UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA
DEPARTAMENTO DE ENGENHARIA ELÉTRICA

FELIPPE FAVATI DE SOUZA

USO DE ALGORITMOS DE MACHINE LEARNING PARA AUXILIAR NO DIAGNÓSTICO DE PACIENTES PORTADORES DE DISFONIA ESPASMÓDICA

SÃO CARLOS - SP
2024

FELIPPE FAVATI DE SOUZA

USO DE ALGORITMOS DE MACHINE LEARNING PARA AUXILIAR NO DIAGNÓSTICO DE PACIENTES PORTADORES DE DISFONIA ESPASMÓDICA

Trabalho de Conclusão de Curso apresentado ao Departamento de Engenharia Elétrica do Centro de Ciências Exatas e de Tecnologia da Universidade Federal de São Carlos, para obtenção do título de Bacharel em Engenharia Elétrica.

Orientador: Prof. Dr. Robson Barcellos

São Carlos - SP
2024

DEDICATÓRIA

Este trabalho é dedicado à minha mãe Fabiana e a toda minha família e amigos. Com todo o apoio e incentivo sou capaz de alcançar meus objetivos.

Agradecimentos

Meus agradecimentos iniciais faço a minha mãe Fabiana, por todo apoio e confiança durante minha graduação. Seu carinho e companheirismo sempre me fortaleceram quando necessário.

Segundo agradeço aos meus avós Maria Elisa e Antônio José, que sempre acreditaram em mim e me apoiaram com imensuráveis esforços para que essa conquista pudesse ser alcançada.

Agradeço também ao meu primo, padrinho e irmão de consideração Otávio. Além de ser para mim um exemplo a ser seguido, sempre me proporcionou direção quando me sentia perdido.

Agradeço as minhas irmãs Bárbara e Maria Clara por toda calma e tranquilidade que me proporcionaram durante todo esse período.

Agradeço a minha namorada Marinara, por nunca ter me deixado desistir, mesmo quando me sentia exausto. Seu carinho e paciência foram primordiais para a conclusão desse trabalho.

Agradeço também meu grande amigo Paulo. Por todos os anos de amizade e história que temos. Todas as nossas conversas sobre construir nosso futuro me inspiram a sempre procurar ser uma pessoa melhor.

Agradeço ao meu amigo Bruno. Quantas madrugadas de estudos nós tivemos, seu apoio foi essencial por toda minha graduação e espero levar nossa amizade para o resto da vida.

Um agradecimento especial ao meu orientador Prof. Dr. Robson Barcellos, cujo conhecimento, dedicação e apoio foram impecáveis durante todo o desenvolvimento desse trabalho.

Meus agradecimentos finais ficam para todos os amigos com quem tive a oportunidade de dividir um lar durante minha graduação. Todas as histórias que vivenciamos juntos são belíssimas memórias que carregarei comigo por toda a vida.

“Lendas não são definidas pelo seu sucesso,
elas são definidas por como se recuperam de
seus fracassos”
(Chris Bosh)

Resumo

A disfonia espasmódica (DE) é um distúrbio vocal raro de etiologia neurológica que acomete pacientes por volta dos 30 anos de idade, dificultando o processo de produção vocal e podendo ser agravado por situações de estresse. Dado sua origem neurológica, seu diagnóstico por profissionais da área da fonoaudiologia é difícil e depende muito da *expertise* do médico avaliador para reconhecer a doença. Isso pode levar o paciente a não receber o tratamento ideal, acarretando a persistência dos sintomas. Com a evolução da tecnologia, a busca por métodos não invasivos para identificação de distúrbios vocais pode ser vantajosa no auxílio do diagnóstico de DE. Este trabalho tem como objetivo avaliar o uso de técnicas de *machine learning* para realizar a classificação de sinal de vozes entre pacientes portadores de DE e pessoas saudáveis.

Palavras-chave: Disfonia Espasmódica; Machine Learning; Aprendizado de Máquina; Random Forest; Decision Tree; Árvore de Decisão; Processamento Vocal; Análise Vocal

Abstract

Spasmodic dysphonia (DE) is a rare vocal disorder of neurological etiology that affects patients around the age of 30, making the vocal production process difficult and can be aggravated by stressful situations. Given its neurological origin, its diagnosis is difficult for professionals in the field of speech therapy, depending largely on the expertise of the evaluating physician to recognize the disease. This can lead to the patient not receiving the ideal treatment, resulting in the persistence of the symptoms. With the evolution of technology, the search for non-invasive methods to identify voice disorders may be advantageous in aiding the diagnosis of DE. This work aims to evaluate the use of machine learning techniques to classify voice signals between patients with DE and healthy people.

Keyword: Spasmodic Dysphonia; Machine Learning; Random Forest; Decision Tree; Voice Processing; Vocal Analysis

Lista de ilustrações

Lista de tabelas

Lista de Abreviaturas

APQ	*Amplitude perturbation quotient*
DCT	*Discrete Cosine Transform*
DE	Disfonia Espasmódica
FFT	*Fast Fourier Transform*
J_1	Jitter Local Absoluto
J_2	Jitter Local
J_3	Rap Jitter
J_4	Jitter ppq5
J_5	Jitter ddp
MFCC	*Mel-Frequency Cepstral Coefficient*
ML	*Machine Learning*
ReLU	*Rectified Linear Unit*
RF	*Random Forest*
RN	Rede Neural
S_1	Shimmer local
S_2	Shimmer local dB
S_3	Shimmer APQ_3
S_4	Shimmer APQ_5
S_5	Shimmer APQ_{11}
S_6	Shimmer dda

Sumário

1. INTRODUÇÃO

A disfonia vocal é um diagnóstico clínico para dificuldade na produção da voz, comumente utilizado quando uma pessoa produz voz com irregularidades, afetando cerca de 30% da população global em algum momento durante sua vida (TULICS, 2019). O foco desse trabalho é a disfonia espasmódica (DE), um distúrbio de origem neurológica que afeta os músculos responsáveis pela produção vocal resultando na dificuldade da fala. A DE é uma patologia considerada rara, com uma prevalência de 3,5 a 7 pessoas a cada 100.000, atingindo predominantemente mulheres, cerca de 4 vezes mais, a partir dos 30 anos de vida (SANUKI, 2023). Entre os sintomas de DE estão: voz tensa e estrangulada, interrupções súbitas da produção vocal, voz com característica soprosa, dificuldade e esforço para produção vocal (BARKMEIER; CASE; LUDLOW, 2001). A DE pode ser dividida em dois fenótipos: disfonia espasmódica de adução, que é a mais comum, e disfonia espasmódica de abdução, que é relativamente rara. Disfonia espasmódica mista envolve características dos dois (SANUKI, 2023). A DE adutora é caracterizada por espasmos dos músculos adutores da laringe, resultando em voz estrangulada ou interrompida, acarretando dificuldade em iniciar ou manter a fala. Já a DE abdutora é caracterizada por músculos abdutores da laringe, resultando em voz fraca ou sussurrada, acarretando a dificuldade de produzir som vocal audível.

Pacientes com DE usualmente buscam tratamento com profissionais da área da fonoaudiologia, devido à dificuldade na fala, porém dado que a etiologia do distúrbio é neurológica, o diagnóstico pode muitas vezes ser impreciso, o que leva a um tratamento incorreto do problema do paciente. Imamura e Tsuji (2006) levantaram o questionamento se seria possível realizar o diagnóstico diferencial entre DE, tremor vocal e disfonia de tensão muscular. Os autores ressaltam,

> Mesmo para laringologistas experientes, acostumados com estas afecções, o diagnóstico pode ser difícil. O diagnóstico incorreto pode conduzir a um tratamento malsucedido. Indicar injeção de toxina botulínica, tratamento preconizado para disfonia espasmódica, para um paciente com disfonia de tensão muscular, por exemplo, pode resultar em um quadro grave de incompetência glótica, com voz extremamente rouca e aspiração a alimentos e saliva.

Nesse contexto o emprego da tecnologia na área médica pode ser de grande

valia para possibilitar o desenvolvimento de um método não invasivo para auxiliar na identificação de distúrbios vocais. Uma das alternativas se dá pelo uso de *machine learning* com algoritmos de classificação para analisar gravações da voz de pacientes e diagnosticar com base nos dados previamente conhecidos.

1.1. OBJETIVOS

O objetivo do trabalho é analisar a viabilidade do emprego de técnicas de *machine learning* juntamente com algoritmos de classificação no auxílio do diagnóstico de pacientes portadores de DE. Para isso, foi adotado como satisfatório uma classificação que alcance pelo menos 90% de acurácia.

2. FUNDAMENTAÇAO TEÓRICA

Neste capítulo serão apresentados: o conceito de *machine learning* (ML) junto com suas abordagens principais, os algoritmos aplicados no desenvolvimento do trabalho: redes neurais artificiais (RNA) e *Random Forest* (RF) e características extraídas do sinal de áudio, sendo elas: Jitter, Shimmer e os coeficientes mel-cepstrais (em inglês *mel-frequency cepstral coefficients* MFCCs).

2.1. MACHINE LEARNING

Machine learning, ou aprendizado de máquina, é um subcampo da inteligência artificial que se concentra no desenvolvimento de algoritmos e técnicas que permitem aos computadores aprenderem a partir de dados. O objetivo principal do ML é permitir que os sistemas computacionais automaticamente melhorem seu desempenho em tarefas específicas à medida que são expostos a mais dados.

O conceito fundamental por trás do ML é a capacidade de identificar padrões nos dados e utilizar esses padrões para fazer previsões ou tomar decisões. Em vez de programar explicitamente um computador para realizar uma tarefa, no machine learning, o computador é treinado usando exemplos de dados para aprender como realizar a tarefa por conta própria.

O ML é uma ferramenta poderosa com uma ampla gama de aplicações em diversas áreas, incluindo reconhecimento de padrões, processamento de linguagem natural, visão computacional, medicina, finanças, entre outros. À medida que os conjuntos de dados e os recursos computacionais continuam a crescer, o potencial do ML para resolver problemas complexos e tomar decisões automatizadas só tende a aumentar.

Existem várias abordagens diferentes para o ML, A seguir estão listadas as mais utilizadas, uma breve explicação de seu conceito teórico e principais aplicações.

2.1.1. Aprendizado Supervisionado

No aprendizado supervisionado, o algoritmo é treinado em um conjunto de dados rotulados, onde cada exemplo de treinamento consiste em um par de entrada e saída desejada. O objetivo é aprender uma função que mapeie os dados de entrada para as saídas desejadas, de modo que o modelo possa fazer previsões precisas em novos dados para os quais as saídas corretas são desconhecidas.

