UNIVERSIDADE FEDERAL DO CEARÁ

CENTRO DE TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE TELEINFORMÁTICA

MESTRADO ACADÊMICO EM ENGENHARIA DE TELEINFORMÁTICA

BRÍGIDA FARIAS CARDOSO OLIVEIRA

CARACTERIZAÇÃO E CLASSIFICAÇÃO DE SINAIS DE VOZ POR COMBINAÇÃO DE VOGAIS SUSTENTADAS: UM ESTUDO BASEADO NA TRANSFORMADA WAVELET HAAR

FORTALEZA

2021

BRÍGIDA FARIAS CARDOSO OLIVEIRA

CARACTERIZAÇÃO E CLASSIFICAÇÃO DE SINAIS DE VOZ POR COMBINAÇÃO DE VOGAIS SUSTENTADAS: UM ESTUDO BASEADO NA TRANSFORMADA WAVELET HAAR

Dissertação apresentada ao Curso de Mestrado Acadêmico em Engenharia de Teleinformática do Programa de Pós-Graduação em Engenharia de Teleinformática do Centro de Tecnologia da Universidade Federal do Ceará, como requisito parcial à obtenção do título de mestre em Engenharia de Teleinformática. Área de Concentração: Sinais e Sistemas

Orientadora: Profª. Drª. Fátima Nelsizeuma Sombra de Medeiros

Coorientadora: Profª. Drª. Deborah Maria Vieira Magalhães

FORTALEZA

2021

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Biblioteca Universitária
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

O45c Oliveira, Brígida.
Caracterização e Classificação de Sinais de Voz por Combinação de Vogais Sustentadas : Um Estudo Baseado na Transformada Wavelet Haar / Brígida Oliveira. – 2021.
41 f. : il. color.

Dissertação (mestrado) – Universidade Federal do Ceará, Centro de Tecnologia, Programa de Pós-Graduação em Engenharia de Teleinformática, Fortaleza, 2021.
Orientação: Prof. Dr. Fátima Nelsizeuma Sombra de Medeiros.
Coorientação: Prof. Dr. Deborah Maria Vieira Magalhães.

1. Wavelet Haar. 2. Análise estatística de Kruskal-Wallis. 3. Detecção de distúrbios na voz. 4. Vogais sustentadas. I. Título.

CDD 621.38

BRÍGIDA FARIAS CARDOSO OLIVEIRA

CARACTERIZAÇÃO E CLASSIFICAÇÃO DE SINAIS DE VOZ POR COMBINAÇÃO DE VOGAIS SUSTENTADAS: UM ESTUDO BASEADO NA TRANSFORMADA WAVELET HAAR

Dissertação apresentada ao Curso de Mestrado Acadêmico em Engenharia de Teleinformática do Programa de Pós-Graduação em Engenharia de Teleinformática do Centro de Tecnologia da Universidade Federal do Ceará, como requisito parcial à obtenção do título de mestre em Engenharia de Teleinformática. Área de Concentração: Sinais e Sistemas

Aprovada em: 21 de junho de 2021

BANCA EXAMINADORA

Profª. Drª. Fátima Nelsizeuma Sombra de Medeiros (Orientadora)
Universidade Federal do Ceará (UFC)

Profª. Drª. Deborah Maria Vieira Magalhães (Coorientadora)
Universidade Federal do Piauí (UFPI)

Prof. Dr. Charles Casimiro Cavalcante
Universidade Federal do Ceará (UFC)

Ialis Cavalcante de Paula Junior
Universidade Federal do Ceará (UFC) - Campus de Sobral

À minha mãe, minha irmã e ao meu pai.

AGRADECIMENTOS

Agradeço aos meus pais e a minha irmã por me apoiarem em todos os momentos da minha vida.

À minha orientadora Profa. Fátima Sombra pela paciência na orientação e incentivo que tornaram possível a conclusão deste trabalho. À minha coorientadora Profa. Deborah Vieira e ao Prof. Daniel Ferreira por estarem sempre presentes quando precisei.

Aos colegas do Grupo de Processamento de Imagens pela amizade adquirida após todo esse tempo e pelo incentivo e pelo apoio constante.

Por fim, meu agradecimento à FUNCAP por apoiar financeiramente este trabalho.

“Ideas and only ideas can light the darkness.”

(Ludwig Von Mises)

RESUMO

Este trabalho investiga o uso de vogais sustentadas isoladas e combinadas na classificação de sinais de voz normais e sinais de voz com disfonia com base na energia dos coeficientes de decomposição da transformada wavelet Haar, considerando os gêneros masculino e feminino e uma combinação de ambos. É realizado ainda o teste estatístico de Kruskal-Wallis onde analisamos os pares de variáveis compostos pelas vogais (isoladas ou combinadas) e o nível de decomposição wavelet. A saída deste teste estatístico permite identificar os níveis de decomposição do sinal que alcançam os melhores valores de acurácia da classificação. O menor nível de decomposição wavelet foi selecionado neste trabalho por estar associado ao menor custo computacional. Embora trabalhos recentes tenham mostrado que vogais sustentadas isoladas permitem a caracterização precisa da voz, a literatura carece de evidências sobre o uso da combinação dessas vogais. A metodologia proposta para caracterização e classificação de vozes normais e anormais considera o cálculo da energia dos coeficientes de aproximação e de detalhes obtidos da decomposição da transformada wavelet Haar. Utilizamos a combinação das vogais sustentadas /a/, /i/ e /u/ na detecção de distúrbios na voz assim como destas vogais isoladas de forma a identificar o cenário que resulta na melhor classificação dos sinais de voz. Conduzimos experimentos em duas bases de dados públicas de origem portuguesa e alemã, *Advanced Voice Function Assessment Database* (AVFAD) e *Saarbrücken Voice Database* (SVD), respectivamente. Analisamos os níveis de decomposição wavelet no intervalo de 4 a 18 nos diferentes cenários, a saber, vozes de falantes dos gêneros feminino e masculino e ambos os gêneros. Os resultados obtidos revelaram que os coeficientes wavelet extraídos da combinação de vogais melhoraram a descrição do sinal e, portanto, a identificação de características sutis de vozes doentes e saudáveis. Também mostramos que as características baseadas na energia dos coeficientes da transformada wavelet Haar extraídas de vogais combinadas alcançaram uma boa classificação de voz com menor número de decomposições. Esta abordagem melhorou a acurácia em pelo menos 2,61% e 15,61% para dados das bases AVFAD e SVD, respectivamente, independentemente do gênero do falante.

Palavras-chave: Wavelet Haar. Análise estatística de Kruskal-Wallis. Detecção de distúrbios na voz. Vogais sustentadas.

ABSTRACT

This work investigates the use of single and combined sustained vowels in the characterization of normal and abnormal voice signals based on the Haar wavelet decomposition coefficients, considering signals from the biological genders male and female and a combination of both. We also perform the Kruskal-Wallis statistical test, in order to analyze the pairs of variables composed of the vowels (single or combined) and the wavelet decomposition levels. The output of this statistical test allows identifying which level reaches (leads to) the best classification accuracy. We selected the lowest level of decomposition because it is associated with the lowest computational cost. Although recent studies have shown that single sustained vowels allow accurate voice characterization, the literature lacks evidence on using the combination of them. The proposed methodology for characterizing and classifying normal and abnormal voices considers the energy calculation of details and approximation coefficients obtained from the decomposition of the Haar wavelet transform. We use the /a/, /i/, and /u/ single vowels and the combination to identify the scenario that results in the best classification of the voice signals. We conducted experiments on two public datasets, one from Portugal named Advanced Voice Function Assessment Database (AVFAD) and another one from Germany named Saarbrücken Voice Database (SVD). We analyzed the wavelet decomposition levels in the range of 4 to 18 in different scenarios: voices of female and male speakers and both genders. Our results revealed that the wavelet coefficients extracted from the combination of vowels improved the signal description and identified subtle features of pathological voices. We also showed that the Haar wavelet-based features extracted from combined vowels achieved accurate voice classification with fewer decomposition levels. This approach enabled accuracy improvements of at least 2,61% e 15,61% for AVFAD and SVD datasets, respectively, regardless of the biological gender.

Keywords: Wavelet Haar. Kruskal-Wallis analysis. Voice disorder detection. Sustained vowels.

LISTA DE FIGURAS

Figura 1 – Representação gráfica da wavelet Haar. Fonte: autoria própria (2021). . . . 22

Figura 2 – Metodologia para avaliar o desempenho da wavelet Haar na detecção de doenças na voz. a) Fase de caracterização e classificação, resultando em b) um conjunto de histogramas de acurácia para cada nível de decomposição n e a vogal sustentada de entrada v usada na análise do teste estatístico para investigar as hipóteses. D_g representa um subconjunto de um dos conjuntos de dados pesquisados D, de acordo com o gênero g. K é definido como $K = n|n = \{2,4,6,8,10,12,14,16,18\}$ e $V = v|v = \{$/*/,/a/,/i/,/u/$\}$. Fonte: autoria própria (2021). 29

Figura 3 – Resultados da avaliação quantitativa para análise de combinação de vogais (H1). a) e b) apresentam os melhores valores de acurácia (Equação 3.2) de acordo com as vogais para AVFAD e SVD, respectivamente. Os números na legenda representam os níveis de decomposição da wavelet Haar, seguidos pela sequência de vogais. Fonte: autoria própria (2021). 33

Figura 4 – Comparações pareadas para todos os gêneros avaliados usando o teste estatístico Kruskal-Wallis com o teste post-hoc de Nemenyi: (a-c) gêneros para o conjunto de dados AVFAD e (d-f) gêneros para o conjunto de dados SVD. As caixas pretas representam os pares com diferenças significativas em $\alpha = 0,05$. Fonte: autoria própria (2021). 35

LISTA DE TABELAS

Tabela 1 – Comparação dos principais parâmetros vocais na infância, na idade adulta e na terceira idade (BEHLAU, 2004). 18
Tabela 2 – Número mínimo de níveis de decomposição da transformada wavelet Haar. . 36

SUMÁRIO

1 INTRODUÇÃO 12
1.1 Produção científica 15
1.2 Organização da dissertação 16
2 FUNDAMENTAÇÃO TEÓRICA 17
2.1 Distúrbios vocais 17
2.1.1 Desenvolvimento da voz ao longo da vida 17
2.1.2 Diferenças entre voz normal e anormal 19
2.2 Extração de características utilizando a transformada wavelet discreta . 20
2.2.1 Análise de Fourier 20
2.2.2 Análise Wavelet 21
2.2.3 Wavelet Haar 22
2.2.4 Cálculo da energia dos coeficientes 23
2.3 Classificação por *random forest* 24
2.4 Análise estatística utilizando o teste de Kruskal-Wallis e teste *post-hoc* Nemenyi 25
2.4.1 Teste de Kruskal-Wallis 25
2.4.2 Teste de Nemenyi 26
3 METODOLOGIA 28
3.1 Bases utilizadas 28
3.2 Caracterização e classificação de vozes doentes 29
3.3 Análise estatística 31
4 RESULTADOS 32
5 CONCLUSÕES E TRABALHOS FUTUROS 37
REFERÊNCIAS 38
ANEXOS 41
ANEXO A – Tabela de distribuição qui-quadrado 42

1 INTRODUÇÃO

A literatura reporta que os distúrbios na voz correspondem a qualquer alteração na qualidade, altura ou amplitude, entre outras características, que divergem de vozes de idade, gênero e grupos sociais semelhantes (HEGDE *et al.*, 2019). Existem diferentes razões para identificar padrões de voz anormais, por exemplo, o tratamento precoce de uma doença e a redução do impacto nas habilidades de comunicação do indivíduo, o que é relevante para o bom desempenho de papéis sociais e profissionais (HEGDE *et al.*, 2019). No entanto, a identificação de padrões anormais em vozes humanas é subjetiva e geralmente depende da experiência do especialista (WU *et al.*, 2018). A percepção do distúrbio na voz geralmente é baseada em uma análise acústica que é suscetível a ruídos, interferindo no diagnóstico. Além disso Fonseca *et al.* (2020) reportam que diferentes doenças podem coexistir com sintomas semelhantes, o que torna difícil a detecção desses distúrbios tanto para a análise da voz pelo clínico, bem como para a videolaringoscopia e para a análise de luz estroboscópica.

