contemporâneas. O autor também apresenta uma nova arquitetura de rede neural, que incorpora as características do sinal em uma representação interna, sem utilizá-las como entrada, além de treinar a rede para saber diferenciar ruído de voz limpa, poluindo a entrada de dados com sinal ruidoso. O autor conclui, experimentalmente através de *benchmarks* específicos para os desafios em questão, que o modelo proposto contribui grandemente para a performance do perfilamento por voz, e que, embora utilizar apenas técnicas orientadas a dados tenha relatado uma alta precisão com as amostras de teste, sozinhas não são generalizáveis o suficiente para cobrir todas as deficiências do estado da arte da área.

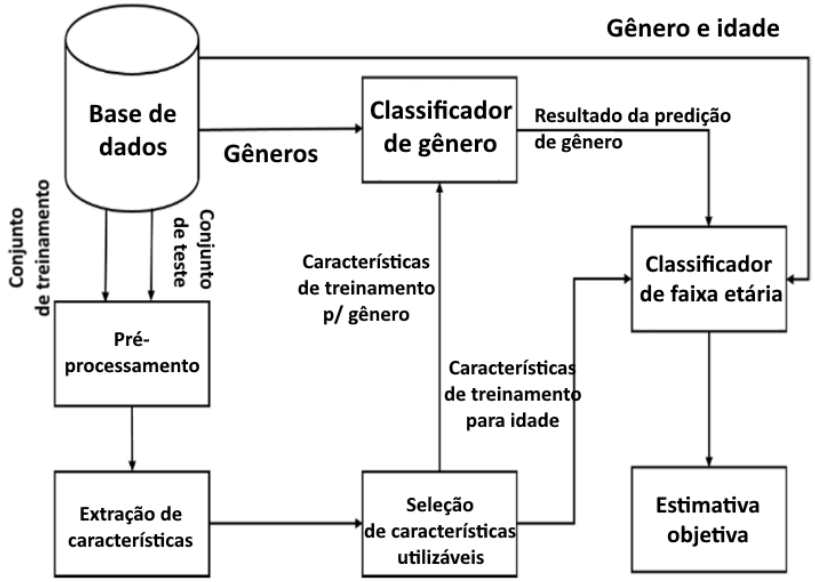
Marster & Schilling (2015) mostram uma pesquisa contando com casos de exemplo de quais parâmetros podem ser descobertos através de perfilamento por voz, e a utilidade da área para fins linguísticos forenses. A dissertação aborda alguns métodos de perfilamento de locutor, e discute brevemente sobre expertise na área. Um dos assuntos mais inquietantes do trabalho, porém, é respeito de disfarce de voz, um aspecto que, segundo os autores, deve sempre ser levado em conta em casos forenses, já que criminosos tendem a desejar manter

suas identidades ocultas. Em casos como esse, o apoio de linguistas forenses profissionais podem vir a calhar.

Em seu artigo Goyal, Patage & Tiwari (2020) discorrem a cerca de dedução de gênero e idade através da voz, construindo um modelo de predição com arquitetura baseada em MLP. Os autores extraíram características como *Mel Frequency Cepstral Coefficients*, formantes, tom, valores de croma, entra outros, de cada uma das amostras de fala, que são então selecionadas utilizando as técnicas de Análise de Componente Principal (do inglês: *Principal Component Analysis*, PCA) e Eliminação de Características Redundantes (do inglês, *Redundant Feature Elimination*). Como base de dados, o estudo utilizou o repositório *The Common Voice* do *Mozilla Speech Database*, um banco de dados público com uma gigante variedade de amostras. Empregando uma abordagem de duas frentes, as quais estão ilustradas na Figura 2.3 e Figura 2.4, respectivamente, as predições realizadas pelo MLP apresentaram uma precisão maior de predição, com 89,58%, em comparação com outros algoritmos classificadores como árvores de decisão, XGBoost e o HMM Gaussiano (*Hidden Markov Model* - Modelo Oculto de Markov). Os dados foram obtidos separando os dados em oito faixas etárias e nos 2 gêneros. Os autores concluem ressaltando que o modelo proposto não foi testado com amostras em que contém mais de um locutor, e comentam sobre o uso de *deep learning* (aprendizado profundo) no estágio de pré-processamento, com o objetivo de conseguir um ganho de performance.

Shen et al. (2021) discutem em seu artigo sobre o uso de gravações de áudio não linguístico para inferir atributos de interesse (AoI) de indivíduos, conhecido de "perfilamento de usuário". Os autores apresentam uma solução nova para a detecção de gênero e personalidade de locutores utilizando este método. Ao invés de características linguísticas e acústicas da fala, o estudo foca em características conversacionais que poderiam refletir nos AoIs. Primeiramente, foi desenvolvido um algoritmo de detecção de atividade vocal capaz de identificar diferenças individuais na voz, e ruído produzido por pessoas próximas. Em seguida, um métodos de aprendizagem de máquina assistido por gênero foi desenvolvido para enfrentar as

Figura 2.3: Abordagem do modelo sequencial para predição de idade e gênero usando a voz, de Goyal, Patage & Tiwari (2020)