Existem dois tipos principais de problemas em aprendizado supervisionado:

- Classificação: Neste tipo de problema, as saídas desejadas são categorias ou classes discretas. Por exemplo, determinar se um e-mail é spam ou não spam, ou se uma imagem contém um gato ou um cachorro.
- Regressão: Aqui, as saídas desejadas são valores contínuos. Por exemplo, prever o preço de uma casa com base em suas características, ou prever a temperatura com base em variáveis meteorológicas.

Algoritmos populares em aprendizado supervisionado incluem regressão linear, regressão logística, árvores de decisão, redes neurais artificiais e máquinas de vetores de suporte.

2.1.2. Aprendizado não Supervisionado

No aprendizado não supervisionado, o algoritmo é treinado em um conjunto de dados não rotulados, onde as saídas desejadas são desconhecidas. O objetivo é descobrir estruturas intrínsecas ou padrões nos dados, como agrupamentos naturais ou relações entre as variáveis.

Os principais tipos de algoritmos em aprendizado não supervisionado incluem:

- Agrupamento (*Clustering*): Agrupar os dados em conjuntos distintos, onde os elementos dentro de um mesmo grupo são mais semelhantes entre si do que com os elementos de outros grupos.
- Redução de Dimensionalidade: Reduzir a dimensionalidade dos dados mantendo o máximo de informações possível. Isso é útil para visualização de dados e para remover redundâncias.

O aprendizado não supervisionado é frequentemente usado em tarefas como segmentação de clientes, análise de texto não estruturado e reconhecimento de padrões.

2.1.3. Aprendizado por Reforço

No aprendizado por reforço, o agente aprende a interagir com um ambiente dinâmico para maximizar uma recompensa cumulativa ao longo do tempo. O agente toma decisões sequenciais em um ambiente incerto e aprende a melhor política de ação por meio de tentativa e erro.

Os principais componentes do aprendizado por reforço são:

- Agente: O sistema de aprendizado que toma decisões.
- Ambiente: O mundo no qual o agente interage.
- Ações: As ações disponíveis que o agente pode executar no ambiente.
- Recompensas: *Feedback* imediato do ambiente que informa ao agente a qualidade de suas ações.
- Política: A estratégia adotada pelo agente para selecionar ações em cada estado do ambiente.

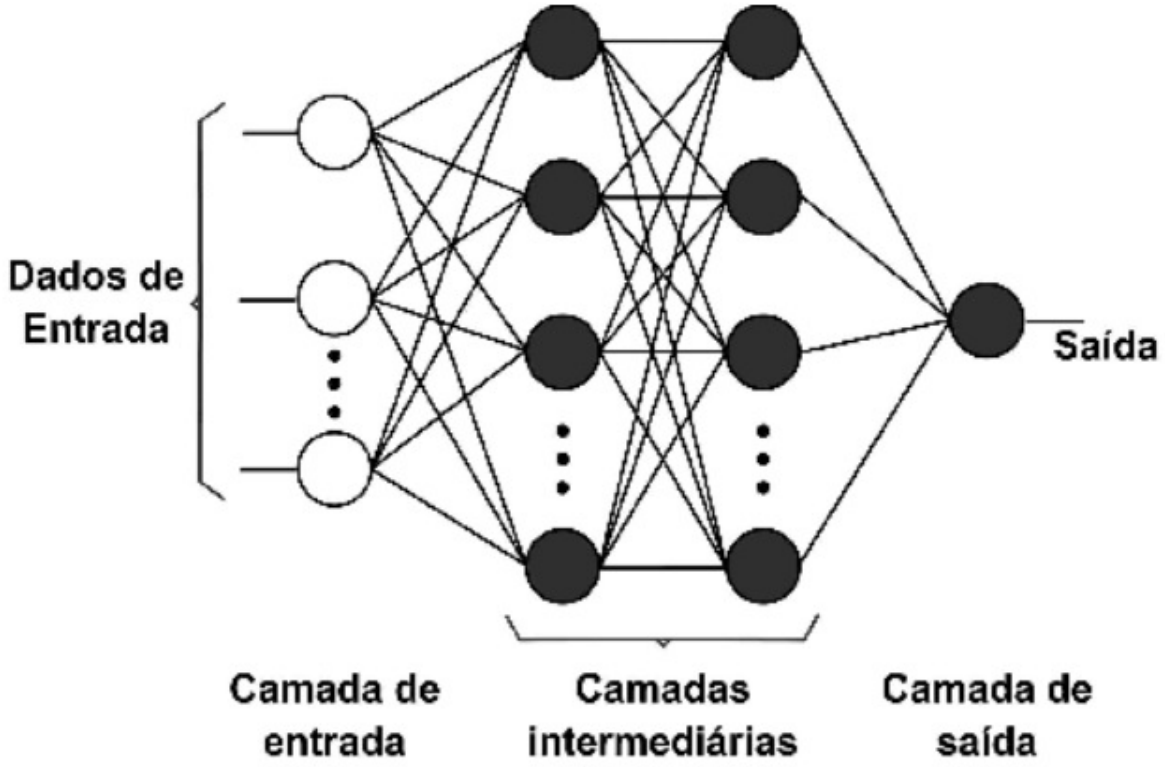
2.2. REDE NEURAL

Uma rede neural (RN) é um modelo computacional composto por unidades de processamento interconectadas, chamadas neurônios artificiais, organizadas em camadas. Cada neurônio recebe entradas, aplica uma função de ativação e produz uma saída. Essas camadas são:

- Camada de Entrada: Recebe os dados brutos do problema.
- Camadas Ocultas: Camadas intermediárias entre a entrada e a saída, onde ocorre o processamento principal. Cada neurônio nessas camadas recebe entradas das camadas anteriores e produz uma saída que é propagada para as camadas seguintes.
- Camada de Saída: Produz a saída final da rede neural.

A Figura 1 exibe uma ilustração de uma rede neural artificial.

Figura 1 - Ilustração de uma rede neural artificial



Fonte: FIORIN, 2011

Durante o treinamento, os pesos das conexões entre os neurônios são ajustados iterativamente usando um algoritmo de otimização, como gradiente descendente, para minimizar uma função de perda que quantifica o erro entre as saídas previstas e as saídas reais. Esse processo de ajuste dos pesos permite que a rede aprenda a melhor representação dos dados para realizar uma determinada tarefa.

RN são especialmente adequadas para lidar com problemas complexos e não lineares, como reconhecimento de imagens, processamento de linguagem natural e previsão de séries temporais. Elas têm se destacado em uma variedade de aplicações devido à sua capacidade de aprender representações hierárquicas dos dados e realizar generalizações robustas.

2.2.1. Funções de Ativação

As funções de ativação são responsáveis por adicionar a não-linearidade nos modelos das RN, sem elas uma RN seria apenas uma regressão linear. Existem diversos tipos de funções de ativação, mas nesse trabalho o foco será em apenas

dois desses tipos, sendo eles: a unidade linear retificada (do inglês *Rectified Linear Unit* ou ReLU) e sigmoide.

A função ReLU é amplamente adotada em modelos de deep learning, visto que sua função matemática é simples e computacionalmente eficiente. Observando a equação 1 que descreve a função ReLU, nota-se que ela leva valores negativos para zero, isso resulta em uma rede esparsa, melhorando a eficiência do modelo e sua interpretabilidade.

$$ReLU(x) = \max(0, x) \tag{1}$$

Já a função sigmoide é comumente utilizada para classificação binária, visto que ela transforma o valor de entrada para um valor entre 0 e 1. Sua curva visa mapear valores grande e negativos para valores próximos a 0 e valores grandes e positivos para valores próximos a 1. A equação 2 descreve a função de ativação sigmoide.

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{2}$$

2.3. ÁRVORE DE DECISÃO

Um modelo de árvore de decisão é uma estrutura hierárquica semelhante a um fluxograma, composta por nós e arestas, onde cada nó representa uma decisão ou um ponto de divisão com base em uma característica específica dos dados. Essa estrutura é construída de forma recursiva dividindo o conjunto de dados em subconjuntos menores com base nas características, de modo a maximizar a pureza das classes em cada subconjunto.

Aqui estão os componentes principais de um modelo de árvore de decisão:

- Nó Raiz: O nó superior da árvore, que representa a característica de divisão inicial. Este nó é dividido em outros nós com base nos valores dessa característica.
- Nós Internos: Nós intermediários na árvore que representam decisões com base nas características dos dados. Cada nó interno possui uma condição que determina qual caminho seguir para o próximo nível da árvore.
- Nós Folha: Os nós finais da árvore, que representam as classes de saída ou as previsões. Cada nó folha contém uma classe ou um valor previsto.

- Arestas: As conexões entre os nós, representando o fluxo de decisão com base nas características dos dados.

A construção de uma árvore de decisão envolve encontrar as melhores características e pontos de divisão para maximizar a separação entre as classes. Isso é feito usando uma métrica de impureza, como o índice de Gini ou a entropia, para medir a homogeneidade dos subconjuntos resultantes. O objetivo é encontrar divisões que levem a subconjuntos mais puros ou mais homogêneos em termos de classes.

Existem várias estratégias para construir árvores de decisão, incluindo:

- Divisão Binária: Cada nó é dividido em dois subconjuntos, um para cada ramo.
- Corte Antecipado (*Pruning*): Remoção de ramos da árvore que não contribuem significativamente para a redução da impureza ou do erro de classificação.
- Seleção de Características: Escolha das melhores características para divisão em cada nó com base em critérios como ganho de informação ou redução de impureza.

Os modelos de árvore de decisão são simples de entender e interpretar, tornando-os populares em muitas aplicações. Eles são frequentemente usados em tarefas de classificação e regressão e podem ser facilmente visualizados para insights sobre como as decisões são tomadas com base nos dados. No entanto, árvores de decisão simples tendem a ter alto viés e baixa variância, o que pode levar ao sobreajuste em conjuntos de dados complexos. Esses problemas podem ser mitigados usando técnicas como *Random Forest* ou poda da árvore.

2.4. RANDOM FOREST

Random Forest (RF) é um algoritmo de aprendizado de máquina baseado em árvores de decisão que opera por meio da construção de uma grande quantidade de árvores de decisão durante o treinamento e agregando suas previsões para obter uma resposta final.

Durante a construção de cada árvore na floresta, um subconjunto aleatório dos dados de treinamento é amostrado com substituição (*bootstrapping*) e um subconjunto aleatório das características (variáveis) é selecionado em cada divisão

da árvore. Essa aleatoriedade introduz diversidade entre as árvores individuais, o que ajuda a reduzir o sobreajuste (*overfitting*) e a aumentar a robustez do modelo.