Recentemente, diferentes metodologias para o pré-diagnóstico baseado em computador aplicado em vozes humanas, como as desenvolvidas por Fonseca *et al.* (2020), Fang *et al.* (2019) e Zhang e Hu (2018), podem auxiliar os especialistas na identificação precoce e no tratamento de várias doenças (HEGDE *et al.*, 2019). Essas metodologias são geralmente baseadas em extração de características, empregando técnicas como *Mel-frequency cepstral coefficients* (MFCCs)[1] (FANG *et al.*, 2019), wavelets (ZHANG; HU, 2018; SHIA; JAYASREE, 2017; FONSECA *et al.*, 2017), redes neurais convolucionais (*Convolutional Neural Networks* - CNNs) (WU *et al.*, 2018; ALHUSSEIN; MUHAMMAD, 2018; FONSECA *et al.*, 2020). Em Fonseca *et al.* (2020), características como energia, entropia e taxas de cruzamento de zeros foram extraídas do sinal de áudio. Essas características foram associadas por meio de uma máquina paraconsistente discriminativa [2] para classificar diferentes doenças com o mesmo sintoma vocal principal. Fonseca *et al.* (2020) relataram uma acurácia de 95%; no entanto, seu estudo é limitado a subconjuntos específicos tirados de uma base de dados maior: 687 amostras de vozes saudáveis, 34 gravações de vozes de indivíduos afetados exclusivamente por Edema

[1] MFCCs são os resultados de uma transformação de cosseno do logaritmo real do espectro de energia de curto prazo expresso em uma escala de frequência de *mel* (ZHENG *et al.*, 2001). A escala de *mels* (do inglês, *melody*) foi construída com base em avaliações subjetivas do *pitch*, envolvendo a determinação de intervalos de frequência que correspondem à metade ou ao dobro do *pitch* (BEHLAU, 2004).

[2] Máquina paraconsistente discriminativa consiste em um algoritmo de aprendizado supervisionado inspirado em Máquinas de Vetores de Suporte (*Support Vector Machines* - SVMs) probabilísticas e técnicas relacionadas, que combinam conceitos importantes para ter a capacidade de tratar contradições e incertezas (GUIDO *et al.*, 2013).

de Reinke[3], 82 amostras de voz de indivíduos afetados apenas por laringite e 10 amostras de voz de indivíduos afetados por Edema de Reinke e laringite, sendo que a base de dados de onde veio essas amostras possui amostras de mais de 2000 vozes e 70 doenças. Essa abordagem baseada em subgrupos de doenças tende a se ajustar a apenas algumas doenças, o que reduz a escalabilidade do modelo para avaliação clínica. Além disso, Fonseca *et al.* (2020) exploraram apenas a vogal sustentada /a/, desconsiderando os traços potenciais de outras vogais sustentadas, como /i/ e /u/.

Algoritmos de aprendizado profundo também têm sido usados para identificar doenças na voz. Em (WU *et al.*, 2018), os espectrogramas são entradas para diferentes arquiteturas da CNN, e foram gerados a partir de áudios referentes a seis doenças, considerando apenas a vogal sustentada /a/ em tom neutro. A partir do treinamento em 724 amostras, esses modelos baseados em CNN alcançaram até 77% de acurácia, que foi calculada a partir de 240 amostras de teste. Esses resultados mostraram que quando a acurácia é calculada no grupo de treino, o valor obtido aumentou 11,5%, revelando a existência de ocorrência de sobre-ajuste aos dados. Wu *et al.* (2018) relataram esse comportamento e sugeriram treinar o modelo em um conjunto de dados maior, o que demanda muito tempo e recursos na fase de coleta de dados.

Visto que, como Chandran e Boashash (2016) afirmam, a maioria dos sinais de voz são processos não estacionários com componentes que são variantes de tempo e frequência, a transformada wavelet discreta (*discrete wavelet transform* - DWT) (HEIL; WALNUT, 1989) realiza uma análise de tempo-frequência desta categoria de sinais e fornece análise de recursos para vozes anormais. Por isso, esta é uma ferramenta amplamente utilizada na detecção de doenças vocais (HEGDE *et al.*, 2019). A wavelet Haar com quatro níveis de decomposição foi usada por Shia e Jayasree (2017) para extrair características de sinais de áudio pelo cálculo da energia dos coeficientes de aproximação e de detalhes para classificação de voz. Os autores selecionaram 49 amostras normais e 49 anormais, referentes a uma frase dita em alemão. Como a fala ocorre em diferentes idiomas, entonações e emoções (PANEK *et al.*, 2015), a detecção de doenças na voz usando frases em vez de vogais sustentadas se torna ainda mais complexa. A acurácia máxima alcançada por Shia e Jayasree (2017) foi de 93,33% usando uma rede neural *feedforward*[4].

[3] Edema de Reinke é uma doença crônica da laringe na qual a camada superficial da lâmina própria é expandida por muco espesso conferindo-lhe aspecto gelatinoso (MARTINS *et al.*, 2009). Ainda de acordo com Martins *et al.* (2009), essa doença se relaciona ao tabagismo e acomete, preferencialmente mulheres, fazendo com que elas apresentam a voz mais grave.

[4] Rede Neural Feedforward (*Feedforward Neural Network* - FNN) é um perceptron multicamadas onde o fluxo de

Fonseca *et al.* (2017) realizaram a classificação da voz usando uma única vogal sustentada com diferentes famílias de wavelets, a saber, Haar, Daubechies, Coiflet e Symmlet. Estes autores alcançaram uma acurácia de até 85,94% usando a wavelet Daubechies com dois níveis de decomposição. No entanto, esta análise está restrita a um conjunto de dados privado, com 32 registros de vozes doentes e 32 de vozes saudáveis da vogal sustentada /a/. Todos os indivíduos afetados apresentam diagnóstico de nódulos nas pregas vocais. Zhang e Hu (2018) propuseram um método de análise multiescala para detectar distúrbios na laringe. Os autores adotaram a wavelet Daubechies para decompor o sinal de voz em 16 sub-bandas de frequência diferentes. Eles extraíram o expoente de Hurst[5] e as sub-bandas de características da entropia de Rényi de segunda ordem para classificar as amostras. Em seguida, realizaram experimentos usando subconjuntos compostos por classes de sinais de voz normal, com paralisia e sem paralisia de dois conjuntos de dados diferentes. Todas as amostras corresponderam às vogais sustentadas /a/ e a acurácia da classificação alcançou 92,83%. A maioria dos modelos mencionados foi desenhada para detectar vozes com distúrbios, relatando resultados experimentais dos subconjuntos privados e/ou públicos. Normalmente, esses subconjuntos são desenhados para modelar um grupo reduzido de doenças relacionadas à vogal sustentada /a/, mesmo quando outras vogais estão disponíveis. Essa preferência vocálica se deve à produção de sons com trato vocal relativamente aberto permitindo o exame de todo o aparelho vocal (GÓMEZ-GARCÍA *et al.*, 2019b). Essa abordagem baseada em uma única vogal tende a limitar a habilidade do sistema de detecção automática. Além disso, o uso de subgrupos de doenças aumenta o ajuste excessivo do modelo de probabilidade a uma doença específica, o que é especialmente crítico para a avaliação clínica.

Apesar de Gómez-García *et al.* (2019a) e Hemmerling *et al.* (2016) terem explorado as vogais sustentadas /a/, /i/ e /u/, não há evidências sobre os impactos de combiná-las. Dentre os diversos trabalhos que encontramos, os efeitos de todas as vogais e entonações combinadas foram estudados apenas por Martinez *et al.* (2012). Os resultados de Martinez *et al.* (2012) mostraram uma área sob a curva ROC (AUC) de 0,804 para todas as entonações /a/, 0,783 para todas as entonações /i/ e 0,797 para todas as entonações /u/, enquanto todas as vogais combinadas mostram uma AUC de 0,879. Essas questões que foram apresentadas por Martinez *et al.* (2012)

decisão é unidirecional, avançando da entrada para a saída em camadas sucessivas, sem ciclos ou loops (URSO *et al.*, 2019).

[5] O expoente de Hurst é um fator adimensional usado para estimar a presença e a magnitude da propriedade de autossimilaridade (FERNANDES *et al.*, 2015). É uma medida não apenas da autossimilaridade, mas também das propriedades estatísticas que a autossimilaridade acarreta (FERNANDES *et al.*, 2015). As estimativas do expoente de Hurst são úteis para entender a estrutura de autocorrelação e a evolução de um processo, e assim atingir os objetivos mencionados nos quais se baseia o estudo da autossimilaridade (FERNANDES *et al.*, 2015).

nos levaram a deduzir que a combinação de diferentes vogais poderia melhorar a caracterização e classificação de vozes doentes. No entanto, a literatura carece de evidências de que esses achados possam ser extrapolados para um modelo baseado em wavelet. Assim, a ideia central do nosso trabalho consiste em focar na combinação de vogais sustentadas na caracterização de doenças na voz usando uma transformada wavelet, em particular, a wavelet Haar. Além disso, argumentamos que um estudo sem especificação de doenças pode revelar resultados próximos à rotina clínica mais geral. Outro aspecto importante é entender como os sistemas baseados em wavelet Haar para reconhecimento de doenças na voz atuam nos coeficientes de detalhes e de aproximação de diferentes níveis de decomposição. Assim, estamos particularmente interessados em confirmar duas hipóteses principais:

- Hipótese 1 (H1): O desempenho de um sistema baseado em wavelet Haar para detecção de vozes doentes melhora quando a entrada do sinal compreende a combinação de vogais sustentadas.
- Hipótese 2 (H2): o processo de detecção de doenças na voz com precisão depende do nível de decomposição da transformada wavelet Haar.

Para abordar essas hipóteses, este trabalho apresenta três contribuições: 1) uma avaliação quantitativa do desempenho da extração de características da wavelet Haar em relação aos diferentes níveis de decomposição para detecção de voz anormal; 2) a identificação do número mínimo de decomposições da wavelet Haar que leva ao melhor desempenho de classificação segundo a análise de Kruskal-Wallis e 3) uma análise quantitativa do desempenho da classificação, por gênero, de alterações na voz. Além disso, consideramos todas as doenças presentes em dois conjuntos de dados públicos, ao contrário da maioria dos estudos da literatura, que selecionam apenas um número limitado delas.

1.1 Produção científica

Artigo Publicado

- **OLIVEIRA, B. F. C**; MAGALHÃES D. M. V; FERREIRA, D. S; MEDEIROS, F. N. S.. *Combined Sustained Vowels Improve the Performance of the Haar Wavelet for Pathological Voice Characterization*. 2020 International Conference on Systems, Signals and Image Processing (IWSSIP), Niterói-RJ, pp. 381-386, 2020. Doi:10.1109/IWSSIP48289.2020.9145258.

1.2 Organização da dissertação

Os capítulos remanescentes deste trabalho estão organizados da seguinte forma:

- **Capítulo 2:** apresenta a fundamentação teórica do estudo.
- **Capítulo 3:** trata dos materiais e métodos usados e apresenta a configuração experimental para investigar as hipóteses propostas.
- **Capítulo 4:** descreve e discute os testes realizados e seus resultados correspondentes.
- **Capítulo 5:** apresenta as nossas conclusões e trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo destacamos conceitos necessários para o entendimento deste trabalho. Descrevemos distúrbios vocais e a importância de identificá-los, as transformadas wavelets discretas e o uso da energia de seus coeficientes de aproximação e detalhes como vetor de características para o uso no classificador escolhido. Também apresentamos o classificador *random forest*, o teste estatístico de Kruskal-Wallis e *post-hoc* de Nemenyi para a validação das hipóteses elaboradas neste trabalho.

2.1 Distúrbios vocais

Esta seção trata aspectos relacionados aos distúrbios ou anomalias na voz. Para isso abordamos o que seria o desenvolvimento de uma voz considerada normal na Subseção 2.1.1 e o que seria uma voz com disfonia na Subseção 2.1.2 .