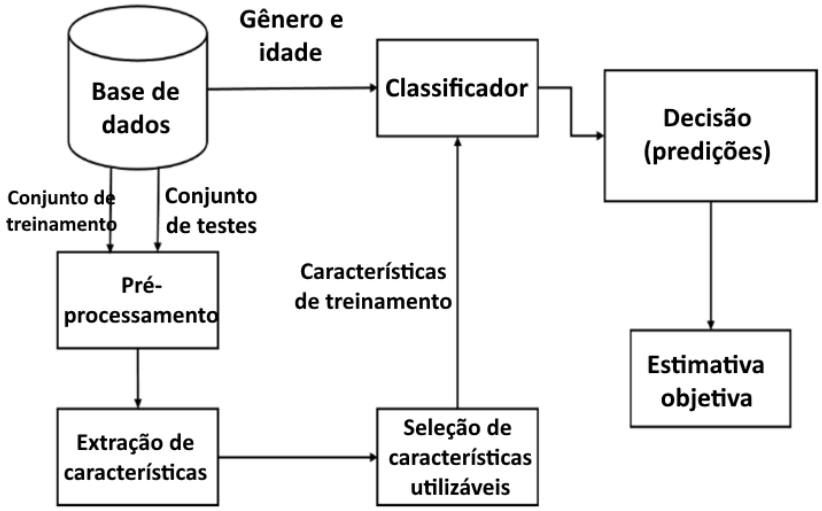


Fonte: adaptado de (GOYAL; PATAGE; TIWARI, 2020).

dinâmicas na voz e comportamentos humanos, através da integração de diferenças de gênero e sua correlação com traços de personalidade. Aplicando-se extensivos experimentos com um grupo de estudos real, os autores introduziram um sistema de perfilamento de usuário através da voz que faz uso eficiente de dados de vozes não linguísticas por meio dos algoritmos implementados durante o projeto. Para identificação de gênero, o método proposto obteve uma precisão de 79,5% para identificação de gênero e cerca de 60% para o reconhecimento dos tipos de personalidades avaliadas.

Batthalla et al. (2016) propõem uma aplicação capaz de identificar comportamentos de usuários através de atividades de VoIP (do inglês, *Voice over IP*, Voz sobre IP). Os autores enumeram algumas condutas que podem ser prejudiciais para outros usuários como congestionamento intencional da rede, mensagens mal formadas, e *spam*, es esperam que sua aplicação possa detectar tais comportamentos e evitá-los. Partindo da definição de um conjunto de parâmetros, o experimento realiza cálculos estatísticos a fim de classificar os indivíduos que utilizam o serviço e, consequentemente, detectar anomalias. Por intermédio de treinamentos utilizando uma ampla base de vozes originadas de gravações de VoIP, os autores conseguiram

Figura 2.4: Abordagem do modelo individual para predição de idade e gênero usando a voz, de Goyal, Patage & Tiwari (2020)



Fonte: adaptado de (GOYAL; PATAGE; TIWARI, 2020).

identificar os diferentes tipos de usuários incluindo mal intencionados, foco da pesquisa, com uma alta taxa de sucesso.

Em seu trabalho, Kalluri, Vijayasenan & Ganapathy (2020) exploram estimativas de variados parâmetros físicos de indivíduos por meio de gravações de curta duração de suas vozes, extraindo características formantes, harmônicas, e espectrogramas de Mel de curto prazo. Os autores também usaram um modelo baseado em Regressão de Vetores de Suporte, que se trata de uma versão de SVM para problemas de regressão. Os experimentos deste trabalho foram aplicados sobre a base de dados de vozes acústicas TIMIT, com um conjunto de treinamento contando com 462 locutores (326 homens e 136 mulheres), mais um conjunto de testes com 168 locutores (56 mulheres e 112 homens), e obteve uma melhora significante nos resultados de dedução de altura e idade dos locutores, quando comparados com estudos anteriores. O foco do projeto foi utilizar a menor duração possível de gravações para fazer as estimativas, e os autores conseguiram atingir uma duração mínima de 0,5 segundos, sem que o erro médio absoluto (EMA) das predições ficasse muito abaixo do conjunto de treinamento. Os resultados obtidos com gravações de 1 a 2 segundos ficaram pareados com os dos trabalhos mais recentes na área. Através da combinação de extração das três categorias de características, cu-

jos erros de estimativa são complementares entre si, os autores alcançaram uma performance com EMA de 5,3 e 4,8 centímetros para nas predições de altura de homens e mulheres, respectivamente. Similarmente, obteve-se um EMA de 5,2 e 5,6 anos para as predições de idade de homens e mulheres, respectivamente. O estudo também se estende para a dedução de outras características físicas, como largura dos ombros, tamanho da cintura e peso.