Durante a fase de previsão, cada árvore na floresta faz uma previsão e a classe mais frequente entre todas as árvores é selecionada como a previsão final.

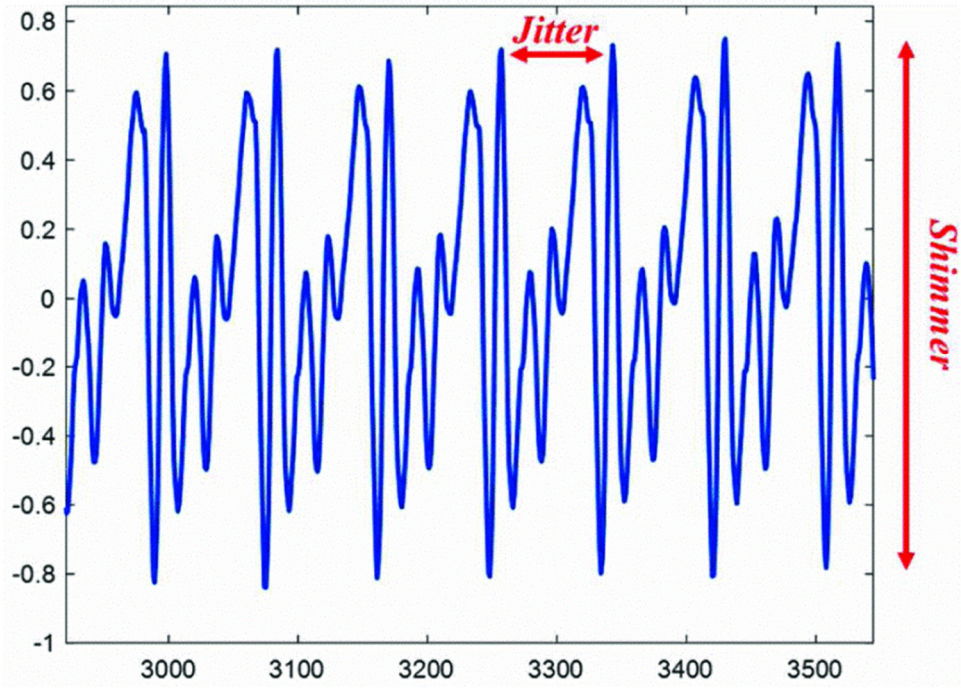
RF é conhecido por sua robustez, escalabilidade e eficácia em uma ampla gama de problemas de classificação e regressão. Ele não requer muita configuração e é menos sensível a dados desbalanceados e outliers em comparação com outros modelos. Além disso, sua capacidade de calcular a importância das características ajuda a identificar quais variáveis são mais relevantes para a previsão.

2.5. JITTER

O jitter é uma medida da variação de frequência entre ciclos sucessivos de uma onda sonora, especialmente no contexto da análise de voz. Em termos simples, ele quantifica a instabilidade temporal do período de uma onda sonora. Essa variação é crucial na análise de voz, pois pode ser indicativa de desordens vocais ou patologias laringianas.

O jitter é utilizado para avaliar a qualidade vocal, onde um alto nível de jitter pode indicar problemas como disfonia ou tremor vocal. Em pesquisas, ele é frequentemente usado em estudos de fonética, fonoaudiologia e engenharia biomédica. A Figura 2 nos dá uma ideia de qual variação é avaliada pelo jitter e pelo shimmer em um sinal de voz.

Figura 2 – Representação gráfica da grandeza medida pelo jitter e shimmer para um dado sinal de voz



Fonte: HADJAIDJI; KORBA; KHELIL, 2021

2.5.1. Jitter Local Absoluto

O jitter local absoluto (J_1), também conhecido como Jitta, representa a diferença média absoluta entre dois períodos consecutivos ($T_i - T_{i-1}$), onde T_i é a duração do período em segundos e N o número de períodos, portanto o Jitta também é medido em segundos. Ele é definidor por:

$$Jitta = \frac{1}{N-1}\sum_{i=1}^{N-1}|T_i - T_{i-1}| \tag{3}$$

2.5.2. Jitter Local

O jitter local (J_2), também conhecido como Jitt, representa a diferença absoluta entre dois períodos consecutivos divido pelo período médio, sendo expresso como porcentagem. Ele é definidor por:

$$Jitt = \frac{Jitta}{\frac{1}{N}\sum_{i=1}^{N} T_i} \tag{4}$$

2.5.3. Rap Jitter

O rap jitter (J_3), também conhecido como rap, representa a diferença média

absoluta de um período e a média desse período com seus dois vizinhos, dividido pelo período médio, sendo expresso como porcentagem. Ele é definidor por:

$$rap = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}|T_i - \left(\frac{1}{3}\sum_{n=i-1}^{i+1} T_n\right)|}{\frac{1}{N}\sum_{i=1}^{N} T_i} \quad (5)$$

2.5.4. Jitter ppq5

O jitter ppq5 (J_4), representa a diferença absoluta média de um período e a média contendo seus quatro vizinhos dividido pelo período médio, sendo expresso como porcentagem. Ele é definidor por:

$$ppq5 = \frac{\frac{1}{N-1}\sum_{i=2}^{N-2}|T_i - \left(\frac{1}{5}\sum_{n=i-2}^{i+2} T_n\right)|}{\frac{1}{N}\sum_{i=1}^{N} T_i} \quad (6)$$

2.5.5. Jitter ddp

O jitter ddp (J_5), representa a média das diferenças absolutas entre variações sucessivas dos períodos, sendo expresso em segundos. Ele é definidor por:

$$ddp = \frac{1}{N-2}\sum_{i=1}^{N-2}|(T_{i+2} - T_{i+1}) - (T_{i+1} - T_i)| \quad (7)$$

2.6. SHIMMER

O shimmer mede a variação da amplitude entre ciclos sucessivos de uma onda sonora. Assim como o jitter, é uma métrica essencial na análise de voz, mas ao invés de focar na frequência, ele avalia a instabilidade na intensidade da voz. Isso é particularmente útil para detectar irregularidades na produção vocal que podem não ser evidentes apenas pela análise de frequência.

Clinicamente, o shimmer é utilizado para avaliar a estabilidade da intensidade vocal, com altos níveis indicando possíveis desordens vocais como a disfonia. Em pesquisas, ele é uma ferramenta crucial na análise acústica da voz, contribuindo para estudos em fonoaudiologia e em tecnologias de reconhecimento de fala.

2.6.1. Shimmer Local

O shimmer local (S_1), também conhecido como Shim, representa a diferença absoluta média entre duas amplitudes A_i e A_{i+1} de dois períodos consecutivos T_i e T_{i+1}, dividido pela amplitude média, sendo expresso em porcentagem. Ele é definidor por:

$$Shim = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}|A_i - A_{i+1}|}{\frac{1}{N}\sum_{i=1}^{N} A_i} \quad (8)$$

Onde N é o número de períodos.

2.6.2. Shimmer Local dB

O shimmer local dB (S_2), também conhecido como *ShdB*, representa a diferença absoluta média do logaritmo base 10 da diferença entre dois períodos consecutivos, sendo expresso em decibéis (dB). Ele é definidor por:

$$ShdB = \frac{1}{N-1}\sum_{i=1}^{N-1}|20 * \log\left(\frac{A_{i+1}}{A_i}\right)| \quad (9)$$

2.6.3. Shimmer Coeficiente de Perturbação de Amplitude

O shimmer coeficiente de perturbação de amplitude (em inglês *amplitude perturbation quotient* APQ), representa a variação relativa na amplitude média de ciclos consecutivos, sendo expresso em porcentagem. O shimmer APQ normalmente é calculado levando em conta três diferentes quantidades de ciclos consecutivos: 3, 5 e 11.

2.6.3.1. Shimmer APQ3

O shimmer APQ3 (S_3) é definido por:

$$APQ3 = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}|A_i - \left(\frac{1}{3}\sum_{n=i-1}^{i+1} An\right)|}{\frac{1}{N}\sum_{i=1}^{N} A_i} \quad (10)$$

2.6.3.2. Shimmer APQ5

O shimmer APQ5 (S_4) é definido por:

$$APQ5 = \frac{\frac{1}{N-1}\sum_{i=2}^{N-2}|A_i - \left(\frac{1}{5}\sum_{n=i-2}^{i+2} An\right)|}{\frac{1}{N}\sum_{i=1}^{N} A_i} \quad (11)$$

2.6.3.3. Shimmer APQ11

O shimmer APQ11 (S_5) é definido por:

$$APQ11 = \frac{\frac{1}{N-1}\sum_{i=5}^{N-5}|A_i - \left(\frac{1}{11}\sum_{n=i-5}^{i+5} An\right)|}{\frac{1}{N}\sum_{i=1}^{N} A_i} \quad (12)$$

2.6.4. Shimmer dda

O shimmer dda (S_6), representa a média das diferenças absolutas entre ciclos sucessivos, sendo expresso em segundos. Ele é definidor por:

$$dda = \frac{1}{N-1}\sum_{i=1}^{N-1}|A_{i+1} - A_i| \quad (13)$$

2.7. COEFICIENTES MEL-CEPSTRAIS

Os coeficientes mel-cepstrais são uma representação compacta do espectro de potência de um sinal. Eles são amplamente utilizados em processamento de sinais de áudio, especialmente em reconhecimento de fala e análise acústica porque capturam as características relevantes do espectro de áudio de maneira consistente com a forma como os humanos percebem o som. O ouvido humano percebe frequências de maneira não linear, sendo mais sensível a diferenças em frequências baixas do que em frequências altas. A escala mel reflete essa propriedade.

2.7.1. Escala Mel

A escala mel é uma escala perceptual de tons que surgiu para identificar como diversas frequências eram percebidas pelo aparelho auditivo humano. Ela foi desenvolvida experimentalmente por Stevens, Volkmann e Newman (1937), onde o objetivo principal do experimento era descrever uma relação entre frequência real e o que era interpretado pelo ouvido humano.

A fórmula atual para conversão da frequência em Hertz para a escala mel foi proposta por Makhoul e Cossel (1976) e é dada pela equação 14, onde f é a

frequência em Hertz e m é a frequência na escala mel.

$$m = 2595log_{10}\left(1 + \frac{f}{700}\right) \quad (14)$$

2.7.2. Cepstrum

O *cepstrum* é uma ferramenta matemática que consiste em calcular a transformada discreta do cosseno (*discrete cosine transform* DCT) no logaritmo da energia e é utilizado para analisar estruturas periódicas em um espectro de frequência.

2.7.3. Cálculo dos MFCCs

Os MFCC são calculados através dos 7 passos listados a seguir.