2.1.1 Desenvolvimento da voz ao longo da vida

O desenvolvimento da voz acompanha e representa o desenvolvimento do indivíduo, tanto do ponto de vista físico como psicológico e social (BEHLAU, 2004). Esses autores afirmam que não existem estudos longitudinais completos sobre os períodos de evolução vocal, mas eles encontraram uma tentativa de classificação feita por Schrager (1966, p. 205-6 apud BEHLAU *et al.*, 2004, p.57) que apresenta seis fases de evolução de acordo com as características vocais:

1. **Neonatal:** do nascimento aos 40 dias de idade, identificam-se emissões de frequências elevadas, ataque vocal[1] brusco e forte intensidade, com modulações reduzidas; a frequência do sinal de voz nos primeiros dias de nascido está ao redor de 400 Hz (Lá3), as emissões chegam a 784 Hz (Sol4) e o grito pode chegar a 1318 Hz (Mi5).
2. **Primeira infância:** do primeiro mês de vida até os seis anos, verifica-se uma redução na presença do ataque vocal brusco e a modulação vocal é mais clara; aos 18 meses aparece a modulação entre 523 e 784 Hz (de Dó4 a Sol4).
3. **Segunda infância:** dos seis anos ao começo da puberdade, as variações na voz chegam a uma oitava e meia de extensão.
4. **Puberdade:** começam a surgir as características vocais de diferenciação sexual, que

[1] O ataque vocal é a maneira como se inicia o som e está relacionado à configuração glótica no momento da emissão (BEHLAU, 2004). Segundo Behlau (2004), o ataque vocal pode ser de três modos: isocrônico (que seria o suave ou normal), brusco e soproso.

são mais notáveis no menino; ocorre a mudança vocal fisiológica ao redor dos 13-14 anos, com a redução da frequência fundamental; a voz do menino passa nessa fase a ser rouca, áspera e soprosa; na menina, a frequência fundamental não se modifica de modo acentuado, porém, gradualmente, ocorre uma diminuição em seu valor, acompanhado por modificações nas características espectrais do som.

5. **Estabilização:** do jovem ao adulto; nesta fase a voz é estável e apresenta características próprias de cada sexo.
6. **Senescência:** período da menopausa e do envelhecimento; o envelhecimento vocal é mais precoce na mulher e pode apresentar um impacto maior na voz cantada; ocorre perda de potência e diminuição dos harmônicos tanto no homem como na mulher, com a diminuição da extensão vocal.

As principais diferenças dos parâmetros vocais na infância, idade adulta e terceira idade encontram-se na Tabela 1. Convém ressaltar que os dados dos indivíduos idosos representam tendências de alteração, e não a realidade vocal de todos os indivíduos nessa faixa etária.

Tabela 1 – Comparação dos principais parâmetros vocais na infância, na idade adulta e na terceira idade (BEHLAU, 2004).

Parâmetros Vocais	**Infância**	**Idade Adulta**	**Terceira Idade**
Qualidade vocal	Delgada	Plena	Tendência a instável e trêmula
F0 média	Acima de 250 Hz	Mulheres: 204 Hz Homens: 113 Hz	Mulheres: 180 Hz Homens: 140 Hz
Pitch[2]	Agudo	Adequado ao sexo	Mulheres: tendência a grave Homens: tendência a agudo
Extensão vocal	Reduzida, com picos extremos ocasionais	Ampla	Perda nos extremos
Gama tonal	Rica à exagerada, mas nos primeiros anos	3 a 5 semitons	Tendência a reduzida
Identificação do sexo	Indiferenciada na voz sustentada	Nitidamente diferenciada	Pode ser comprometida
Intensidade	Moderada para elevada	Extensão ampla	Tendência reduzida
Loudness[3]	Tendência a elevada	Adequada	Tendência reduzida
Estabilidade vocal	Inconstante a reduzida	Adequada	Reduzida
Ataque vocal	Brusco	Isocrônico	Tendência ao soproso
Padrão respiratório	Superior	Médio	Superficial
Coordenação Pneumofonoarticulatória[4]	Tendência à incoordenação por imaturidade neurológica	Adequada	Tendência à incoordenação por falta de suporte respiratório
Tempos máximos de fonação	Abaixo de 12 s	Mulheres: acima de 15 s Homens: acima de 20 s	Mulheres: acima de 10 s Homens: acima de 15 s

[2] *Pitch* é a sensação subjetiva de frequência, de acordo com LOPES FILHO *et al.* (2013). A sensação de frequência é um atributo da impressão auditiva que mostra uma elevação ou diminuição na percepção da escala musical e está sujeita, primeiramente, altura, tonalidade das ondas sonoras, ou seja, a sensação auditiva de que sons podem ser ordenados, variando de graves a agudos (LOPES FILHO *et al.*, 2013)

[3] Segundo Behlau (2004), *loudness* é a sensação psicofísica relacionada à intensidade, ou seja, como julgamos um som, considerando-o forte ou fraco.

[4] A coordenação pneumofonoarticulatória é o resultado da interrelação harmônica das forças expiratórias, mioelásticas da laringe e musculares da articulação (BEHLAU, 2004).

2.1.2 Diferenças entre voz normal e anormal

De acordo com Behlau (2004) para uma voz ser considerada normal, ela tem que possuir um som dito de boa qualidade para quem a escuta e ser produzida sem dificuldade ou desconforto para o falante. Por outro lado, Behlau (2004) afirmam que uma voz dita com distúrbio, também chamada de disfônica, ocorre quando atributos mínimos de harmonia e conforto não são respeitados.

Outros autores também definem o que seria uma voz normal, como Greene & Mathieson (1989 apud BEHLAU *et al.*, 2001), definem a voz normal como uma voz comum, que não apresenta nada especial em seu som. Eles também afirmam que para uma voz ser aceita ela precisa ser forte o suficiente para ser ouvida e apropriada para o sexo e a idade do falante. Segundo os mesmos autores a voz também precisa ser razoavelmente agradável para o ouvinte, modulada e clara, apropriada ao contexto e não muito intensa, não possuindo nenhum desvio acentuado de ressonância.

Desse modo, a disfonia é conceituada por Behlau (2004) como um distúrbio de comunicação oral no qual a voz não consegue cumprir seu papel básico de transmissão da mensagem verbal e emocional. Esses autores afirmam que uma disfonia representa toda e qualquer dificuldade ou alteração na emissão vocal que impede a produção natural da voz. Assim, uma disfonia pode se manifestar através de uma série ilimitada de alterações, tais como: desvios na qualidade vocal, esforço à emissão, fadiga vocal, perda de potência vocal, variações descontroladas da frequência fundamental, falta de volume e projeção, perda de eficiência vocal, baixa resistência vocal e sensações desagradáveis à emissão (BEHLAU, 2004).

Os distúrbios na voz, ou disfonias, são queixas comuns que, de acordo com Cohen *et al.* (2012), Reiter *et al.* (2015), Cohen (2010) e Stachler *et al.* (2018) afetam quase um terço da população em algum momento de sua vida. A disfonia pode afetar pacientes de todas as idades e sexos, mas tem prevalência aumentada em professores, idosos e outras pessoas com demandas vocais significativas (JONES *et al.*, 2002); (LONG *et al.*, 1998); (SMITH *et al.*, 1998); (DAVIDS *et al.*, 2012). De acordo com Stachler *et al.* (2018), os distúrbios na voz são frequentemente causados por doenças benignas ou autolimitadas[5], assim como também podem estar associados a um sintoma inicial de doença mais séria ou progressiva que requer diagnóstico e tratamento imediatos.

[5] Doença ou processo que tem um decurso específico e limitado, com começo, meio e fim, e que pode terminar sem tratamento (AUTOLIMITADO, 2003-2021)

2.2 Extração de características utilizando a transformada wavelet discreta

Nesta seção abordaremos a técnica de extração de características utilizada nesse projeto, que foi a obtenção da energia dos coeficientes de aproximação e de detalhes da wavelet Haar. Antes de apresentar a formulação adotada para a energia dos coeficientes, introduzimos de forma breve a teoria de wavelets nas subseções seguintes.

2.2.1 Análise de Fourier

A Análise de Fourier é um método de definição de formas de onda periódicas em termos de funções trigonométricas (PRAKASH, 2018). De acordo com Prakash (2018), esse método foi estudado e desenvolvido pelo matemático e físico francês Jean-Baptiste Joseph Fourier no século XVIII. Segundo o mesmo autor, a Análise de Fourier afirma que qualquer função $f(x)$ com periodicidade pode ser representada como

$$f(x) = a_0 + \sum_{k=1}^{\infty} \left(a_k cos kx + b_k sen kx\right), \tag{2.1}$$

em que os coeficientes a_0, a_k e b_k são definidos como:

$$a_0 = \frac{1}{2\pi} \int_0^{2\pi} f(x) dx, \tag{2.2}$$

$$a_k = \frac{2}{2\pi} \int_0^{2\pi} f(x) cos kx dx, \tag{2.3}$$

e

$$b_k = \frac{2}{2\pi} \int_0^{2\pi} f(x) sen kx dx. \tag{2.4}$$

Ainda segundo Prakash (2018), a transformada de Fourier tem uma visão muito específica e limitada da frequência. Esse autor afirma que em termos de sinal, os métodos de Fourier são apenas uma coleção de frequências individuais de sinais periódicos dos quais um determinado sinal é composto. Esse autor também afirma que a transformada de Fourier é essencialmente uma integral ao longo do tempo e que, por conta disso, todas as informações que

variam com o tempo são perdidas e que tudo o que pode ser dito da análise de Fourier é que um sinal tem vários componentes de frequência distintos.

De acordo com Prakash (2018), as propriedades matemáticas da matriz de transformada em wavelet e Fourier são semelhantes.O autor ainda afirma que ambos são apenas as rotações no espaço funcional. O mesmo autor ainda afirma que, para Fourier, as funções básicas são senos e cossenos e, para a transformada wavelet, as funções básicas escolhidas são chamadas wavelets, wavelets-mãe ou wavelets de análise. O autor citado nesse parágrafo ainda afirma que observação que as funções wavelet estão localizadas no espaço, enquanto as funções básicas na transformada de Fourier não.

2.2.2 Análise Wavelet

De acordo com Prakash (2018) podemos considerar a Análise de Fourier como primeiro passo para a Análise Wavelet. Como senos e cossenos na análise de Fourier, o autor citado nesse parágrafo utiliza wavelets como funções básicas para representar outras funções. Digamos que a wavelet seja $\psi(x)$, também chamada de wavelet mãe, Prakash (2018) afirma que pode-se formar translações e dilatações da wavelet mãe $\psi(x)$:

$$\left\{ \psi\left(\frac{x-b}{a}\right), (a,b)\varepsilon\mathbb{R}^{+}x\mathbb{R} \right\}, \tag{2.5}$$

onde $a = 2^{-j}$ e $b = k2^{-j}$ e k e j são inteiros. A escolha de a e b requer amostragem crítica, o que dá uma matriz esparsa (uma matriz com a maioria de seus elementos nulos) (PRAKASH, 2018).

Segundo Oliveira *et al.* (2010), as principais propriedades das wavelets para o domínio do tempo são: (I) uma wavelet é uma função finita e é admissível, ou seja, apesar de oscilar, tem média zero; (II) uma wavelet é uma função regular, com as propriedades derivadas de suavidade e continuidade; (III) uma wavelet é uma função com suporte compacto, ou seja, está localizada no espaço.

Essas características permitem a aproximação das wavelets pela superposição nas funções das wavelets-mãe, resultando em um conjunto de representações do sinal em escala de tempo, cada uma com uma resolução diferente, ou seja, uma análise multi-resolução (OLIVEIRA *et al.*, 2010).