Capítulo 3

Desenvolvimento

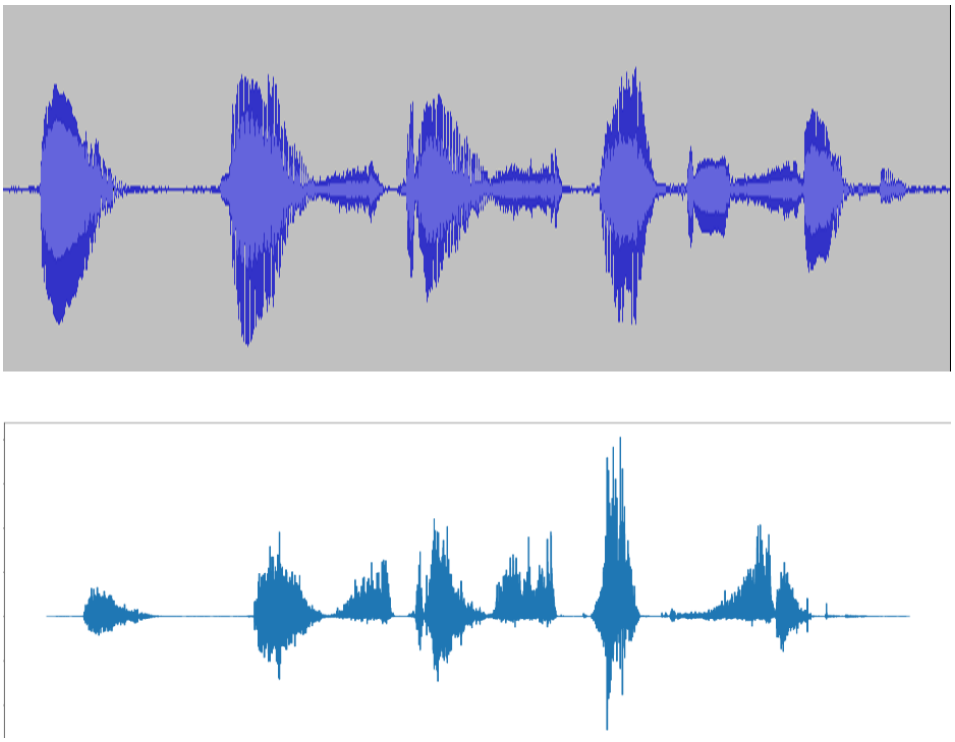
Neste capítulo serão discutidos os detalhes da metodologia utilizada para o desenvolvimento da aplicação proposta para o presente trabalho. Após todo o discorrimento feito no capítulo anterior acerca dos conceitos técnicos relacionados ao projeto em questão, bem como uma breve apresentação do estado da arte da área, o conteúdo a seguir pode ser introduzido com mais convicção. O capítulo está dividido em seções que apresentam cada etapa do estudo realizado.

3.1 Pré-processamento

Na primeira etapa do desenvolvimento, foram colhidas 32 amostras de vozes gravadas digitalmente, em ambiente silencioso, e convertidas para o formato WAV. Em cada amostra, o locutor foi instruído a pronunciar uma contagem de 1 até 5, em seu próprio ritmo e timbre natural. Após a coleta, cada amostra passou por um processamento a fim de manipular e extrair seus respectivos os dados brutos, em valores numéricos, para arquivos de texto externos, referidos daqui em diante no texto como **sinais brutos**.

Para efeito de comparação, foi gerado um gráfico para cada arquivo de texto contendo os dados brutos, que em seguida, foram comparados com seus respectivos gráficos de forma de onda no domínio temporal gerados pelo editor de áudio *opensource Audacity*.

Figura 3.1: Comparação do gráfico de forma de onda no domínio temporal gerado pelo software *Audacity* de um sinal de exemplo com o gráfico traçado a partir dos dados brutos extraídos do mesmo sinal



De forma a facilitar as futuras manipulações, executou-se um processo de normalização, onde foram realizadas as seguintes operações sobre cada sinal bruto:

- Cálculo da média dos valores;
- divisão de cada valor pela média calculada;
- cálculo do valor máximo;
- divisão da cada valor pelo valor máximo calculado.

Dessa forma, é possível garantir que todos os valores de cada sinal bruto estará entre 0 e 1. Os novo sinais foram gravados em novos arquivos de textos, os quais serão referidos daqui em diante por **sinais normalizados**.

3.2 Definição das classes

O presente trabalho explora os seguintes aspectos fisiológicos humanos: **faixa etária, altura, peso e sexo**. Sendo assim, foram definidas as seguintes classes, separadas por categorias, baseadas em intervalos de valores que uma pessoa pode apresentar para esses traços em questão:

Tabela 3.1: Classes definidas para o estudo do presente projeto

Sexo	Faixa Etária	Altura	Peso
Masculino	Jovem (1 a 29 anos)	Baixo (< 170cm)	Mais leve (< 80Kg)
Feminino	Mais Velho (30 anos ou mais)	Alto (>= 170cm)	Mais pesado (>= 80Kg)

Dessa forma, um sinal específico se encaixa em uma classe por categoria, de modo que cada um terá exatas 4 classes possíveis. Ou seja, um sinal obtido, por exemplo, de um homem de 40 anos, com 167cm de altura, pesando 82Kg, será um sinal Masculino, Mais Velho, Baixo e Mais Pesado.

A distribuição dos 32 sinais colhidos entre as classes ficou da seguinte forma: 17 Masculinos e 15 Femininos, 19 Mais Leves e 13 Mais Pesados, 19 Baixos e 13 Altos, e 15 Jovens e 17 Mais Velhos. Procurou-se colher sinais de maneira a abranger o maior número possível de combinações entre as classes de cada categoria.

3.3 Extração de características dos sinais

Após a etapa de pré-processamento, iniciou-se a extração de características dos sinais normalizados. Para tal procedimento, empregou-se a estratégia de detecção de período de *pitch* da voz, através do cálculo do *Cepstrum* do sinal.

Foram selecionadas 16 características, no total, para extração, cujas fórmulas matemáticas estão apresentadas no Apêndice A, onde $F0$ corresponde ao vetor de valores de *pitch* detectado, N corresponde ao tamanho do vetor de $F0$, e $T0$ corresponde ao vetor cujos valores se

dão por $T0 = \frac{1}{F0}$

O cálculo das características escolhidas necessitam, como pode ser observado pela suas respectivas fórmulas, de vários valores de $F0$ como entrada. Tais valores são obtidos da seguinte forma:

- Divide-se o sinal em janelas;
- Aplica-se o *Cepstrum* sobre os valores da janela;
- Obtêm-se o valor da frequência fundamental $F0$ de cada janela, buscando por picos, em um intervalo de valores relevantes;
- Coleciona-se cada valor de $F0$ em um vetor com tamanho relativo ao tamanho do sinal e da janela, que será utilizado para a extração de características.