1. Pré-ênfase

Antes do processamento, aplica-se um filtro de pré-ênfase ao sinal de áudio para amplificar as frequências altas.

$$y(t) = x(t) - \alpha x(t-1) \quad (15)$$

Onde α é tipicamente entre 0.95 e 1, $y(t)$ é o valor resultante do filtro e $x(t)$ é o sinal de voz discretizado.

2. Segmentação em janelas

O sinal de áudio é divido em pequenos segmentos temporais (*frames*) usualmente de 20 a 40 ms de duração, com sobreposição entre as janelas para garantir continuidade, tipicamente a sobreposição é de 50%.

3. Aplicação da transformada rápida de Fourier (FFT)

Para cada frame calcula-se a FFT de modo a obter o espectro de frequência.

$$X_k[m] = \sum_{n=0}^{N-1} x_k[n] e^{-\frac{j2\pi mn}{N}}, \qquad k = 0, 1, \dots, N-1 \quad (16)$$

Onde N é o número de amostras na janela, $x_k[n]$ o sinal de áudio segmentado e ponderado pela k-ésima janela, m o índice bin de frequência na FFT, j a unidade imaginária, n o índice da amostra no domínio do tempo, k o índice da janela e $X_k[m]$ o resultado da FFT no bin de frequência m para janela k.

4. Conversão para escala mel

As magnitudes do espectro obtido são então mapeadas para a escala mel

utilizando a fórmula (14).

5. Banco de filtros triangulares

Aplica-se um banco de filtros triangulares na escala mel ao espectro de potência. Cada filtro é centrado em uma frequência mel diferente.

$$H_m(k) = \begin{cases} 0 & se\ k < f(m-1) \\ \dfrac{k - f(m-1)}{f(m) - f(m-1)} & se\ f(m-1) \le k \le f(m) \\ \dfrac{f(m+1) - k}{f(m+1) - f(m)} & se\ f(m) \le k \le f(m+1) \\ 0 & se\ k > f(m+1) \end{cases} \quad (17)$$

Onde $f(m)$ são as frequências centrais dos filtros, m o índice do filtro triangular, k o índice do bin de frequência na FFT e $H_m(k)$ a resposta em frequência do m-ésimo filtro triangular.

6. Logaritmo da energia

Para cada filtro triangular, calcula-se a energia logarítmica.

$$E_m = \log\left(\sum_{k=f(m-1)}^{f(m+1)} |X_k[m]|^2 H_m(k)\right) \quad (18)$$

Onde E_m é a energia logarítmica do m-ésimo filtro triangular.

7. Cepstrum

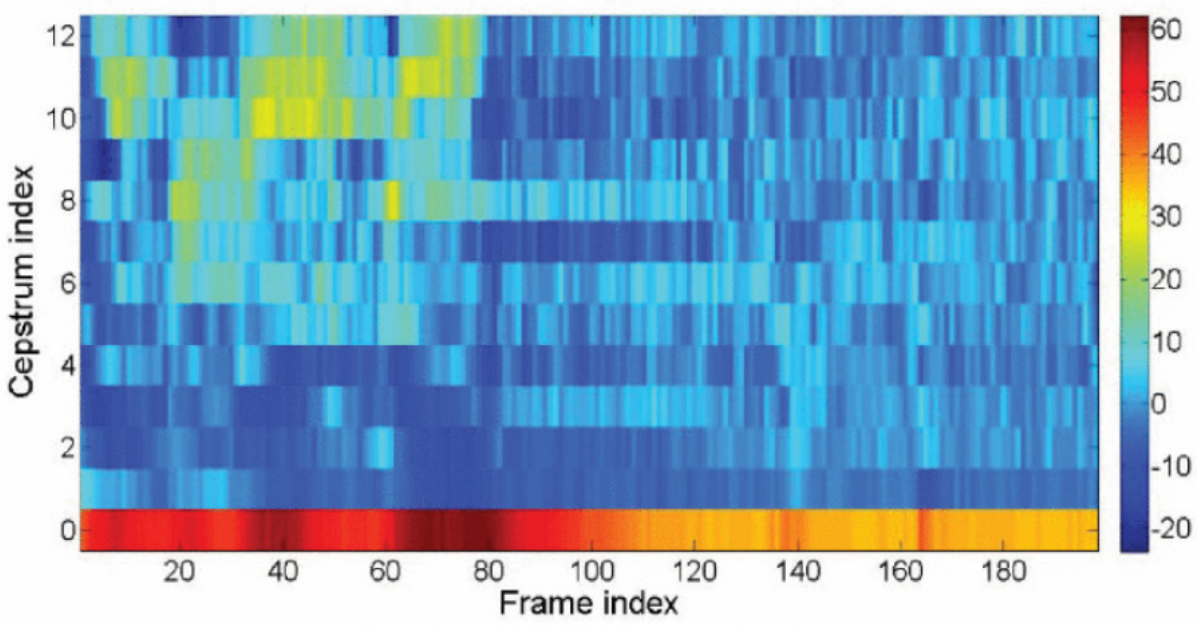
Por último calcula-se o *cepstrum* aplicando a DCT.

$$C_n = \sum_{m=1}^{M} E_m \cos\left[n\left(m - \frac{1}{2}\right)\frac{\pi}{M}\right], \qquad n = 0, 1, \dots, L-1 \quad (19)$$

Onde C_n é o n-ésimo coeficiente Mel Cepstral (MFCC), Mo número de filtros no banco de filtros Mel e n o índice do coeficiente cepstral.

O resultado são os MFCC, que representam a forma geral do espectro de potência do sinal de áudio. A Figura 3 mostra uma representação gráfica dos MFCC.

Figura 3 – Representação gráfica dos MFCC para um dado sinal de voz



Fonte: ISLAM; TARIQUE; ABDEL-RAHEEM, 2020

3. DESENVOLVIMENTO

Nesse capítulo serão detalhadas as etapas realizadas durante a realização do trabalho.

3.1. OBTENÇÃO DOS DADOS

Para realização do trabalho decidiu-se procurar um banco de dados preexistente, visto que fazer a coleta de vozes para posteriormente realizar a classificação desejada seria inviável devido ao fato de que a DE é uma doença rara, se torna impraticável encontrar um número razoável de pessoas portadora dessa doença para treinamento e teste dos algoritmos utilizados.

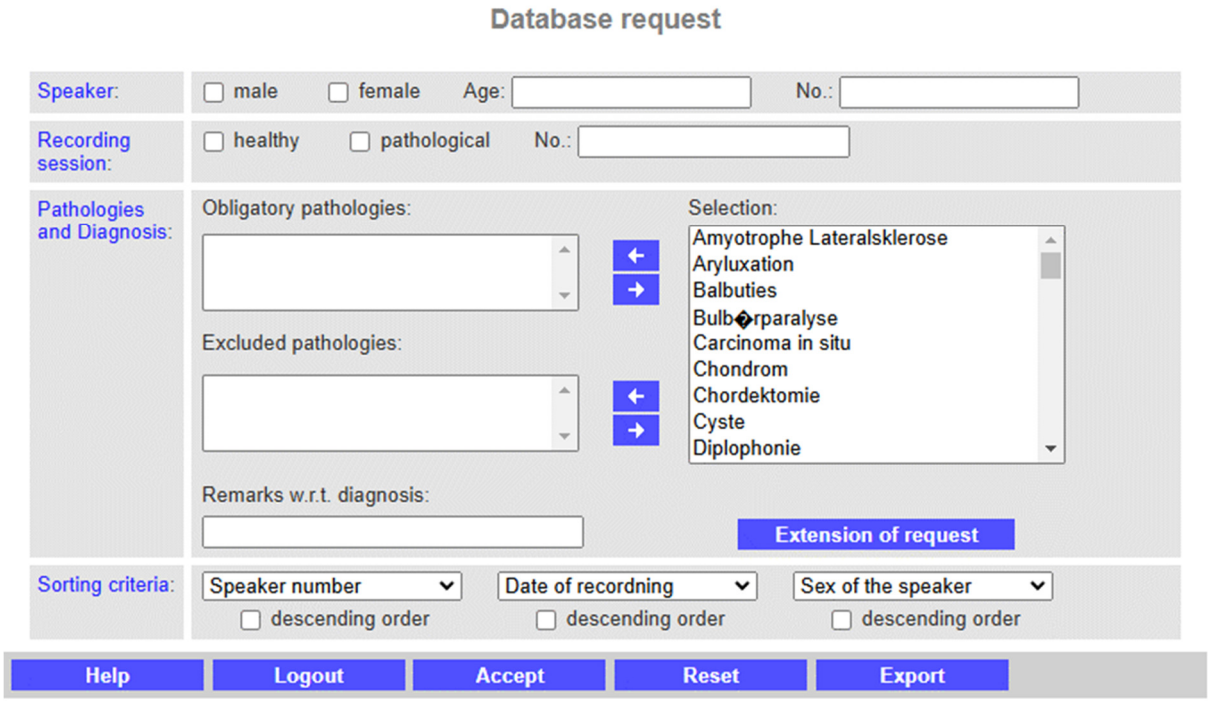
Desse modo, diversos trabalhos na área foram consultados de maneira a identificar uma base de dados que poderia ser utilizada. Diversos trabalhos utilizam a base de dados vocais da Universidade do Sarre, a Saarbruecken Voice Database. Essa base possui os dados públicos, podendo ser acessados por qualquer pessoa.

A base conta com uma interface gráfica onde é possível filtrar os dados que serão retornados. A Figura 4 contém os filtros presentes nessa base dados. Esses filtros estão divididos em 4 categorias. A primeira delas diz respeito ao paciente, os filtros são: gênero, idade e número do paciente. A segunda categoria se refere a sessão da gravação e são: saudável ou patológica e número da sessão. A terceira está relacionada com as patologias, onde é apresentada uma lista com diversas patologias encontradas nas gravações e pode-se escolher patologias obrigatórias, excluir patologias além de estender o filtro para consultas mais elaboradas. A quarta categoria de filtros são apenas os critérios de ordenação.

Além dos filtros a interface apresenta botões com diversas funcionalidades, sendo eles o botão *help*, onde é possível ter acesso a um manual de uso da base escrito em inglês e alemão, o botão *logout* que retorna para a página inicial da base, o botão de *accept*, que serve para aceitar as informações preenchidas no filtro e fazer a requisição de dados para a base, o botão *reset*, que serve para limpar os dados retornados e os filtros configurados, e o botão *export*, que serve para aceitar os

retornos selecionados e seguir para a tela de exportação dos dados de voz.