2.2.3 Wavelet Haar

Dentre as diversas wavelets existentes escolhemos a wavelet Haar por ela ser ortogonal e permitir uma análise multi-resolução. Segundo (PRAKASH, 2018) a wavelet Haar é considerada a mais simples e antiga wavelet. Esse mesmo autor afirma que ela é uma função escalonada que assume valores de unidade positiva e negativa para algum intervalo definido e desaparece fora desse intervalo. Nickolas (2017) define a wavelet Haar como:

$$\phi(x) = \begin{cases} 1 & 0 \leq x \leq \frac{1}{2} \\ -1 & \frac{1}{2} \leq x \leq 1 \\ 0 & \text{caso contrário.} \end{cases} \tag{2.6}$$

A representação gráfica da wavelet Haar é mostrada na Figura 1. Segundo Nickolas

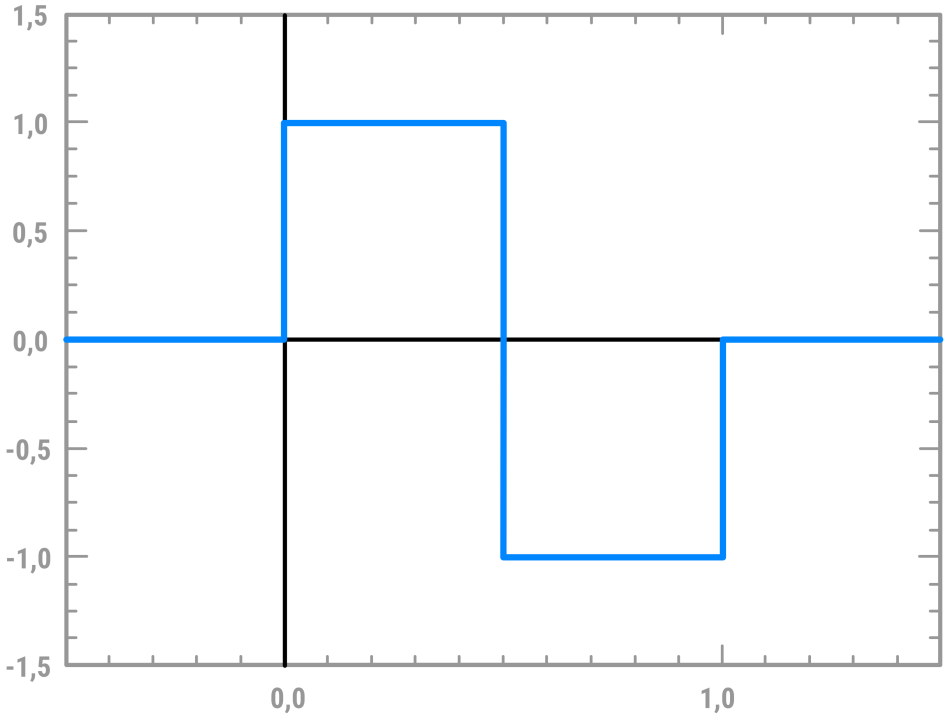


Figura 1 – Representação gráfica da wavelet Haar. Fonte: autoria própria (2021).

(2017), se f for qualquer função, chamamos o conjunto $x\varepsilon\mathbb{R} : f(x) \neq 0$ o apoio de f. Assim, o suporte da wavelet Haar é o intervalo semiaberto [0, 1).

Para fazer uso da wavelet $\phi(x)$ da Equação 2.6, Nickolas (2017) trabalha com suas

'cópias' escalonadas, dilatadas e transladadas de $\phi_{j,k}$, definidas pela fórmula

$$\phi_{j,k}(x) = 2^{j/2}\phi(2^j x - k), \tag{2.7}$$

para todo $j,k\varepsilon\mathbb{Z}$. O termo $2^{j/2}$ é definido como o fator de escala, 2^j corresponde ao fator de dilatação e k indica a translação.

- De acordo com Nickolas (2017), o fator de escala $2^{j/2}$ é o mais simples: esse autor afirma que o efeito desse fator é meramente expandir ou comprimir o gráfico na vertical ou direção y, expandindo no caso $j > 0$ e comprimindo no caso $j < 0$, e não tendo efeito no caso $j = 0$.
- O fator de dilatação 2^j tem o efeito expandir ou comprimir o gráfico na horizontal ou na direção x, comprimindo no caso $j > 0$ e alongando no caso $j < 0$ e não tendo nenhum efeito no caso $j = 0$ (NICKOLAS, 2017).
- O fator de translação k tem o efeito de deslocar o gráfico na horizontal ou na direção x, para a direita no caso $k > 0$ e para a esquerda no caso $k < 0$, e não tendo efeito no caso $k = 0$ (NICKOLAS, 2017).

Neste trabalho extraímos a energia dos coeficientes de aproximação e de detalhes da wavelet Haar. A fórmula utilizada para o cálculo da energia será detalhada na próxima subseção.

2.2.4 Cálculo da energia dos coeficientes

De acordo com Shia e Jayasree (2017) a energia é a característica mais informativa de sinais não estacionários como a fala. Os autores citados acima afirmam que do ponto de vista do processamento de sinal, a energia é frequentemente considerada na forma de densidade espectral de potência. Esses autores afirmam que para um sinal em tempo discreto $x(n)$, a energia associada aos coeficientes da transformada de Fourier é dada pelo teorema de Parseval (OPPENHEIM *et al.*, 2010) no domínio da frequência como:

$$\frac{1}{N}\sum_{n=<N>}|x[n]|^2 = \sum_{k=<N>}|a_k|^2, \tag{2.8}$$

sendo a_k os coeficientes da série de Fourier de $x[n]$, e N, o período (OPPENHEIM *et al.*, 2010). Da mesma forma, Shia e Jayasree (2017) afirmam que a energia associada aos coeficientes DWT de um sinal de voz $x(t)$ pode ser dada pelo teorema de Parseval como:

$$\frac{1}{N}\sum_{t}|x(t)|^2 = \frac{1}{N_j}\sum_{k}|a_j(k)|^2 + \sum_{j}\frac{1}{N_j}\sum_{k}|d_j(k)|^2, \tag{2.9}$$

onde N é o período de amostragem, $a_j(k)$ são os coeficientes de aproximação e $d_j(k)$ são os coeficientes de detalhes. Logo, vemos que o primeiro termo da Equação 2.9 é dado por:

$$\frac{1}{N_j}\sum_{k}|a_j(k)|^2, \tag{2.10}$$

em que este termo corresponde à energia dos coeficientes de aproximação. O segundo termo da Equação 2.9 é dado por:

$$\sum_{j}\frac{1}{N_j}\sum_{k}|d_j(k)|^2, \tag{2.11}$$

e o mesmo corresponde à energia dos coeficientes de energia dos detalhes.

2.3 Classificação por *random forest*

Florestas aleatórias ou *random forests* são descritos por Breiman (2001) como uma combinação de árvores de decisão de modo que cada árvore depende dos valores de um vetor aleatório amostrado independentemente e com a mesma contribuição para todas as árvores na floresta. De acordo com Ali *et al.* (2012), o vetor de características é selecionado no processo de indução e a predição é feita agregando as previsões do conjunto, por meio de voto majoritário para classificação ou média para regressão.

O algoritmo *random forest* (BREIMAN; CUTLER, 2004) pode ser descrito de modo que seja inicialmente colocado o vetor de entrada em cada uma das árvores de decisão na floresta. Cada árvore fornece uma classificação, indicando assim que a árvore 'vota' para aquela classe. Dessa forma, o algoritmo *random forest* escolhe a classificação com mais votos entre as árvores na floresta.

De acordo com Breiman e Cutler (2004) cada árvore é gerada da seguinte maneira:

1. Se o número de casos no conjunto de treinamento for N, tem que ser feita uma amostra de N casos aleatoriamente, mas com substituição, a partir dos dados originais. Este será o conjunto de treinamento para o crescimento da árvore.
2. Se houver M variáveis de entrada, um número $m \ll M$ é especificado de forma que em cada nó, m variáveis sejam selecionadas aleatoriamente de M e a melhor divisão nessas m seja usada para dividir o nó. O valor de m é mantido constante durante o crescimento da floresta.
3. Cada árvore é cultivada na maior extensão possível. Não há poda.

Breiman (2001) em seu trabalho original sobre *random forests* mostrou que a taxa de erro da floresta depende de dois fatores:

- A correlação entre quaisquer duas árvores na floresta. Em outras palavras, se aumentarmos a correlação aumentamos a taxa de erro da floresta.
- A força de cada árvore individual da floresta. Isso significa que uma árvore com baixa taxa de erro representa um classificador forte. Logo, aumentar a força das árvores individuais diminui a taxa de erro da floresta.

2.4 Análise estatística utilizando o teste de Kruskal-Wallis e teste *post-hoc* Nemenyi

Nesta seção abordaremos os testes estatísticos utilizados para verificar as hipóteses delineadas neste trabalho. Esses testes são conhecidos na literatura como Kruskal-Wallis e *post-hoc* de Nemenyi. Nós os utilizamos para comparar os valores de acurácia encontrados a cada nível de decomposição e verificar se há diferença estatística entre eles.

2.4.1 Teste de Kruskal-Wallis

O teste de Kruskal-Wallis, segundo Corder e Foreman (2014), é um procedimento estatístico não paramétrico utilizado para comparar mais de dois grupos amostrais independentes ou não relacionadas. No caso deste trabalho o teste foi utilizado para identificar se há diferença estatística entre as acurácias obtidas por cada nível de decomposição.

De acordo com Kruskal e Wallis (1952), a estatística H desse teste avalia a hipótese nula de que todas as amostras vêm de populações idênticas. Assim, quando o teste de Kruskal-Wallis leva a resultados significativos, então pelo menos um dos grupos de amostras é diferente dos outros grupos.

Segundo Kruskal e Wallis (1952), a estatística H do teste de Kruskal-Wallys, a ser calculada se não houver empates (ou seja, se não houver duas observações iguais) é dada por:

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{K} \frac{R_i^2}{n_i} - 3(N+1), \tag{2.12}$$

onde K é o número de amostras, n_i é o número de observações na amostra de posição i, $N = \sum n_i$ é o número de observações em todas as amostras combinadas e R_i é a soma dos postos na amostra de posição i.

Se o ranqueamento resultar em empate, é necessária uma correção de empate. Nesse caso, Liu e Chen (2012) afirmam que devemos encontrar uma nova estatística H dividindo a estatística H original pelo fator de correção dado por:

$$C = 1 - \frac{\sum(t_i^3 - t_i)}{N^3 - N}, \tag{2.13}$$

onde K é o número de agrupamentos de diferentes postos empatados, t_i é o número de valores empatados no grupo i e N o número de observações em todas as amostras combinadas.

Assim Corder e Foreman (2014) argumentam que se a estatística H não for significativa, não existem diferenças entre quaisquer amostras, mas se a estatística H for significativa, existe uma diferença entre pelo menos duas das amostras. Portanto, os autores citados aconselham a usar testes *post-hoc*, para determinar qual dos pares de amostra é significativamente diferente. Por esse motivo utilizamos o teste *post-hoc* de Nemenyi, que será abordado na subseção seguinte.

2.4.2 Teste de Nemenyi

O teste de Nemenyi é um teste *post-hoc* que foi utilizado neste trabalho após a aplicação do teste de Kruskal-Wallis para encontrar os níveis de decomposição em que houve diferença estatística entre as acurácias.

De acordo com Liu e Chen (2012), no teste de Nemenyi há duas hipóteses. A primeira é a hipótese nula, que assume que as duas amostras são da mesma população. Por outro lado, a hipótese alternativa assume que as duas amostras vêm de diferentes populações.

Assim, segundo Liu e Chen (2012), o valor do teste de Nemenyi é calculado aplicando a seguinte fórmula:

$$X^2 = \frac{(\bar{r}_i - \bar{r}_j)^2}{\frac{n(n+1)}{12}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)C}, \tag{2.14}$$

onde x^2 é estatística qui-quadrado, r_i e r_j são as fileiras médias de grupos i e j, respectivamente, n é o número total de observações em todos os grupos, n_i e n_j são o número de observações em grupos i e j, respectivamente e C é o fator de correlação para valores empatados.

Quando o valor de X^2 é menor que o valor da tabela do qui-quadrado (Anexo A) com um $df = K - 1$, a hipótese nula será aceita; caso contrário, a hipótese nula será rejeitada (LIU; CHEN, 2012).

3 METODOLOGIA

Nesse capítulo apresentamos os materiais e métodos utilizados no projeto. Na primeira seção introduzimos as bases de dados utilizadas nesse trabalho, onde selecionamos duas bases de vozes, disponibilizadas publicamente, com uma grande quantidade de amostras e diversidade de doenças. Na segunda seção descrevemos a etapa de extração de características e de classificação, que foi onde extraímos a energia dos coeficientes de aproximação e detalhes da transformada wavelet Haar e os utilizamos para classificar as vozes entre saudáveis e doentes. A última seção trata da etapa da análise estatística, que aplica o teste estatístico de Kruskal-Wallis e o teste *post-hoc* de Nemenyi para averiguar as hipóteses levantadas no trabalho.