O tamanho da janela escolhida foi o mesmo para todos os sinais: 30ms. Uma vez que todos os sinais possuem uma taxa de amostragem de 44100 Hz, cada janela ficou com 1440 valores. Aplicou-se também um *overlay* de 50% em cada janela.

O intervalo de valores relevantes para a busca de picos após o *Cepstrum* foi $[60, 400]$, visto que são valores suficientes para determinar o *pitch* em uma amostra de fala humana.

Obtidos os vetores de $F0$ para cada um dos sinais, executou-se o computo das características mencionadas anteriormente, de modo que, ao final do processo, cada sinal era representado por um vetor de características com tamanho 16 (um valor para cada característica calculada).

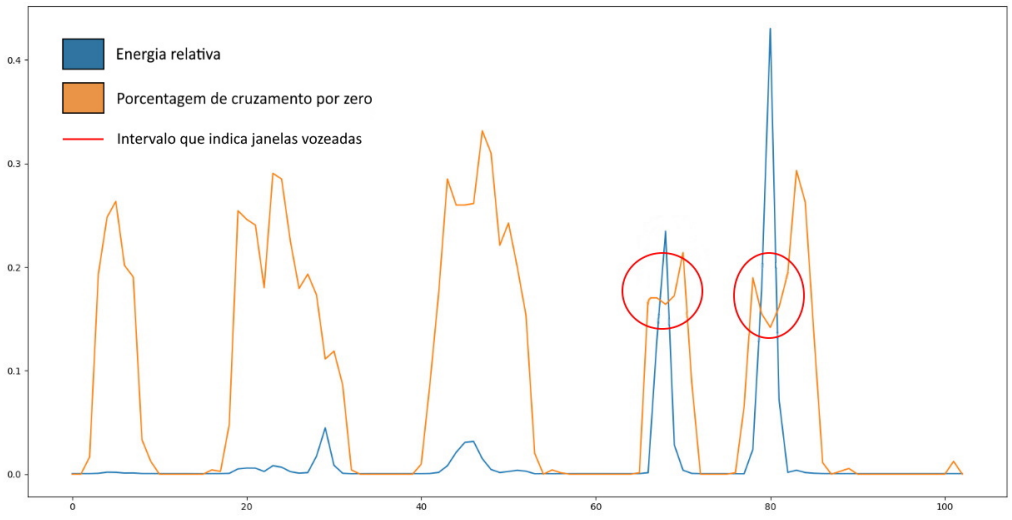
Com o propósito de analisar o quanto os vetores se diferenciavam entre as classes definidas, foi construído um gráfico para cada classe, traçando os vetores de características correspondentes às respectivas classes de cada gráfico. Pode-se observar que não havia diferenças notáveis entre os gráficos em termos de valores.

Para contornar esse fato, o *Cepstrum* foi aplicado novamente, e novos vetores de $F0$ foram obtidos para cada sinal, desta vez apenas sobre janelas **vozeadas**. Para determinar tais

janelas, o seguinte método foi empregado, utilizando o mesmo tamanho de janela definido anteriormente:

- Calcula-se a energia relativa de cada janela, dado por $E = \frac{\sum_{i=1}^{n}(s[i])^2}{n}$, onde n é o tamanho da respectiva janela;
- Calcula-se a porcentagem de cruzamentos por zero (PCZ) de cada janela, dada pelo número de cruzamentos por zero dividido pelo número de valores da respectiva janela;
- Procura-se janelas cuja energia seja relativamente alta, e a (PCZ) seja relativamente baixa, definindo-se um limiar para tais valores;
- Janelas cuja energia e (PCZ) ambos estejam acima e abaixo, respectivamente, de seus limiares, são classificadas como "vozeadas".

Figura 3.2: Intervalo de valores de energia e cruzamentos por zero, calculados sobre um sinal de exemplo, que corresponde a janelas vozeadas



Após a repetição dos procedimentos utilizando apenas janelas vozeadas, obteve-se novos vetores de características para cada sinal, os quais, ao serem traçados em gráficos da mesma maneira de anteriormente, apresentaram diferenças mais consideráveis em termos de valores numéricos. Sendo assim iniciou-se o método de classificação dos sinais.

3.4 Classificação

Para este projeto, foi realizada uma comparação entre um classificador mais simples, baseado no cálculo da distância euclidiana entre os vetores de características, e um mais robusto, utilizando o modelo *Random Decision Forest*.

Primeiramente, para cada categoria, os sinais são separados em suas devidas classes. Em seguida são sorteados de maneria aleatória e divididos em conjuntos de teste e modelo, numa proporção de 50% para cada. Testes e modelos só foram comparados entre classes da mesma categoria, uma vez que não faria sentido, por exemplo, testar **Peso Mais Leve** com **Faixa Etária Jovem**. Isso foi feito para ambos os classificadores utilizados.

3.4.1 Distância Euclidiana

A distância euclidiana d entre dois vetores q e p, de tamanho n, é dada por:

$$d(p,q) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$

A classificação foi executada da seguinte forma: para cada vetor de características de teste, calcula-se a distância euclidiana entre este e cada um dos vetores de modelos, e toma-se o menor resultado, ou seja, a menor distância euclidiana entre os em questão. Verifica-se a qual classe pertence o modelo que deu a menor distância. Se pertence a mesma classe do teste, conta-se um acerto, caso contrário, conta-se um erro. Procede-se até passar por todos os vetores de teste e, em seguida, faz-se o mesmo para os vetores de teste da outra classe da mesma categoria. Esse método se repete para todas as categorias restantes.