Figura 4 – Página inicial para requisição de dados da base Saarbruecken Voice Database

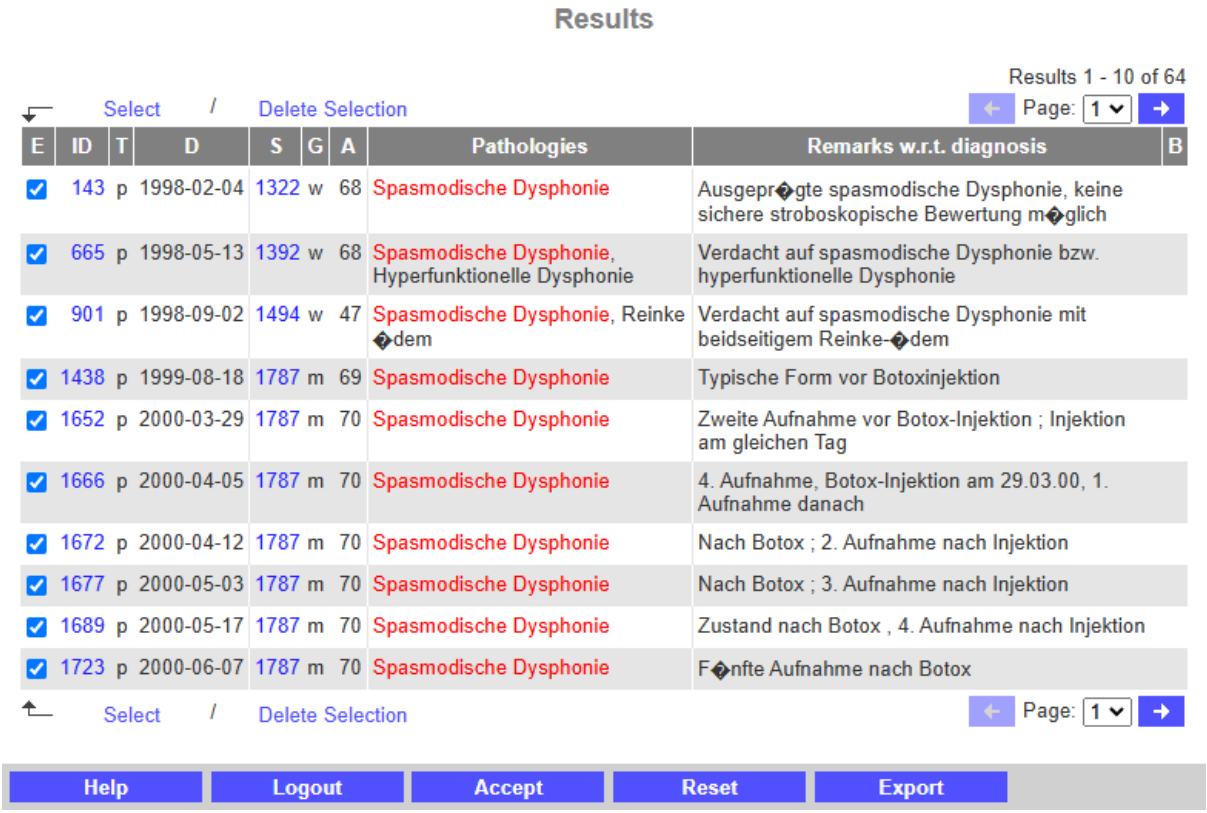


Fonte: PÜTZER; BARRY, 2007

Ainda na tela de requisição, após preencher o filtro e clicar no botão aceitar, é retornado uma lista contendo informações sobre os dados consultados, onde é possível selecionar quais das gravações deseja-se baixar. A lista vem inicialmente com todos os dados selecionados e é possível excluir manualmente cada uma das entradas. A Figura 5 ilustra esse retorno.

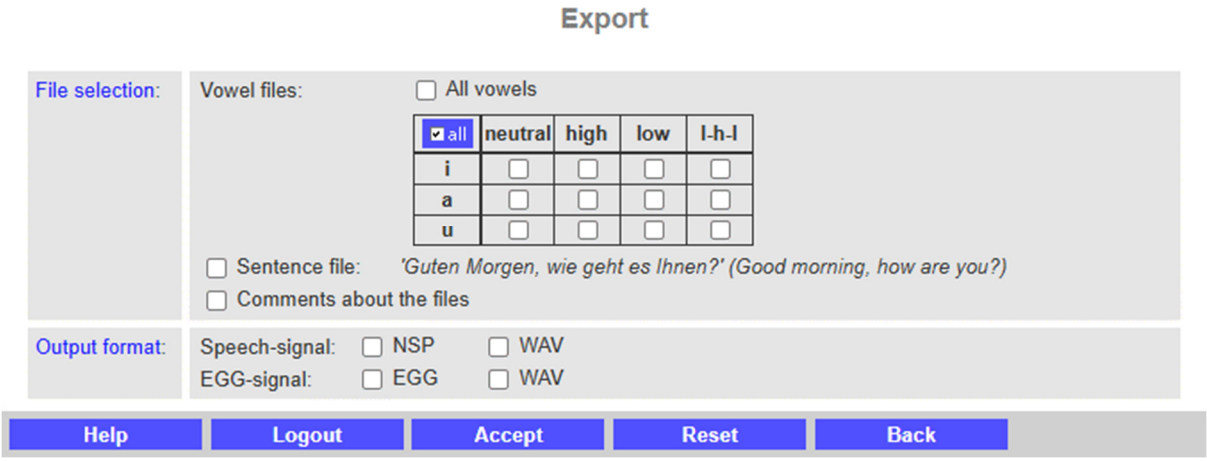
Depois de selecionar os dados e clicar no botão de exportar, a base nos leva para a página de exportação de dados, onde pode-se escolher o que será baixado de cada gravação. Estão disponíveis para seleção as vogais [a, i, u] disponíveis em tom neutro, alto, baixo e ascendente-descendente, e a frase "*Guten Morgen, wie geht es Ihnen?*" ("Bom dia, como você está?" em alemão). Também é possível escolher qual sinal será exportado, vocal e/ou eletroglotógrafo, além de escolher o formato exportado, que pode ser o formato original, NSP para vocal e EGG para eletroglotógrafo, além do formato WAV para ambos. Esse menu está representado na Figura 6.

Figura 5 – Página de retorno da consulta da base para seleção das vozes que serão baixadas



Fonte: PÜTZER; BARRY, 2007

Figura 6 – Página para seleção dos dados que serão baixados referentes aos registros vocais pré-selecionados



Fonte: PÜTZER; BARRY, 2007

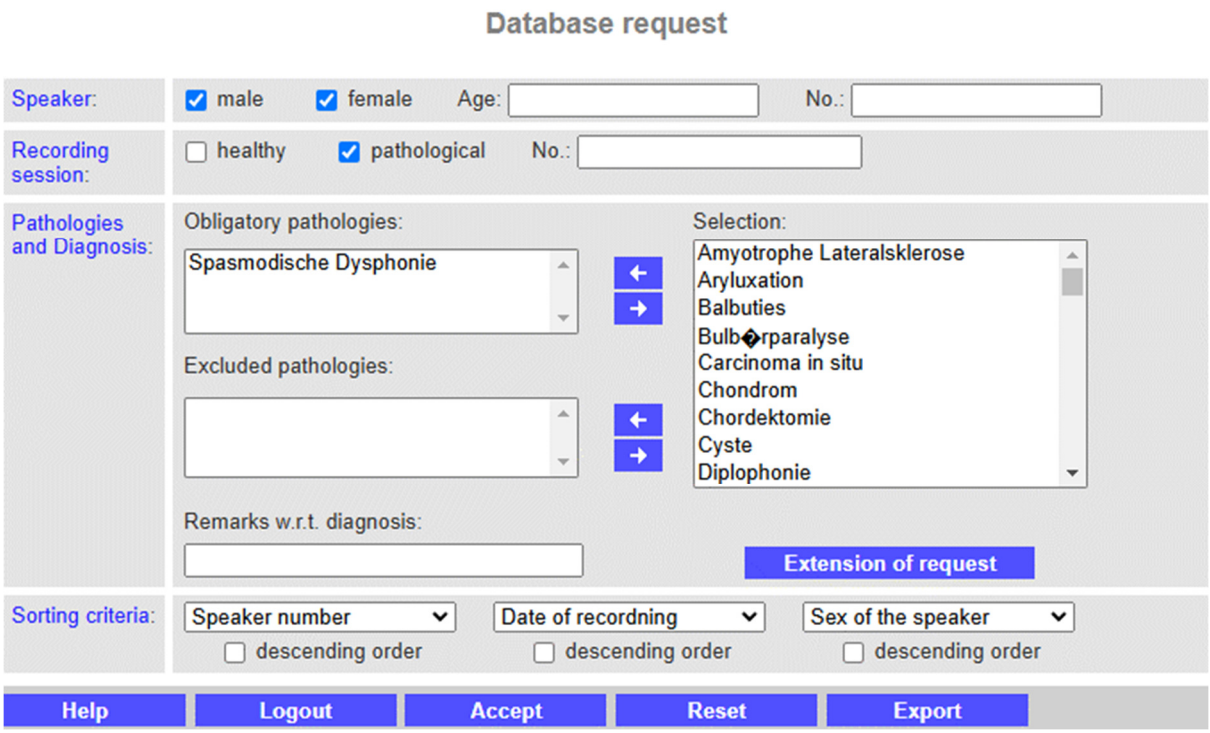
Após seleção dos dados para exportação e clicar no botão *accept*, os dados são preparados pelo servidor e é apresentado um link para realizar baixar um arquivo zipado contendo os registros.

Neste trabalho foi escolhido o sinal vocal em formato WAV e para todas as seleções discutidos posteriormente serão baixados os sinais da vogal /a/ em tom alto e a gravação da frase.

3.1.1. Seleção de Vozes Portadoras de DE

Para este trabalho é necessário obter os registros de vozes portadores de DE. Para isso foi utilizado o filtro representado na Figura 7 na tela de seleção inicial da base.

Figura 7 – Filtro utilizado para selecionar as vozes de pacientes portadores de DE



Fonte: PÜTZER; BARRY, 2007

A busca com esse filtro apresenta um retorno de 64 registros de pacientes que possuem DE.