3.1 Bases utilizadas

As bases públicas utilizadas nos experimentos são intituladas *Advanced Voice Function Assessment Database* (AVFAD) (JESUS *et al.*, 2017) e *Saarbrücken Voice Database* (SVD) de (PÜTZER; BARRY, 2007).

A base de dados portuguesa AVFAD contém amostras de vozes de 709 pessoas, divididas entre 363 vozes saudáveis (253 mulheres e 110 homens) e 346 vozes anormais (247 mulheres e 99 homens). Para cada uma dessas pessoas foi solicitado que falassem as vogais /a/, /i/ e /u/ cada vogal sendo repetida três vezes na gravação. Da mesma forma foi solicitado que repetissem cada frase três vezes na gravação sendo elas "A Marta e o avô vivem naquele casarão rosa velho", "Sofia saiu cedo da sala", "A asa do avião andava avariada", "Agora é hora de acabar", "A minha mãe mandou-me embora", "O Tiago comeu quatro peras". Além dessas solicitações foi pedido que lessem um texto chamado "O Vento Norte e o Sol"e que fizessem um discurso espontâneo com no mínimo 30 segundos. Essas amostras foram gravadas a 48 kHz e com resolução de 16 bits.

A base alemã SVD contém dados com gravações de vozes de mais de 2000 mil pessoas, divididas em 687 vozes saudáveis (428 mulheres e 259 homens) e 1354 vozes doentes (727 mulheres e 627 homens). Para cada uma dessas pessoas foi solicitado que falassem as vogais sustentadas /a/, /i/ e /u/ e a frase "Guten Morgen, wie geht es Ihnen?", que significa "Bom dia, como você está?", em tons de voz normal, alto, baixo e baixo-alto-baixo, gravadas a 50 kHz e com resolução de 16 bits.

As amostras de vozes anormais da base AVFAD estão divididas entre 24 doenças,

enquanto que na base SVD há registros de 70 doenças e ambas as bases possuem amostras com oito doenças em comum.

Em nossos experimentos, adotamos vogais sustentadas ao invés de frases devido ao uso frequente de vogais na rotina clínica e para evitar vieses linguísticos (PANEK *et al.*, 2015). Além disso, também utilizamos apenas as amostras de vogais sustentadas com tons normais para a base SVD, pois na base AVFAD as amostras só foram gravadas com um único tom de voz. Com isso, utilizamos todas as vogais sustentadas da base AVFAD e todas as vogais sustentadas em tom normal da base SVD.

3.2 Caracterização e classificação de vozes doentes

Para abordar as hipóteses H1 e H2, realizamos experimentos usando a transformada wavelet Haar discreta não-escalonada como um extrator de características para reconhecimento de voz patológica (Figura 2a).

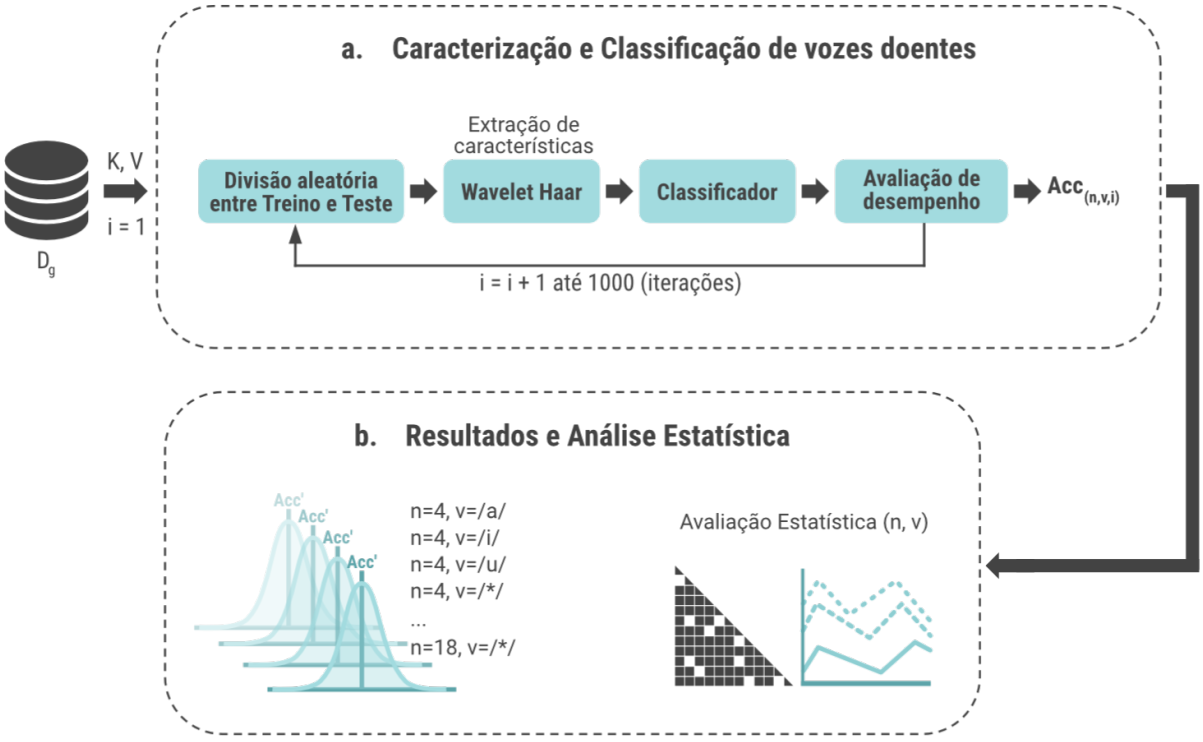


Figura 2 – Metodologia para avaliar o desempenho da wavelet Haar na detecção de doenças na voz. a) Fase de caracterização e classificação, resultando em b) um conjunto de histogramas de acurácia para cada nível de decomposição n e a vogal sustentada de entrada v usada na análise do teste estatístico para investigar as hipóteses. D_g representa um subconjunto de um dos conjuntos de dados pesquisados D, de acordo com o gênero g. K é definido como $K = n|n = \{2,4,6,8,10,12,14,16,18\}$ e $V = v|v = \{/*/,/a/,/i/,/u/\}$. Fonte: autoria própria (2021).

Adotamos vários níveis de decomposição, onde os níveis mínimo e máximo foram 4 e 18, de 2 em 2 níveis. Observamos que as diferenças entre as amostras normais e anormais em ambos os conjuntos de dados surgem no 4º nível de decomposição. Além disso, o valor máximo de 18 níveis foi estabelecido com base na alta ocorrência de coeficientes wavelet nulos em níveis de decomposição acima de 18. A escolha dos níveis de decomposição não modificou o número de amostras (vozes).

Os resultados obtidos por Fonseca *et al.* (2020) e Shia e Jayasree (2017) mostraram que o uso da energia dos coeficientes de decomposição wavelet levou a resultados promissores para detecção de voz patológica. Isso indica que a energia é uma característica relevante em sinais não estacionários. Assim, empregamos o cálculo da energia dos coeficientes de aproximação e decomposição da wavelet de Haar como um descritor.

Visto que as bases AVFAD e SVD têm classes normais e com disfonia desbalanceadas, subamostramos aleatoriamente as classes principais, ou seja, normal (ou saudável) e anormal (ou com disfonia), para balanceá-las. Inspirado no algoritmo de *bootstrap estratificado* (PONS, 2007; EFRON, 1983), realizamos este procedimento selecionando, com substituição, Q amostras das classes principais e secundárias, onde Q é o número de amostras na classe secundária. Utilizamos as amostras selecionadas de ambas as classes para treinamento e as demais para teste. O tamanho do conjunto de teste da classe secundária foi usado para estabelecer o número de amostras no conjunto de teste da classe principal. Executamos esse procedimento de classificação para iterações independentes em cada conjunto de dados D_i de entrada, seguindo o fluxo de trabalho T na Figura 2a. Para todos os experimentos, adotamos o classificador *random forest* (BREIMAN, 2001) com 910 árvores de decisão, critério escolhido sendo Impureza de Gini[1] e profundidade máxima das árvores igual a 95 para classificar as vozes entre saudáveis e doentes.

Dado D_g como um subconjunto de $D \quad | \quad D = \{\text{AVFAD,SVD}\}$ e $g = \{\text{Masculino, Feminino, Gêneros\_Combinados}\}$, propomos dois cenários experimentais C_1 e C_2 em nossa investigação. Descrevemos C_1 e C_2 como $C_1 = T(D_{masculino}) + T(D_{feminino})$ e $C_2 = T(D_{masculino} + T_{feminino}) = T(D_{generos\_combinados})$. Em ambos os cenários, realizamos nossos experimentos nas vogais sustentadas /a/, /i/ e /u/, incluindo as vogais combinadas.

[1] De acordo com So *et al.* (2020) Impureza de Gini é uma métrica usada para medir a aleatoriedade de uma distribuição. Sua fórmula é $Gini = 1 - \sum_{i=0}^{n} p_i^2$ onde p_i aqui representa a probabilidade de um dos valores possíveis da variável de destino ocorrer (SO *et al.*, 2020)

3.3 Análise estatística

Empregamos uma métrica de avaliação bem estabelecida, a acurácia (*Acc*), para avaliar quantitativamente nossos resultados. De acordo com Baratloo *et al.* (2015), essa métrica pode ser descrita como:

$$Acc = \frac{VP + VN}{VP + VN + FP + FN}, \tag{3.1}$$

onde VP, VN, FP, e FN representam verdadeiro positivo, verdadeiro negativo, falso positivo e falso negativo, respectivamente.

Realizamos a classificação da voz aplicando $N = 1000$ iterações (Figura 2a), portanto, produzindo Acc_i para cada iteração (i). Em seguida, obtemos um histograma de acurácias, do qual extraímos a acurácia média (*Acc*) definida por:

$$Acc' = \frac{1}{N} \sum_{i=1}^{N} Acc_i, \tag{3.2}$$

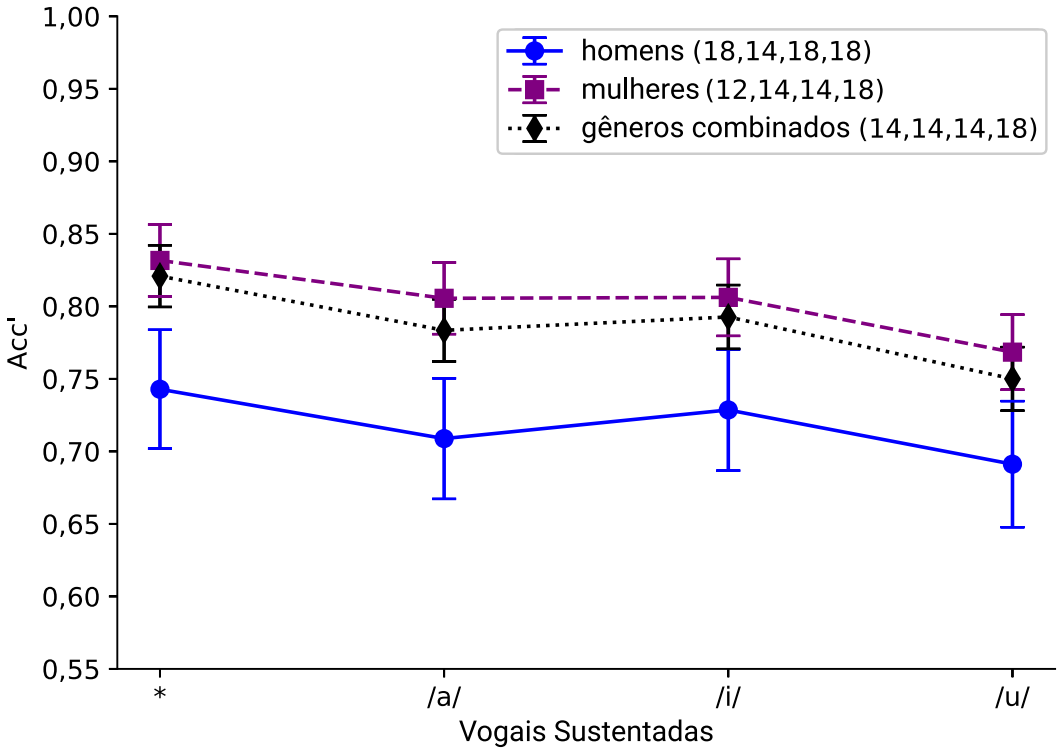
onde N é o número de iterações de classificação. Além disso, calculamos o desvio padrão (σ) da população de Acc_i revelando o intervalo de confiança dos resultados.