Os resultados de erros e acertos são anotados em matrizes de confusão, uma para cada categoria. Como os conjuntos de teste e modelos são escolhidos aleatoriamente, de modo a realizar uma análise mais precisa do classificador, os processos descritos acima foram repetidos por 500, 1000, 5000 e 10000 iterações, obtendo-se, desse modo, para cada iteração, uma matriz de confusão para cada categoria. Ou seja, no cenário de 500 iterações, obteve-se 500

matrizes para cada categoria e assim por diante.

Para fins de análise, em cada cenário de iterações, manteve-se apenas a melhor e a pior matriz de confusão por categoria, resultando em 8 matrizes para cada cenário. Por fim, calculou-se, sobre cada matriz, os seguintes valores estatísticos: **acurácia**, ***recall***, **precisão** e **f-score**. A fórmula matemática para cada um desses cálculos são dadas a seguir:

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

Fórmula da acurácia

$$recall = \frac{TP}{TP + FN}$$

Fórmula do recall

$$precision = \frac{TP}{TP + FP}$$

Fórmula da precisão

$$fscore = 2 * \frac{precision * recall}{precision + recall}$$

Fórmula do *f-score*

Os resultados serão apresentados no próximo capítulo.

3.4.2 Modelo *Random Forest*

De forma similar ao classificador por distância euclidiana, os vetores de características também foram divididos em suas respectivas classes, dentro das categorias, e também separou-se em conjuntos de modelos e testes, com proporção de 50%. Após isso, os vetores foram submetidos a uma implementação do modelo *Random Forest* na linguagem *Python*.

Os conjuntos de modelos e teste são inseridos como entrada para treinamento e teste do algoritmo *Random Forest*, que calcula, então, as probabilidades de cada vetor de teste ser classificado para a sua classe real. Com isso, é possível traçar uma **curva ROC** e sua respectiva ***area under the curve*** (AUC - em português, área sob a curva). Obteve-se uma curva ROC e AUC para cada categoria, contrastando as respectivas classes. Os resultados deste classificador também serão apresentados no próximo capítulo.

Capítulo 4

Testes e Resultados

Neste capítulo são apresentados os resultados obtidos a partir da implementação do método proposto no presente trabalho. Além disso, especifica-se a plataforma de testes e discutem-se os resultados.

4.1 Plataforma de Desenvolvimento e Testes

Todos os algoritmos foram escritos em linguagem Python e C++, e desenvolvidos e testados em um computador com processador Intel Core i5 9300H, 8GB de RAM 2666Mhz, placa de vídeo Nvidia GTX 1050 3GB, configurado com sistema operacional Windows 11 Pro e Ubuntu 22.04 LTS, tendo sido testados em ambos SOs.

4.2 Resultados Obtidos

Distância Euclidiana

Como mencionado no capítulo anterior, o algoritmo de classificação baseado em distância euclidiana foi executado em diferentes números de iterações: 500, 1000, 5000 e 10000. Na Tabela 4.1, 4.2, 4.3 e 4.4 estão contidas a melhor e pior matriz de confusão obtidas em cada categoria de classes, para a execução de 10 mil iterações do classificador.

Tabela 4.1: **Categoria peso**: Melhor e pior matriz de confusão, respectivamente

		Classe real	
		Mais Leve	Mais Pesado
Classe predita	Mais Leve	10	0
	Mais Pesado	0	7

		Classe real	
		Mais Leve	Mais Pesado
Classe predita	Mais Leve	2	8
	Mais Pesado	6	1

Tabela 4.2: **Categoria altura**: Melhor e pior matriz de confusão, respectivamente

		Classe real	
		Baixo	Alto
Classe predita	Baixo	10	0
	Alto	0	7

		Classe real	
		Baixo	Alto
Classe predita	Baixo	3	7
	Alto	6	1

Tabela 4.3: **Categoria sexo**: Melhor e pior matriz de confusão, respectivamente

		Classe real	
		Feminino	Masculino
Classe predita	Feminino	7	1
	Masculino	1	8

		Classe real	
		Feminino	Masculino
Classe predita	Feminino	0	8
	Masculino	8	1

Tabela 4.4: **Categoria faixa etária**: Melhor e pior matriz de confusão, respectivamente

		Classe real	
		Jovem	Mais Velho
Classe predita	Jovem	8	0
	Mais Velho	2	7

		Classe real	
		Jovem	Mais Velho
Classe predita	Jovem	1	7
	Mais Velho	7	2

Nas Tabelas 4.5 e 4.6 estão apresentados os valores estatísticos calculados sobre a melhor

e pior matriz de confusão, respectivamente, obtidas em cada categoria de classes com o algoritmo classificador baseado em distância euclidiana, após 10 mil iterações (números expressos em porcentagens (%) quando o valor calculado só pode assumir resultados no intervalo $[0,1]$):

Tabela 4.5: Valores estatísticos calculados com a **melhor** matriz de confusão em cada categoria.

Aspecto humano	**Acurácia**	***Recall***	**Precisão**	***F-Score***
Peso	100,00%	100,00%	100,00%	4,00
Altura	100,00%	100,00%	100,00%	4,00
Gênero	88,20%	87,50%	87,50%	3,50
Faixa Etária	88,20%	88,00%	100,00%	3,20

Tabela 4.6: Valores estatísticos calculados com a **pior** matriz de confusão em cada categoria.