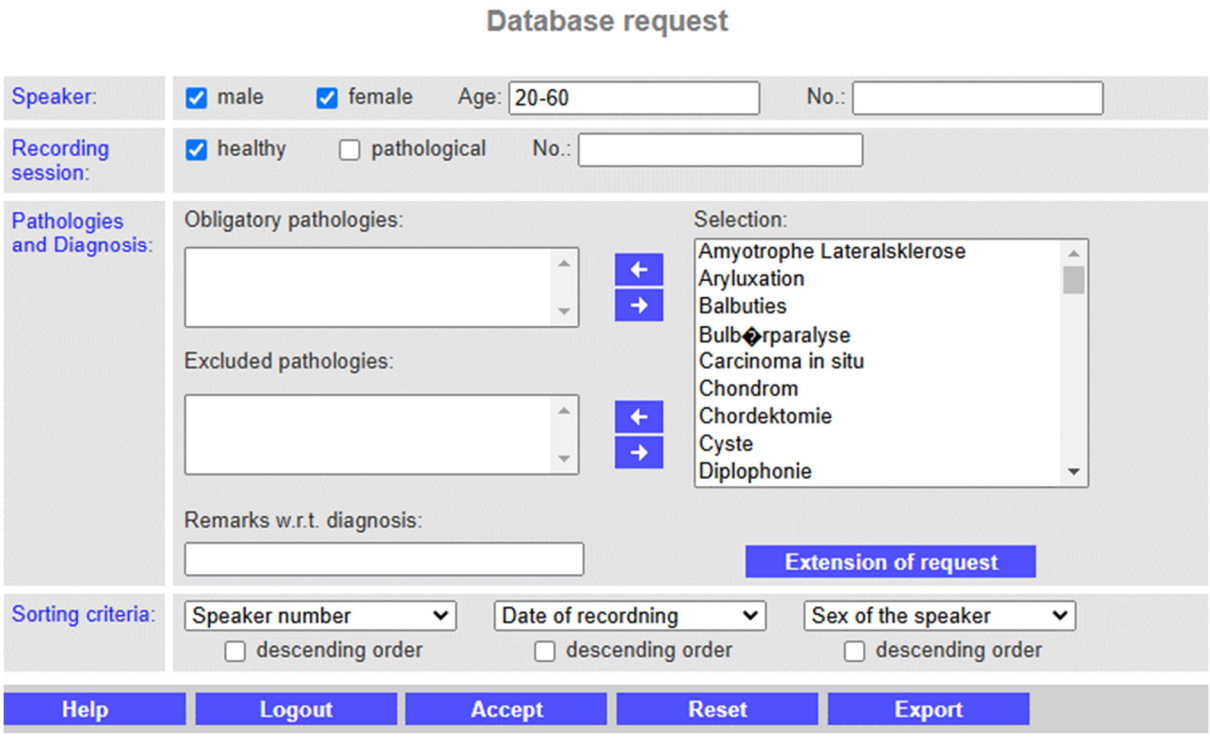
3.1.2. Seleção de Vozes Saudáveis

Também é necessário obter dados de pacientes saudáveis. Para isso foi utilizado o filtro representado na Figura 8 na tela de seleção inicial da base. Nele optou-se por incluir apenas registros de pessoas na faixa de idade dos 20 aos 60 anos, de maneira a reduzir os impactos da idade na característica vocal dos dados coletados.

A busca com esse filtro apresenta um retorno de 703 registros de pacientes saudáveis, porém ao ser feito o *download* dos arquivos, apenas 566 são retornados

quando selecionado a vogal /a/ em tom alto e 523 quando selecionada a frase.

Figura 8 – Filtro utilizado para selecionar as vozes de pacientes saudáveis



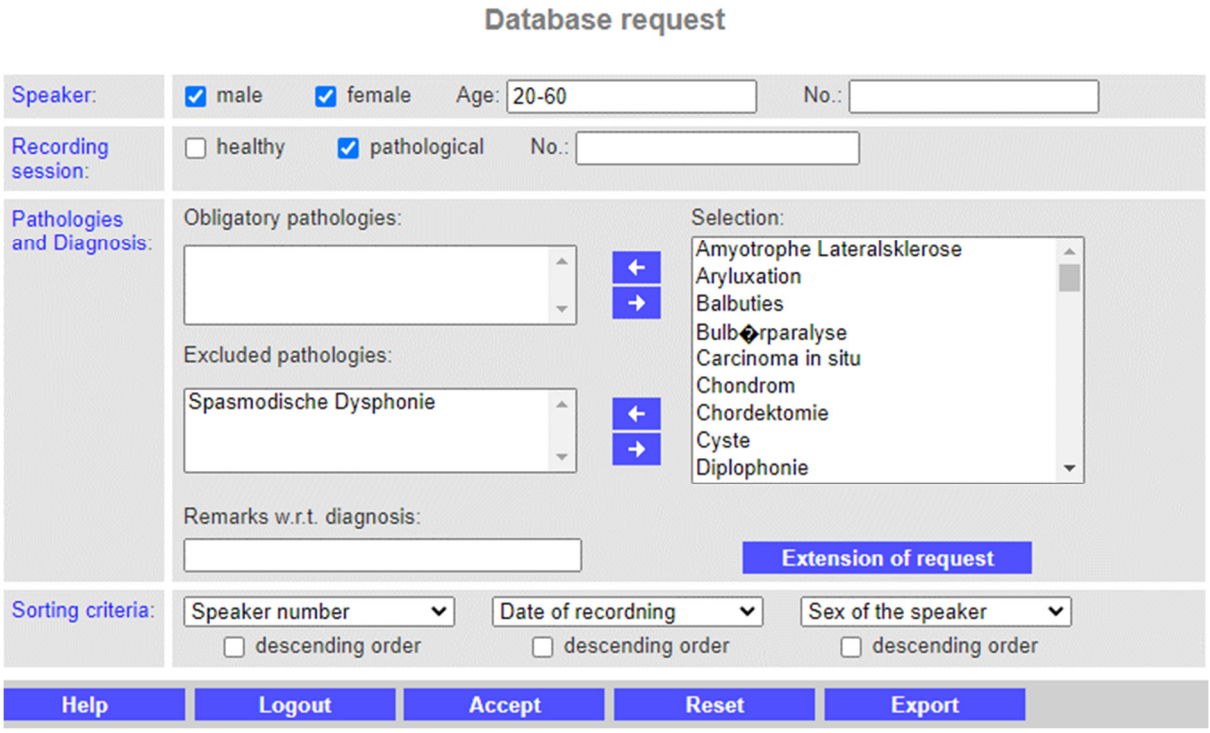
Fonte: PÜTZER; BARRY, 2007

3.1.3. Seleção de Vozes não Saudáveis e não Portadoras de DE

O último tipo de voz que foi utilizado no desenvolvimento desse trabalho é de pacientes que possuem alguma patologia diferente de DE. Para isso foi utilizado o filtro representado na Figura 9.

A busca com esse filtro apresenta um retorno de 881 registros de pacientes não saudáveis e não portadores de DE.

Figura 9 – Filtro utilizado para selecionar as vozes de pacientes portadores de alguma patologia diferente de DE



Fonte: PÜTZER; BARRY, 2007

3.2. EXTRAÇÃO DO VETOR DE CARACTERÍSTICAS

A definição do vetor de características que irão compor a entrada para o algoritmo de classificação é vital para que se obtenha um resultado satisfatório. Diversas características comumente utilizadas na identificação de patologias vocais são apresentadas por Islam, Tarique e Abdel-Raheem (2020). Para este trabalho foram selecionados os seguintes conjuntos de características: jitter, shimmer e MFCC. Essas características foram escolhidas baseadas em trabalhos semelhantes, onde Hadjaidji, Korba e Khelil utilizaram os coeficientes de jitter e shimmer para detectar pacientes que possuem DE e Dhamani e Guerti utilizaram os MFCC para classificar três grupos distintos: pacientes saudáveis, pacientes portadores de DE e pacientes com paralisia das pregas vocais.

3.2.1. Obtenção dos Valores de Jitter e Shimmer

Para obtenção dos valores de jitter e shimmer, optou-se pela utilização da biblioteca Parselmouth, uma biblioteca python para processamento de áudio, essa biblioteca faz a implementação do *software* Praat em python, facilitando a obtenção

dos valores de jitter e shimmer.

Islam, Tarique e Abdel-Raheem (2020) apresentam valores para alguns dos coeficientes de jitter e shimmer que seriam o limiar para diferenciar uma voz saudável de uma voz patológica, com isso, antes de apresentar os coeficientes obtidos para os algoritmos de classificação, decidiu-se fazer uma análise manual dos dados obtidos, porém o resultado foi uma taxa de acerto que variava entre 42 e 57% das vozes analisadas.

O resultado insatisfatório levantou a suspeita de que os coeficientes poderiam estar incorretos, mas analisando a documentação do *software* Praat encontra-se uma observação relatando sobre esses coeficientes para patologias e de como eles não podem ser aplicados aos dados encontrados por este software, isso devido ao método de cálculo abordado. Há um trabalho feito por Boersma (2009) onde é explorado o método de cálculo utilizado pelo *software* Praat e o MDVP, que é o *software* de onde os limiares para patologia foram retirados. Com isso decidiu-se seguir com o desenvolvimento do trabalho.

3.2.2. Obtenção dos MFCCs

Para obtenção dos MFCC, decidiu-se utilizar a biblioteca Librosa, que nos disponibiliza todo o cálculo dos MFCCs encapsulado em uma única chamada para a biblioteca.

3.3. CLASSIFICAÇÃO POR REDE NEURAL

O primeiro método de classificação utilizado foi RN. Para esse algoritmo decidiu-se utilizar para o vetor de características os coeficientes de jitter e shimmer. Para essa classificação foram usados os registros vocais da vogal /a/ em tom alto, essa escolha foi arbitrária, visto que não há um padrão adotado por outros trabalhos na área. Inicialmente foram escolhidas 64 vozes aleatoriamente do conjunto de dados de pacientes saudáveis e 64 vozes de pacientes com outra patologia que não de DE, de modo a criar um universo de dados balanceado com os 64 registros de vozes de pacientes portadores de DE.

É necessário separar os dados vocais em dois subconjuntos, um deles para treinamento do algoritmo e outro para verificar a performance do algoritmo. SILVA, SPATTI e FLAUZINO (2010) mencionam que cerca de 60 a 90% dos dados

compostos de maneira aleatória sejam destinados ao treinamento do algoritmo, enquanto os 10 a 40% restantes fiquem reservados para testes. A divisão adotada para os dados entre treino e teste foi de 70/30.

A RN foi montada com 2 camadas ocultas com 10 neurônios cada e função de ativação ReLU, 2 neurônios na camada de saída com função de ativação sigmoide. Dos neurônios da camada de saída o primeiro seria ativado caso a voz analisada pertencesse a um paciente portador de qualquer patologia (0 – saudável e 1 – patológico), já o segundo neurônio indicaria se a voz analisada é portadora de DE (0 – não possui DE e 1 – portador de DE). A Tabela 1 ilustra as saídas possíveis do sistema.