Para investigar os impactos das vogais sustentadas e os níveis de decomposição da wavelet de Haar para detecção de voz patológica, realizamos a análise estatística de teste como mostra a Figura 2b. Com base na população de Acc_i, empregamos o teste estatístico Kruskal-Wallis (KRUSKAL; WALLIS, 1952) com o teste *post-hoc* Nemenyi ($\sigma = 0,05$) (HOLLANDER *et al.*, 2013) para encontrar e relatar as combinações de pares entre vogais e nível de decomposição que diferem significativamente entre si. Nossa motivação para usar esta análise estatística foi produzir uma comparação sistemática entre as vogais pesquisadas e encontrar o nível mínimo de decomposição para a caracterização patológica da voz com base na transformada wavelet.

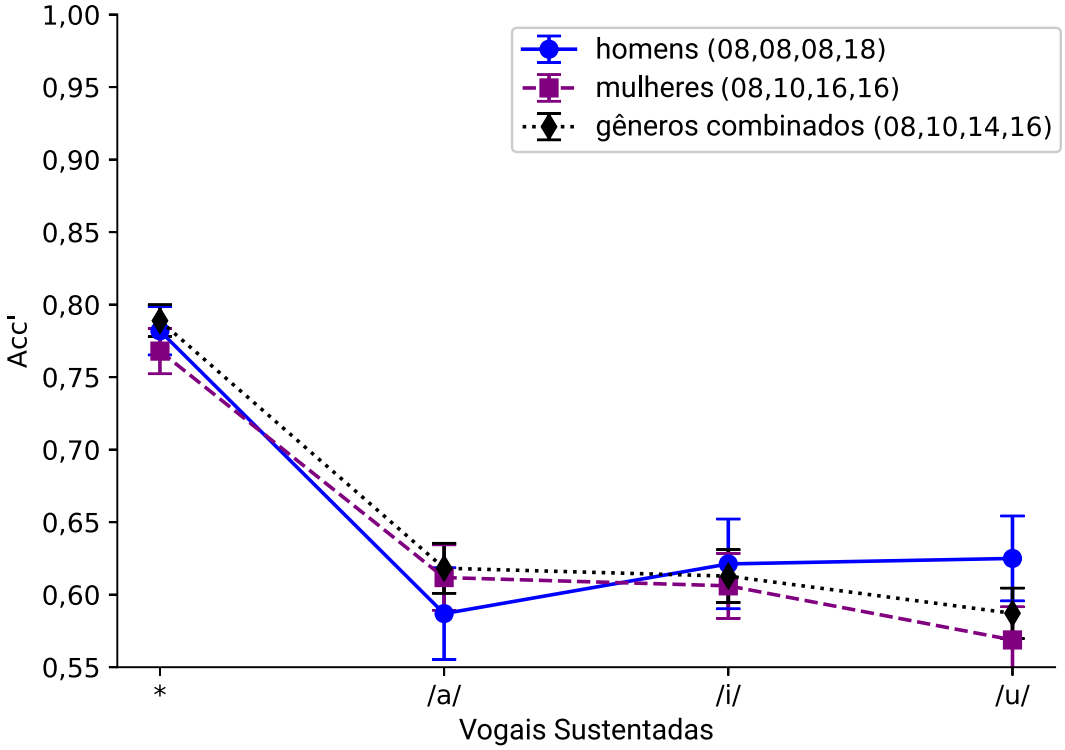
4 RESULTADOS

Analisamos H1 com base nos resultados da Figura 3, onde mostra os melhores valores de acurácia de acordo com as vogais separada e todas as vogais combinadas ($*$) para os cenários C_1 e C_2. É importante notar que as diferentes vogais separadas, por exemplo, /a/, /i/ e /u/ e vogais combinadas ($*$) na Figura 3, foram pronunciadas pelos mesmos indivíduos. A Figura 3 mostra que nossa abordagem alcançou os melhores resultados para AVFAD usando todas as vogais ($*$) para masculino (74,28%), feminino (83,16%) e gêneros combinados (82,08%), respectivamente. Esse resultado foi seguido pela vogal sustentada /i/ para masculino (72,85%), feminino (80,61%) e gêneros combinados (79,26%). Na sequência, a vogal /a/ alcançou valores de acurácia iguais a 70,88%, 80,55% e 78,34% para os gêneros masculino, feminino e combinados, respectivamente. Contrastando com os melhores resultados, a vogal /u/ obteve os seguintes valores de acurácia: masculino (69,11%), feminino (76,84%) e gêneros combinados (75,00%). A Figura 3a também mostra que as vozes femininas alcançaram os melhores resultados, seguidas dos gêneros combinados. Nossa abordagem teve um desempenho significativamente inferior em experimentos com amostras masculinas e produziu o maior desvio padrão. De acordo com o trabalho de (PANEK *et al.*, 2015), o gênero masculino também produziu as maiores variações para as métricas de Sensibilidade e Precisão, considerando as classes saudável e doente, o que explicaria essa diferença encontrada entre os resultados dos gêneros obtidos no nosso trabalho. Esse resultado mostra que os coeficientes wavelet Haar extraídos de vozes do sexo masculino são menos robustos para a caracterização de doenças na voz do que aqueles extraídos do sexo feminino ou dos dois sexos combinados.

A Figura 3b mostra os melhores resultados para o conjunto de dados SVD. Em relação às vogais combinadas ($*$), os valores de acurácia observados foram 78,19%, 76,79% e 78,89% para os gêneros masculino, feminino e combinado, respectivamente. Além disso, as vogais sustentadas produziram os seguintes resultados: a acurácia da vogal /a/ atingiu os valores 58,69%, 61,18% e 61,18% para as vozes masculinas, femininas e vozes de ambos os gêneros combinados respectivamente, enquanto que a acurácia de /i/ atingiu os valores 62,12%, 60,61% e 61,28% paras as vozes masculinas, femininas e de ambos os gêneros combinados respectivamente. Os piores resultados foram obtidos com a vogal sustentada /u/, exceto para o sexo masculino. Os gêneros combinados alcançaram os melhores resultados para as vogais (*) e /a/. As vozes masculinas produziram os piores resultados para a vogal sustentada /a/, e as vogais /i/ e /u/ produziram os melhores resultados.



(a)



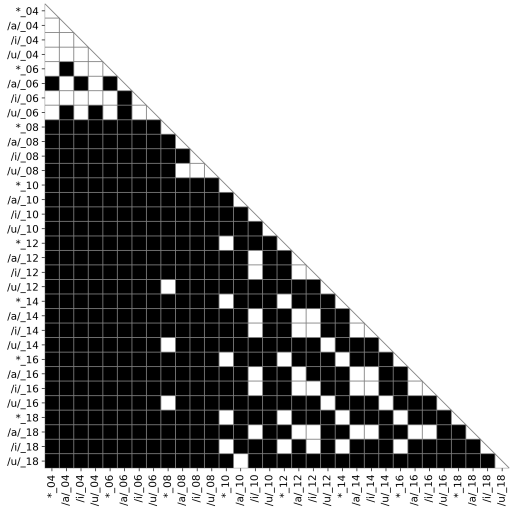
(b)

Figura 3 – Resultados da avaliação quantitativa para análise de combinação de vogais (H1). a) e b) apresentam os melhores valores de acurácia (Equação 3.2) de acordo com as vogais para AVFAD e SVD, respectivamente. Os números na legenda representam os níveis de decomposição da wavelet Haar, seguidos pela sequência de vogais. Fonte: autoria própria (2021).

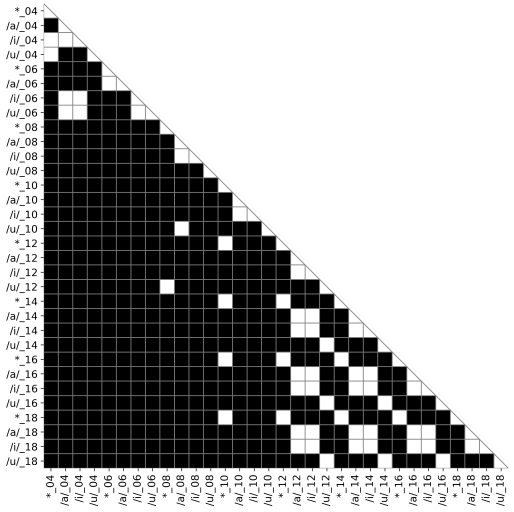
Em resumo, as evidências relatadas na Figura 3 revelam que a extração de características de ($*$) pela metodologia proposta teve melhor desempenho nos cenários C_1 e C_2, confirmando a hipótese H1. Usando /a/ como referência, a combinação de vogais sustentadas melhorou o desempenho da detecção de voz doente da seguinte forma: 15,61% e 2,61% para mulheres, 19,5% e 3,4% para homens, e 17,07% e 3,74% para gêneros combinados em conjuntos de dados SVD e AVFAD, respectivamente. Esses resultados podem ser explicados por nossa etapa de treinamento robusta, que compreende amostras mais discriminantes. Além disso, argumentamos que diferentes componentes de frequência intrínsecos das vogais melhoram a separabilidade entre as classes normais e anormais. Nossos achados reforçam os resultados reportados em (MARTINEZ *et al.*, 2012) e os extrapolam para conjuntos de dados com diferentes métodos de captura do sinal de áudio, doenças, gênero e distribuições de idade (adultos e crianças) usando nossa metodologia baseada em wavelet Haar. Eles podem suportar novas abordagens para análise automática de voz, uma vez que os sistemas amplamente existentes usam uma única vogal, ou seja, /a/, como entrada.

A Figura 4 mostra a análise estatística para os cenários C_1 e C_2. Nesta análise, consideramos pares de variáveis compostas por uma vogal (única ou combinada) e um nível de decomposição wavelet. Assim, indicamos as diferenças significativas entre eles. A comparação par a par nos permitiu identificar os níveis mínimos de decomposição de um sinal vocálico para atingir o melhor desempenho do sistema para detecção precisa de voz doente (hipótese H2). Os pares com diferença significativa em $\alpha = 0,05$ foram sinalizados como caixas pretas na Figura 4. Observamos que o gênero masculino não apresenta diferença estatística notável quanto ao maior número de pares, quando comparado ao feminino.

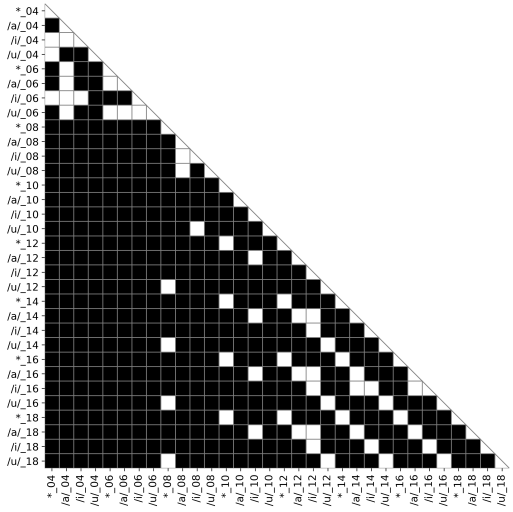
A Tabela 2 mostra o nível mínimo de decomposição da wavelet Haar que fornece a melhor discriminação entre vozes normais e anormais, apoiando a análise H2. Os resultados são apresentados para cada D_g, considerando vogais simples e combinadas. A legenda da Figura 3a mostra 18 níveis de decomposição para o gênero masculino e todas as vogais ($*$). A Tabela 2 apresenta 10 níveis de decomposição para o mesmo caso. Esta discrepância é explicada na Figura 4a, onde o cenário (linha $*$_18) com a melhor acurácia apresenta caixas brancas nas colunas $*$_10, $*$_12, $*$_14 e $*$_16. Consequentemente, o teste de Kruskal-Wallis indica que não há diferença estatisticamente significativa entre esses níveis. Portanto, não observamos melhora significativa na acurácia da classificação para o conjunto de dados AVFAD a partir da decomposição da wavelet Haar de nível 10.