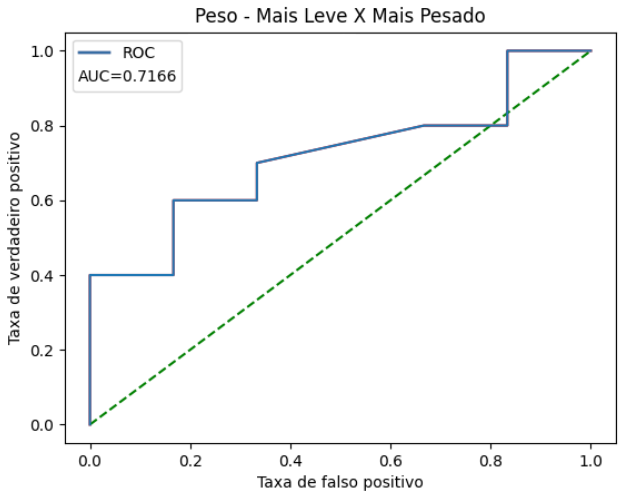
Aspecto humano	**Acurácia**	***Recall***	**Precisão**	***F-Score***
Peso	17,65%	25,00%	20,00%	1,00
Altura	23,50%	33,30%	30,00%	1,30
Gênero	5,90%	0,00%	0,00%	0,00
Faixa Etária	17,65%	12,50%	12,50%	0,50

Observou-se que o classificador em questão atingiu 100% de performance, baseado nos valores estatísticos computados, nos dois aspectos físicos humanos: peso e altura. Nos aspectos fisiológicos: sexo e faixa etária, essa marca não foi atingida, embora, em todo caso, a performance tenha sido satisfatória. Em execuções com um número menor de iterações, percebeu-se que os valores ficaram próximos a 80% para sexo e faixa etária, crescendo lentamente, enquanto que para peso e altura, os valores cresceram mais rapidamente. Em relação aos piores resultados atingidos pelo classificador, verifica-se que as categorias que representam aspectos fisiológicos foram, também, as que demonstraram o pior desempenho.

Random Decision Forest

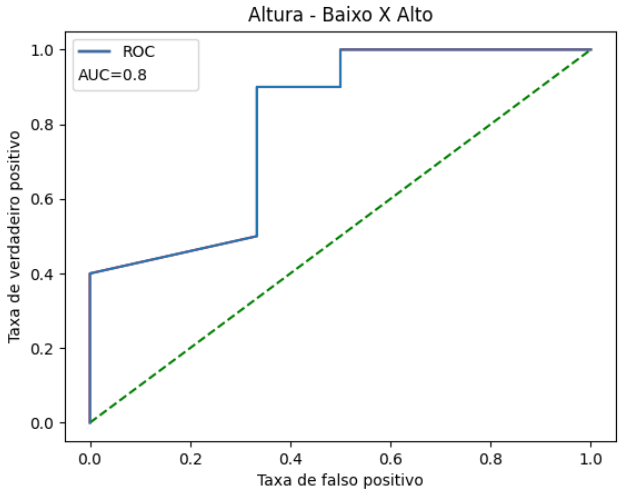
Nas Figuras 4.1, 4.2, 4.3 e 4.4 estão apresentados as curvas ROC e suas respectivas AUCs para cada categoria de classes, calculadas pelo algoritmo de classificação baseados em *Random Decision Forest*.

Figura 4.1: **Categoria peso**: Curva ROC e sua AUC gerados pelo algoritmo



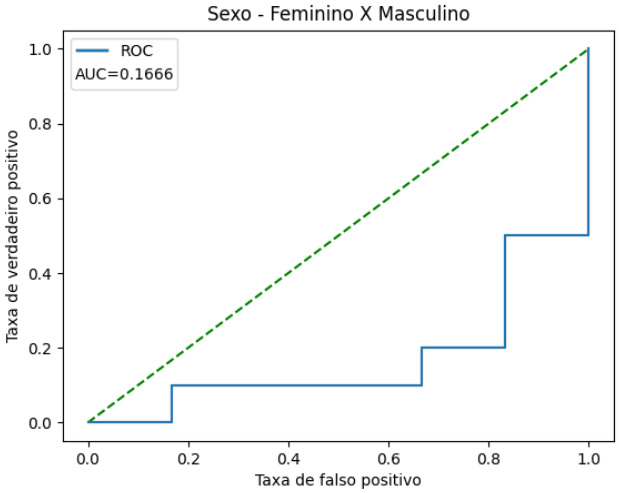
Fonte: elaborado pelo autor.

Figura 4.2: **Categoria altura**: Curva ROC e sua AUC gerados pelo algoritmo



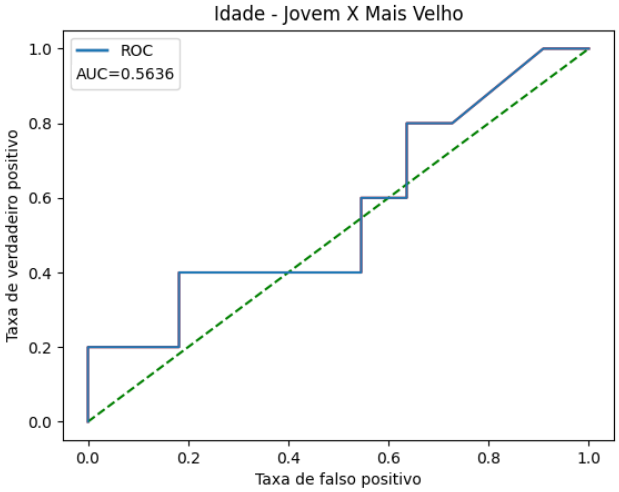
Fonte: elaborado pelo autor.

Figura 4.3: **Categoria sexo**: Curva ROC e sua AUC gerados pelo algoritmo



Fonte: elaborado pelo autor.