Tabela 1 – Tabela com as combinações possíveis resultantes da RN camada de saída

	Primeiro neurônio	**Segundo neurônio**
Voz saudável	0	0
Patológica, mas não DE	1	0
Portador de DE	1	1

Fonte: Autoria própria

A Tabela 2 apresenta os resultados dessa primeira classificação, apresentado os diversos vetores de entrada juntamente com a performance da RN para cada entrada.

Tabela 2 – Classificação utilizando RN com 2 camadas ocultas e 2 neurônios na camada de saída

Coeficientes	**Acurácia**	**Precisão**	**Revocação**
J_1, J_3, S_6	33.33%	66.10%	65%
J_1, J_3, J_4, S_6	33.33%	66.66%	66.66%
J_2, J_5, S_2, S_4	31.66%	65.52%	63.33%
$Todos\ os\ coeficientes$	33.33%	66.66%	66.66%

Fonte: Autoria própria

Com a baixa performance da RN, decidiu-se modificar sua estrutura,

adicionando outras 4 camadas ocultas, todas com 10 neurônios e ativação ReLU. A Tabela 3 exibe o resultado da nova RN para os mesmos vetores de entrada.

Tabela 3 – Classificação utilizando RN com 6 camadas ocultas e 2 neurônios na camada de saída

Coeficientes	Acurácia	Precisão	Revocação
J_1, J_3, S_6	33.33%	66.66%	66.66%
J_1, J_3, J_4, S_6	33.33%	66.66%	66.66%
J_2, J_5, S_2, S_4	33.33%	66.66%	66.66%
$Todos\ os\ coeficientes$	33.33%	66.66%	66.66%

Fonte: Autoria própria

Para compreender melhor o resultado obtido, decidiu-se analisar como estava sendo o comportamento da camada de saída da RN que, embora seja representado por 0 ou 1, quando observado diretamente o seu valor, nota-se que o mesmo fica entre 0 e 1, denotando o grau de certeza que a RN possui que aquele neurônio deveria ser 0 ou 1. Nesse caso, a saída da RN estava próxima aos 50% (0,50) em ambos os neurônios, sendo um pouco acima no primeiro neurônio e um pouco abaixo no segundo. Isso indica que a RN não está conseguindo classificar bem os dados.

Após essa análise, decidiu-se remover as vozes patológicas mas não portadoras de DE, restando apenas 2 grupos de dados de entrada e uma classificação binária simples. O modelo da RN também foi alterado, possuindo agora apenas 1 neurônio na camada de saída. Repetiu-se então a classificação utilizando tanto 2 quando 6 camadas ocultas. Os resultados encontram-se na Tabela 4 e Tabela 5.

Tabela 4 – Classificação utilizando RN com 2 camadas ocultas e 1 neurônio na camada de saída

Coeficientes	Acurácia	Precisão	Revocação
J_1, J_3, S_6	57.50%	100%	15%
J_1, J_3, J_4, S_6	52.50%	51.28%	100%
J_2, J_5, S_2, S_4	57.50%	54.05%	100%
$Todos\ os\ coeficientes$	50%	50%	100%

Fonte: Autoria própria

Tabela 5 – Classificação utilizando RN com 6 camadas ocultas e 1 neurônio na camada de saída

Coeficientes	Acurácia	Precisão	Revocação
J_1, J_3, S_6	72.50%	80%	60%
J_1, J_3, J_4, S_6	60%	83.33%	25%
J_2, J_5, S_2, S_4	57.50%	54.05%	100%
$Todos\ os\ coeficientes$	50%	50%	100%

Fonte: Autoria própria

Os resultados ainda são insatisfatórios para uma aplicação real de RN na identificação de DE. Uma possível causa da baixa performance da RN se dá ao fato de que o universo de dados disponíveis para treinamento da rede é insuficiente. Com isso decidiu-se adotar outro algoritmo de classificação que atendesse melhor o cenário do projeto.

3.4. CLASSIFICAÇÃO POR *RANDOM FOREST*

3.4.1. Classificação Utilizando Jitter e Shimmer

Para classificação dos registros vocais por RF optou-se por seguir utilizando apenas os dados vocais de pessoas saudáveis e de portadores de DE, resultando em uma classificação binária. Para o vetor de entrada também decidiu-se seguir com os coeficientes de jitter e shimmer e realizando a comutação entre os mesmos.

O único parâmetro utilizado para especificação do algoritmo RF é o número de estimadores (árvores de decisão) que serão utilizados. A princípio foram utilizados 150 estimadores para realizar a classificação. O resultado encontra-se na Tabela 6.

Tabela 6 – Classificação utilizando RF com 150 classificadores

Coeficientes	Acurácia	Precisão	Revocação
J_1, J_3, S_6	77.50%	73.91%	85%
J_1, J_3, J_4, S_6	77.50%	82.35%	70%
J_2, J_5, S_2, S_4	67.50%	70.59%	60%
$Todos\ os\ coeficientes$	77.50%	78.95%	75%

Fonte: Autoria própria

Embora o algoritmo tenha apresentado uma melhora nos resultados obtidos, os mesmos ainda não satisfazem o objetivo do trabalho, então experimentou-se utilizar uma função que busca otimizar o número de estimadores para uma classificação. Para essa função foram utilizados três parâmetros, primeiro o número de estimadores que deveria estar entre 20 e 200, segundo o tamanho máximo da profundidade de cada árvore variando entre 1 e 50 e por último o número de iterações para encontrar o melhor resultado, que foi utilizado como sendo 10. O resultado encontra-se na Tabela 7.

Tabela 7 – Classificação utilizando RF com função de otimização para buscar a melhor configuração

Coeficientes	Acurácia	Precisão	Revocação
J_1, J_3, S_6	72.50%	80%	60%
J_1, J_3, J_4, S_6	80%	92.86%	65%
J_2, J_5, S_2, S_4	75%	72.73%	80%
$Todos\ os\ coeficientes$	82.50%	78.26%	90%

Fonte: Autoria própria

A função de otimização da RF trouxe uma melhoria de -5% no pior caso e +7.5% no melhor, o que não justifica sua utilização dado que o tempo de execução do algoritmo aumentou consideravelmente ao executar a função de otimização, sendo

assim, optou-se por não a utilizar.

3.4.2. Classificação Utilizando MFCC

Como o algoritmo de RF performa melhor com muitas variáveis no vetor de características, decidiu-se utilizar os dados de MFCC das gravações vocais como vetor de características de entrada do algoritmo. Um problema decorrente dessa decisão é que o vetor possuía tamanho diferente a depender da duração da gravação de voz. Para resolver esse problema, foi feito a quebra do sinal vocal em segmentos menores de tamanho fixo, calculado os MFCC para cada segmento, preenchido os coeficientes faltantes do último segmento com 0 (*padding*) e calculada a média dos MFCC para os segmentos. Com isso foi possível o uso dos MFCC como vetor de entrada para o algoritmo RF.

Inicialmente escolheu-se o intervalo de 1s para quebra do sinal vocal. O algoritmo de RF foi novamente ajustado para utilizar 150 estimadores para realizar a classificação dos dados. O resultado encontra-se na Tabela 8.

Tabela 8 – Classificação utilizando RF com os MFCC com quebra de 1s e *padding* no último segmento

Acurácia	Precisão	Revocação
80%	83.33%	75%

Fonte: Autoria própria

Com a melhora no resultado da classificação utilizando MFCC, decidiu-se então partir para uma abordagem mais próxima do caso de uso real, onde a maioria das entradas seriam de pessoas saudáveis. Para isso foi usado então todos os registros vocais de pessoas saudáveis junto com os 64 de portadores de DE. Seguiu-se com a mesma divisão de 30% dos dados para teste do algoritmo. O resultado encontra-se na Tabela 9.

Tabela 9 – Classificação utilizando RF com os MFCC com quebra de 1s e *padding* no último segmento e todos os registros de pacientes saudáveis

Acurácia	Precisão	Revocação
90%	57.14%	20%

Fonte: Autoria própria

Analisando os resultados obtidos, o baixo percentual de revocação obtido

indica que os registros classificados erroneamente pelo algoritmo eram, em sua maioria, de portadores de DE. Como a RF utiliza um subconjunto aleatório dos dados de treinamento para cada um dos classificadores internamente, fornecer um conjunto de dados desbalanceado para treinamento do RF pode acarretar com que vários classificadores sejam expostos a apenas uma das saídas, nesse caso a de pessoas saudáveis. Para evitar tal comportamento adotou-se uma estratégia diferente onde o número de vozes de pessoas saudáveis expostas para o treinamento do RF é igual ao número de vozes de pessoas portadoras de DE. O resultado encontra-se na Tabela 10.

Tabela 10 – Classificação utilizando RF com os MFCC com quebra de 1s e *padding* no último segmento e todos os registros de pacientes saudáveis com conjunto de treinamento balanceado

Acurácia	Precisão	Revocação
73.43%	11.25%	90%

Fonte: Autoria própria

Agora decidiu-se variar o intervalo de quebra da gravação vocal para cálculo do MFCC e o número de estimadores. A Tabela 11 contém o resultado para cada combinação de parâmetros utilizados.

Tabela 11 – Classificação utilizando RF com os MFCC com quebra variável e *padding* no último segmento e todos os registros de pacientes saudáveis com conjunto de treinamento balanceado

Intervalo de quebra	Acurácia	Precisão	Revocação
1s	73.43%	11.25%	90%
0.5s	78.41%	9.24%	55%
0.25s	72.88%	10.06%	80%

Fonte: Autoria própria

De modo a evitar ruído nos dados de MFCC, decidiu-se evitar o preenchimento do último segmento vocal com 0 e apenas descartá-lo da análise. Repetiu-se então as classificações anteriores e os resultados encontram-se na Tabela 12.

Tabela 12 – Classificação utilizando RF com os MFCC com quebra variável descartando o último segmento e todos os registros de pacientes saudáveis com conjunto de treinamento balanceado

Intervalo de quebra	Acurácia	Precisão	Revocação
1s	67.90%	9.47%	90%
0.5s	73.25%	10.69%	85%
0.25s	77.31%	11.85%	80%

Fonte: Autoria própria

Finalmente, decidiu-se trocar as gravações da vogal /a/ e utilizar a frase para realizar a extração dos MFCC, com a intenção de verificar se o fato de uma gravação mais longa evidenciaria ainda mais as características vocal para identificação dos portadores de DE. Para essa classificação foi utilizado intervalo de quebra de 1s para a análise dos MFCC e a classificação foi feita utilizando 150 classificadores. O resultado dessa classificação serão exibidos no capítulo 4.