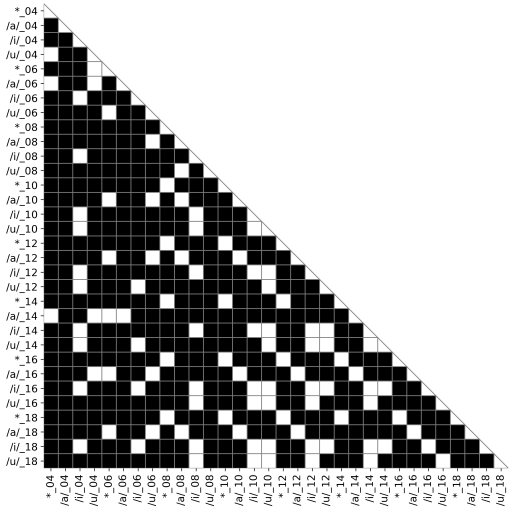
(a) homens - base AVFAD



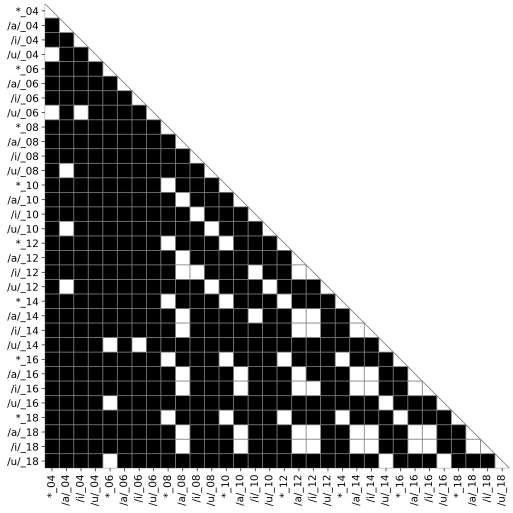
(b) mulheres - base AVFAD



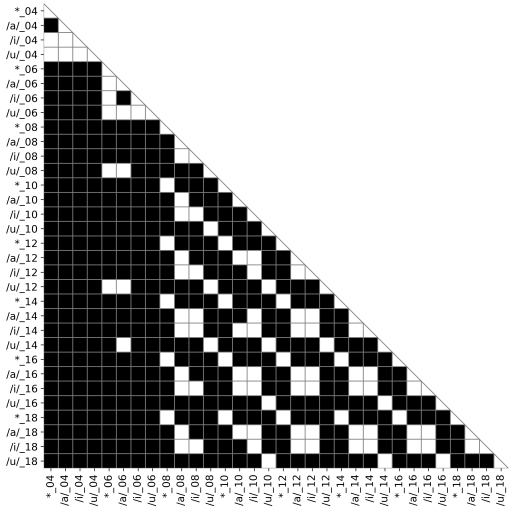
(c) gêneros combinados - base AVFAD



(d) homens - base SVD



(e) mulheres - base SVD



(f) gêneros combinados - base SVD

Figura 4 – Comparações pareadas para todos os gêneros avaliados usando o teste estatístico Kruskal-Wallis com o teste post-hoc de Nemenyi: (a-c) gêneros para o conjunto de dados AVFAD e (d-f) gêneros para o conjunto de dados SVD. As caixas pretas representam os pares com diferenças significativas em $\alpha = 0,05$. Fonte: autoria própria (2021).

Considerando todas as vogais ($*$) para o gênero feminino, a legenda da Figura 3a mostra que a melhor acurácia é obtida com 12 níveis de decomposição; em contraste, a Tabela 2 apresenta 10 níveis de decomposição para o mesmo caso. Essa diferença ocorre porque a linha $*$ 12 da Figura 4b apresenta caixas brancas na coluna $*$ 10. Portanto, o teste de Kruskal-Wallis sugere que não há diferença significativa entre esses níveis, o que significa que 10 níveis de decomposição oferecem resultados semelhantes com a melhor acurácia.

Tabela 2 – Número mínimo de níveis de decomposição da transformada wavelet Haar.

Níveis de decomposição da wavelet Haar	**AVFAD**				**SVD**			
	/a/	***/i/***	***/u/***	***vogais combinadas***	***/a/***	***/i/***	***/u/***	***vogais combinadas***
homens	12	12	18	10	8	4	10	8
mulheres	14	14	12	10	10	14	14	8
gêneros combinados	12	12	12	10	8	8	10	8

A análise da Figura 3a indica que a melhor acurácia pode ser alcançada com 14 níveis de decomposição quando consideramos os gêneros combinados e todas as vogais ($*$). Da mesma forma, a Figura 4c aponta que 10 níveis são estatisticamente satisfatórios para obter resultados próximos aos melhores valores de acurácia. A linha $*$_14 é estatisticamente semelhante às outras duas: $*$_10 e $*$_12. Portanto, $*$_10 é o menor número de níveis de decomposição necessários para obter resultados próximos da melhor acurácia. Os resultados dos outros cenários apresentados na Tabela 2 foram obtidos de forma semelhante.

A Tabela 2 mostra que nossa abordagem converge estatisticamente para menos níveis de decomposição quando treinada em todas as vogais ($*$) e amostras de gêneros combinados. Mostramos que resultados precisos de amostras masculinas exigiram menos níveis de decomposição do que amostras femininas, exceto para a vogal /u/. Esse achado pode ser devido às propriedades da voz masculina, como largura de banda estreita e frequência fundamental inferior, em comparação com as vozes femininas (CHILDERS; WU, 1991). Assim, as distorções na voz causadas por doenças têm um impacto diferente nas vozes masculinas e femininas e tornam-se observáveis nas maiores faixas de largura de banda dos primeiros níveis de decomposição para os homens. Por meio da avaliação cruzada do conjunto de dados apresentada na Figura 4 e na Tabela 2, identificamos que do nível 8 ao 10 a decomposição da wavelet Haar forneceu recursos notáveis para realizar detecção de voz doente precisa, confirmando a hipótese H2.

5 CONCLUSÕES E TRABALHOS FUTUROS

Neste trabalho realizamos experimentos empregando a transformada wavelet Haar discreta como um extrator de características para reconhecimento de voz doente em duas bases de dados públicas, uma portuguesa e outra alemã. Separamos os dados por vogais e por gêneros, assim como também adotamos vários níveis de decomposição, onde os níveis mínimo e máximo foram 4 e 18, de 2 em 2 níveis. Observamos que as diferenças entre as amostras normais e anormais em ambos os conjuntos de dados surgem no 4º nível de decomposição. Além disso, o valor máximo de 18 níveis foi estabelecido com base na alta ocorrência de coeficientes wavelet nulos em níveis acima de 18 de decomposição. Com isso empregamos o cálculo da energia dos coeficientes de aproximação e decomposição da wavelet de Haar como um descritor. Em seguida realizamos a classificação utilizando o classificador *Random Forest*. Por fim, para investigar os impactos das vogais sustentadas e os níveis de decomposição da wavelet Haar para detecção de voz doente empregando o teste estatístico Kruskal-Wallis com o teste post-hoc Nemenyi ($\sigma = 0,05$) e relatamos as combinações de pares entre vogais e nível de decomposição que diferem significativamente entre si.

Nossos resultados mostraram que as vogais combinadas proporcionam maior capacidade de caracterizar e detectar distúrbios vocais independentemente dos gêneros. Além disso, eles também indicam que os sinais das vozes femininas requerem níveis de decomposição mais elevados, exceto para a vogal sustentada /u/, provavelmente devido às propriedades da voz masculina, que diferem das vozes femininas, o que faz com que as distorções na voz causadas por doenças tenham um impacto diferente nas vozes masculinas e femininas e tornando-as observáveis nas maiores faixas de largura de banda dos primeiros níveis de decomposição para os homens.

Uma vez que demonstramos que as vogais combinadas melhoram a eficiência do modelo baseado em wavelet Haar para detectar vozes anormais, outros fatores de interferência potenciais devem ser investigados, como a faixa de frequência em que cada doença opera predominantemente. Como trabalhos futuros investigaremos descritores baseados em CNNs para avaliar diferentes doenças por gênero e vogal.

REFERÊNCIAS

ALHUSSEIN, M.; MUHAMMAD, G. Voice pathology detection using deep learning on mobile healthcare framework. **IEEE Access**, v. 6, p. 41034–41041, jul. 2018.

ALI, J.; KHAN, R.; AHMAD, N.; MAQSOOD, I. Random forests and decision trees. **International Journal of Computer Science Issues(IJCSI)**, v. 9, p. 273–278, set. 2012.

AUTOLIMITADO. in DICIONÁRIO infopédia da Língua Portuguesa. Porto: Porto Editora, 2003–2021. Acesso em: 19 abr. 2021. Disponível em: <https://www.infopedia.pt/dicionarios/lingua-portuguesa/autolimitado>.

BARATLOO, A.; HOSSEINI, M.; NEGIDA, A.; ASHAL, G. E. Evidence based emergency medicine; Part 1: Simple definition and calculation of accuracy, sensitivity and specificity. **Emergency**, v. 3, p. 48–49, maio 2015.

BEHLAU, M. **Voz: o livro do especialista**. Rio de Janeiro: Editora RevinteR Ltda., 2004. v. 1.

BREIMAN, L. Random forests. **Machine Learning**, v. 45, n. 1, p. 5–32, 2001.

BREIMAN, L.; CUTLER, A. **Random Forests**. 2004. Disponível em: <https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm>. Acesso em: 01 maio 2021.

CHANDRAN, V.; BOASHASH, B. Time-frequency methods in radar, sonar, and acoustics. **Time-Frequency Signal Analysis and Processing: A comprehensive reference**, Academic Press, p. 793–856, 2016.

CHILDERS, D. G.; WU, K. Gender recognition from speech. Part II: Fine analysis. **The Journal of the Acoustical Society of America**, v. 90, n. 4, p. 1841–1856, 1991.

COHEN, S. Self-reported impact of dysphonia in a primary care population: An epidemiological study. **The Laryngoscope**, v. 120, p. 2022–32, out. 2010.

COHEN, S.; KIM, J.; ROY, N.; ASCHE, C.; COUREY, M. Prevalence and causes of dysphonia in a large treatment-seeking population. **The Laryngoscope**, v. 122, p. 343–8, fev. 2012.

CORDER, G. W.; FOREMAN, D. I. **Nonparametric statistics : a step-by-step approach**. 2. ed. Hoboken: John Wiley & Sons, Inc., 2014.

DAVIDS, T.; KLEIN, A.; JOHNS, M. Current dysphonia trends in patients over the age of 65: Is vocal atrophy becoming more prevalent? **The Laryngoscope**, v. 122, p. 332–5, fev. 2012.

EFRON, B. Estimating the error rate of a prediction rule: improvement on cross-validation. **Journal of the American Statistical Association**, v. 78, n. 382, p. 316–331, 1983.

FANG, S.; TSAO, Y.; HSIAO, M.; CHEN, J.; LAI, Y.; LIN, F.; WANG, C. Detection of pathological voice using cepstrum vectors: A deep learning approach. **Journal of Voice**, v. 33, n. 5, p. 634–641, 2019.

FERNANDES, D. A.; NETO, M.; SOARES, L. F.; FREIRE, M. M.; INÁCIO, P. R. On the self-similarity of traffic generated by network traffic simulators. In: OBAIDAT, M. S.; NICOPOLITIDIS, P.; ZARAI, F. (Ed.). **Modeling and Simulation of Computer Networks and Systems: Methodologies and Applications**. Boston: Elsevier, 2015. cap. 10, p. 285–311.

FONSECA, E. S.; GUIDO, R. C.; JUNIOR, S. B.; DEZANI, H.; GATI, R. R.; PEREIRA, D. C. M. Acoustic investigation of speech pathologies based on the discriminative paraconsistent machine (DPM). **Biomedical Signal Processing and Control**, v. 55, p. 101615, 2020.