Figura 4.4: **Categoria faixa etária**: Curva ROC e sua AUC gerados pelo algoritmo



Fonte: elaborado pelo autor.

Pôde-se observar que, em comparação com o classificador por distância euclidiana, o presente não obteve bons resultados, chegando a 80% de precisão para o aspecto altura, e um pouco mais que 70% para o aspecto peso (analisando as curvas e AUCs produzidas). Também demonstrou um desempenho terrível para o aspecto sexo, e ficando bem próximo do limiar para faixa etária (o que não demonstra eficiência, de qualquer forma). Aqui, pode-se reparar,

também, que o classificador obteve melhores resultados para os aspectos físicos do que para os fisiológicos.

Pode-se concluir que, para o escopo do experimentos realizado neste trabalho, o cálculo da distância euclidiana como parâmetro de classificação atuou melhor do que o modelo *Random Forest*. Porém, é possível que a quantidade pequena de amostras de sinal tenha sido um fator que prejudicou o algoritmo mais robusto. No próximo capítulo, são dadas mais conclusões sobre o projeto, bem como sugestões para pesquisas futuras.

Capítulo 5

Conclusões

Os estudos realizados durante a execução deste projeto buscaram explorar o uso de métodos computacionais para processamento de sinais de fala, utilizando-se da extração de características baseadas em detecção de *pitch* da voz para o perfilamento de aspectos físicos e fisiológicos dos seres humanos: peso, altura, idade e gênero através do sinal produzido pela fala.

Os resultados para o classificador mais simples baseado em distância euclidiana e para o mais robusto baseado no modelo *Random Decision Forest*, foram apresentados e comparados, em termos de performance. O primeiro se mostrou mais eficiente para o escopo deste projeto, enquanto que o último não obteve um desempenho favorável, atingindo resultados muitos ruins para categorias de classes que representam os aspectos fisiológicos explorados.

Existe um número de fatores que podem ter levado a esta conclusão. Os recursos limitados a este projeto não contribuíram para que a base de amostras fosse muito grande, apesar do esforço para torná-la o mais variada possível, em termos das atributos humanos estudados. Sugere-se que trabalhos futuros busque-se trabalhar com uma base de sinais maior. O trabalho também focou apenas na extração de características de *pitch* vocal do sinal, e, portanto, futuras pesquisas podem procurar extração de características temporais como energia , ZCR e volume, ou qualitativas como respiração e irregularidades na voz, por exemplo.Ademais, recomenda-se examinar o desempenho de outros classificadores como *Support Vector Machi-*

nes ou *Perceptron*.

5.1 Nota Pessoal

Em geral, o projeto contribuiu muito para meu desenvolvimento pessoal e amadurecimento. Possibilitou-me explorar assuntos mais aprofundados no campo de processamento de sinais, os quais, certamente, corroboraram para um ganho de conhecimento por minha parte. As dificuldades encontradas foram principalmente na fase de coleta dos sinais de fala dos diversos interlocutores. A maioria das pessoas geralmente fica bastante suspeita ao ser inquirida por suas medidas físicas como peso e altura, oferecendo bastante resistência. Espero que o presente trabalho possa ter colaborado, mesmo que minimamente, para o enriquecimento da área, bem como inspirar novos estudantes a adquirirem interesse pelo assunto.

Referências

AGHA, A. Voice, footing, enregisterment. *Journal of Linguistic Anthropology*, v. 15, p. 38–59, 2005.

BATTHALLA, S.; SWARNKAR, M.; HUBBALLI, N.; NATU, M. Voip profiler: Profiling voice over ip user communication behavior. In: IEEE. *2016 11th International Conference on Availability, Reliability and Security (ARES)*. [S.l.], 2016. p. 312–320.

BENTO, B. M. P.; KUX, H. J. H.; KÖRTING, T. S. Assessment of classifiers through decision tree and regression tree algorithms in urban area using worldview-2 image. 2019.

GOVE, R.; FAYTONG, J. Machine learning and event-based software testing: Classifiers for identifying infeasible gui event sequences. In: . [S.l.]: Elsevier, 2012, (Advances in Computers, v. 86). p. 109–135.

GOYAL, S.; PATAGE, V. V.; TIWARI, S. Gender and age group predictions from speech features using multi-layer perceptron model. In: IEEE. *2020 IEEE 17th India Council International Conference (INDICON)*. [S.l.], 2020. p. 1–6.

GUIDO, R. C. A tutorial review on entropy-based handcrafted feature extraction for information fusion. *Information Fusion*, v. 41, p. 161–175, 2018. ISSN 1566-2535. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1566253517303846>.

HOLLIEN, H. *The acoustics of crime: the new science of forensic phonetics*. Berlin, Germany: Springer Science & Business Media, 2013.

KALLURI, S. B.; VIJAYASENAN, D.; GANAPATHY, S. Automatic speaker profiling from short duration speech data. *Speech Communication*, Elsevier, v. 121, p. 16–28, 2020.

KEANE, W. Voice. *Journal of Linguistic Anthropology*, v. 9, p. 271–273, 1999.

KHAN, D. A. *Representation Learning for Voice Profiling*. 2019. Submitted in partial fulfillment of the requirements for the degree of Master of Science in Computer Science.

MALINOW, K. L.; LYNCH, J. J.; THOMAS, S. A.; FRIEDMANN, E.; LONG, J. M. . Automated blood pressure recording: The phenomenon of blood pressure elevations during speech. *Angiology*, v. 33 (7), p. 474–479, 1982.