4. RESULTADOS

O foco deste trabalho era analisar a viabilidade do emprego de técnicas de *machine learning* juntamente com algoritmos de classificação no auxílio do diagnóstico de pacientes portadores de DE. Inicialmente decidiu-se utilizar RN para realizar a classificação dos registros vocais, porém a RN não apresentou resultado satisfatório. Uma possível explicação para a baixa performance pode ser o baixo número de amostras disponíveis para treinamento da rede. Posteriormente adotou-se RF como algoritmo classificador.

4.1. RESULTADOS OBTIDOS POR *RANDOM FOREST*

A Tabela 13 exibe os resultados da classificação feita utilizando o algoritmo RF com 150 classificadores, utilizados os coeficientes de MFCC dos registros vocais da frase como vetor de entrada e 1s para intervalo de quebra do registro vocal na análise dos MFCC.

Tabela 13 – Classificação utilizando RF com os MFCC com quebra de 1s e descartando o último segmento e todos os registros de pacientes saudáveis com conjunto de treinamento balanceado e registros vocais da frase

Acurácia	Precisão	Revocação
90.38%	29.41%	100%

Fonte: Autoria própria

Como o número de amostras de pacientes saudáveis é muito maior que o número de portadores de DE, isso resulta em um número maior de registros negativos, o que justifica a baixa precisão do algoritmo, no entanto o recall chegou a 100%, indicando que o modelo classificou corretamente todas as amostras de pacientes portadores de DE.

A Tabela 14 traz uma comparação do resultado obtido com outros trabalhos

na área. Por ela é possível concluir que o resultado do trabalho é satisfatório para classificação de pacientes portadores de DE.

Tabela 14 – Comparação dos resultados obtidos com os de outros autores

Autor	Acurácia	Precisão	Revocação
Souza	90.38%	29.41%	100%
Murthy, et. al.	93.5%	-	-
Dahmani e Guerti	90%	-	85%
Hadjaidji, Korba e Khelil	86.66%	-	87.5%

Fonte: Autoria própria

5. CONCLUSÃO

Com os resultados exibidos no capítulo 4, é possível concluir que o algoritmo de RF conseguiu classificar satisfatoriamente os registros vocais de pacientes portadores de DE. Com isso, pode-se afirmar que o uso da tecnologia para auxiliar no diagnóstico de patologias vocais é de grande valia na busca por métodos não invasivos para tais diagnósticos.

Dada a natureza do comportamento do algoritmo de RF, é possível observar como o uso de um vetor de entrada com um número maior de características (MFCC) apresenta melhores resultados quando comparado com um vetor de características menor (coeficientes de jitter e shimmer).

Nota-se também que o uso dos registros vocais da gravação da frase apresentou um resultado melhor quando comparado ao obtido pela gravação da vogal /a/, uma possível explicação para isso é de que o tempo necessário para produção vocal da vogal /a/ possa não ser o suficiente para os sintomas de DE se manifestarem, enquanto os mesmos devam ser bem mais presentes em uma fala prolongada como a de uma frase.

Embora os resultados da classificação por RN tenham sidos muito abaixo do ideal, aplicações de RN para identificação de DE não devem ser descartadas, uma vez que sua baixa performance possa ser explicada pelo baixo número de amostras vocais disponíveis para treinamento da rede. Outra tentativa seria a de usar os registros da frase para avaliar a performance do algoritmo, assim como foi feito com RF.

Para trabalhos futuros, recomenda-se avaliar o desempenho dos algoritmos classificadores em um universo de dados de entrada que contenha registros de vários tipos de patologia, para aproximar o modelo do uso prático. Adicionalmente, avaliar técnicas para redução do número de erros da classificação pode contribuir positivamente para acelerar o uso dos algoritmos classificadores na prática.

REFERÊNCIA

BARKMEIER, J. M.; CASE, J. L.; LUDLOW, C. L. Identification of symptoms for spasmodic dysphonia and vocal tremor: a comparison of expert and nonexpert judges. **Journal of Communication Disorders**, v. 34, n. 1-2, p. 21–37, jan. 2001.

BOERSMA, P. Should jitter be measured by peak picking or by waveform matching? **Folia Phoniatrica et Logopaedica**, v. 61, n. 5, p. 305-308, 2009. DOI: 10.1159/000245159. Epub 10 out. 2009. PMID: 19828997.

DAHMANI, M.; GUERTI, M. Vocal folds pathologies classification using Naïve Bayes Networks. **2017 6th International Conference on Systems and Control (ICSC)**, Batna, 2017, pp. 426-432, maio 2017.

FIORIN, D. V. et al. Aplicações de redes neurais e previsões de disponibilidade de recursos energéticos solares. **Revista Brasileira de Ensino de Física**, v. 1, n. 33, p. 1309, 2011. Disponível em: <https://www.researchgate.net/publication/235932975_Aplicacoes_de_redes_neurais_e_previsoes_de_disponibilidade_de_recursos_energeticos_solares>.

FLECK, L. et al. Redes neurais artificiais: princípios básicos. **Revista Eletrônica Científica Inovação e Tecnologia**, v. 7, n. 15, p. 47–57, 2016.

HADJAIDJI, E.; KORBA, M. C. A.; KHELIL, K. Spasmodic dysphonia detection using machine learning classifiers. **2021 International Conference on Recent Advances in Mathematics and Informatics (ICRAMI)**. Anais...IEEE, p 1-5, 2021.

IMAMURA, R.; TSUJI, D. H. Disfonia espasmódica de adução, tremor vocal e disfonia de tensão muscular: é possível fazer o diagnóstico diferencial? **Revista brasileira de otorrinolaringologia**, v. 72, n. 4, p. 434–434, ago. 2006.

ISLAM, R; TARIQUE, M.; ABDEL-RAHEEM, E. A Survey on Signal Processing Based Pathological Voice Detection Techniques. **IEEE Access**, v. 8, p. 66749-66776, 2020. DOI: 10.1109/ACCESS.2020.2985280

MAKHOUL, J.; COSELL, L. LPCW: An LPC vocoder with linear predictive spectral warping. **ICASSP '76 - IEEE International Conference on Acoustics, Speech, and Signal Processing**, 1976, Philadelphia, PA, USA. Proceedings [...]. Philadelphia: IEEE, 1976. p. 466-469.

MURTHY, N. et al. A Novel Algorithm for Detecting Spasmodic Dysphonia Voice Pathology using Random Forest Frame Work. **2022 International Conference on Edge Computing and Applications (ICECAA)**. Anais...IEEE, p. 800-804, 2022.

PÜTZER, M; BARRY; W. J. **Saarbruecken voice database**, 2007. Disponível em: <https://stimmdb.coli.uni-saarland.de/index.php4#target>. Acesso em: 26 jan. 2024

SANUKI, T. Spasmodic dysphonia: An overview of clinical features and treatment options. **Auris, nasus, larynx**, v. 50, n. 1, p. 17–22, fev. 2023.

SILVA, I. N.; SPATTI, D. H.; FLAUZINO, R. A. **Redes neurais artificiais para engenharia e ciências aplicadas: curso prático**. São Paulo: Artliber, 2010.

STEVENS, S. S.; VOLKMANN, J.; NEWMAN, E. B. A scale for the measurement of the psychological magnitude pitch. **Journal of the Acoustical Society of America**, v. 8, n. 3, p. 185–190, 1937.

TULICS, M. G. et al. Artificial Neural Network and SVM based Voice Disorder Classification. **2019 10th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)**, Naples, Italy, 2019, pp. 307-312, out. 2019.

Apêndice A – Filtro Amplificador De Frequências Altas

O filtro de pré-ênfase utilizado no cálculo do MFCC pode ser melhor entendido quando analisamos seu comportamento no domínio da frequência. Para isso aplica-se a transformada Z na equação 15 que descreve o filtro.

$$Y(z) = X(z) - \alpha z^{-1} X(z) \tag{20}$$

$$Y(z) = X(z)(1 - \alpha z^{-1}) \tag{21}$$

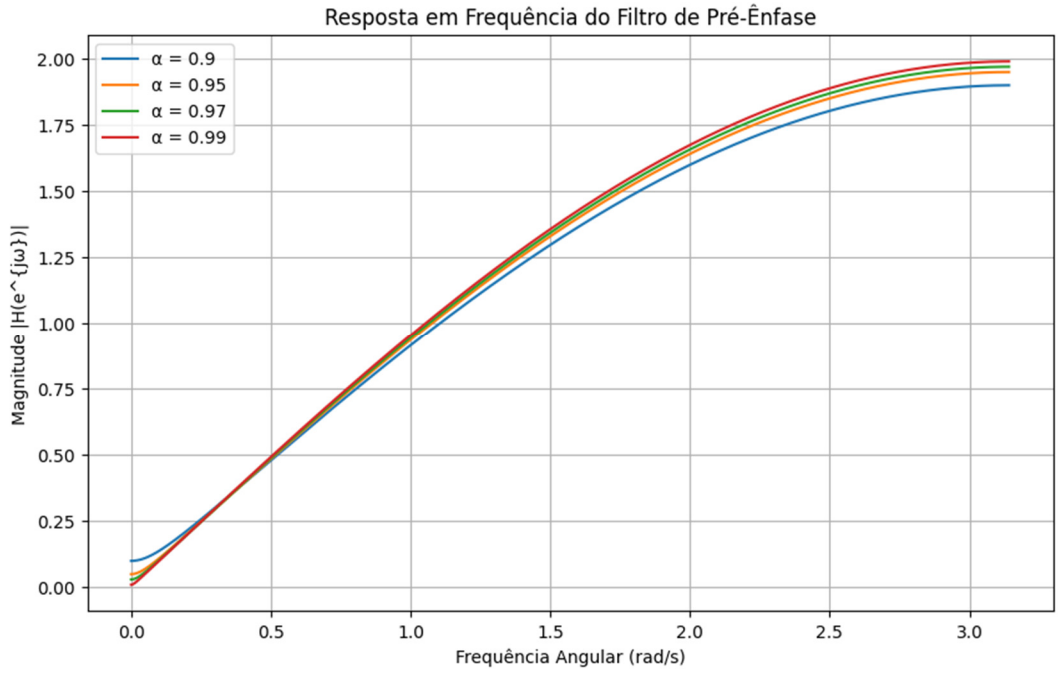
Portanto a função de transferência do filtro é dada por:

$$H(z) = 1 - \alpha z^{-1} \tag{22}$$

$$H\left(e^{j\omega}\right) = 1 - \alpha e^{-j\omega} \tag{23}$$

A Figura 10 exibe a resposta do filtro de acordo com o aumento da frequência angular (ω) para diversos valores de α. Nela é possível verificar como frequências baixas resultarão em uma atenuação do sinal, enquanto frequências mais altas serão amplificadas.

Figura 10 – Representação gráfica da resposta do filtro de pré-ênfase



Fonte: Autoria própria