FONSECA, E. S.; PEREIRA, D. C. M.; MASCHI, L. F. C.; GUIDO, R. C.; PAULO, K. C. S. Linear prediction and discrete wavelet transform to identify pathology in voice signals. In: **2017 Signal Processing Symposium (SPSympo)**. Jachranka: IEEE, 2017. v. 1, p. 1–4.

GÓMEZ-GARCÍA, J.; MORO-VELÁZQUEZ, L.; GODINO-LLORENTE, J. I. On the design of automatic voice condition analysis systems. Part II: Review of speaker recognition techniques and study on the effects of different variability factors. **Biomedical Signal Processing and Control**, v. 48, p. 128 – 143, 2019.

GÓMEZ-GARCÍA, J. A.; MORO-VELÁZQUEZ, L.; GODINO-LLORENTE, J. I. On the design of automatic voice condition analysis systems. Part I: Review of concepts and an insight to the state of the art. **Biomedical Signal Processing and Control**, v. 51, p. 181 – 199, 2019.

GUIDO, R. C.; BARBON, S.; SOLGON, R. D.; PAULO, K. C. S.; RODRIGUES, L. C.; SILVA, I. N. da; ESCOLA, J. P. L. Introducing the discriminative paraconsistent machine (DPM). **Information Sciences**, v. 221, p. 389 – 402, 2013.

HEGDE, S.; SHETTY, S.; RAI, S.; DODDERI, T. A survey on machine learning approaches for automatic detection of voice disorders. **Journal of Voice**, v. 33, n. 6, p. 947.e11 – 947.e33, 2019.

HEIL, C. E.; WALNUT, D. F. Continuous and discrete wavelet transforms. **SIAM Review**, v. 31, n. 4, p. 628–666, 1989.

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

HOLLANDER, M.; WOLFE, D.; CHICKEN, E. **Nonparametric statistical methods**. 3. ed. Nova Iorque: John Wiley & Sons, 2013. v. 751.

JESUS, L.; BELO, I.; MACHADO, J.; HALL, A. The advanced voice function assessment databases (AVFAD): Tools for voice clinicians and speech engineering research. **Advances in Speech-language Pathology**, IntechOpen, p. 237–255, set. 2017.

JONES, K.; SIGMON, J.; HOCK, L.; NELSON, E.; SULLIVAN, M.; OGREN, F. Prevalence and Risk Factors for Voice Problems Among Telemarketers. **Archives of Otolaryngology–Head Neck Surgery**, v. 128, n. 5, p. 571–577, maio 2002.

KRUSKAL, W. H.; WALLIS, W. A. Use of ranks in one-criterion variance analysis. **Journal of the American Statistical Association**, Taylor Francis, v. 47, n. 260, p. 583–621, 1952.

LIU, Y.; CHEN, W. A sas macro for testing differences among three or more independent groups using kruskal-wallis and nemenyi tests. **Journal of Huazhong University of Science and Technology [Medical Sciences]**, v. 32, p. 130–4, fev. 2012.

LONG, J.; WILLIFORD, H. N.; OLSON, M. S.; WOLFE, V. Voice problems and risk factors among aerobics instructors. **Journal of Voice**, v. 12, n. 2, p. 197–207, 1998.

LOPES FILHO, O.; CAMPIOTTO, A. R.; LEVY, C. C. A. da C.; REDONDO, M. do C.; ANELLI, W. **Novo Tratado de Fonoaudiologia**. 3. ed. Barueri: Editora Manole Ltda., 2013.

MARTINEZ, D.; LLEIDA, E.; ORTEGA, A.; MIGUEL, A.; VILLALBA, J. Voice pathology detection on the saarbrücken voice database with calibration and fusion of scores using multifocal toolkit. In: **Communications in Computer and Information Science**. Berlin: Springer, 2012. v. 328, p. 99–109.

MARTINS, R. H. G.; DOMINGUES, M. A.; FABRO, A. T.; DIAS, N. H.; SANTANA, M. F. Edema de Reinke: estudo da imunoexpressão da fibronectina, da laminina e do colágeno IV em 60 casos por meio de técnicas imunoistoquímicas. **Brazilian Journal of Otorhinolaryngology**, v. 75, p. 821 – 825, dez. 2009.

NICKOLAS, P. **Wavelets: A Student Guide**. 1. ed. Cambridge: Cambridge University Press, 2017. (Australian Mathematical Society Lecture Series).

OLIVEIRA, E. A. F. A.; BIANCHI, A. G. C.; MARTINS-FILHO, L. d. S.; MACHADO, R. F. Granulometric analysis based on the energy of Wavelet Transform coefficients. **Rem: Revista Escola de Minas**, v. 63, p. 347 – 354, jun. 2010.

OPPENHEIM, A. V.; WILLSKY, A. S.; NAWAB, S. H. **Sinais e sistemas**. 2. ed. São Paulo: Pearson Prentice Hall, 2010.

PANEK, D.; SKALSKI, A.; GAJDA, J.; TADEUSIEWICZ, R. Acoustic analysis assessment in speech pathology detection. **International Journal of Applied Mathematics and Computer Science**, Sciendo, Berlin, v. 25, n. 3, p. 631 – 643, 2015.

PONS, O. Bootstrap of means under stratified sampling. **Electronic Journal of Statistics**, v. 1, p. 381–391, 2007.

PRAKASH, A. Wavelet and its applications. **International Journal of Scientific Research in Computer Science, Engineering and Information Technology**, v. 3, p. 95–104, nov. 2018.

PÜTZER, M.; BARRY, W. J. **Saarbrücken Voice Database**. 2007. Disponível em: <http://www.stimmdatenbank.coli.uni-saarland.de.>

REITER, R.; HOFFMANN, T. K.; PICKHARD, A.; BROSCH, S. Hoarseness. **Dtsch Arztebl International**, v. 112, n. 19, p. 329–337, 2015.

SHIA, S. E.; JAYASREE, T. Detection of pathological voices using discrete wavelet transform and artificial neural networks. In: **2017 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS)**. Srivilliputtur: IEEE, 2017. p. 1–6.

SMITH, E.; KIRCHNER, H. L.; TAYLOR, M.; HOFFMAN, H.; LEMKE, J. H. Voice problems among teachers: Differences by gender and teaching characteristics. **Journal of Voice**, v. 12, n. 3, p. 328–334, 1998.

SO, A.; SO, W.; NAGY, Z. **The Applied Artificial Intelligence Workshop: Start working with AI today, to build games, design decision trees, and train your own machine learning models**. 1. ed. Birmingham: Packt Publishing, 2020.

STACHLER, R. J.; FRANCIS, D. O.; SCHWARTZ, S. R.; DAMASK, C. C.; DIGOY, G. P.; KROUSE, H. J.; MCCOY, S. J.; OUELLETTE, D. R.; PATEL, R. R.; REAVIS, C. C. W.; SMITH, L. J.; SMITH, M.; STRODE, S. W.; WOO, P.; NNACHETA, L. C. Clinical practice guideline: Hoarseness (dysphonia) (update) executive summary. **Otolaryngology–Head and Neck Surgery**, v. 158, n. 3, p. 409–426, 2018.

URSO, A.; FIANNACA, A.; ROSA, M. L.; RAVÌ, V.; RIZZO, R. Data mining: Classification and prediction. In: RANGANATHAN, S.; GRIBSKOV, M.; NAKAI, K.; SCHÖNBACH, C. (Ed.). **Encyclopedia of Bioinformatics and Computational Biology**. Oxford: Elsevier, 2019. p. 384–402.

WU, H.; SORAGHAN, J.; LOWIT, A.; CATERINA, G. D. Convolutional neural networks for pathological voice detection. In: **2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)**. Honolulu: IEEE, 2018. p. 1–4.

ZHANG, X.; HU, W. Pathological voice classification based on wavelet packet multiscale analysis. In: **Proceedings of the 2018 International Conference on Algorithms, Computing and Artificial Intelligence**. Sanya, China: Association for Computing Machinery, 2018. p. 1–6.

ZHENG, F.; ZHANG, G.; SONG, Z. Comparison of different implementations of mfcc. **J. Comput. Sci. Technol.**, v. 16, p. 582–589, nov. 2001.

ANEXO A – TABELA DE DISTRIBUIÇÃO QUI-QUADRADO

Graus de	α									
liberdade	**0,995**	**0,99**	**0,975**	**0,95**	**0,90**	**0,10**	**0,05**	**0,025**	**0,01**	**0,005**
1	–	–	0,001	0,004	0,016	2,706	3,841	5,024	6,635	7,879
2	0,010	0,020	0,051	0,103	0,211	4,605	5,991	7,378	9,210	10,597
3	0,072	0,115	0,216	0,352	0,584	6,251	7,815	9,348	11,345	12,838
4	0,207	0,297	0,484	0,711	1,064	7,779	9,488	11,143	13,277	14,860
5	0,412	0,554	0,831	1,145	1,610	9,236	11,070	12,833	15,086	16,750
6	0,676	0,872	1,237	1,635	2,204	10,645	12,592	14,449	16,812	18,548
7	0,989	1,239	1,690	2,167	2,833	12,017	14,067	16,013	18,475	20,278
8	1,344	1,646	2,180	2,733	3,490	13,362	15,507	17,535	20,090	21,955
9	1,735	2,088	2,700	3,325	4,168	14,684	16,919	19,023	21,666	23,589
10	2,156	2,558	3,247	3,940	4,865	15,987	18,307	20,483	23,209	25,188
11	2,603	3,053	3,816	4,575	5,578	17,275	19,675	21,920	24,725	26,757
12	3,074	3,571	4,404	5,226	6,304	18,549	21,026	23,337	26,217	28,300
13	3,565	4,107	5,009	5,892	7,042	19,812	22,362	24,736	27,688	29,819
14	4,075	4,660	5,629	6,571	7,790	21,064	23,685	26,119	29,141	31,319
15	4,601	5,229	6,262	7,261	8,547	22,307	24,996	27,488	30,578	32,801
16	5,142	5,812	6,908	7,962	9,312	23,542	26,296	28,845	32,000	34,267
17	5,697	6,408	7,564	8,672	10,085	24,769	27,587	30,191	33,409	35,718
18	6,265	7,015	8,231	9,390	10,865	25,989	28,869	31,526	34,805	37,156
19	6,844	7,633	8,907	10,117	11,651	27,204	30,144	32,852	36,191	38,582
20	7,434	8,260	9,591	10,851	12,443	28,412	31,410	34,170	37,566	39,997
21	8,034	8,897	10,283	11,591	13,240	29,615	32,671	35,479	38,932	41,401
22	8,643	9,542	10,982	12,338	14,041	30,813	33,924	36,781	40,289	42,796
23	9,260	10,196	11,689	13,091	14,848	32,007	35,172	38,076	41,638	44,181
24	9,886	10,856	12,401	13,848	15,659	33,196	36,415	39,364	42,980	45,559
25	10,520	11,524	13,120	14,611	16,473	34,382	37,652	40,646	44,314	46,928
26	11,160	12,198	13,844	15,379	17,292	35,563	38,885	41,923	45,642	48,290
27	11,808	12,879	14,573	16,151	18,114	36,741	40,113	43,195	46,963	49,645
28	12,461	13,565	15,308	16,928	18,939	37,916	41,337	44,461	48,278	50,993
29	13,121	14,256	16,047	17,708	19,768	39,087	42,557	45,722	49,588	52,336
30	13,787	14,953	16,791	18,493	20,599	40,256	43,773	46,979	50,892	53,672
40	20,707	22,164	24,433	26,509	29,051	51,805	55,758	59,342	63,691	66,766
50	27,991	29,707	32,357	34,764	37,689	63,167	67,505	71,420	76,154	79,490
60	35,534	37,485	40,482	43,188	46,459	74,397	79,082	83,298	88,379	91,952
70	43,275	45,442	48,758	51,739	55,329	85,527	90,531	95,023	100,425	104,215
80	51,172	53,540	57,153	60,391	64,278	96,578	101,879	106,629	112,329	116,321
90	59,196	61,754	65,647	69,126	73,291	107,565	113,145	118,136	124,116	128,299
100	67,328	70,065	74,222	77,929	82,358	118,498	124,342	129,561	135,807	140,169

Fonte: LARSON, R; FARBER, B. **Estatística Aplicada**.4.ed. São Paulo: Pearson Prentice Hall, 2010.