MIA, M. M. A.; BISWAS, S. K.; URMI, M. C.; SIDDIQUE, A. An algorithm for training multilayer perceptron (mlp) for image reconstruction using neural network without overfitting. *International Journal of Scientific & Technology Research*, v. 4, n. 02, p. 271–275, 2015.

MISRA, S.; LI, H. Chapter 9 - noninvasive fracture characterization based on the classification of sonic wave travel times. In: *Machine Learning for Subsurface Characterization*. [S.l.]: Gulf Professional Publishing, 2020. p. 243–287.

NAKARAT, T.; LäßIG, A. K.; LAMPE, C.; KEILMANN, A. Alterations in speech and voice in patients with mucopolysaccharidoses. *Journal of Linguistic Anthropology*, v. 39, p. 30–37, 2014.

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

SAFAVIAN, S. R.; LANDGREBE, D. A survey of decision tree classifier methodology. *IEEE transactions on systems, man, and cybernetics*, IEEE, v. 21, n. 3, p. 660–674, 1991.

SANCHEZ, F. L.; BARBON, S.; VIEIRA, L. S.; GUIDO, R. C.; FONSECA, E. S.; SCALASSARA, P. R.; MACIEL, C. D.; PEREIRA, J. C.; CHEN, S.-H. Wavelet-based cepstrum calculation. *Journal of Computational and Applied Mathematics*, v. 227, n. 2, p. 288–293, 2009. ISSN 0377-0427. Special Issue on Emergent Applications of Fractals and Wavelets in Biology and Biomedicine. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0377042708001118>.

SCHILLING, N.; MARSTERS, A. Unmasking identity: Speaker profiling for forensic linguistic purposes. *Annual Review of Applied Linguistics*, Cambridge University Press, v. 35, p. 195–214, 2015.

SHANNON, C. E. A mathematical theory of communication. *The Bell System Technical Journal*, v. 27, n. 3, p. 379–423, 1948.

SHEN, J.; CAO, J.; LEDERMAN, O.; TANG, S.; PENTLAND, A. ser profiling based on nonlinguistic audio data. *ACM Transactions on Information Systems (TOIS)*, ACM New York, NY, v. 40, n. 1, p. 1–23, 2021.

SILVA, L. M. Uma aplicação de árvores de decisão, redes neurais e knn para a identificação de modelos arma não sazonais e sazonais. *Rio de Janeiro. 145p. Tese de Doutorado-Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro*, 2005.

SINGH, R. *Profiling Humans From Their Voice*. 1. ed. Pittsburgh, PA, USA: Springer, 2019.

ZEMLIN, W. R. Speech and hearing science: Anatomy and physiology. *Otology & Neurotology*, v. 4 (2), p. 186–187, 1982.

Apêndice A

Fórmulas matemáticas das características extraídas

$$\overline{F0} = \frac{1}{N}\sum_{i=1}^{N} F0_i$$

Média

$$\overline{F0}_g = \sqrt[N]{\prod_{i=1}^{N} F0_i}$$

Média geométrica

$$F0_{rms} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} F0_i^2}$$

Raiz da média dos quadrados (RMS)

$$\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(F0_i - \overline{F0}_i)^2}$$

Dispersão

$$\frac{\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(F0_i - \overline{F0}_i)^2}}{\overline{F0}}$$

Coeficiente de variação

$$\frac{\frac{1}{N}\sum_{i=1}^{N}|F0_i - \overline{F0}_i|}{\overline{F0}}$$

Variabilidade percentual

$$\frac{1}{N-1}\sum_{i=1}^{N-1}|T0_i - T0_{i+1}|$$

Jitter

$$100 \times \frac{\frac{1}{N-1}\sum_{i=1}^{N}|T0_i - T0_{i+1}|}{\frac{1}{N}\sum_{i=1}^{N}T0_i}$$

***Jitter* percentual**

$$\frac{1}{max_j(T0_j)}\frac{1}{N}\sum_{i=1}^{N-1}|T0_{i+1} - T0_i|$$

Média normalizada do *jitter* absoluto

$$\frac{\frac{1}{N-4}\sum_{i=1}^{N-4}\frac{1}{5}\sum_{r=0}^{4}|T0_{i+r} - T0_{i+2}|}{\sum_{i=1}^{N}T0_i}$$

Perturbação do período de *pitch*

$$DLT = \frac{1}{N-5}\sum_{i=1}^{N-2}\left|\frac{T0_{i-2} + T0_{i+2}}{2} - T0_i\right|$$

Desvio da tendência linear (DLT)

$$\frac{\frac{1}{N-2}\sum_{i=2}^{N-1}\left|\frac{T0_{i-1}+T0_{i+1}}{2}-T0_i\right|}{\frac{1}{N}\sum_{i=1}^{N}T0_i}$$

Fator de perturbação sem tendência linear

$$PVI = 1000 \times \frac{\frac{1}{N}\sum_{i=1}^{N}(T0_i-\overline{T0})^2}{\overline{T0^2}}$$

Índice de variabilidade de *pitch* (PVI)

$$FPQ = \frac{\frac{1}{N-2}\sum_{i=1}^{N-2}\left|\frac{F0_{i-1}+F0_i+F0_{i+1}}{3}-F0_i\right|}{\sum_{i=1}^{N}F0_i}$$

Quociente de perturbação de frequência (FPQ)

$$RAP = \frac{\frac{1}{N-2}\sum_{i=1}^{N-2}\left|\frac{T0_{i-1}+T0_i+T0_{i+1}}{3}-T0_i\right|}{\sum_{i=1}^{N}T0_i}$$

Perturbação média relativa (RAP)

$$\frac{max(T0)-min(T0)}{max(T0)+min(T0)}$$

Modulação de $F0$